LEARNING AND UNLEARNING WITH NOISY GRADIENT DESCENT

by

RISHAV CHOURASIA
(B.Tech. in Computer Science and Engineering, IIT Guwahati)

A THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

in the

GRADUATE DIVISION

of the

NATIONAL UNIVERSITY OF SINGAPORE

2024

Supervisor:
Professor Xiaokui Xiao

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Rishav Chourasia

December 19, 2024

Acknowledgments

This thesis would not have been possible without the unwavering support, encouragement, and love of my family and friends. I owe profound gratitude to my parents, whose early nurturing of my curiosity and steady belief in my abilities laid the foundation for my academic journey.

To my partner, *Bhawna Sharma*, your boundless patience and understanding—especially during the most demanding phases of this work—kept me grounded and motivated. I am equally indebted to my supervisor, **Dr. Xiaokui Xiao**, whose kindness, guidance, and unwavering confidence in my potential helped me navigate many challenging moments.

I am grateful to my friends—Martin, Bob, Hannah, Hongyan, and Duy—who offered uplifting distractions when needed and heartfelt encouragement when perseverance seemed difficult. My university peers have also been invaluable, contributing thoughtful discussions, shared experiences, and supportive camaraderie that made this research endeavor both achievable and enjoyable.

I extend my sincere thanks to the team at Betterdata—**Dr. Uzair Javaid**, **Kevin Yee**, and **Prof. Biplak Sikdar**—for recognizing the promise of my research, entrusting me as their data privacy expert, and providing the financial support that enabled my work.

Finally, I wish to acknowledge the members of my thesis committee, **Dr. Jonathan Scarlett** and **Dr. Prashant Nalini Vasudevan**, for their constructive feedback and thoughtful guidance. Their input has greatly enriched the quality, clarity, and presentation of this thesis.

Contents

Abstract

Learning and Unlearning
with Noisy Gradient Descent

by

Rishav Chourasia

Doctor of Philosophy in Computer Science

National University of Singapore

Stochastic Gradient Descent (SGD) is one of the classical optimization algorithms in machine learning (ML). SGD and its variants are the dominant training techniques used in deep learning today, powering complex neural networks for tasks such as natural language processing, image generation, and recommendation systems. In recent years, heightened awareness of personal data protection, stricter regulations, and a surge in privacy attacks have significantly increased the importance of protecting the privacy of training data in ML models. There is a growing interest in ML algorithms that satisfy *differential privacy* (DP), a mathematical guarantee that bounds the worst-case information leakage about the training data points through the algorithm's output. Most notable among such *privacy-preserving ML* algorithms is *Noisy Gradient Descent* (Noisy-GD).

However, first-order optimization methods like Noisy-GD have a complex interaction with the training dataset, making it challenging to derive sharp privacy guarantees. This issue has received considerable attention in recent years. This thesis develops contemporary mathematical tools and techniques for customizing the privacy analysis of complex algorithms, specifically Noisy-GD and its variants. Tailoring the privacy analysis to an algorithm directly benefits its utility as it reduces the amount of noise necessary to meet the privacy constraints.

First, we study the *privacy dynamics of Noisy-GD* algorithm by formalizing the *interplay between noise and gradients during training*. Under certain loss function and hyperparameters, we prove a privacy bound for Noisy-GD that converges exponentially fast in the number of iterations. In contrast, prior privacy analyses based on composition yield bounds that keep growing proportionally with the number

of iterations, eventually becoming vacuous. We also show that our *privacy bound is tight* and that Noisy-GD achieves the *optimal utility that can be achieved under DP constraints*.

Second, we analyze Noisy-GD as a *machine unlearning* algorithm, which deals with *selective erasure* of specific training data points from a trained ML model at a cost lower than full retraining. Unfortunately, the nascent notion of machine unlearning does not align with the legal guidelines of the *Right to be Forgotten*. We highlight two unique flaws in present definitions that may lead to catastrophic breach of privacy in compliant algorithms and remedy this by proposing a *data deletion* guarantee that is *provably sound*. Then, we analyze the four-way trade-off between Noisy-GD's differential privacy, deletion privacy, utility and computational saving guarantees. A remarkable result in this regard is that for convex losses, unlearning using Noisy-GD is able to maintain the Pareto-optimal DP/utility trade-off for an indefinite number of deletion requests with a computation footprint that is negligible in comparison to the cost of a full retraining.

Finally, we present a novel interpretation of DP through the lens of Laplace transformations. We highlight a set of expressions that appear in several related works on analyzing DP, either as an integral or an expectation, and show that recognizing these expressions as a Laplace transform unlocks a new way to reason about DP properties by exploiting the duality between time and frequency domains. This interpretation provides a formal approach to reason about relevant properties such as adaptive composition, amplification on sub-sampling, and interconversion between alternate formulations of DP, such as (ε, δ)-DP, Rényi DP, and PLD formalism. With our interpretation, we strengthen the equivalences across functional notions of differential privacy. We also show that all functional definitions of differential privacy include a reversal theorem, which addresses the issue of asymmetry in information leakage (i.e., differences under "add" vs. "remove" adjacency) while ensuring precise privacy accounting under composition.

In summary, this thesis provides a comprehensive exploration of the information-theoretic properties of Noisy-GD with applications in both privacy-preserving machine learning and unlearning. Additionally, mathematical techniques presented in this thesis offer novel expressions, manipulations, and interpretations of differential privacy that befit contemporary explorations on privacy in machine learning.

List of Figures

List of Tables

Chapter 1

Introduction

Discussions about privacy issues have existed throughout human history, with early references dating back to Aristotle's philosophies in ancient Greece [82]. The right to privacy has been acknowledged as a fundamental element of human dignity and has been legally protected in nations like Italy and France for at least 200 years [98, 14]. Societal perceptions of privacy however have evolved with changing societal needs [95, 48, 53]. Extensive academic discourse traces this evolution of philosophies on privacy [91, 94, 85]. At the turn of the 21st century, the explosion of digital data and its rapid dissemination on the internet created an urgent need to again adapt privacy concepts to address new threats from rapid technological and socio-economic changes. Subsequently, the advent of the internet era prompted the enactment of several new privacy policies like the European Union Data Protection Directive in 1995 and the 1998 Children's Online Privacy Protection Act (COPPA) in the USA.

At the time, many scholars noted that both literature and legislations on privacy treated "information" as a commodity whose movement can be regulated as if it were a tangible object with defined boundaries [16, 94]. In reality, information is a conceptual object that is difficult to govern; information does not conform to a fixed structure (e.g., a written document can be dictated over a phone) and cannot be redacted once revealed (e.g., documents published on platforms like Wikileaks). Crucially, information does not combine or decompose like physical goods; it neither adds nor divides in the same manner. Two separate pieces of information, which might seem trivial on their own, can provide significant insights when merged. The pizza index exemplifies this, revealing how the innocuous detail of pizza orders from

Domino's franchises in the Washington area can indicate a crisis at the U.S. Capitol. This occurs when paired with the insight that government staff members tend to order more pizzas during prolonged crises.

By Sarah Schafer
Washington Post Staff Writer
Saturday, December 19, 1998; Page D1

Call it the Washington pizza index: The bigger the crisis and the more time that government staffers hole up in their offices, the more pizza they eat.

This week – with the confluence of impeachment hearings in Congress and war management over at the Pentagon – the index is shattering records.

"It's going haywire," said Frank Meeks, owner of 59 Domino's franchises in the Washington area and the capital's unofficial

An aide to House Majority Whip Tom DeLay (R-Texas) carries a stack of pizzas during the day-long impeachment debate on Friday. (Reuters)

Figure 1.1: Snippet from a Washington Post article on the *pizza index*.

Based on the same principle, several reconstruction attacks were brought to light that could link multiple sources of information to reveal sensitive information about the contributors [100, 101, 33], as shown in Figure 1.2.

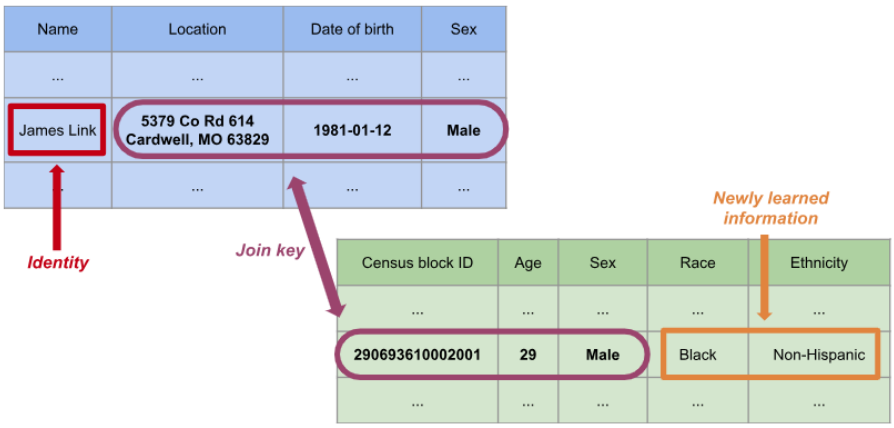


Figure 1.2: Demonstration of a linkage attack using two different datasets that might contain overlapping information about an individual.

Hence, there was a need for a mathematically grounded definition that addresses not only how society views privacy norms but also the nuanced nature of information itself.

1.1 Why Differential Privacy works

In the early 2000s, several *quantitative notions of information privacy* emerged based on intuitive views, such as k-Anonymity [102], t-Closeness [63], and L-diversity [66]. These notions were based on a database-like perception of information where personal data of an individual corresponds to a distinct record entry. These notions attempted to quantify privacy by leveraging redundancies among entries in the database to obscure individual identities when the database is queried (e.g., when every query on the database returns k-matches). While these notions could be effective in many situations, it was soon realized that access to *side information* can easily invalidate the privacy protection by making it easy to eliminate the redundant entries that fail to match. In hindsight, these notions failed because adding two pieces of information satisfying these notions can still reveal significantly more than is acceptable.

In contrast, differential privacy [37], which emerged around the same time, provided a form of privacy quantification that added or *composed* gracefully and was unaffected by any external side information that could be obtained. True to its name, differential privacy quantifies a person's "information" as the difference including or excluding it makes on any downstream outcome (e.g., a query performed on the database). The notion was based on adding carefully calibrated noise to the output of a data-processing mechanism so that any observer would be uncertain about the influence a single data-point might have had.

Definition 1 ((ε, δ)-Differential Privacy [37])**.** Specifically, an algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ is considered (ε, δ)-differentially private (or simply (ε, δ)-DP) if, for every pair of neighboring databases $D, D' \in \mathcal{X}^n$ differing by a single individual's data (denoted as $D \simeq D'$), and for every possible outcome set $S \subset \mathcal{Y}$,

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(D') \in S] + \delta. \tag{1.1}$$

It turned out that DP guarantees *compose* akin to addition; two different sources of information that satisfy DP individually will also satisfy DP collectively, albeit with larger ε, δ values. That is why within two decades since its inception, differential privacy has emerged as the gold standard quantification of data protection.

The success of differential privacy can be partially attributed to two other reasons. Firstly, the (ε, δ)-statistical-indistinguishability property in Equation 1.1 is kind of a quirky constraint on distributional similarity that forms a barrier that freely allows up to ε nats[1] of information to be passed within the outcome $Y \sim M(D)$ about the presence of differing record among $D \simeq D'$. But an outcome Y that carries information beyond ε nats must be *exponentially rare*, proportional to the excess of information, in the order $O(\delta/(1 - e^{-h}))$ where h is the excess[2] (e.g., for an $h \geq 0.1$ and $\delta \leq 10^{-6}$, failure probability is $\leq 10.51 \times \delta \approx 10^{-5}$). This type of quirk is not exhibited by other similarity notions, like Total Variation distance or Kullback-Liebler divergence, making differential privacy a particularly well-suited quantification of privacy.

Secondly, despite concerns that differential privacy might be too restrictive of a constraint for algorithms to be useful, one of the great success of the field is showing that the opposite is true—most tasks achievable without differential privacy guarantees can also be accomplished with it. Problems in this collection include virtually all statistical tasks from computing descriptive statistics (e.g., mean, median, standard deviation) [38, 57, 21] to hypothesis and distribution testing (e.g. determining if samples are distributed in a particular way) [3, 22, 23]. Additionally, with minor privacy-preserving interventions, most machine learning (ML) algorithms can be made differentially private, including support vector machines [86], principal component analysis [42], backpropagation for deep learning [1], boosted trees [65], etc. This comes with only a minor reduction in utility, especially when working with large datasets, which are common in practice today.

1.2 Background and Outline of Contributions

However, one of the challenges in crafting differentially private algorithms is that it involves providing a pen-and-paper proof of its statistical indistinguishability property, which is often challenging. To facilitate this, a *privacy algebra* has been devised, consisting of rules like data-processing inequality [31], rules for composing

[1]The *natural unit of information*, or simply *nat*, is the unit of the information content of an event in terms of its probability of occurring. If event E occurs with probability p, its occurrence is said to convey $\log(1/p)$ nats of information.

[2]See Theorem 29 for the exact statement.

private mechanisms [37, 56], sources of amplification from sub-sampling [103, 8] and shuffling [28, 9], etc., that helps in evaluating the privacy of a wide variety of complex algorithms.

Unfortunately, these basic techniques only go so far; maximizing utility of algorithms under a differential privacy constraint requires a precise analysis that carefully accounts for all the internal features that protect personal information from leaking out through the algorithm's output, which often cannot be achieved using the basic algebra. There has been a recent trend in customizing the analysis specific to an algorithm, especially those in machine learning, with the goal of tightening the analysis as much as possible [1, 8, 5, 44]. New analysis techniques based on a variety of classic probability concepts like moment generating function [1], maximal couplings [8], and characteristic functions [122] have been explored in this regard.

1.2.1 Privacy Analysis in Machine Learning

Present methods for privacy analysis of gradient-based optimization algorithms *rely on composition theorems*. The basic principle of these approaches is to bound the privacy of the final model as a cumulation of privacy bounds on uniformly contributing blocks, such as epochs or gradient steps [1, 44]. For instance, consider the *Noisy Stochastic Gradient Descent* algorithm (aka. Noisy-SGD or DP-SGD) described in the Algorithm 1.

Algorithm 1 Noisy-SGD: Noisy Stochastic Gradient Descent [97]

Input: Dataset $D \in \mathcal{X}^n$, loss function $\ell : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$, learning rate η, noise variance $\hat{\sigma}^2$, sampling probability q, initial parameter vector $\Theta_0 \in \mathbb{R}^d$.

1: **for** $k = 0, 1, \cdots, K-1$ **do**

2: Sample indices $U \sim \mathcal{S}^{\text{po}}([n])$ with probability q

3: Update model $\Theta_{k+1} \leftarrow \Theta_k - \frac{\eta}{n} \sum_{i \in U} \nabla \ell(\Theta_k; D[i]) + \sqrt{2\eta \left(\frac{\hat{\sigma}}{n}\right)^2} \mathcal{N}(0, I_d)$

4: Output Θ_K

By using *Advanced Composition* [40], it is possible, with basic privacy algebra, to certify that K iterations of Algorithm 1 with an $O(1)$-Lipschitz gradient (or with gradient clipping) and Gaussian noise of $O(\hat{\sigma}^2)$ variance, where each training data point is included independently with a probability q (i.e., Poisson sub-samping), satisfies $\left(O\left(\frac{q}{\hat{\sigma}}\sqrt{K \log K} \cdot \log(1/\delta)\right), \delta\right)$-DP. Note that for a fixed δ, the bound on ε

grows at an order $O(\sqrt{K \log K})$ with the number of iterations. This dependency is not ideal as it greatly reduces the number of training iterations permissible under a given (ε, δ)-budget. This $O(\sqrt{K \log K})$ dependence was improved by Abadi et al. [1] to $O(\sqrt{K})$ by noting that gradients are privatized using Gaussian mechanism, which composes slightly better than the worst-case (ε, δ)-DP mechanism. Several later works have focused on improving this dependency [44, 8, 45, 99]. Yet, $O(\sqrt{K})$ dependency on the number of iterations has persisted, which begs the question:

Question 1. *How much can we improve the dependency in the number of iterations?*

We answer this question by showing that the dependency on the number of iterations can be improved to $O(1)$ under certain *isoperimetric* conditions, such as when the loss function is Lipschitz, convex and smooth. This improvement comes from a *privacy amplification effect due to convergence*—under the proper setup, the training of iterative gradient-based algorithms eventually converges to an invariant model distribution. This results in the information content in the model reaching an equilibrium state, where the contributions from subsequent gradients-descent steps are nullified by the gradient noise. To reason about this behavior in information flow, we develop a novel privacy analysis technique based on tracing the discrete-time stochastic process of Noisy-GD's parameter distributions with a continuous-time Langevin diffusion process [88]. Our result is informally stated as follows (with an extension to Noisy-SGD discussed in later sections).

Theorem 1 (Privacy of Noisy-GD)**.** *Let $\ell(\theta; \mathbf{x})$ be any strongly convex, Lipschitz, and β-smooth loss function on a closed convex set $\mathcal{Y}$. Then, for any $\Theta_0 \in \mathcal{Y}$, step-size $\eta < \frac{1}{\beta}$, and sampling probability $q = 1$, Algorithm 1 is $\left(O\left(\frac{1}{\hat{\sigma}}\sqrt{\log(1/\delta)}\right), \delta\right)$-DP.*

For the formal version, see Corollary 1 in Chapter 3.

A privacy analysis is incomplete without a utility analysis because no matter how private an algorithm is, the criteria of it being useful in practice depends on how well it actually performs. So we ask the following question.

Question 2. *What is the privacy-utility trade-off of Noisy-GD algorithm?*

In statistical learning theory, the go-to measure of utility is the *excess empirical risk*, which is defined as the expected difference between the training loss of a model

produced by the ML algorithm and the best model possible for the training dataset. Specifically, the excess empirical risk of a (random) model $\Theta \in \mathbb{R}^d$ on a given dataset $D \in \mathcal{X}^n$ is defined as

$$\operatorname{err}(\Theta; D) \stackrel{\text{def}}{=} \mathbb{E}\left[\mathcal{L}_D(\Theta) - \min_{\theta \in \mathbb{R}^d} \mathcal{L}_D(\theta)\right], \tag{1.2}$$

where the *empirical risk* (aka. training error) is given by

$$\mathcal{L}_D(\theta) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in D} \ell(\theta; \mathbf{x}). \tag{1.3}$$

Showing that Noisy-GD is an optimal private algorithm requires proving that no other algorithm exists that can attain a strictly better excess empirical risk (aside from improving the constant factors) under an (ε, δ)-DP constraint. The following theorem provides a lower-bound on the excess empirical risk that any (ε, δ)-DP algorithm must satisfy (i.e., the inherent cost in utility necessary for satisfying DP).

Theorem 2 (Lower bound[3] for (ε, δ)-DP algorithms [12])**.** *Let $n, d \in \mathbb{N}$, $\delta = o(\frac{1}{n})$, and $\mathcal{X} = \mathcal{Y} = \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$. For every (ε, δ)-DP algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$, there is a dataset $D \in \mathcal{X}^n$ such that for some 1-Lipschitz and 1-strongly convex loss function $\ell(\theta; \mathbf{x})$, we must have*

$$\operatorname{err}(\Theta; D) = \Omega\left(\min\left(1, \frac{d}{\varepsilon^2 n^2}\right)\right). \tag{1.4}$$

Our refined privacy analysis allows us to show an upper-bound on Noisy-GD's excess empirical risk under (ε, δ)-DP that matches the lower bound in Theorem 2.

Theorem 3. *For 1-Lipschitz, $O(1)$-smooth and 1-strongly convex loss function $\ell(\theta; \mathbf{x})$ over a bounded closed convex set $\mathcal{Y} \subset \mathbb{R}^d$ and dataset $D \in \mathcal{X}^n$, by setting the appropriate sampling probability q, step-size η and distribution for initial model Θ_0, Algorithm 1 satisfies (ε, δ)-DP, and simultaneously achieves an excess empirical risk*

$$\operatorname{err}(\Theta_K; D) = O\left(\frac{d \log(1/\delta)}{\varepsilon^2 n^2}\right), \tag{1.5}$$

with noise variance $\hat{\sigma}^2 = O(\frac{\log(1/\delta)}{\varepsilon^2})$ and number of updates $K = O(\log(\frac{n^2\varepsilon^2}{d \log(1/\delta)}))$.

[3] Bassily et al. [13] prove $\Pr_{\Theta \sim \mathcal{A}(D)}\left[\mathcal{L}_D(\Theta) - \min_{\theta \in \mathbb{R}^d} \mathcal{L}_D(\theta) \geq \Omega(\min(1, \frac{d}{\varepsilon^2 n^2}))\right] \geq \frac{1}{3}$, which implies the stated theorem.

Refer to Theorem 17 in Chapter 3 for the formal version.

Together, these three theorems provide the complete characterization of privacy-utility trade-off of Noisy-GD (for strongly convex, smooth and Lipschitz loss functions) and answers the open question raised by Bassily et al. [12] on finding a tight upper-bound for the excess-risk of (ε, δ)-differentially private ERM. Additionally, our analysis shows that optimal utility can be achieved with only $O(n \log n)$ gradient computations, where n is the dataset size. In comparison, most other works on private ERM under convexity could only show near-optimal excess risk with $O(n^2)$ gradient computations [12, 13].

1.2.2 Reasoning about Data Deletion in Machine Learning

Another type of privacy that is becoming increasingly relevant is *deletion privacy*. The need for deletion arises from an evolved understanding that data privacy requires a strong commitment to self-determinism, which is the power of individuals to decide if, to what extent, and how their personal information is collected, used, and shared [15, 83, 32]. Modern day privacy policies, such as General Data Protection Regulation (GDPR) in the European Union and California Consumer Protection Act (CCPA) in the United States, enforces the principle of self-determination by granting individuals the right to *rectify or erase their data*. Since retraining fresh ML models from scratch every time the dataset changes is computationally expensive, there is a growing interest in designing cheap *machine unlearning* algorithms for erasing the influence of deleted data from already trained models. This thesis also delves into Noisy-GD as a machine unlearning algorithm, expanding on the theory of its information flow.

To quantify how well an unlearning algorithm erases the requested information in the worst-case, Ginart et al. [46] propose a DP like (ε, δ)-indistinguishability between the unlearning algorithm's output and that of fresh retraining.

Definition 2 ((ε, δ)-unlearning)**.** For a fixed dataset $D \in \mathcal{X}^*$, remove set $u \subset D$, and a randomized learning algorithm $\mathcal{A}$, an unlearning algorithm $\bar{\mathcal{A}}$ is (ε, δ)-unlearning with respect to $(D, u, \mathcal{A})$ if for all $S \subset \mathcal{Y}$ where $\mathcal{Y}$ denotes the output space, we

(a) (b) (c)

Figure 1.3: A real-life example where even retraining does not ensure total deletion. (a) An original artwork titled "Night scene with Kara" by Sam Yang. (b) Stylistic copy generated using the SamDoesArt-V2 model on Huggingface, trained on Sam Yang's artworks. (c) Stylistic copy produced by a second-generation model trained only on synthetic images imitating Sam Yang's art style.

have:

$$\Pr\left[\bar{\mathcal{A}}(D,u,\mathcal{A}(D)) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\mathcal{A}(D \setminus u) \in S\right] + \delta, \quad \text{and}$$
$$\Pr\left[\mathcal{A}(D \setminus u) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\bar{\mathcal{A}}(D,u,\mathcal{A}(D)) \in S\right] + \delta.$$

The idea behind this definition is that if there is no way to distinguish an unlearned model from a fresh model trained without the deleted data-points, then the unlearned model must not contain any information about the deleted records anymore. Several unlearning certifications, that follow the same intuition, have since been proposed and used to certify numerous unlearning algorithms that belong to the family of iterative first-order gradient methods [50, 55, 72, 105]. Before we analyze Noisy-GD's unlearning capabilities, we must ask a very fundamental question.

Question 3. *Is indistinguishability to retraining a reliable guarantee of deletion?*

We respond negatively to this question and provide three technical reasons for our stance (with two on soundness and one on completeness of this notion as a measure deletion). First, deletion requests in the real world are *reactive* (aka. *adaptive*); individuals typically seek to have their data erased in response to specific events triggered by a model that has been trained on their data. For example, in a recent controversy, a Redditor used the Stability AI platform to train a stable-diffusion model to replicate the art style of a renowned canadian artist. This triggered the

artist to file a class-action lawsuit against Stability AI and Midjourney for violating copyright infringement [67]. Unfortunately, this reactive nature of the real world complicates efforts to censor information and can potentially backfire, resulting in the *opposite* effect[4]. A similar situation occurred with the artist when his actions prompted many others to create their own imitation models to replicate his style and post them out of spite [84]. The internet now hosts numerous stylistic copies of the artist's signature style. Even when retrained from scratch without his original works, a second-hand diffusion model can still copy his style, as shown in Figure 1.3. This indicates that the artist's information doesn't get removed completely even on retraining. We prove this vulnerability formally in Theorem 18 of Chapter 4, showing that under adaptive deletion requests, deleted data can still be recovered from the output after unlearning, even when the algorithm used satisfies a $(0, 0)$-unlearning guarantee, i.e. perfect indistinguishability!

Secondly, approximate indistinguishability from retraining ensures that removed data cannot be accurately recovered from a *single* unlearned model. However, a robust guarantee of deletion privacy should prevent deleted records from being recovered by an observer with access to *multiple* (potentially all) model versions after the deletion request. We demonstrate that algorithms satisfying existing unlearning guarantees can create models that are indistinguishable from retraining beyond a very small threshold, yet still expose deleted data over several releases. This vulnerability arises from algorithms that retain partial computations as internal states to speed up future deletions (e.g., if we apply Principal Component Analysis in the initial dataset for dimensionality reduction but don't update these principal components when data-points get deleted). These internal states can carry information about previously deleted records, impacting several subsequent releases and potentially revealing the deleted data in the long run. Our Theorem 19 in Chapter 4 formalizes this problem of maintaining such partial computations.

Finally, (ε, δ)-unlearning in Definition 2 disregards perfectly valid deletion algorithms. For instance, an algorithm $\bar{\mathcal{A}}$ that outputs a fixed untrained model in response to any deletion request u is a valid deletion algorithm because its output

[4]This phenomenon, known as the *Streisand effect*, is named after Barbara Streisand's attempt to censor a photo of her Malibu residence, taken to document coastal erosion. Before her lawsuit, the photo had only six downloads, two by Streisand's attorneys. However, after the lawsuit gained attention, the site got over 420,000 views in the following month [2].

contains no information about the deleted records in u. However, since its output is easy to tell apart from that of retraining done by any sensible learning algorithm $\mathcal{A}$, the algorithm $\bar{\mathcal{A}}$ cannot satisfy Definition 2. Discussions in § 4.2.3 further elaborates on this incompleteness issue.

To address these issues, this thesis introduces a new definition of deletion privacy, accompanied by a rigorous proof demonstrating the *soundness* of its certification.

Definition 3 ((ε, δ)-deletion privacy)**.** We say that an algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies (ε, δ)-deletion privacy if for all datasets D and all remove sets $u \subset D$ (potentially chosen reactively after seeing $\mathcal{A}(D)$), we have that for all deleted records $\mathbf{x} \in u$ there exists a random variable $\Theta_{\mathbf{x}}$ that is independent of the data-point $\mathbf{x}$ such that

$$\forall S \in \mathcal{Y}, \quad \Pr\left[\bar{\mathcal{A}}(D, u, \mathcal{A}(D)) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\Theta_{\mathbf{x}} \in S\right] + \delta. \tag{1.6}$$

This definition packs multiple subtle, but crucial, features that go a long way towards a reliable guarantee on data deletion, which is discussed in detail in § 4.3. Most notably, Definition 3 patches the flaws of (ε, δ)-unlearning in Definition 2 while being less restrictive in many ways at the same time. For instance, Definition 3 requires indistinguishability in only one direction and allows the freedom to construct an appropriate $\Theta_{\mathbf{x}}$ rather than restricting it to the output of the learning algorithm $\mathcal{A}(D \setminus u)$.

Question 4. *How is deletion privacy different from differential privacy?*

In literature, there is a considerable confusion about the relationship between unlearning and differential privacy. Works such as Sekhari et al. [92] and Huang and Canonne [54] show that DP learning algorithms provide *some* deletion guarantees out-of-the-box, even when the unlearning algorithm is a no-op mechanism that returns the model as is. In this thesis, we emphasize that deletion privacy and differential privacy are distinct concepts that protect different aspects of data. Differential privacy limits information about all data points, regardless of whether they are still in the dataset or deleted. Deletion privacy, however, only limits information about data points that have been deleted. Therefore, treating differential privacy as a quantifier of deletion privacy is basically saying that there is no need to erase information because it was never sufficiently learned in the first place, which is of

little interest. This thesis argues that a machine unlearning algorithm should ensure both: an $(\varepsilon_{\text{dp}}, \delta)$-differential privacy guarantee for existing data points and a much smaller $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy guarantee for removed data points. Together, these two types of protection exhibit a powerful synergy as the same noise needed to ensure differential privacy also helps with deletion privacy at no extra cost.

Our results and discussions in Section § 4.4 formalizes the connections between Differential and Deletion privacy in detail.

1.2.3 Trade-offs: Privacy, Utility, Deletion and Compute

This thesis provides an in-depth study of Noisy-GD, not only as a private machine learning algorithm, but also as an algorithm for data deletion in the streaming setting introduced by Neel et al. [72]. In this setting, a stream of *edit requests*, comprised of a batch of addition, deletion or replacement operations, arrives sequentially. The *data curator*, characterized by a pair of learning and unlearning algorithms $(\mathcal{A}, \bar{\mathcal{A}})$, executes the learning algorithm $\mathcal{A}$ on the initial dataset D_0 during the setup stage before the arrival of the first edit request to generate the initial model $\hat{\Theta}_0 = \mathcal{A}(D_0)$. Thereafter, at any edit step $i \geq 1$, to reflect an edit request u_i that transforms the dataset $D_{i-1} \circ u_i \to D_i$, the curator runs the unlearning algorithm $\bar{\mathcal{A}}$ to get the next model $\hat{\Theta}_i = \bar{\mathcal{A}}(\hat{\Theta}_{i-1}, u_i, D_{i-1})$. This setting is visualized in Figure 1.4.

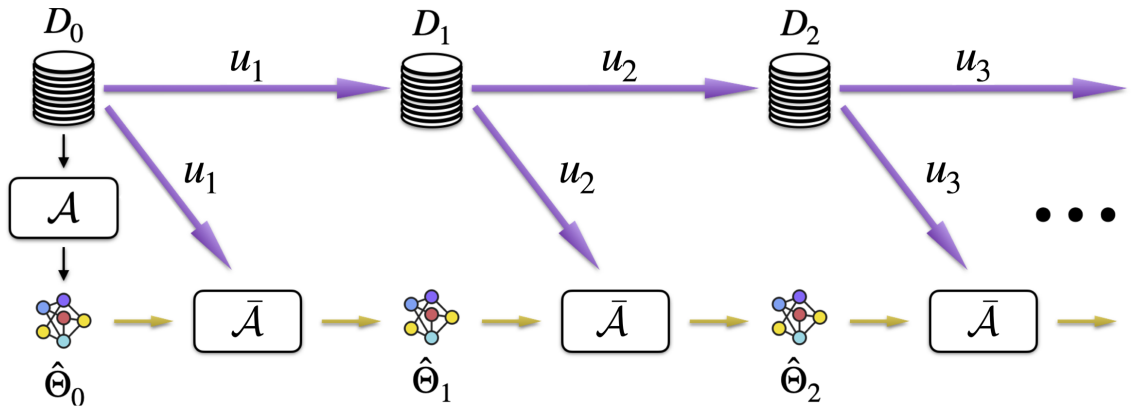


Figure 1.4: Our setup demonstrating the initial learning step followed by a sequence of unlearning steps in response to edit (add, remove or replace) requests on the dataset.

We consider Noisy-GD as the learning algorithm $\mathcal{A}_{\text{Noisy-GD}}$, where for any $D \in \mathcal{X}^n$,

$$\mathcal{A}_{\text{Noisy-GD}}(D) = \text{Noisy-GD}(D, \Theta_0, K_{\mathcal{A}}),$$

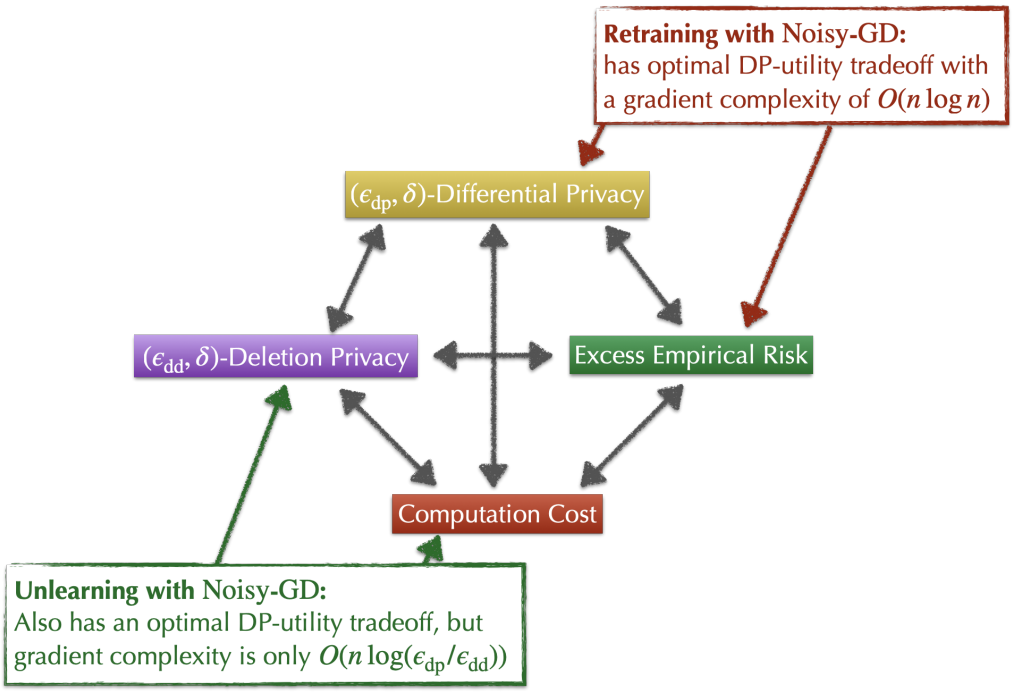


Figure 1.5: Visualizing the four-way trade-off between privacy, utility, deletion and computation complexity and the benefit of unlearning with Noisy-GD instead of retraining with Noisy-GD in the setting of streaming edit requests.

with the initial model Θ_0 randomly sampled from a Gaussian distribution. And, for any edit request u on a database D and any model $\Theta \in \mathcal{Y}$, our unlearning algorithm $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ is also based on Noisy-GD as follows:

$$\bar{\mathcal{A}}_{\text{Noisy-GD}}(\Theta, S, D) = \text{Noisy-GD}(D \circ u, \Theta, K_{\bar{\mathcal{A}}}).$$

Question 5. *What kind of trade-off between the number of unlearning iterations $K_{\bar{\mathcal{A}}}$ and $(\varepsilon_{\text{dd}}, \delta)$-deletion guarantee can the algorithm pair $(\mathcal{A}_{Noisy\text{-}GD}, \bar{\mathcal{A}}_{Noisy\text{-}GD})$ achieve when*

1. *both $\mathcal{A}_{Noisy\text{-}GD}$ and $\bar{\mathcal{A}}_{Noisy\text{-}GD}$ must satisfy $(\varepsilon_{\text{dp}}, \delta)$-DP, and*
2. *for all $D_0 \in \mathcal{X}^n$ and any sequence of edit requests, $\text{err}(\hat{\Theta}_i; D_i) \leq \alpha$ for $i \geq 1$?*

By building on our analysis of information flow in Noisy-GD, we show a substantial computational savings on unlearning using $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ instead of fresh retraining using $\mathcal{A}_{\text{Noisy-GD}}$. For strongly-convex, Lipschitz and smooth losses, we certify that under $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy and $(\varepsilon_{\text{dp}}, \delta)$-differential privacy constraints, we can achieve learning and unlearning of models in $\mathbb{R}^d$ on n-sized datasets such that the excess

empirical risk $\text{err}(\hat{\Theta}_i; D_i) = O\left(\frac{d\log(1/\delta)}{\varepsilon_{\text{dp}}^2 n^2}\right)$ for all $i > 1$, while requiring only $K_{\bar{\mathcal{A}}} = O\left(\log \frac{\varepsilon_{\text{dp}}}{\varepsilon_{\text{dd}}}\right)$ number of unlearning iterations. Note from Theorem 2 that this DP-utility tradeoff is optimal across all unlearning steps, but the computation of $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ in terms of number of iterations could be considerably smaller than $(\log(n^2{\varepsilon_{\text{dp}}}^2/d))$ required for retraining using $\mathcal{A}_{\text{Noisy-GD}}$ (cf. Theorem 3). This computational benefit on unlearning using Noisy-GD for convex losses is described as Theorem 24 in § 4.5.2 and summarized in Figure 1.5.

In § 4.5.3, we also present an analysis of Noisy-GD's unlearning capabilities for non-convex loss functions under L_2-regularization. For the purpose, we develop and improve results that show rapid convergence of Noisy-GD algorithm's output distribution $\mathcal{A}_{\text{Noisy-GD}}(D)$ to something that is very close to the Gibbs distribution $\pi(D) \propto e^{-\mathcal{L}/\sigma^2}$. We study this convergence in terms of *Rényi divergence*, a notion that is strictly stronger than (ε, δ)-indistinguishability used in differential privacy (called (q, ρ)-Rényi differential privacy [69]). By building on these convergence results, we show that not only can Noisy-GD provide (q, ρ_{dp})-Rényi differential privacy and (q, ρ_{dd})-Rényi data deletion guarantees under non-convexity with L_2 regularization (with the additional constraint that the loss function is bounded), but it also achieves *near-optimal empirical risk* of $\text{err}(\hat{\Theta}_i; D_i) = \tilde{O}\left(\frac{dq}{\rho_{\text{dp}} n^2} + \frac{1}{n}\sqrt{\frac{q\rho_{\text{dd}}}{\rho_{\text{dp}}}}\right)$. See Theorem 27 in Section § 4.5.3 for the formal statement.

1.2.4 Laplace transform interpretation of Privacy

Being a statistical property, the design of differentially private algorithms involves a *pen-and-paper* analysis of any randomness internal to the processing of the algorithm that obscures the influence a data-point might have. Therefore, an important direction in differential privacy research is to create adaptable analytical tools that helps in tailored privacy analysis of any given algorithm. That is why, throughout its development, various interpretations of the differential privacy have emerged. These include the privacy-profile curve $\delta(\epsilon)$ [10] that traces the (ϵ, δ)-DP point guarantees, the f-DP [34] view of worst-case *trade-off curve* between type I and type II errors for hypothesis testing membership [56, 11], the Rényi DP [69] function of order q that admits a natural analytical composition [1, 69], the view of the *privacy loss distribution (PLD)* [96] that allows for approximate numerical composition [59, 47],

and the recent *characteristic function formulation* of the dominating privacy loss random variables Zhu et al. [122]. Each of these formalisms have their own properties and use-cases, and none of them seem to be superior in all aspects.

In this thesis, we propose a new interpretation of differential privacy that not only augments existing interpretations but also provides a flexible analytical toolkit that greatly extends our cognitive reach in reasoning about differential privacy and its properties. This interpretation is based on recognizing that the privacy profile curve $(\varepsilon, \delta(\varepsilon))$-DP can be seen as a Laplace transform of the privacy loss distribution[5].

To build the basic intuition of this expression, suppose P and Q represent the probability mass/density function of an algorithm $\mathcal{A}$'s output distribution on neighboring databases $D \simeq D'$ respectively. We can express δ as a function of ε as

$$\delta(\varepsilon) \stackrel{\text{def}}{=} \sup_{S \subset \mathcal{Y}} P(S) - e^{\varepsilon} Q(S) = \int_{\theta \in \mathcal{Y}} \left[1 - e^{\varepsilon} \frac{Q(\theta)}{P(\theta)}\right]_+ P(\theta) \mathrm{d}\theta,$$

where $[\bullet]_+ \stackrel{\text{def}}{=} \max\{0, \bullet\}$, because the supremum set contains only those $\theta \in \mathcal{Y}$ for which $P(\theta) \geq e^{\varepsilon} Q(\theta)$. Let $S_{\geq \varepsilon} \stackrel{\text{def}}{=} \{\theta \in \mathcal{Y} : P(\theta) \geq e^{\varepsilon} Q(\theta)\}$. We can further express this in terms of the privacy loss $L_{P|Q}(\theta) \stackrel{\text{def}}{=} \log(P(\theta)/Q(\theta))$ as

$$\delta(\varepsilon) = \int_{S_{\geq \varepsilon}} \left(1 - e^{\varepsilon - L_{P|Q}(\theta)}\right) P(\theta) \mathrm{d}\theta = \int_{\varepsilon}^{\infty} (1 - e^{\varepsilon - z}) f_Z(z) \mathrm{d}z, \tag{1.7}$$

where $f_Z(z) \stackrel{\text{def}}{=} \int_{L_{P|Q}(\theta) = z} P(\theta) \mathrm{d}\theta$ is the probability density of the privacy loss random variable $Z = L_{P|Q}(\Theta)$ with $\Theta \sim P$. Equation 1.7 is a popular way to express the $(\varepsilon, \delta(\varepsilon))$-DP curve in literature [24, Lemma 9], [99, Proposition 7], and [7, Theorem 5]. In this thesis, we highlight that this expression also has a Laplace transform form as follows:

$$\begin{aligned}
\forall \varepsilon \in \mathbb{R} \ : \ \delta(\varepsilon) &= \int_{\varepsilon}^{\infty} (1 - e^{\varepsilon - z}) f_Z(z) \mathrm{d}z \\
&= 1 - F_Z(\varepsilon) - \int_0^{\infty} e^t f_Z(t + \varepsilon) \mathrm{d}t && \text{(Substitute } z = t + \varepsilon\text{)} \\
&= \int_0^{\infty} e^{-t} (1 - F_Z(t + \varepsilon)) \mathrm{d}t && \text{(Integration by parts)} \\
&= \mathcal{L}\{1 - F_Z(t + \varepsilon)\}(1),
\end{aligned}$$

[5] Laplace transform maps time-domain function $g(t)$ to frequency domain function $\mathcal{L}\{g\}(s) \stackrel{\text{def}}{=} \int_0^{\infty} e^{-st} g(t) \mathrm{d}t$. Complicated operations in the time-domain like convolution or differentiation reduces to simple operations like multiplication and division in the frequency domain.

where $F_Z(t) = \Pr[Z \leq t]$ is the cumulative distribution function of the privacy loss random variable Z.

The Rényi-divergence $\mathrm{R}_q\left(P\|Q\right) := \frac{1}{q-1}\int P^q Q^{1-q}\mathrm{d}\theta$ between any two distributions P, Q on the same space is also expressible as a bilateral Laplace transform of the privacy loss random variable Z:

$$\forall q \in \mathbb{C} \ : \ e^{(q-1)\cdot \mathrm{R}_q(P\|Q)} = \mathop{\mathbb{E}}_{Z\leftarrow \mathrm{PLD}(P\|Q)}\left[e^{(q-1)\cdot Z}\right] = \mathcal{B}\left\{f_Z(t)\right\}(1-q).$$

Using them, we show that the privacy-profile and Rényi divergence between any two distributions have the following equivalence.

$$\forall q \in \mathbb{C} \ : \ e^{(q-1)\cdot \mathrm{R}_q(P\|Q)} = q(q-1)\cdot \mathcal{B}\left\{\delta_{P|Q}(t)\right\}(1-q), \tag{1.8}$$

which again is a Laplace transform expression. Furthermore, Zhu et al. [122]'s characteristic function of the privacy loss $\phi_{P|Q}(q) := \mathop{\mathbb{E}}_{P}\left[e^{iq\log(P/Q)}\right]$ also turns out to be a Fourier transform, which is a special case of the bilateral Laplace transform:

$$\forall q \in \mathbb{R} \ : \ \phi_{P|Q}(q) = \mathop{\mathbb{E}}_{Z\leftarrow \mathrm{PLD}(P\|Q)}\left[e^{iqZ}\right] = \mathcal{B}\left\{f_Z\right\}(-iq). \tag{1.9}$$

Question 6. *How is Laplace transform interpretation of Differential Privacy useful?*

Laplace-transform view of the DP curve unlocks a formal approach to perform a wide-variety of manipulations and transformations using the properties of Laplace transforms (see Table C.1 in Appendix Appendix C), some of which are often seen being performed in literature in complicated and explicit ways within proofs related to DP (e.g., proofs of [10, Theorem 8], [24, Proposition 12] and [70, Lemma 10]). With the Laplace transform expression, performing these complex manipulations become as easy as invoking one of its many time-frequency duality properties. This not only simplifies many proofs, but also provides a powerful tool to prove new properties of significance. To showcase the impact of this interpretation, we prove an *exact composition theorem* using it that applies to the entire $(\varepsilon, \delta(\varepsilon))$-DP curves rather than point DP-guarantees.

Theorem 4. *If algorithms $\mathcal{A}_1 : \mathcal{X}^n \to \mathcal{Y}_1$ and $\mathcal{A}_2 : \mathcal{Y}_1 \times \mathcal{X}^n \to \mathcal{Y}_2$ tightly satisfy $(\varepsilon, \delta_1(\varepsilon))$-DP and $(\varepsilon, \delta_2(\varepsilon))$-DP for all $\varepsilon \in \mathbb{R}$, then the* composed *mechanism defined as $\mathcal{A}(D) \overset{def}{=} (X, Y)$ where $X \sim \mathcal{A}_1(D)$ and $Y \sim \mathcal{A}_2(X, D)$ satisfies $(\varepsilon, \delta(\varepsilon))$-DP with*

$$\delta(\varepsilon) \leq \left(\delta_1 \circledast \left(\ddot{\delta}_2 - \dot{\delta}_2\right)\right)(\varepsilon) = \int_{-\infty}^{\infty} \delta_1(\varepsilon - \tau)\cdot\left(\ddot{\delta}_2(\tau) - \dot{\delta}_2(\tau)\right)\mathrm{d}\tau, \tag{1.10}$$

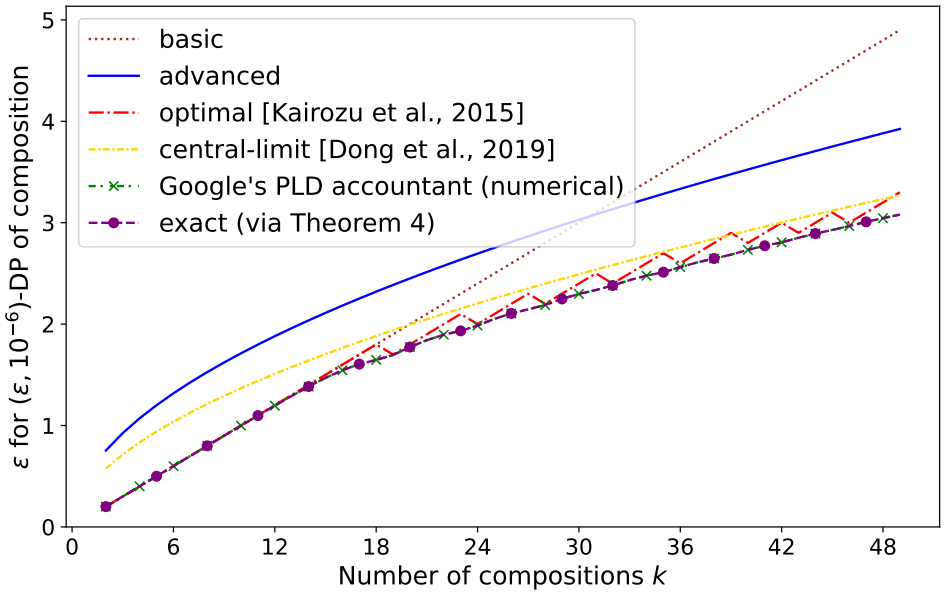


Figure 1.6: Comparison between $(\varepsilon, 10^{-6})$-DP bounds obtained by different composition theorems for k independent Laplace mechanisms where each exactly satisfies 0.1-DP. To obtain the exact DP bound, we recursively alternate between applying Theorem 4 and differentiating subsequent recursion as follows: $\delta_{1:k} = \delta_{1:k-1} \circledast (\ddot{\delta} - \dot{\delta})$, where the privacy profile for 0.1-DP Laplace mechanism is $\delta(\varepsilon) = \max\{0, 1 - e^{\frac{\varepsilon - 0.1}{2}}, 1 - e^{\varepsilon}\}$.

where $\dot{\delta}_2(\varepsilon)$ *and* $\ddot{\delta}_2(\varepsilon)$ *are the first and second order gradients of* $\delta_2(\varepsilon)$ *respectively.*

Theorem 4 is a *tight* composition theorem for $(\varepsilon, \delta(\varepsilon))$-DP curves. In comparison, existing composition guarantees like Advance composition theorem by Dwork and Lei [38] and optimal composition theorem by Kairouz et al. [56] apply to *point* (ε, δ)-DP guarantees only, and can be significantly lossy. Figure 1.6 compares Theorem 4 with other composition bounds in literature.

Additionally, we show a number of results of interest using this Laplace transform interpretation:

1. We note that the Laplace transform expression of Rényi divergence permits the order q to be a complex number in $\mathbb{C}$. In fact, Rényi divergence for complex orders plays a crucial role in interconversion between Rényi DP and (ε, δ)-DP—the privacy profile $\delta_{P|Q}(\varepsilon)$ is the inverse Laplace transform of Rényi

divergence $\mathrm{R}_q\left(P\|Q\right)$:

$$\delta_{P|Q}(\varepsilon)=\mathcal{L}^{-1}\left\{\frac{e^{-s\mathrm{R}_{1-s}(P\|Q)}}{s(s-1)}\right\}(\varepsilon)=\frac{1}{2\pi i}\lim_{\omega\to\infty}\int_{\gamma-i\omega}^{\gamma+i\omega}e^{s\varepsilon}\cdot\frac{e^{-s\mathrm{R}_{1-s}(P\|Q)}}{s(s-1)}ds,$$

where γ could be *any real number in* $\mathbb{R}\setminus\{0,1\}$ under absolute continuity (i.e., $P\ll Q$ and $Q\ll P$).

2. We establish an unexpected result that, although Rényi divergence $\mathrm{R}_q\left(P\|Q\right)$ and privacy profile $\delta_{P|Q}(\varepsilon)$ are equivalent functional representations under absolute continuity, they do not have a compatible notion of *dominance*. More precisely, for two pairs of distributions (P_1,Q_1) and (P_2,Q_2), while the relationship $\delta_{P_1|Q_1}(\varepsilon)\leq\delta_{P_2|Q_2}(\varepsilon)$ for all ε implies that $\mathrm{R}_q\left(P_1\|Q_1\right)\leq\mathrm{R}_q\left(P_2\|Q_2\right)$ for all $q>1$, the *converse does not hold*. This is because being inverse Laplace transforms of $\mathrm{R}_q\left(P_1\|Q_1\right)$ and $\mathrm{R}_q\left(P_2\|Q_2\right)$, the privacy profiles $\delta_{P_1|Q_1}(\varepsilon)$ and $\delta_{P_2|Q_2}(\varepsilon)$ have a dominance relationship that depends on how the Rényi divergences behave along the complex line $\{q\in\mathbb{C}:\mathfrak{Re}(q)=\gamma\}$ at any $\gamma\in\mathbb{R}\setminus\{0,1\}$; not along $\{q\in\mathbb{R}:q>1\}$ that all prior works focus on [11, 122, 5, 24].

3. We provide an *exactly-tight and adaptive theorem for composing any two privacy profiles*, $\delta_{P_1|Q_1}(\varepsilon)$ and $\delta_{P_2|Q_2}(\varepsilon)$, that mirrors the lossless composition of Rényi divergence. For product distributions $P=P_1\times P_2$ and $Q=Q_1\times Q_2$, the composed privacy profile $\delta_{P|Q}$ is

$$\forall\varepsilon\in\mathbb{R}:\delta_{P|Q}(\varepsilon)=\left(\delta_{P_1|Q_1}\circledast\left(\ddot{\delta}_{P_2|Q_2}-\dot{\delta}_{P_2|Q_2}\right)\right)(\varepsilon),\tag{1.11}$$

where $\dot{\delta}_{P_2|Q_2}$ and $\ddot{\delta}_{P_2|Q_2}$ are the first and second order gradient functions of $\delta_{P_2|Q_2}$ and '$\circledast$' denotes the convolution operation. Our composition also applies under adaptivity.

4. We apply our composition theorem for privacy profiles (i.e., $(\varepsilon,\delta_i(\varepsilon))$-DP curves) to derive an optimal composition theorem for (ε_i,δ_i)-DP point guarantees. Our bound surpasses the optimal bound by Kairouz et al. [56, Theorem 3.3], as their result gives only a discrete set of (ϵ,δ) pairs, while ours provides a continuous curve, enabling the tightest ϵ for any δ budget.

5. Working with functional notions have a symmetry problem—in many cases $\delta_{P|Q}(\varepsilon) \neq \delta_{Q|P}(\varepsilon)$ or $\mathrm{PLD}(P\|Q) \neq \mathrm{PLD}(Q\|P)$. Existing works handle asymetry either by maintaining functional representation in both directions, or by enforcing symmetry at the cost of overestimating the curve. We highlight a *reversal property enjoyed by all function notions of DP* and use it to eliminate the asymmetry problem.

1.3 Outline of Thesis

Chapter 2 covers the preliminaries on loss function properties, on relevant divergence notions (including differential privacy notions), and on the theory of diffusion processes.

In Chapter 3 we study the DP dynamics of diffusion processes and extend it to Noisy-GD algorithm. We also explore the privacy-utility trade-off for Noisy-GD.

Chapter 4 introduces the concept of data deletion, identifies flaws in prior attemts to quanitify it and presents corrections to fix those flaws. This chapter also explores the data deletion properties of Noisy-GD algorithm in an online setting of streaming edit (i.e., insert, delete, or modify) requests.

Chapter 5 introduces a frequency-domain interpretation of differential privacy and presents results on composition, conversion and subsampling for functional notions of differential privacy.

Finally, Chapter 6 completes this thesis with a conclusion.

1.4 Finished Works

The research presented in this thesis is primarily based on the following first-authored publications. The * represents (joint) first authorship where applicable.

- Rishav Chourasia*, Jiayuan Ye*, and Reza Shokri. "Differential privacy dynamics of langevin diffusion and noisy gradient descent." Advances in Neural Information Processing Systems 34 (2021).

- Rishav Chourasia*, and Neil Shah. “Forget unlearning: Towards true data-deletion in machine learning.” International Conference on Machine Learning. PMLR, 2023.
- Rishav Chourasia*, Uzair Javaid, Biplap Sikdar. “Laplace Transform Interpretations of Differential Privacy.” **Currently under review at FORC 2025.**

The following publications were also made during this Ph.D. journey:

- Rishav Chourasia, Batnyam Enkhtaivan, Kunihiro Ito, Junki Mori, Isamu Teranishi, and Hikaru Tsuchida. "Knowledge Cross-Distillation for Membership Privacy." Proceedings on Privacy Enhancing Technologies 2022(2).
- Rishav Chourasia*, and Adish Singla. "Unifying ensemble methods for q-learning via social choice theory." Learning with Rich Experience (LIRE) Workshop, NeurIPS 2019.

Chapter 2

Preliminaries

In this chapter we introduce the necessary preliminaries, definitions and notations that will be useful throughout the thesis. Specifically, we introduce various relevant *divergence measures for probability distribution*, *differential privacy* and their properties in § 2.1. Then, in § 2.2 we formalize the statistical problem of *empirical risk minimization* (ERM), provide a refresher on calculus covering the notations and provide definitions of some common assumptions on loss functions made to study the ERM problem. In § 2.3 we introduce *Markov diffusion semigroups* in context of modeling gradient-descent based optimization algorithms and discuss *isoperimetric properties* that make analysis easier. Finally, in § 2.3.2 we provide a brief introduction to Laplace transforms and their properties.

2.1 Divergence Measures and Their Properties

Let P, Q be two probability measures on a Lebesgue space $\mathcal{Y}$. We abuse the notations to denote respective probability densities with P, Q as well. We say that measure P is absolutely continuous with respect to Q (denoted by $P \ll Q$) if for all measurable sets $S \subset \mathcal{Y}$, $P(S) = 0$ whenever $Q(S) = 0$.

Following are some divergence notions and some of their useful properties.

Definition 4 (Hockey-stick divergence [87])**.** The *hockey-stick divergence* of P w.r.t.

Q for any $\gamma > 0$[1] is defined as

$$\mathrm{H}_\gamma\left(P\|Q\right) \stackrel{\text{def}}{=} \sup_{S\subset\mathcal{Y}} P(S) - \gamma\cdot Q(S) = \mathop{\mathbb{E}}_{\Theta\sim P}\left[1-\gamma\frac{P(\Theta)}{Q(\Theta)}\right]_+. \tag{2.1}$$

Definition 5 (Rényi divergence [80])**.** *Rényi divergence* of P w.r.t. Q of order $q > 1$ is defined as

$$\mathrm{R}_q\left(P\|Q\right) \stackrel{\text{def}}{=} \frac{1}{q-1}\log \mathrm{E}_q\left(P\|Q\right), \quad \text{where} \quad \mathrm{E}_q\left(P\|Q\right) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\Theta\sim Q}\left[\left(\frac{P(\Theta)}{Q(\Theta)}\right)^q\right], \tag{2.2}$$

when $P \ll Q$. If $P \not\ll Q$, we'll say $\mathrm{R}_q\left(P\|Q\right) = \infty$.

A bound on Rényi divergence implies a bound on hockey-stick divergence.

Theorem 5 ([69, Proposition 3])**.** *Let $q > 1$ and $\varepsilon > 0$. If distributions P, Q satisfy $\mathrm{R}_q\left(P\|Q\right) < \rho$, then for any $\delta \in (0, 1]$,*

$$P(S) \le \exp\left(\rho + \frac{\log(1/\delta)}{q-1}\right)\cdot Q(S) + \delta, \tag{2.3}$$

for all $S \in \mathcal{Y}$. Equivalently, if $\mathrm{R}_q\left(P\|Q\right) < \rho$, then for any $\varepsilon > \rho$,

$$\mathrm{H}_{e^\varepsilon}\left(P\|Q\right) \le e^{(q-1)(\varepsilon-\rho)}. \tag{2.4}$$

Both Rényi divergence and Hockey stick divergences are monotonic functions.

Theorem 6 (Monotonicity of Rényi divergence [69, Proposition 9])**.** *For any $q' > q > 1$, and any two distributions P and Q, $\mathrm{R}_q\left(P\|Q\right) \le \mathrm{R}_{q'}\left(P\|Q\right)$.*

Theorem 7 (Monotonicity of Hockey-stick divergence)**.** *For any $\varepsilon < \varepsilon'$, and arbitrary probability measures P and Q, $\mathrm{H}_{e^{\varepsilon'}}\left(P\|Q\right) \le \mathrm{H}_{e^\varepsilon}\left(P\|Q\right)$.*

Proof. If $\varepsilon < \varepsilon'$, it holds that $[1 - e^\varepsilon\cdot x]_+ \ge [1 - e^{\varepsilon'}\cdot x]_+$ for all $x > 0$. Therefore,

$$\mathrm{H}_{e^\varepsilon}\left(P\|Q\right) = \int_{\theta\in\mathcal{Y}}\left[1-e^\varepsilon\frac{P(\theta)}{Q(\theta)}\right]_+ P(\theta)\mathrm{d}\theta \ge \int_{\theta\in\mathcal{Y}}\left[1-e^{\varepsilon'}\frac{P(\theta)}{Q(\theta)}\right]_+ P(\theta)\mathrm{d}\theta = \mathrm{H}_{e^{\varepsilon'}}\left(P\|Q\right).$$

□

Rényi divergence satisfies the following triangle inequality.

[1]Original definition in [87] only considers $\gamma \ge 1$ which has become the norm in literature. We instead consider $\gamma > 0$ because it is both economical and efficient as $\gamma \in (0, 1]$ yields the hockey-stick divergence with P and Q switched: $\mathrm{H}_\gamma\left(Q\|P\right) = \sup_{S\subset\mathcal{Y}} Q(S) - \gamma\cdot P(S) = 1 - \gamma + \gamma[\sup_{S'\subset\mathcal{Y}} P(S') - \frac{1}{\gamma}\cdot Q(S')] = 1 - \gamma + \gamma\mathrm{H}_{\frac{1}{\gamma}}\left(P\|Q\right)$.

Theorem 8 (Weak triangle inequality [69, Proposition 12])**.** *For any distributions P, M, Q on $\mathcal{Y}$ such that $P \ll M \ll Q$, the Rényi divergence of P w.r.t. Q satisfies the following weak triangle inequality. For all $1 < q < q'$,*

$$\mathrm{R}_q\left(P\|Q\right) \leq \frac{q'}{q-1}\mathrm{R}_{q\cdot\frac{q'-1}{q'-q}}\left(P\|M\right) + \mathrm{R}_{q'}\left(M\|R\right). \tag{2.5}$$

In particular, $\mathrm{R}_q\left(P\|Q\right) \leq \mathrm{R}_q\left(P\|M\right) + \mathrm{R}_\infty\left(M\|R\right)$. Also, if $\mathrm{R}_q\left(P\|M\right) \leq \kappa_1 \cdot q$ and $\mathrm{R}_q\left(M\|Q\right) \leq \kappa_2 \cdot q$ for all $q \in (1,\infty)$, then $\mathrm{R}_q\left(P\|Q\right) \leq (\sqrt{\kappa_1} + \sqrt{\kappa_2})^2 \cdot q$.

Another popular information divergence is the Kullback-Leibler divergence.

Definition 6 (Kullback-Leibler divergence [60])**.** *Kullback-Leibler* (KL) divergence $\mathrm{KL}\left(P\|Q\right)$ of P w.r.t. Q is defined as

$$\mathrm{KL}\left(P\|Q\right) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\Theta\sim P}\left[\log\frac{P(\Theta)}{Q(\Theta)}\right]. \tag{2.6}$$

Rényi divergence reduces to Kullback-Leibler divergence as the order $q \to 1$, that is $\lim_{q\to 1}\mathrm{R}_q\left(P\|Q\right) = \mathrm{KL}\left(P\|Q\right)$ [107].

Some other divergence notions that we will use are the following.

Definition 7 (Total variation distance)**.** The *total variation distance* between two distributions P and Q on $\mathcal{Y}$ is defined as

$$\mathrm{TV}\left(P,Q\right) \stackrel{\text{def}}{=} \sup_{S\subset\mathcal{Y}}|P(S) - Q(S)|. \tag{2.7}$$

Definition 8 (Wasserstein distance [108])**.** *Wasserstein distance* between two distributions P and Q over $\mathcal{Y}$ is defined as

$$\mathrm{W}_2\left(P,Q\right) \stackrel{\text{def}}{=} \inf_{\Pi}\mathop{\mathbb{E}}_{\Theta,\Theta'\sim\Pi}\left[\|\Theta-\Theta'\|_2^2\right]^{\frac{1}{2}}, \tag{2.8}$$

where Π is any joint distribution on $\mathcal{Y}\times\mathcal{Y}$ such that P and Q as its marginals.

Definition 9 (Relative Fisher information [78])**.** If $P \ll Q$ and $\frac{P(\theta)}{Q(\theta)}$ is differentiable w.r.t θ, then *relative Fisher information* of P w.r.t. Q is defined as

$$\mathrm{I}\left(P\|Q\right) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\Theta\sim P}\left[\left\|\nabla\log\frac{P(\Theta)}{Q(\Theta)}\right\|_2^2\right]. \tag{2.9}$$

The following relative Rényi information generalizes relative Fisher information.

Definition 10 (Relative Rényi information [110])**.** Let $q > 1$. If $P \ll Q$ and $\frac{P(\theta)}{Q(\theta)}$ is differentiable, then *relative Rényi information* of P w.r.t. Q is defined as

$$\mathrm{I}_q\left(P\|Q\right) \stackrel{\text{def}}{=} \frac{4}{q^2}\mathop{\mathbb{E}}_{\Theta\sim Q}\left[\left\|\nabla\left[\left(\frac{P(\Theta)}{Q(\Theta)}\right)^{\frac{q}{2}}\right]\right\|_2^2\right]. \tag{2.10}$$

2.1.1 Differential Privacy and Composition Theorems

For a data universe $\mathcal{X}$, we consider datasets D of maximum size n: $D = (\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathcal{X}^n$, where $\perp \in \mathcal{X}$ denotes a special "null" record[2]. Two datasets D and D' are considered neighboring (denoted by $D \simeq D' \in \mathcal{X}^n$) if by replacing a single entry, we can transform D to D' and vice-versa. For a randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$, where $\mathcal{Y}$ is the outcome space, we denote the distributions of the output random variables $\mathcal{A}(D)$ and $\mathcal{A}(D')$ as P and Q, respectively, throughout the thesis unless stated otherwise.

Definition 1 ((ε, δ)-Differential Privacy [41])**.** *Let $\varepsilon \geq 0$ and $0 \leq \delta \leq 1$. A randomized algorithm $\mathcal{A}$ is* (ε, δ)-differentially private *(hereon (ε, δ)-DP) if for all $D \simeq D'$ and for all $S \subset \mathcal{Y}$, we have*

$$\Pr\left[\mathcal{A}(D) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\mathcal{A}(D') \in S\right] + \delta. \tag{2.11}$$

Equivalently, $\mathcal{A}$ is (ε, δ)-DP if the hockey-stick divergence (Definition 4) between output distributions P and Q satisfies $\mathrm{H}_{e^{\varepsilon}}\left(P\|Q\right) \leq \delta$.

Definition 11 (Rényi Differential privacy [69])**.** A randomized algorithm $\mathcal{A}$ is said to be (q, ρ)-*Rényi differentially private* (hereon (q, ρ)-Rényi DP) if, for all neighboring datasets $D \simeq D' \in \mathcal{X}^n$, we have $\mathrm{R}_q\left(P\|Q\right) \leq \rho$.

We say that algorithm $\mathcal{A}$ *tightly* satisfies (ε, δ)-DP if $\delta = \sup_{D\simeq D'} \mathrm{H}_{e^{\varepsilon}}\left(P\|Q\right)$ and that $\mathcal{A}$ *tightly* satisfies (q, ρ)-Rényi DP if $\rho = \sup_{D\simeq D'} \mathrm{R}_q\left(P\|Q\right)$.

[2]Allowing a "null" record $\perp$ within the universe $\mathcal{X}$, replacement operation can simulate both addition or removal by replacing a record with $\perp$. As such, if an algorithm is differentially private for replacement, it is also differentially private with the same parameters for addition or removal.

Composition Theorems. In differential privacy, composition theorems describe how the privacy guarantees of individual algorithms combine when they are executed on the same dataset sequentially and adaptively (i.e., when the subsequent algorithm's input can depend on preceding algorithms' outputs). These theorems are crucial for understanding the loss in privacy when multiple analyses are performed on the same dataset.

Theorem 9 (Advanced composition [40, Theorem III.3])**.** *For any $\varepsilon > 0$, $\delta \in [0, 1]$, and $\delta' \in (0, 1]$, the class of (ε, δ)-DP mechanisms satisfies $(\varepsilon', k\delta + \delta')$-DP under k-fold adaptive composition, for*

$$\varepsilon' = k\varepsilon(e^{\varepsilon} - 1) + \varepsilon\sqrt{2k\log(1/\delta')}. \tag{2.12}$$

The advanced composition theorem is not optimal. The following optimal composition theorem was proved by Kairouz et al. [56].

Theorem 10 (Optimal composition [56, Theorem 3.3])**.** *For any $\varepsilon \geq 0$ and $\delta \in [0, 1]$, the class of (ε, δ)-DP algorithms satisfies $((k-1)\varepsilon, 1-(1-\delta)^k\delta_i))$-DP under k-fold adaptive composition, for all $i \in \{0, 1, \cdots, \lfloor k/2 \rfloor\}$, where*

$$\delta_i = \frac{\sum_{l=0}^{i-1} \binom{k}{l} \left(e^{(k-l)\varepsilon} - e^{(k-2i+l)\varepsilon}\right)}{(1+e^{\varepsilon})^k}. \tag{2.13}$$

Theorem 10 is optimal in the sense that there exists an (ε, δ)-DP mechanism for which the k-fold composition *tightly satisfies* the theorems composed DP guarantee.

Theorem 11 (Rényi DP composition [69, Proposition 1])**.** *If randomized algorithms $\mathcal{A}_1, \cdots, \mathcal{A}_k$ satisfy Rényi DP with respective parameters $(q, \rho_1), \cdots, (q, \rho_k)$, then their composed mechanism defined as $(\mathcal{A}_1(D), \cdots, \mathcal{A}_k(D))$ is $(q, \rho_1 + \cdots + \rho_k)$-Rényi DP. Moreover, i^{th} algorithm can be adaptively chosen on the basis of the outputs of algorithms $\mathcal{A}_1, \cdots, \mathcal{A}_{i-1}$.*

Theorem 11 is also optimal in the same sense as Theorem 10, but for Rényi DP.

2.2 Empirical Risk Minimization

Given a dataset $D = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ of n data-points from a data universe $\mathcal{X}$, and a closed, convex set $\mathcal{Y}$ denoting the space of model parameters (can be $\mathbb{R}^d$), the

problem of *empirical risk minimization* seeks to

$$\text{minimize } \mathcal{L}(\theta; D) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{\mathbf{x} \in D} \ell(\theta; \mathbf{x}) \text{ over } \theta \in \mathcal{Y}. \tag{2.14}$$

The map $\ell(\theta; \mathbf{x}) : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ defines the loss of a given model parameter vector $\theta \in \mathcal{Y}$ on any data point $\mathbf{x} \in \mathcal{X}$. For instance, given a dataset of input-output pairs $\mathbf{x}_i = (\vec{x}_i, y_i) \in \mathcal{X}$, the *ordinary least squares estimator* is the empirical risk minimizer for the loss $\ell(\theta; \mathbf{x}_i) = (\langle \vec{x}_i, \theta \rangle - y_i)^2$, where $\langle \vec{z}_1, \vec{z}_2 \rangle$ gives the dot product between two vectors $\vec{z}_1, \vec{z}_2$. Similarly, the solution to *support vector machine* (SVM) is the minimizer of the hinge loss function $\ell(\theta; \mathbf{x}_i) = [1 - y_i \langle \theta, \vec{x}_i \rangle]_+$.

To avoid the problem of overfitting, sometimes a data-independent convex regularizer $r(\theta) : \mathcal{Y} \to \mathbb{R}$ is added to the minimization objective. To handle such cases, we fold the regularizer into the loss function itself by defining $\ell(\theta; \mathbf{x}) = \ell_0(\theta; \mathbf{x}) + r(\theta)$ in Equation 2.14, where $\ell_0(\theta; \mathbf{x})$ is the base loss function.

The success of an *optimization algorithm* $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ is measured in terms of the expected *excess empirical risk* for the worst-case dataset $D \in \mathcal{X}^n$, defined as

$$\text{err}(\Theta; D) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\Theta} \left[\mathcal{L}_D(\Theta) - \min_{\theta \in \mathcal{Y}} \mathcal{L}_D(\theta) \right], \tag{2.15}$$

where Θ is a random variable denoting the output $\mathcal{A}(D)$.

Remark 1. *We only focus on the expected empirical risk as one can derive high probability excess risk bounds from it by simply running the algorithm multiple times; with k runs of $\mathcal{A}(D)$ producing models $\Theta_1, \cdots, \Theta_k$, Markov inequality*[3] *gives that*

$$\textit{for all } l \in [k], \quad \Pr \left[\mathcal{L}_D(\Theta_l) - \min_{\theta \in \mathcal{Y}} \mathcal{L}_D(\theta) \geq 2\text{err}(\Theta_l; D) \right] \leq \frac{1}{2}. \tag{2.16}$$

Therefore, by applying the union bound, we have that at least one of the k models will have a competitive excess empirical risk with high probability:

$$\Pr \left[\min_{l \in [k]} \mathcal{L}_D(\Theta_l) \leq \min_{\theta \in \mathcal{Y}} \mathcal{L}_D(\theta) + 2\text{err}(\Theta_l; D) \right] \geq 1 - \frac{1}{2^k}. \tag{2.17}$$

Note that for getting a $1 - \beta$ probability guarantee on excess risk from the expected excess risk, the cost of multiple executions is only $O(\log(1/\beta))$ which is acceptable.

[3] We need to make a mild assumption that the loss function $\ell(\theta; \mathbf{x})$ is a non-negative function.

The problem of empirical risk minimization is generally studied under certain assumptions on the loss function. In § 2.2.2, we provide the definitions of some loss function properties that we will require in this thesis. But first, we provide a short refresher on calculus in the following § 2.2.1 to familiarize with the notations.

2.2.1 Calculus Refresher

Given a twice continuously differentiable function $\mathcal{L} : \mathcal{Y} \to \mathbb{R}$, where $\mathcal{Y}$ is a closed convex subset of $\mathbb{R}^d$, its gradient $\nabla\mathcal{L} : \mathcal{Y} \to \mathbb{R}^d$ is the vector of partial derivatives

$$\nabla\mathcal{L}(\theta) = \left(\frac{\partial\mathcal{L}(\theta)}{\partial\theta_1}, \cdots, \frac{\partial\mathcal{L}(\theta)}{\partial\theta_2}\right). \tag{2.18}$$

Its Hessian $\nabla^2\mathcal{L} : \mathcal{Y} \to \mathbb{R}^{d\times d}$ is the matrix of second partial derivatives

$$\nabla^2\mathcal{L}(\theta) = \left(\frac{\partial^2\mathcal{L}(\theta)}{\partial\theta_i\theta_j}\right)_{1\le i,j\le d}. \tag{2.19}$$

Its Laplacian $\Delta\mathcal{L} : \mathcal{Y} \to \mathbb{R}$ is the trace of its Hessian $\nabla^2\mathcal{L}$, i.e.,

$$\Delta\mathcal{L}(\theta) = \mathsf{Tr}\left(\nabla^2\mathcal{L}(\theta)\right). \tag{2.20}$$

Given a differentiable vector field $\mathbf{v} = (\mathbf{v}_1, \cdots, \mathbf{v}_d) : \mathcal{Y} \to \mathbb{R}^d$, its divergence $\nabla\cdot(\mathbf{v}) : \mathcal{Y} \to \mathbb{R}$ is

$$\nabla\cdot(\mathbf{v})(\theta) = \sum_{i=1}^{d}\frac{\partial\mathbf{v}_i(\theta)}{\partial\theta_i}. \tag{2.21}$$

Some identities that we would rely on:

1. Divergence of gradient is the Laplacian, i.e.,

$$\nabla\cdot(\nabla\mathcal{L})(\theta) = \sum_{i=1}^{d}\frac{\partial^2\mathcal{L}(\theta)}{\partial\theta_i^2} = \Delta\mathcal{L}(\theta). \tag{2.22}$$

2. For any function $f : \mathcal{Y} \to \mathbb{R}$ and a vector field $\mathbf{v} : \mathcal{Y} \to \mathbb{R}^d$ with sufficiently fast decay at the border of $\mathcal{Y}$,

$$\int_{\mathcal{Y}}\langle\mathbf{v}(\theta), \nabla f(\theta)\rangle\,\mathrm{d}\theta = -\int_{\mathcal{Y}} f(\theta)(\nabla\cdot(\mathbf{v}))(\theta)\mathrm{d}\theta. \tag{2.23}$$

3. For any two functions $f, g : \mathcal{Y} \to \mathbb{R}$, out of which at least for one the gradient decays sufficiently fast at the border of $\mathcal{Y}$, the following also holds.

$$\int_{\mathcal{Y}} f(\theta)\Delta g(\theta)\mathrm{d}\theta = -\int_{\mathcal{Y}}\langle\nabla f(\theta), \nabla g(\theta)\rangle\,\mathrm{d}\theta = \int_{\mathcal{Y}} g(\theta)\Delta f(\theta)\mathrm{d}\theta. \tag{2.24}$$

4. Based on Young's inequality, for two vector fields $\mathbf{v}_1, \mathbf{v}_2 : \mathcal{Y} \to \mathbb{R}^d$, and any $a, b \in \mathbb{R}$ such that $ab = 1$, the following inequality holds.

$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle (\theta) \leq \frac{1}{2a} \|\mathbf{v}_1(\theta)\|_2^2 + \frac{1}{2b} \|\mathbf{v}_2(\theta)\|_2^2 . \tag{2.25}$$

Wherever it is clear, we would drop (θ) for brevity. For example, we would represent $\nabla \cdot (\mathbf{v}) (\theta)$ as only $\nabla \cdot (\mathbf{v})$.

2.2.2 Loss function properties

In this section, we provide the formal definition of various properties that we assume in the thesis. Let $\ell(\theta; \mathbf{x}) : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$, for each data point $\mathbf{x} \in \mathcal{X}$, be a loss function on $\mathcal{Y}$.

Definition 12 (Lipschitzness)**.** A function $\ell(\theta; \mathbf{x})$ is said to be *L Lipschitz continuous* if for all $\theta, \theta' \in \mathcal{Y}$ and any $\mathbf{x} \in \mathcal{X}$,

$$|\ell(\theta; \mathbf{x}) - \ell(\theta'; \mathbf{x})| \leq L \|\theta - \theta'\|_2 . \tag{2.26}$$

If $\ell(\theta; \mathbf{x})$ is differentiable, then it is L-Lipschitz if and only if $\nabla \ell(\theta; \mathbf{x}) \leq L$ for all $\theta \in \mathcal{Y}$.

Definition 13 (Boundedness)**.** A function $\ell(\theta; \mathbf{x})$ is said to be *B-bounded* if for all $\mathbf{x} \in \mathcal{X}$, its output takes values in range $[-B, B]$.

Definition 14 (Convexity)**.** A continuous differential function $\ell(\theta; \mathbf{x})$ is said to be *convex* if for all $\theta, \theta' \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}$,

$$\ell(\theta'; \mathbf{x}) \geq \ell(\theta; \mathbf{x}) + \langle \nabla \ell(\theta; \mathbf{x}), \theta' - \theta \rangle , \tag{2.27}$$

and is said to be λ-strongly convex if

$$\ell(\theta'; \mathbf{x}) \geq \ell(\theta; \mathbf{x}) + \langle \nabla \ell(\theta; \mathbf{x}), \theta' - \theta \rangle + \frac{\lambda}{2} \|\theta' - \theta\|_2^2 . \tag{2.28}$$

Theorem 12 ([74, Theorem 2.1.4])**.** *A twice continuously differentiable function $\ell(\theta; \mathbf{x})$ is* convex *if and only if for all $\theta \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}$, its hessian matrix $\nabla^2 \ell(\theta; \mathbf{x})$ is positive semidefinite, i.e., $\nabla^2 \ell(\theta; \mathbf{x}) \succcurlyeq 0$ and is λ-strongly convex if its hessian matrix satisfies $\nabla^2 \ell(\theta; \mathbf{x}) \succcurlyeq \lambda I_d$.*

Definition 15 (Smoothness)**.** A continuously differentiable function $\ell(\theta; \mathbf{x})$ is said to be *β-smooth* if for all $\theta, \theta' \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}$,

$$\|\nabla \ell(\theta; \mathbf{x}) - \nabla \ell(\theta'; \mathbf{x})\|_2 \leq \beta \|\theta - \theta'\|_2 . \tag{2.29}$$

Theorem 13 ([74, Theorem 2.1.6])**.** *A twice continuously differentiable convex function $\ell(\theta; \mathbf{x})$ is β-smooth if and only if for all $\theta \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}$,*

$$\nabla^2 \ell(\theta; \mathbf{x}) \preccurlyeq \beta I_d. \tag{2.30}$$

2.3 Langevin Diffusion Semigroup

Langevin diffusion process on $\mathbb{R}^d$ with noise variance σ^2 under the influence of a potential $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is characterized by the Stochastic Differential Equation (SDE)

$$\mathrm{d}\Theta_t = -\nabla \mathcal{L}(\Theta_t)\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}W_t, \tag{2.31}$$

where $\mathrm{d}W_t = W_{t+\mathrm{d}t} - W_t \sim \sqrt{\mathrm{d}t}\mathcal{N}(0, I_d)$ is the d-dimensional Wiener process.

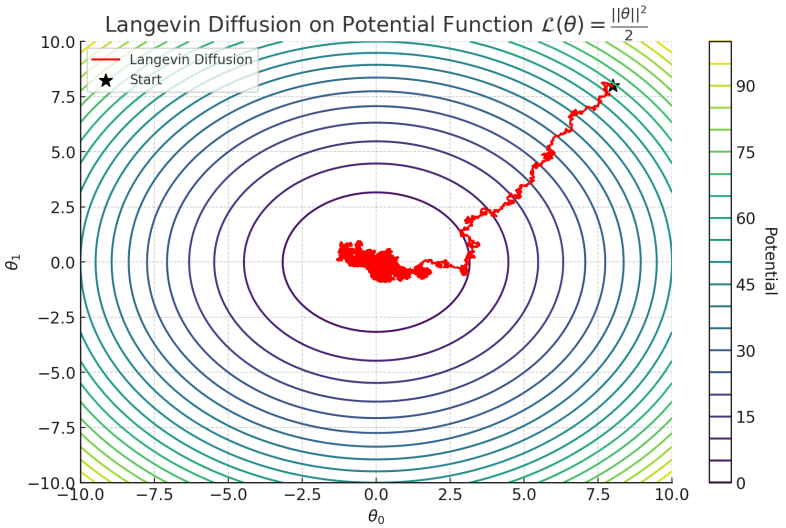


Figure 2.1: Visualizing Langevin diffusion Equation 2.31 on potential $\mathcal{L}(\theta) = \|\theta\|_2^2/2$.

We present some preliminaries on the diffusion theory used in our analyses in this thesis. Let $P_t(\theta_0, \theta_t)$ denote the probability density function describing the distribution of Θ_t, on starting from $\Theta_0 = \theta_0$ at time $t = 0$. For SDE Equation 2.31,

the associated Markov semigroup is defined as a family of operators $(\mathcal{P}_t)_{t\geq 0}$, such that an operator $\mathcal{P}_t$ sends any real-valued measurable function $f : \mathbb{R}^d \to \mathbb{R}$ to

$$\mathcal{P}_t f(\theta_0) = \mathbb{E}\left[f(\Theta_t)|\Theta_0 = \theta_0\right] = \int f(\theta_t) P_t(\theta_0, \theta_t) \mathrm{d}\theta_t. \tag{2.32}$$

When $\mathcal{L}$ is a twice continuously differentiable function in $\mathbb{R}^d$, the infinitesimal generator $\mathcal{G} \stackrel{\text{def}}{=} \lim_{s\to 0} \frac{1}{s}\left[\mathcal{P}_{t+s} - \mathcal{P}_s\right]$ for the Langevin diffusion semigroup, characterized by SDE Equation 2.31, is well-known [61] to be

$$\mathcal{G} f = \sigma^2 \Delta f - \langle \nabla \mathcal{L}, \nabla f \rangle . \tag{2.33}$$

This generator $\mathcal{G}$, when applied on a function $f(\theta_t)$, gives the infinitesimal change in the value of a function f when θ_t undergoes diffusion as per Equation 2.31 for dt time. That is,

$$\partial_t \mathcal{P}_t f(\theta_0) = \int \partial_t P_t(\theta_0, \theta_t) f(\theta_t) \mathrm{d}\theta_t = \int P_t(\theta_0, \theta_t) \mathcal{G} f(\theta_t) \mathrm{d}\theta_t = \mathcal{P}_t(\mathcal{G} f)(\theta_0). \tag{2.34}$$

The dual operator of $\mathcal{G}$ is the Fokker-Planck operator $\mathcal{G}^*$, which is defined as the adjoint of generator $\mathcal{G}$, in the sense that

$$\int f(\mathcal{G}^* g) \mathrm{d}\theta = \int g(\mathcal{G} f) \mathrm{d}\theta, \tag{2.35}$$

for all real-valued measurable functions $f, g : \mathbb{R}^d \to \mathbb{R}$. Note from Equation 2.34 that, this operator provides an alternative way to represent the rate of change of function f at time t:

$$\partial_t \mathcal{P}_t f(\theta_0) = \int f(\theta_t) \partial_t P_t(\theta_0, \theta_t) \mathrm{d}\theta_t = \int f(\theta_t) \mathcal{G}^* P_t(\theta_0, \theta_t) \mathrm{d}\theta_t. \tag{2.36}$$

To put it simply, Fokker-Planck operator gives the infinitesimal change in the distribution of Θ_t with respect to time. For the Langevin diffusion SDE Equation 2.31, the Fokker-Planck operator can be derived as follows. Suppose $f, g : \mathbb{R}^d \to \mathbb{R}$ are functions that decay to 0 at infinity. Then,

$$\begin{aligned}
\int g(\mathcal{G} f) \mathrm{d}\theta &= \int g\left(\sigma^2 \Delta f - \langle \nabla \mathcal{L}, \nabla f \rangle\right) \mathrm{d}\theta \\
&= \sigma^2 \int g \Delta f \mathrm{d}\theta - \int \langle g \nabla \mathcal{L}, \nabla f \rangle \, \mathrm{d}\theta \\
&= \sigma^2 \int f \Delta g \mathrm{d}\theta + \int f \cdot \nabla \cdot (g \nabla \mathcal{L}) \, \mathrm{d}\theta \\
&\qquad\qquad \text{(from Equation 2.24 and Equation 2.23)} \\
&= \int f \left(\sigma^2 \Delta g + \nabla \cdot (g \nabla \mathcal{L})\right) \mathrm{d}\theta \\
&= \int f(\mathcal{G}^* g) \mathrm{d}\theta,
\end{aligned}$$

hence we get the Fokker-Plank operator as $\mathcal{G}^* g = \sigma^2 \Delta g + \nabla \cdot (g \nabla \mathcal{L})$. Fokker-Planck equation gives us quite a convenient way to track changes in the distribution of Θ_t because if we note the identity in Equation 2.36, we get the following relationship (we can ignore θ_0 for brevity).

$$\partial_t P_t(\theta) = \mathcal{G}^* P_t(\theta) = \nabla \cdot (P_t(\theta) \nabla \mathcal{L}(\theta)) + \sigma^2 \Delta P_t(\theta). \tag{2.37}$$

From this Fokker-Planck equation, we can show that the stationary or invariant distribution P_∞ of Langevin diffusion (i.e., when $t \to \infty$ for any starting $\theta_0 \in \mathbb{R}^d$), which is the solution of $\partial_t P_t = 0$ (for all $\theta \in \mathbb{R}^d$). If P_∞ is the stationary distribution, we have

$$\mathcal{G}^* P_\infty = \nabla \cdot \left(P_\infty \nabla \mathcal{L} + \sigma^2 \nabla P_\infty \right) = \nabla \cdot (J) = 0, \qquad \text{(Using Equation 2.22)}$$

where $J(\theta)$ is the *probability flux*. This equation implies that $J(\theta)$ is constant. However $P_\infty(\theta)$ and $\nabla P_\infty(\theta)$ must satisfy certain boundary condition that forces $J(\theta) = 0$ at infinity, which means $J(\theta) = 0$ everywhere. This gives us that

$$J(\theta) = P_\infty(\theta) \nabla \mathcal{L}(\theta) + \sigma^2 \nabla P_\infty(\theta) = 0, \tag{2.38}$$

the solution for which is the following Gibbs distribution.

$$P_\infty(\theta) \propto \exp\left(-\frac{\mathcal{L}(\theta)}{\sigma^2} \right). \tag{2.39}$$

One last thing to note is that since P_∞ is the stationary distribution, for any measurable function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathop{\mathbb{E}}_{P_\infty} [\mathcal{G} f] = \int f \mathcal{G}^* P_\infty \mathrm{d}\theta = 0. \tag{2.40}$$

2.3.1 Isoperimetric Inequalities and Their Properties

Convergence properties of various diffusion semigroups have been extensively analyzed in literature under certain isoperimetric assumptions on the stationary distribution P_∞ [6]. One such property of interest is the *logarithmic Sobolev (*LSI*) inequality* [49], which we define next.

The *carré du champ* operator Γ of a diffusion semigroup with invariant measure P_∞ is defined using its infinitesimal generator $\mathcal{G}$ as

$$\Gamma(f, g) = \frac{1}{2} \left[\mathcal{G}(fg) - f \mathcal{G} g - g \mathcal{G} f \right], \tag{2.41}$$

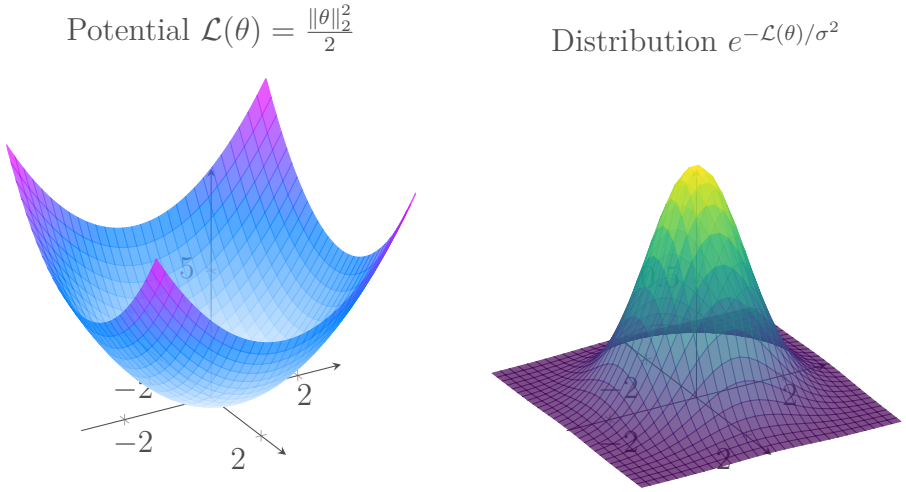


Figure 2.2: Visualizing the potential function $\mathcal{L}$ and the resulting Gibbs distribution $e^{-\mathcal{L}/\sigma^2}$ as per Equation 2.39.

for every $f, g \in \mathbb{L}^2(P_\infty)$. Carré du champ operator represent fundamental properties of a Markov semigroup that affect its convergence behavior. One can verify that Langevin diffusion semigroup's carré du champ operator (on differentiable f, g) is

$$\Gamma(f, g) = \sigma^2 \langle \nabla f, \nabla g \rangle . \tag{2.42}$$

We use shorthand notation $\Gamma(f) = \Gamma(f, f) = \sigma^2 \|\nabla f\|_2$.

Definition 16 (Logarithmic Sobolev Inequality (see Bakry et al. [6, p. 24]))**.** A distribution with probability density P is said to satisfy a *logarithmic Sobolev inequality* (LSI(c)) (with respect to Γ in Equation 2.42) if for all functions $f \in \mathbb{L}^2(P)$ (i.e., $\mathop{\mathbb{E}}_{\Theta \sim P}[f^2(\Theta)] < \infty$) with continuous derivatives ∇f,

$$\mathrm{Ent}_P(f^2) \leq \frac{1}{2c} \int \frac{\Gamma(f^2)}{f^2} P \mathrm{d}\theta = \frac{2\sigma^2}{c} \int \|\nabla f\|_2^2 P \mathrm{d}\theta, \tag{2.43}$$

where entropy Ent_P is defined as

$$\mathrm{Ent}_P(f^2) = \mathop{\mathbb{E}}_P \left[f^2 \log f^2 \right] - \mathop{\mathbb{E}}_P \left[f^2 \right] \log \mathop{\mathbb{E}}_P \left[f^2 \right] . \tag{2.44}$$

Logarithmic Sobolev inequality is a very non-restrictive assumption and is satisfied by a large class of distributions. The following well-known result show that Gaussians satisfy LSI inequality.

Lemma 1 (LS inequality of Gaussian distributions (see Bakry et al. [6, p. 258]))**.** *Let P be a Gaussian distribution on $\mathbb{R}^d$ with covariance σ^2/λ (i.e., the Gibbs distribution Equation 2.39 with $\mathcal{L}(\cdot)$ being the L_2 regularizer $r(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$). Then P satisfies* LSI(λ) *tightly (with respect to Γ in Equation 2.42), i.e.*

$$\mathrm{Ent}_P(f^2) = \frac{2\sigma^2}{\lambda}\int \|\nabla f\|_2^2 \, P\mathrm{d}\theta. \tag{2.45}$$

Additionally, if Q is a distribution on $\mathbb{R}^d$ that satisfies LSI(c)*, then the convolution $Q \circledast P$, defined as the distribution of $\Theta + W$ where $\Theta \sim Q$ and $W \sim P$, satisfies* LSI *inequality with constant $\left(\frac{1}{c} + \frac{1}{\lambda}\right)^{-1}$.*

Bobkov [17] show that like Gaussian distribution, all strongly log concave distributions (or more generally, log-concave distributions with finite second order moments) satisfy LSI inequality (e.g. Gibbs distribution in Equation 2.39 with any strongly convex $\mathcal{L}$). LSI inequality is also satisfied under non-log-concavity too. For example, LSI inequality is stable under Lipschitz maps, although such maps can destroy log-concavity.

Lemma 2 (LS inequality under Lipschitz maps (see Ledoux [62]))**.** *If P is a distribution on $\mathbb{R}^d$ that satisfies* LSI(c)*, then for any L-Lipschitz map $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d$, the pushforward distribution $\mathbf{T}_{\#P}$, representing the distribution of $T(\Theta)$ where $\Theta \sim P$, satisfies* LSI(c/L^2)*.*

LSI inequality is also stable under bounded perturbations to the distribution, as shown in the following lemma by Holley and Stroock [52].

Lemma 3 (LSI inequality under bounded perturbations (see Holley and Stroock [52]))**.** *If P is the probability density of a distribution that satisfies* LSI(c)*, then any probability distribution with density Q such that $\frac{1}{\sqrt{B}} \le \frac{P(\theta)}{Q(\theta)} \le \sqrt{B}$ everywhere in $\mathbb{R}^d$ for some constant $B > 1$ satisfies* LSI(c/B)*.*

Logarithmic Sobolev inequality is of interest to us due to its equivalence to the following inequalities on Kullback-Leibler and Rényi divergence.

Lemma 4 (LSI inequality in terms of KL divergence [110])**.** *The distribution P satisfies* LSI(c) *inequality (with respect to Γ in Equation 2.42) if and only if for all*

distributions Q on $\mathbb{R}^d$ such that $\frac{Q}{P} \in \mathbb{L}^2(P)$ with continuous derivatives $\nabla \frac{Q}{P}$,

$$\mathrm{KL}\left(Q \| P\right) \leq \frac{\sigma^2}{2c} \mathrm{I}\left(Q \| P\right). \tag{2.46}$$

Proof. Set f^2 in Equation 2.43 to $\frac{Q}{P}$ to obtain Equation 2.46. Alternatively, set $Q = \frac{f^2 P}{\mathbb{E}_P[f^2]}$ in Equation 2.46 to obtain Equation 2.43. □

Lemma 5 (Wasserstein distance bound under LSI inequality [78, Theorem 1])**.** *If distribution P satisfies* LSI(c) *inequality (with respect to Γ in Equation 2.42) then for all distributions Q on $\mathbb{R}^d$,*

$$\mathrm{W}_2\left(Q, P\right)^2 \leq \frac{2\sigma^2}{c} \mathrm{KL}\left(Q \| P\right). \tag{2.47}$$

Lemma 6 (LSI inequality in terms of Rényi Divergence [110])**.** *The distribution P satisfies* LSI(c) *inequality (with respect to Γ in Equation 2.42) if and only if for all distributions Q on $\mathbb{R}^d$ such that $\frac{Q}{P} \in \mathbb{L}^2(P)$ with continuous derivatives $\nabla \frac{Q}{P}$, and any $q > 1$,*

$$\mathrm{R}_q\left(Q \| P\right) + q(q-1)\partial_q \mathrm{R}_q\left(Q \| P\right) \leq \frac{q^2\sigma^2}{2c} \frac{\mathrm{I}_q\left(Q \| P\right)}{\mathrm{E}_q\left(Q \| P\right)}. \tag{2.48}$$

Proof. For brevity, let the functions $R(q) = \mathrm{R}_q\left(Q \| P\right)$, $E(q) = \mathrm{E}_q\left(Q \| P\right)$, and $I(q) = \mathrm{I}_q\left(Q \| P\right)$. Let function $f^2(\theta) = \left(\frac{Q(\theta)}{P(\theta)}\right)^q$. Then,

$$\mathop{\mathbb{E}}_P\left[f^2\right] = \mathop{\mathbb{E}}_P\left[\left(\frac{Q}{P}\right)^q\right] = E(q), \qquad \text{(From Definition 5)}$$

and,

$$\begin{aligned}
\mathop{\mathbb{E}}_P\left[f^2 \log f^2\right] &= \mathop{\mathbb{E}}_P\left[\left(\frac{Q}{P}\right)^q \log\left(\frac{Q}{P}\right)^q\right] \\
&= q\partial_q \mathop{\mathbb{E}}_P\left[\int_q \left(\frac{Q}{P}\right)^q \log\left(\frac{Q}{P}\right) \mathrm{d}q\right] && \text{(From Lebniz rule)} \\
&= q\partial_q \mathop{\mathbb{E}}_P\left[\left(\frac{Q}{P}\right)^q\right] && \text{(From Definition 5)} \\
&= q\partial_q E(q).
\end{aligned}$$

Moreover,

$$\mathop{\mathbb{E}}_P\left[\|\nabla f\|_2^2\right] = \mathop{\mathbb{E}}_P\left[\left\|\nabla\left(\frac{Q}{P}\right)^{\frac{q}{2}}\right\|_2^2\right] = \frac{q^2}{4} I(q) \qquad \text{(From Equation 2.10)}$$

On substituting Equation 2.43 with the above equalities, we get:

$$\begin{aligned}
&\mathrm{Ent}_P(f^2) \leq \frac{2\sigma^2}{c}\mathop{\mathbb{E}}_{P}\left[\|\nabla f\|_2^2\right] \\
\iff & q\partial_q E(q) - E(q)\log E(q) \leq \frac{q^2\sigma^2}{2c} I(q) \\
\iff & q\partial_q \log E(q) - \log E(q) \leq \frac{q^2\sigma^2}{2c}\frac{I(q)}{E(q)} \\
\iff & q\partial_q\left((q-1)R(q)\right) - (q-1)R(q) \leq \frac{q^2\sigma^2}{2c}\frac{I(q)}{E(q)} \qquad \text{(From Definition 5)} \\
\iff & R(q) + q(q-1)\partial_q R(q) \leq \frac{q^2\sigma^2}{2c}\frac{I(q)}{E(q)}
\end{aligned}$$

□

2.3.2 Background on Laplace Transforms

Laplace transform is an integral transform that maps time domain functions with real arguments ($t \in \mathbb{R}$) to frequency domain functions with complex arguments ($s \in \mathbb{C}$). The one-sided and two-sided Laplace transformations of a function $g(t)$ at complex frequency s is defined respectively as

$$\mathcal{L}\left\{g(t)\right\}(s) \stackrel{\text{def}}{=} \int_{0^+}^{\infty} e^{-st} g(t)dt, \quad \text{and} \quad \mathcal{B}\left\{g(t)\right\}(s) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} e^{-st} g(t)dt, \tag{2.49}$$

where 0^+ is the shorthand notation for limit approaching 0 from positive side. We can express two-sided Laplace transforms using one-sided transform as

$$\mathcal{B}\left\{f(t)\right\}(s) = \mathcal{L}\left\{f(t)\right\}(s) + \int_{0^-}^{0^+} f(t)dt + \mathcal{L}\left\{f(-t)\right\}(-s), \tag{2.50}$$

where $\int_{0^-}^{0^+} f(t)dt$ may not be 0 if f has an *impulse* (aka. *integrable singularity*) at 0, informally defined to be an infinitely dense point but with a finite mass.

Remark 2. *Conventionally, one-sided Laplace transform is defined to include point mass located at* 0 *entirely (i.e., integration is from* 0^- *instead of* 0^+*). For aligning with the conventions in differential privacy, our definition entirely excludes the point mass located at* 0 *by convention. The Laplace transform properties presented in this thesis are adjusted to reflect the same.*

The values of $s \in \mathbb{C}$ for which the integrals in Equation 2.49 converges conditionally[4] is referred to as the *region of (conditional) convergence* (ROC) of the respective transforms. This region is always a strip parallel to the imaginary axis ωi where $i = \sqrt{-1}$ and $\omega \in \mathbb{R}$, which follows from dominated convergence theorem [76]. We denote the region of convergence for a function g as $\mathrm{ROC}_{\mathcal{L}}\{g\}$ in case of one-sided Laplace transform and as $\mathrm{ROC}_{\mathcal{B}}\{g\}$ for two-sided Laplace transforms.

Uniqueness and Inversion. Laplace transforms are unique [30]: if two continuous functions g and h share the same Laplace transform, then, $g(t) = h(t)$ holds for all $t \in \mathbb{R}^+$ in the case of a one-sided Laplace transform and for all $t \in \mathbb{R}$ for a two-sided Laplace transform. As a consequence of uniqueness, Laplace transform $\bar{g}(s) = \mathcal{L}\{g(t)\}(s)$ can be inverted to get back $g(t)$ by applying an *Inverse Laplace transform* [77], defined as

$$g(t) = \mathcal{L}^{-1}\{\bar{g}(s)\}(t) \stackrel{\text{def}}{=} \frac{1}{2\pi i} \lim_{\omega \to \infty} \int_{\gamma - i\omega}^{\gamma + i\omega} e^{st} \bar{g}(s) ds = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{st} \bar{g}(s) ds, \quad (2.51)$$

where the integral is taken over the line consisting of all points s with $\mathfrak{Re}(s) = \gamma$ for any γ lying in the ROC of $\bar{g}(s)$. The formula Equation 2.51 also inverts two-sided Laplace transform $\mathcal{B}\{g(t)\}$, as long as we choose γ in the ROC of the two-sided transform [76].

Remark 3. *Uniqueness of Laplace transforms extends to discontinuous functions as well. If g and h are continuous* almost everywhere*, i.e. the set where either isn't continuous has a total Lebesgue measure of zero, then $g(t) = h(t)$ almost everywhere in the respective time-domains. In such cases, it can be shown [43] that the inversion formula Equation 2.51 gives*

$$\frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{st} \bar{g}(s) ds = \frac{1}{2}[g(t^-) + g(t^+)]. \quad (2.52)$$

Relation to Fourier transform. The Fourier transform $G(\omega)$ of a function g is defined as

$$G(\omega) = \int_{-\infty}^{\infty} e^{-i2\pi\omega t} g(t) dt, \quad (2.53)$$

[4] Conditional convergence for $\mathcal{L}\{g\}$ at $s \in \mathbb{C}$ means that the limit $\lim_{\gamma \to \infty} \int_{0+}^{\gamma} e^{-st} g(t)\, \mathrm{d}t$ exists. Similarly, the limit $\lim_{\gamma \to \infty} \int_{-\gamma}^{\gamma} e^{-st} g(t)\, \mathrm{d}t$ should exist for $\mathcal{B}\{g\}$ to be conditionally convergent at s.

which is the same as the two-sided laplace transform $\mathcal{B}\{g(t)\}(s)$ for a purely imaginary $s = i2\pi\omega$. As such, Fourier transform is seen as a *special case* of Laplace transforms.

Properties of Laplace transforms. The Laplace transformation is a very useful tool because a lot of operations in the time-domain correspond to simpler operations in the frequency domain and vice versa. For a detailed exposition on these properties, refer to Cohen [30] for one-sided Laplace transform and to Oppenhiem et al. [76] for the two-sided counterpart. In Appendix C.1, we provide a table summarizing all the properties that we rely on in this thesis. We reference properties of Table C.1 throughout the thesis using the notation $\overset{(m)}{=}$, where (m) is the equation number of the used property.

Chapter 3

Privacy of Noisy Gradient Descent

In this chapter, we aim to analyze the information leakage of an iterative randomized learning algorithm concerning its training data, given that the internal state of the algorithm remains private. Specifically, we focus on noisy gradient descent algorithms and model the dynamics of Rényi differential privacy throughout the training process. We derive a tight bound on the Rényi divergence between the probability distributions over the final parameters of models trained on neighboring datasets. Our results demonstrate that the privacy risk converges exponentially fast for smooth and strongly convex loss functions, representing a significant improvement over composition theorems that typically overestimate privacy loss by summing the total value across all intermediate gradient computations. Furthermore, we establish the optimality of noisy gradient descent algorithms and show that this optimality can be achieved with gradient complexity that is almost linear in the training dataset size. Consequently, our work offers a

> *...tight analysis of the trade-off between privacy and utility for noisy gradient descent.*

Related works. The study of differential privacy guarantee for Noisy-GD was initiated by Bassily et al. [12] that applied Dwork et al. [40]'s advanced composition theorem to derive an (ε, δ)-DP bound where ε has an $O(\sqrt{K \log K})$ dependence on number of iterations K. Abadi et al. [1] improved this dependence to $O(\sqrt{K})$ for subsampled variant of Noisy-GD, calling it DP-SGD, by introducing moments accountant, which later Mironov [69] resolved to be equivalent to Rényi DP composition. Balle et al. [8] tightened the amplification effect due to subsampling for DP-SGD, resulting in an improved composition guarantee. Feldman et al. [45]

attempted to refine the DP bound for convex and smooth losses by accounting for the amplification-via-iteration effect—data points seen earlier within an epoch of training enjoy more privacy than data points seen later. However, their amortized privacy bounds over multiple epochs grows at the same order as Rényi composition.

Following our work, Altschuler and Talwar [4] extend Feldman et al. [45]'s approach to strongly-convex losses and show a bounded privacy when parameters are clipped within an L2 ball, using a very different argument than ours. Ye and Shokri [118] extends our bounds on Noisy-GD to subsampled setting of Noisy-SGD, but could only provide Rényi DP amplification of the order $O(b/n)$ for dataset size n and batch size of b, whereas the actual amplification is expected to be $O(b^2/n^2)$. Recent work of Bok et al. [18] closes this gap by extending works of Feldman et al. [45], Altschuler and Talwar [4] onto f-DP notion by introducing a construction of a shifted-interpolated process that allows for a tighter characterization.

3.1 Problem Description

Let $D = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ be a dataset of size n with records taken from some data universe $\mathcal{X}$. Let $\ell(\theta; \mathbf{x}) : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ be a loss function of a parameter vector $\theta \in \mathcal{Y}$ on the data point $\mathbf{x}$, where $\mathcal{Y}$ is a closed convex set. Recall from § 2.2 that a generic formulation of the optimization problem to learn the model parameters, is in the form of empirical risk minimization (ERM) with the following objective, where $\mathcal{L}_D(\theta)$ is the empirical loss of the model, with parameter vector θ, on a dataset D.

$$\theta^* = \arg\min_{\theta \in \mathcal{Y}} \mathcal{L}_D(\theta), \quad \text{where} \quad \mathcal{L}_D(\theta) = \frac{1}{n} \sum_{\mathbf{x} \in D} \ell(\theta; \mathbf{x}). \tag{3.1}$$

Releasing this optimization output (i.e., θ^*) can leak information about the dataset D, hence violating data privacy. To mitigate this risk, there exist randomized algorithms to ensure that the (q-Rényi) differential privacy of the ERM algorithm is upper-bounded by ρ, i.e., the algorithm satisfies (q, ρ)-Rényi DP. One such algorithm is the noisy stochastic gradient descent introduced as Algorithm Algorithm 1 in § 1.2.1. In this chapter we focus on its full-batch variant.

Remark 4. *For convenience, we denote $\hat{\sigma}/n = \sigma$ thought this chapter, which helps in expressing line 2 in Algorithm 2 as $\theta_{k+1} \leftarrow \theta_k - \eta \mathcal{L}_D(\theta) + \sqrt{2\sigma^2}\mathcal{N}(0, \eta I_d)$ which*

Algorithm 2 Noisy-GD: Noisy Gradient Descent [12]

Input: Dataset $D \in \mathcal{X}^n$, loss function $\ell : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$, learning rate η, noise variance $\hat{\sigma}^2$, initial parameter vector $\theta_0 \in \mathbb{R}^d$.

1: **for** $k = 0, 1, \cdots, K-1$ **do**

2: Update model $\theta_{k+1} \leftarrow \theta_k - \frac{\eta}{n} \sum_{i=1}^{n} \nabla \ell(\theta_k; \mathbf{x}_i) + \sqrt{2\eta \left(\frac{\hat{\sigma}}{n}\right)^2} \mathcal{N}(0, I_d)$

3: Output θ_K

resembles the Langevin diffusion PDE: $\mathrm{d}\Theta_t = -\nabla \mathcal{L}(\Theta_t)\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}W_t$ *introduced in § 2.3, where* $\mathrm{d}W_t \sim \mathcal{N}(0, \mathrm{d}tI_d)$.

In this thesis, our objective is to analyze privacy loss of *Noisy Gradient Descent* (Noisy-GD Algorithm 2), which is a randomized ERM algorithm. Let Θ_k, Θ'_k be the parameter vectors at the k'th iteration of Noisy-GD on neighboring datasets D and D', respectively. We denote by $\Theta_{\eta k}$ and $\Theta'_{\eta k}$ the corresponding random variables that model θ_k and θ'_k. We denote their probability distributions by $P_{\eta k}$ and $Q_{\eta k}$. We model and analyze the *dynamics of differential privacy* of this algorithm. More precisely, we focus on the following.

1. Compute an Rényi-DP bound (i.e., the worst case Rényi divergence $\mathrm{R}_q\left(P_{\eta K} \| Q_{\eta K}\right)$ between the output distributions of two neighboring datasets) for Algorithm 2, and analyze its tightness.

2. Compute the contribution of each iteration to the privacy loss. As we go from step $k = 0$ to $K - 1$ in Algorithm 2, we investigate how the algorithm's privacy loss changes as it runs the k^{th} iteration (computed as $\mathrm{R}_q\left(P_{\eta k} \| Q_{\eta k}\right) - \mathrm{R}_q\left(P_{\eta(k-1)} \middle\| Q_{\eta(k-1)}\right)$).

In the end, we aim to provide a Rényi DP bound that is tight, thus facilitating optimal excess empirical risk [12]. We emphasize that our goal is to construct a theoretical framework for analyzing privacy loss of releasing the output θ_K of the algorithm, assuming *private* internal states (i.e., $\theta_1, \cdots, \theta_{K-1}$).

3.2 Privacy analysis of Noisy-GD

3.2.1 Tracing Diffusion for Noisy-GD

To analyze the privacy loss of Noisy-GD, which is a *discrete-time stochastic process*, we first interpolate each discrete update from θ_k to θ_{k+1} with a piece-wise continuously differentiable diffusion process. Let D and D' be a pair of arbitrarily chosen neighboring datasets. Given step-size η and initial parameters $\theta_0 = \theta'_0$, the respective k^{th} discrete updates in Algorithm 2 on neighboring datasets D and D' are

$$\begin{cases} \theta_{k+1} = \theta_k - \eta \nabla \mathcal{L}_D(\theta_k) + \sqrt{2\eta\sigma^2}\mathcal{N}\left(0, I_d\right), \\ \theta'_{k+1} = \theta'_k - \eta \nabla \mathcal{L}_{D'}(\theta'_k) + \sqrt{2\eta\sigma^2}\mathcal{N}\left(0, I_d\right). \end{cases} \tag{3.2}$$

These two discrete jumps can be interpolated with two stochastic processes Θ_t and Θ'_t over time $\eta k < t \leq \eta(k+1)$ respectively. At the start of each step, $t = \eta k$, suppose the random variables $\Theta_{\eta k}$ and $\Theta'_{\eta k}$ model the distribution of the θ_k and θ'_k in the noisy GD processes respectively. During time $\eta k < t \leq \eta(k+1)$, we model the respective gradient updates on D and D' with the following stochastic processes.

$$\begin{cases} \Theta_t = \mathbf{T}(\Theta_{\eta k}) - (t - \eta k)\mathbf{U}(\Theta_{\eta k}) + \sqrt{2\sigma^2}(W_t - W_{\eta k}) \\ \Theta'_t = \mathbf{T}(\Theta'_{\eta k}) + (t - \eta k)\mathbf{U}(\Theta'_{\eta k}) + \sqrt{2\sigma^2}(W'_t - W'_{\eta k}) \end{cases} \tag{3.3}$$

where W_t, W'_t are two independent Wiener processes and the vector maps $\mathbf{T}(\theta) \stackrel{\text{def}}{=} \theta - \frac{\eta}{2}\left(\nabla \mathcal{L}_D(\theta) + \nabla \mathcal{L}_{D'}(\theta)\right)$ and $\mathbf{U}(\theta) \stackrel{\text{def}}{=} \frac{1}{2}\left(\nabla \mathcal{L}_D(\theta) - \nabla \mathcal{L}_{D'}(\theta)\right)$ represent the average gradient step and the contribution of differing record between $D \simeq D'$ respectively.

Note that Equation 3.3 is identical to Equation 3.2 when $t = \eta(k+1)$. Repeating the construction for $k = 0, \cdots, K-1$, we define two piece-wise continuous stochastic processes $\{\Theta_t\}_{t \geq 0}$ and $\{\Theta'_t\}_{t \geq 0}$ whose distributions at time $t = \eta k$ are consistent with θ_k and θ'_k in the Noisy-GD update Equation 3.2 for any $k \in \{0, \cdots, K-1\}$.

Definition 17 (Coupled tracing diffusions)**.** Let $\Theta_0 = \Theta'_0$ be two identically distributed random variables. We refer to the stochastic processes $\{\Theta_t\}_{t \geq 0}$ and $\{\Theta'_t\}_{t \geq 0}$ that undergo identical transformation $\lim_{s \to 0^+} \Theta_{\eta k + s} = \mathbf{T}(\Theta_{\eta k})$ and $\lim_{s \to 0^+} \Theta'_{\eta k + s} = \mathbf{T}(\Theta'_{\eta k})$ at the start of k^{th} step and evolve along diffusion processes Equation 3.4 in the duration $\eta k < t \leq \eta(k+1)$, as coupled tracing diffusions for Noisy-GD on

neighboring datasets $D \simeq D'$. We denote their probability distribution of Θ_t and Θ'_t for any $t \geq 0$ as P_t and Q_t respectively.

The Rényi divergence $\mathrm{R}_q\left(P_{\eta K} \| Q_{\eta K}\right)$ reflects the Rényi privacy risk of Noisy-GD with K steps. Conditioned on observing θ_k and θ'_k, the probability distributions $\{P_t\}_{\eta k<t\leq\eta(k+1)}$ and $\{Q_t\}_{\eta k<t<\eta(k+1)}$ in Equation 3.4 form a Langevin diffusion semi-group along vector fields $-\mathbf{U}(\theta_k)$ and $\mathbf{U}(\theta'_k)$ respectively, for duration η. Therefore, conditioned on observing θ_k and θ'_k, recall from § 2.3 that this diffusion processes in Equation 3.4 have the following stochastic differential equations (SDEs) within the time period $\eta k < t \leq \eta(k+1)$

$$\begin{cases} \mathrm{d}\Theta_t = -\mathbf{U}(\theta_k)\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}W_t \\ \mathrm{d}\Theta'_t = +\mathbf{U}(\theta'_k)\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}W'_t, \end{cases} \quad \text{where } \mathbf{U}(\theta) = \frac{1}{2}\left(\nabla\mathcal{L}_D(\theta) - \nabla\mathcal{L}_{D'}(\theta)\right), \quad (3.4)$$

and recall that $\mathrm{d}W_t = W_{t+\mathrm{d}t} - W_t \sim \sqrt{\mathrm{d}t}\mathcal{N}\left(0, I_d\right)$ describes the Wiener process. Therefore, the conditional probability density functions $P_{t|\eta k}(\theta|\theta_k)$ and $Q_{t|\eta k}(\theta|\theta'_k)$ follow the following Fokker-Planck equation. For brevity, we use $P_{t|\eta k}(\theta|\theta_k)$ and $Q_{t|\eta k}(\theta|\theta'_k)$ to represent the conditional probability density function $p(\Theta_t = \theta|\Theta_{\eta k} = \theta_k)$ and $p(\Theta'_t = \theta|\Theta'_{\eta k} = \theta'_k)$ respectively.

$$\begin{cases} \partial_t P_{t|\eta k}(\theta|\theta_k) = \nabla \cdot \Big(P_{t|\eta k}(\theta|\theta_k)\mathbf{U}(\theta_k)\Big) + \sigma^2\Delta P_{t|\eta k}(\theta|\theta_k), \\ \partial_t Q_{t|\eta k}(\theta|\theta'_k) = -\nabla \cdot \Big(Q_{t|\eta k}(\theta|\theta'_k)\mathbf{U}(\theta'_k)\Big) + \sigma^2\Delta Q_{t|\eta k}(\theta|\theta'_k). \end{cases} \quad (3.5)$$

By taking expectations over probability density function $P_{\eta k}(\theta_k)$ or $Q_{\eta k}(\theta'_k)$ on both sides of Equation 3.5, we obtain the partial differential equation that models the evolution of (unconditioned) probability density function $P_t(\theta)$ and $Q_t(\theta)$ in the coupled tracing diffusions.

Lemma 7. *For coupled tracing diffusion processes Equation 3.4 in time $\eta k < t \leq \eta(k+1)$, the equivalent Fokker-Planck equations are*

$$\begin{cases} \partial_t P_t(\theta) = \nabla \cdot (P_t(\theta)\mathbf{V}_t(\theta)) + \sigma^2\Delta P_t(\theta), \\ \partial_t Q_t(\theta) = \nabla \cdot (Q_t(\theta)\mathbf{V}'_t(\theta)) + \sigma^2\Delta Q_t(\theta), \end{cases} \quad (3.6)$$

where $\mathbf{V}_t(\theta) = \mathop{\mathbb{E}}_{\Theta_{\eta k}\sim P_{\eta k|t}}[\mathbf{U}(\Theta_{\eta k})|\Theta_t = \theta]$ *and* $\mathbf{V}'_t(\theta) = \mathop{\mathbb{E}}_{\Theta'_{\eta k}\sim Q_{\eta k|t}}\left[-\mathbf{U}(\Theta'_{\eta k})|\Theta'_t = \theta\right]$ *are time-dependent maps, and* $\mathbf{U}(\theta) = \frac{1}{2}(\nabla\mathcal{L}_D(\theta) - \nabla\mathcal{L}_{D'}(\theta))$.

See § A.1 for the proof of Lemma 7. By this density evolution equation, we model the noisy gradient descent updates with coupled tracing diffusions. The tracing diffusion process is similar to Langevin diffusion. Therefore, we first study the privacy dynamics in coupled tracing (Langevin) diffusions.

3.2.2 Privacy erosion in tracing (Langevin) diffusion

The Rényi divergence (privacy risk) $\mathrm{R}_q\left(P_t \| Q_t\right)$ between coupled tracing diffusion processes increases over time, as the vector fields $\mathbf{V}_t, \mathbf{V}'_t$ underlying two processes are different. We refer to this phenomenon as *privacy erosion*. This increase is determined by the amount of change in the probability density functions for coupled tracing diffusions, characterized by the Fokker-Planck Equation 3.6 for diffusions under different vector fields.

Using equation Equation 3.6, we compute a bound on the rate (partial derivative) of $\mathrm{R}_q\left(P_t \| Q_t\right)$ over time in the following lemma, to model privacy erosion between two different diffusion processes. We refer to *coupled diffusions* as respective diffusion processes under different vector fields $\mathbf{V}_t$ and $\mathbf{V}'_t$.

Lemma 8 (Rate of Rényi privacy loss)**.** *Let $\mathbf{V}_t$ and $\mathbf{V}'_t$ be two vector fields on $\mathbb{R}^d$ corresponding to a pair of arbitrarily chosen neighboring datasets D and D' with $\max_{\theta \in \mathbb{R}^d} \left\|\mathbf{V}_t(\theta) - \mathbf{V}'_t(\theta)\right\|_2 \leq S_v$ for all $t \geq 0$. Then, for corresponding coupled diffusions $\{P_t\}_{t\geq 0}$ and $\{Q_t\}_{t\geq 0}$ under vector fields $\mathbf{V}_t$ and $\mathbf{V}'_t$ and noise variance σ^2, the Rényi privacy loss rate at any $t \geq 0$ is upper bounded by*

$$\partial_t \mathrm{R}_q\left(P_t \| Q_t\right) \leq \frac{1}{\gamma} \frac{q S_v^2}{4\sigma^2} - (1-\gamma)\sigma^2 q \frac{\mathrm{I}_q\left(P_t \| Q_t\right)}{\mathrm{E}_q\left(P_t \| Q_t\right)}, \tag{3.7}$$

where $\gamma > 0$ is a tuning parameter that we can fix arbitrarily according to our need.

See § A.2 for the proof of Lemma 8. We provide a discussion on our interpretation of the terms in Lemma 8 in § A.2. Although Lemma 8 bounds the Rényi privacy loss rate, the term $\mathrm{I}_q\left(P_t \| Q_t\right)$ depends on unknown distributions P_t, Q_t, and is intractable to compute. Even with explicit expressions for distributions P_t, Q_t, the calculation would involve integration in $\mathbb{R}^d$ which is computationally prohibitive for large d. Note that, however, the ratio $\mathrm{I}_q/\mathrm{E}_q$ is always positive by definition. Therefore, the Rényi divergence (privacy loss) rate in Equation 3.7 is bounded by its first component (a constant w.r.t. t).

Theorem 14 (Linear Rényi divergence bound)**.** *Let* $\mathbf{V}_t$ *and* $\mathbf{V}'_t$ *be two vector fields on* $\mathbb{R}^d$*, with* $\max_{\theta\in\mathbb{R}^d} \|\mathbf{V}_t(\theta) - \mathbf{V}'_t(\theta)\|_2 \le S_v$ *for all* $t \ge 0$*. Then, the coupled diffusions under vector fields* $\mathbf{V}_t$ *and* $\mathbf{V}'_t$ *with noise variance* σ^2 *for time* ηK *has q-Rényi divergence bounded by* $\rho = \frac{qS_v^2}{4\sigma^2} \cdot \eta K$*.*

Proof. Setting $\gamma = 1$ in Lemma 8 gives constant privacy loss rate. Integrating over t suffices. □

Definition 18 (Gradient sensitivitiy)**.** For a differentiable loss function $\ell(\theta; \mathbf{x})$, we define its *gradient sensitivity* S_ℓ to be the L_2-sensitivity of its gradient

$$S_\ell \stackrel{\text{def}}{=} \max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \max_{\theta\in\mathbb{R}^d} \|\nabla\ell(\theta;\mathbf{x}) - \nabla\ell(\theta;\mathbf{x}')\|_2 . \tag{3.8}$$

When the vector fields are $\mathbf{V}_t = -\nabla\mathcal{L}_D$ and $\mathbf{V}'_t = -\nabla\mathcal{L}_{D'}$, the coupled diffusions follow Langevin diffusion. Note that in this case, for all neighboring datasets $D \simeq D' \in \mathcal{X}^n$, we have $\max_{\theta\in\mathbb{R}^d} \|\mathbf{V}_t(\theta) - \mathbf{V}'_t(\theta)\|_2 \le S_\ell/n$. Therefore, this naïve privacy analysis in Theorem 14 gives a linear Rényi-DP guarantee of $(q, \frac{qS_\ell^2}{4\hat{\sigma}^2} \cdot \eta K)$ for Langevin diffusion i.e., when step-size η in Algorithm 1 is infinitesimally small while K is infinitely large so that ηK is a constant. This bound resembles the moment accountant analysis [1] for Noisy-GD which does not require a vanishingly small step-size (described in detail in § A.3).

However, a tighter bound of Rényi privacy loss is possible with finer control of the ratio $\mathrm{I}_q(P_t\|Q_t)/\mathrm{E}_q(P_t\|Q_t)$, which by definition depends on the likelihood ratio between P_t and Q_t, thus is connected with Rényi privacy loss itself. When this ratio grows, the Rényi privacy loss rate decreases, thus slowing down privacy loss accumulation, and leading to tighter privacy bound.

3.2.3 Controlling Rate of Divergence under Isoperimetry

We control the I_q/E_q term in Lemma 8 by making a *log-Sobolev* [6] type *isoperimetric* assumption, as introduced in Definition 16, which we restate here (simplified for Langevin diffusion SDE Equation 2.31).

Definition 16 (Logarithmic Sobolev Inequality (see Bakry et al. [6, p. 24]))**.** *A distribution with probability density* Q *on* $\mathbb{R}^d$ *satisfies* logarithmic Sobolev inequality*,*

i.e. Q *is* LSI(c)*, if for all functions* $f : \mathbb{R}^d \to R$ *such that* ∇f *is continuous, and* $\mathbb{E}_Q[f(\Theta)^2] < \infty$*, we have*

$$\mathop{\mathbb{E}}_{Q}\left[f^2 \log f^2\right] - \mathop{\mathbb{E}}_{Q}\left[f^2\right] \log \mathop{\mathbb{E}}_{Q}\left[f^2\right] \leq \frac{2\sigma^2}{c} \int \|\nabla f\|_2^2 \, Q \mathrm{d}\theta. \tag{3.9}$$

LSI was introduced by Gross [49] as a necessary and sufficient condition for rapid convergence of a diffusion processes. Recently, Vempala and Wibisono [110] showed that this isoperimetric condition is sufficient for rapid convergence of Langevin diffusion in Rényi divergence. Under LSI, they provide the following useful lower bound on I_q/E_q for an arbitrary pair of distributions (which we proved in § 2.3.1).

Lemma 6 (LSI inequality in terms of Rényi Divergence [110])**.** *The distribution* Q_t *satisfies* LSI(c) *inequality if and only if for all distributions* P_t *on* $\mathbb{R}^d$ *such that* $\frac{P_t}{Q_t} \in \mathbb{L}^2(Q_t)$ *with continuous derivatives* $\nabla \frac{P_t}{Q_t}$*, and any* $q > 1$*,*

$$\mathrm{R}_q\left(P_t \| Q_t\right) + q(q-1)\partial_q \mathrm{R}_q\left(P_t \| Q_t\right) \leq \frac{q^2\sigma^2}{2c} \frac{\mathrm{I}_q\left(P_t \| Q_t\right)}{\mathrm{E}_q\left(P_t \| Q_t\right)}. \tag{3.10}$$

Note that $\partial_q \mathrm{R}_q\left(P_t \| Q_t\right)$ is always positive, as from Theorem 6, $\mathrm{R}_q\left(P_t \| Q_t\right)$ monotonically increases with $q > 1$. This lemma shows that $\mathrm{I}_q\left(P_t \| Q_t\right)/\mathrm{E}_q\left(P_t \| Q_t\right)$ grows monotonically with the Rényi privacy loss $\mathrm{R}_q\left(P_t \| Q_t\right)$. By Lemma 8, this implies a throttling privacy loss rate as privacy loss accumulates. Combining Lemma 8 and Lemma 6, we therefore model the **dynamics for Rényi privacy risk under** LSI(c) with the following partial differential inequality (PDI), which describes the relation between privacy risk, its changes over time, and its change over Rényi parameter q. For brevity, let $R(q,t)$ represent $\mathrm{R}_q\left(P_t \| Q_t\right)$.

$$\partial_t R(q,t) \leq \frac{1}{\gamma} \frac{q S_v^2}{4\sigma^2} - 2c(1-\gamma)\left[\frac{R(q,t)}{q} + (q-1)\partial_q R(q,t)\right] \tag{3.11}$$

The initial privacy loss $R(q,0) = 0$, as $P_0 = Q_0$. The solution for this PDI increases with time $t \geq 0$, and models the erosion of Rényi privacy loss in coupled tracing diffusions P_t and Q_t.

3.2.4 Privacy Guarantee for Noisy-GD

We now use the privacy dynamics Equation 3.11 of coupled tracing diffusions to analyze the privacy dynamics for Noisy-GD. We first bound the difference between

the underlying vector fields $\mathbf{V}_t$ and $\mathbf{V}'_t$ for coupled tracing diffusions for Noisy-GD on neighboring datasets D and D'.

Lemma 9. *Let $\ell(\theta;\mathbf{x})$ be a loss function on closed convex set $\mathcal{Y}$ with gradient sensitivity S_ℓ (defined in Definition 18). Let $\{P_t\}_{t\geq 0}$ and $\{Q_t\}_{t\geq 0}$ be the coupled tracing diffusions for Noisy-GD on neighboring datasets $D \simeq D' \in \mathcal{X}^n$, under loss $\ell(\theta;\mathbf{x})$ and noise variance σ^2. Then the difference between underlying vector fields $\mathbf{V}_t$ and $\mathbf{V}'_t$ for coupled tracing diffusions is bounded by*

$$\max_{\theta\in\mathbb{R}^d} \|\mathbf{V}_t(\theta) - \mathbf{V}'_t(\theta)\|_2 \leq \frac{S_\ell}{n}, \tag{3.12}$$

where $\mathbf{V}_t(\theta)$ and $\mathbf{V}'_t(\theta)$ are time-dependent vector fields on $\mathbb{R}^d$, defined in Lemma 7.

Proof. By triangle inequality, for any $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
\|\mathbf{V}_t(\theta) - \mathbf{V}'_t(\theta)\|_2 &\leq \|\mathbf{V}_t(\theta)\|_2 + \|\mathbf{V}'_t(\theta)\|_2 \\
&\leq \mathop{\mathbb{E}}_{\Theta_{\eta k}\sim P_{\eta k|t}} \left[\|\mathbf{U}(\Theta_{\eta k})\|_2 \middle| \Theta_t = \theta\right] + \mathop{\mathbb{E}}_{\Theta'_{\eta k}\sim Q_{\eta k|t}} \left[\left\|-\mathbf{U}(\Theta'_{\eta k})\right\|_2 \middle| \Theta'_t = \theta\right] \\
&\qquad\qquad \text{(From Jensen's inequality)} \\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\Theta_{\eta k}\sim P_{\eta k|t}} \left[\|\nabla\mathcal{L}_D(\Theta_{\eta k}) - \nabla\mathcal{L}_{D'}(\Theta_k)\|_2 \middle| \Theta_t = \theta\right] \\
&\quad + \frac{1}{2}\mathop{\mathbb{E}}_{\Theta'_{\eta k}\sim Q_{\eta k|t}} \left[\left\|\nabla\mathcal{L}_{D'}(\Theta'_{\eta k}) - \nabla\mathcal{L}_D(\Theta'_{\eta k})\right\|_2 \middle| \Theta'_t = \theta\right].
\end{aligned}$$

By Definition 18 of gradient sensitivity, for any neighboring datasets $D \simeq D'$ and any $\theta \in \mathcal{Y}$, we have

$$\|\nabla\mathcal{L}_D(\theta) - \nabla\mathcal{L}_{D'}(\theta)\|_2 \leq \frac{S_\ell}{n}, \quad \|\nabla\mathcal{L}_{D'}(\theta) - \nabla\mathcal{L}_D(\theta)\|_2 \leq \frac{S_\ell}{n}.$$

Therefore, by plugging this inequality, we obtain Equation 3.12. □

We then compute the following PDI modelling Rényi privacy loss dynamics of tracing diffusion at $\eta k < t \leq \eta(k+1)$, under LSI($c$) condition.

$$\partial_t R(q,t) \leq \frac{1}{\gamma}\frac{qS_\ell^2}{4\hat{\sigma}^2} - 2c(1-\gamma)\left(\frac{R(q,t)}{q} + (q-1)\partial_q R(q,t)\right), \tag{3.13}$$

where we've substituted $S_v = S_\ell/n$ and $\sigma = \hat{\sigma}/n$ from Remark 4. We solve this PDI under $\gamma = \frac{1}{2}$ for each time piece, and combine multiple pieces by seeing the map $\mathbf{T}$ as a privacy-preserving post-processing step. We derive the following Rényi-DP guarantee for the Noisy-GD algorithm.

Theorem 15 (Rényi-DP for Noisy-GD under LSI(c))**.** *Let $\{P_t\}_{t\geq 0}$ and $\{Q'_t\}_{t\geq 0}$ be the tracing diffusion for Noisy-GD on neighboring datasets $D \simeq D' \in \mathcal{X}^n$, under noise variance $\hat{\sigma}^2$ and loss function $\ell(\theta; \mathbf{x})$. Let $\ell(\theta; \mathbf{x})$ be a loss function on $\mathcal{Y} = \mathbb{R}^d$, with a finite gradient sensitivity S_ℓ. If for any neighboring datasets D and D', the corresponding coupled tracing diffusions P_t and Q_t satisfy* LSI(c) *throughout $0 \leq t \leq \eta K$, then for all $q > 1$, Noisy-GD satisfies (q, ρ) Rényi Differential Privacy for*

$$\rho = \frac{qS_\ell^2}{2c\hat{\sigma}^2}(1 - e^{-c\eta K}). \tag{3.14}$$

We provide a proof of Theorem 15 in § A.4. This theorem offers a strong converging privacy guarantee, on the condition that LSI(c) is satisfied throughout the Noisy-GD process. This behavior is visualized in Figure 3.1.

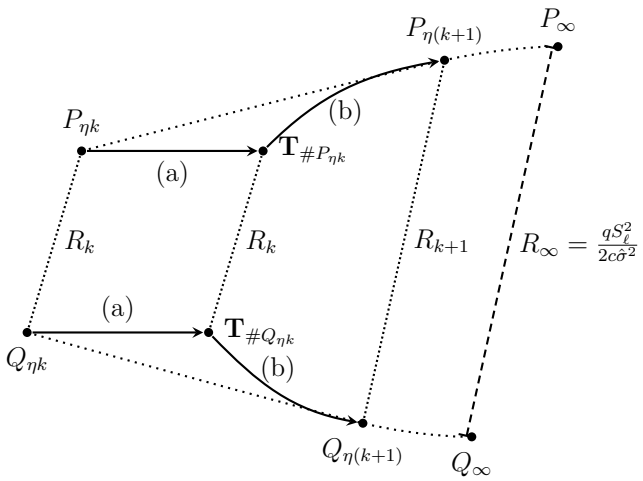


Figure 3.1: Caricature of the converging behavior of the Rényi DP bound for Noisy-GD algorithm as per Theorem 15. Part (a) represents the post-processing transformation $\lim_{s\to 0^+} \Theta_{\eta k+s} = \mathbf{T}(\Theta_{\eta k})$ at the start of step k (denoted by the respective push-forward distributions $\mathbf{T}_{\#P_{\eta k}}$ and $\mathbf{T}_{\#Q_{\eta k}}$). Part (b) shows the coupled diffusions during time $\eta k < t \leq \eta(k+1)$ under vector fields $\mathbf{V}_t$ and $\mathbf{V}'_t$ respectively.

Isoperimetric constants for Noisy-GD. To realize this converging DP guarantee, we need to provide conditions where the LSI constant c is bounded for given Noisy-GD process.

When the loss function is strongly convex and smooth, we prove that tracing diffusion of Noisy-GD satisfies LSI. This is because the gradient descent update

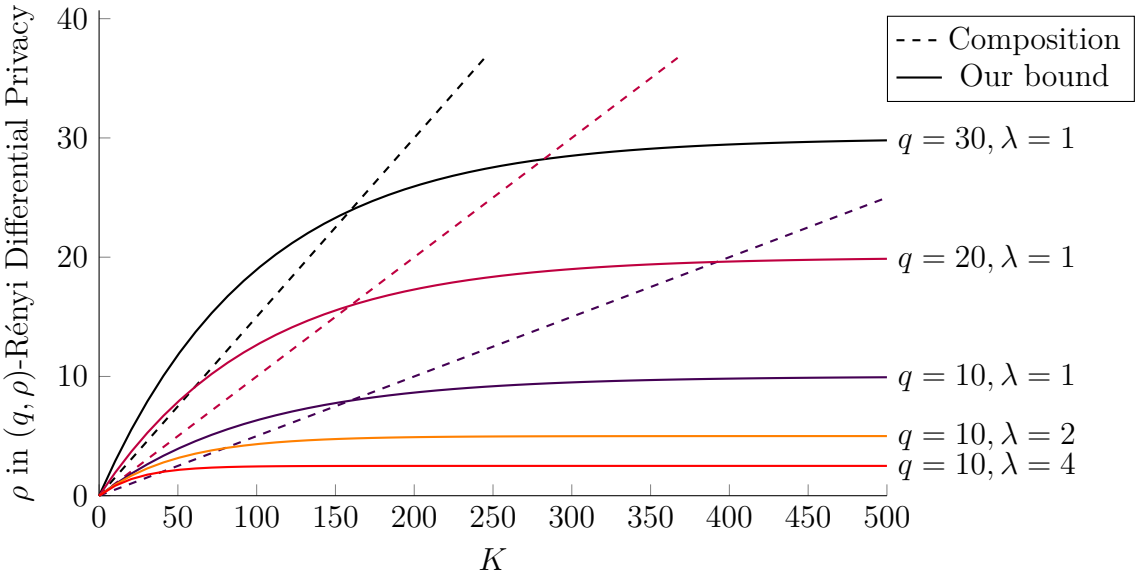


Figure 3.2: Rényi privacy risk of Noisy-GD Algorithm 2 over K iterations, quantified using our DP Dynamics Analysis. We show ρ in the (q, ρ)-RDP guarantee derived in Corollary 1 (bold lines), and the Baseline composition analysis of prior works [1, 69], derived in Theorem 39 (dashed lines). We evaluate under the following setting: Rényi-DP order $q \in \{10, 20, 30\}$; λ-strongly convex loss function with $\lambda \in \{1, 2, 4\}$; β-smooth loss function with $\beta = 4$; gradient sensitivity S_ℓ for loss $\ell(\theta; \mathbf{x})$ with $S_\ell = 1$; arbitrary size of the data set $n \in \mathbb{N}$; step-size $\eta = 0.02$; noise standard deviation $\hat{\sigma} = 1$. The expressions for computing the privacy loss are: our analysis: $\rho = \frac{qS_\ell^2}{\lambda \hat{\sigma}^2} \cdot \left(1 - e^{-\lambda \eta K/2}\right)$; and baseline analysis (derived by Rényi-DP composition with details in § A.3): $\rho = \frac{qS_\ell^2}{4\hat{\sigma}^2} \cdot \eta K$.

is Lipschitz under smooth loss, and the Gaussian noise preserves LSI (follows from Lemma 2 and Lemma 1 introduced in Chapter 2).

Lemma 10 (LSI for Noisy-GD)**.** *If loss function $\ell(\theta; \mathbf{x})$ is λ-strongly convex and β-smooth over $\mathbb{R}^d$, the step-size is $\eta < \frac{1}{\beta}$, and for any initla $\theta_0 \in \mathbb{R}^d$, the coupled tracing diffusion processes $\{P_t\}_{t \geq 0}$ and $\{Q_t\}_{t \geq 0}$ for Noisy-GD on any neighboring datasets D and D' satisfy* LSI(c) *for any $t \geq 0$ with $c = \frac{\lambda}{2}$.*

Appendix § A.4 contains the proof of Lemma 10. Using the LSI constant proved by this lemma, we immediately prove the following RDP bound for noisy GD on Lipschitz smooth strongly convex loss, as a corollary of Theorem 15.

Corollary 1 (Privacy Guarantee for Noisy-GD)**.** *Let $\ell(\theta; \mathbf{x})$ be a λ-strongly convex, and β-smooth loss function on $\mathbb{R}^d$, with a gradient sensitivity S_ℓ, then the Noisy-GD*

algorithm (Algorithm 2) with any start parameter $\theta_0 \in \mathbb{R}^d$*, and step-size* $\eta < \frac{1}{\beta}$*, satisfies* (q, ρ) *Rényi Differential Privacy with*

$$\rho = \frac{qS_\ell^2}{\lambda\hat{\sigma}^2}(1 - e^{-\lambda\eta K/2}) \leq \frac{qS_\ell^2}{\lambda\hat{\sigma}^2}. \tag{3.15}$$

Furthermore, for any $\delta \in (0, 1]$*, it satisfies* $(\frac{2S_\ell\sqrt{\log(1/\delta)}}{\hat{\sigma}\sqrt{\lambda}}, \delta)$*-differential privacy.*

Proof. The first part follows by plugging in the LSI constant proved in Lemma 10 in Theorem 15. For the (ε, δ)-DP part, we use the fact (which is formally proved in Chapter 5) that for any two distributions P, Q, if $\mathrm{R}_q(P\|Q) \leq q \cdot \kappa$ for all $q > 1$, then for all S, we have $P(S) \leq e^\varepsilon Q(S) + \delta$ where $\varepsilon = 2\sqrt{\kappa \log(1/\delta)}$. □

This privacy bound has quadratic dependence on the gradient sensitivity S_ℓ of $\ell(\theta; \mathbf{x})$, which is upper bounded by $S_\ell \leq 2L$ for L-Lipschitz loss functions. The smoothness condition β restricts the step-size and ensures Lipschitz gradient mapping, thus facilitating LSI by Lemma 10. Figure 3.2 demonstrates how this Rényi-DP guarantee for Noisy-GD converges with the number of iterations K. Through y-axis, we show the ρ guaranteed for Noisy-GD under various Rényi divergence orders q and strong convexity constant λ. The RDP order q linearly scales the asymptotic guarantee, but does not affect the convergence rate of Rényi-DP guarantee. However, the strong convexity parameter λ positively affects the asymptotic guarantee as well as the convergence rate; the larger the strong convexity parameter λ is, the stronger the asymptotic Rényi-DP guarantee and the faster the convergence.

3.3 Lower bound on privacy risk of Noisy-GD

Differential privacy guarantees reflect a *bound* on privacy loss on an algorithm; thus, it is very crucial to also have an analysis of their tightness (i.e., how close they are to the exact privacy risk). We prove that our Rényi-DP guarantee in Theorem 15 is tight. To this end, we construct an instance of the ERM problem and two neighboring datasets $D \simeq D'$, for which we show that the Rényi divergence of the Noisy-GD algorithm grows at an order matching our guarantee in Theorem 15.

It is very challenging to lower bound the exact Rényi privacy loss $\mathrm{R}_q(P_{\eta K}\|Q_{\eta K})$ in general. This might require having an explicit expression for the probability distribution over the last-iterate parameters θ_k. Computing a close-form expression

is, however, feasible when the loss gradients are linear. This is due to the fact that, after a sequence of linear transformations and Gaussian noise additions, the parameters follow a Gaussian distribution. Therefore, we construct such an ERM objective, compute the exact privacy risk, and prove the following lower bound.

Theorem 16 (Lower bound on Rényi DP of Noisy-GD)**.** *There exist neighboring datasets $D \simeq D' \in \mathcal{X}^n$, a start parameter $\theta_0 \in \mathbb{R}^d$, and a smooth loss function $\ell(\theta; \mathbf{x})$ on $\mathbb{R}^d$, with a gradient sensitivity S_ℓ, such that for any step-size $\eta < 1$, noise variance $\hat{\sigma}^2 > 0$, and iteration $K \in \mathbb{N}$, the Rényi divergence of the output distributions, $P_{\eta K}$ and $Q_{\eta K}$, of Noisy-GD Algorithm 2 on respective datasets D, D' is lower-bounded by*

$$\mathrm{R}_q \left(P_{\eta K} \| Q_{\eta K} \right) \geq \frac{q S_\ell^2}{4\hat{\sigma}^2} \left(1 - e^{-\eta K} \right). \tag{3.16}$$

The proof of Theorem 16 is provided in § A.5. We prove this lower bound using the L_2-squared norm loss in the ERM objective: $\ell(\theta; \mathbf{x}) = \frac{1}{2} \left\| \theta - \mathbf{x}_i \right\|_2^2$. We assume bounded data domain $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \left\| x \right\|_2 \leq \frac{S_\ell}{2}\}$, due to which the gradient $\nabla \ell(\theta; \mathbf{x})$ has finite sensitivity S_ℓ. With start parameter $\theta_0 = 0^d$, the k^{th} step parameter θ_k is distributed as Gaussian with mean $\mu_k = \eta \bar{\mathbf{x}} \sum_{i=0}^{k-1} (1-\eta)^i$ and variance $\hat{\sigma}_k^2 = \frac{2\eta\hat{\sigma}^2}{n^2} \sum_{i=0}^{k-1} (1-\eta)^{2i}$ in each dimension, where $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ is the empirical dataset mean. We explicitly compute the privacy loss at any step K, which is lower bounded by $\frac{qS_\ell^2}{4\hat{\sigma}^2}(1 - e^{-\eta K})$.

Meanwhile, Corollary 1 establishes an (q, ρ)-Rényi-DP upper bound where $\rho = \frac{qS_\ell^2}{\hat{\sigma}^2} \left(1 - e^{-\eta K/2} \right)$ for this same ERM objective. This upper bound matches the lower bound at every step K, up to a small constant of 4. This constant can be improved further. We can upper bound the Rényi-DP better than the one stated in Corollary 1 by directly using Theorem 15 as for this construction, we can compute the exact LSI constant of $c = \frac{2-\eta}{2}$ for the corresponding coupled diffusion. Doing so, Theorem 15 gives the following tighter Rényi-DP upper-bound for Noisy-GD.

Corollary 2 (Rényi-DP guarantee of Noisy-GD on L_2-norm squared loss)**.** *For any two neighboring datasets $D \simeq D' \in \mathcal{X}^n$, start parameter $\theta_0 \in \mathbb{R}^d$, step-size $\eta \leq 1$, noise variance $\hat{\sigma}^2$, and $K \in \mathbb{N}$, if the loss function is L_2-norm squared loss $\ell(\theta; \mathbf{x}) = \frac{1}{2} \left\| \theta - \mathbf{x} \right\|_2^2$ on $\mathbb{R}^d$, with a gradient sensitivity S_ℓ, the Rényi divergence of Noisy-GD's outputs on D, D' is upper-bounded by*

$$\mathrm{R}_q \left(P_{\eta K} \| Q_{\eta K} \right) \leq \frac{q S_\ell^2}{(2-\eta)\hat{\sigma}^2} (1 - e^{-\frac{2-\eta}{2}\eta K}). \tag{3.17}$$

See § A.5 for the proof of Corollary 2. This Rényi-DP guarantee converges fast to $O(\frac{qS_\ell^2}{\hat{\sigma}^2})$, which matches the lower bound at every step K, up to a constant of $\frac{4}{2-\eta} \approx 2$. This immediately shows tightness of our converging RDP guarantee *throughout* the training process, for a converging Noisy-GD algorithm.

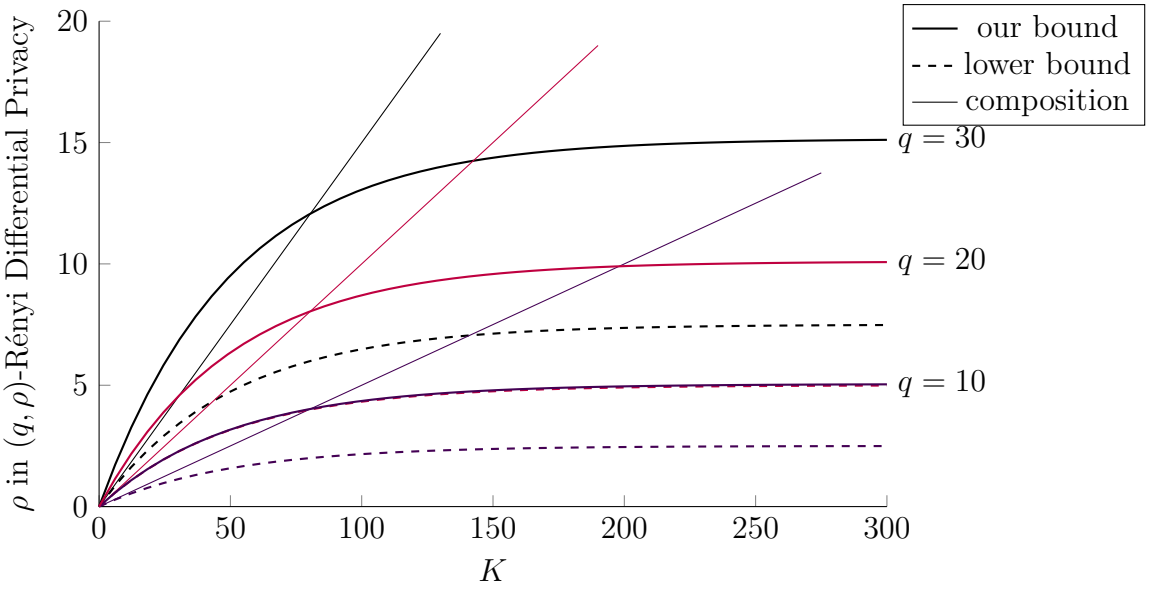


Figure 3.3: Tightness analysis of our Rényi-DP guarantee for the Noisy-GD algorithm. We show the changes of q-Rényi-DP guarantee computed using Corollary 2, over K iterations (number of full passes over the dataset) versus the lower-bounds (dashed lines) which are computed using Theorem 16. The loss function is the L_2-norm squared function $\ell(\theta; \mathbf{x}) = \frac{1}{2} \|\theta - \mathbf{x}\|_2^2$ with a gradient sensitivity of $S_\ell = 1$ (by restricting data domain to $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d | \, \|\mathbf{x}\|_2 \leq S_\ell/2\}$), noise standard deviation is $\hat{\sigma} = 1$, the step size is $\eta = 0.02$ and any arbitrary dataset size n. The Baseline is composition-based (derived in § A.3): $\rho = \frac{qS_\ell^2}{4\hat{\sigma}^2} \cdot \eta K$.

In contrast, composition-based privacy bound grows linearly as training proceeds, as shown in Figure 3.3. When the number of iterations K is small, however, composition-based bound can be better as it grows at the same rate with the lower bound (discussed further in § A.5). Therefore, to conclude whether our Rényi-DP guarantee is superior to composition-based bound, we need to understand the number of iterations Noisy-GD needs, to achieve optimal utility. We discuss this in the following section.

3.4 Utility Analysis for Noisy Gradient Descent

The randomness, required for satisfying differential privacy, can adversely affect the utility of the trained model. The standard way to measure the utility of a randomized ERM algorithm is to quantify its worst case *excess empirical risk* (preliminaries covered in § 2.2), which is defined as

$$\max_{D \in \mathcal{X}^n} \mathbb{E}\left[\mathcal{L}_D(\Theta) - \mathcal{L}_D(\theta^*)\right], \tag{3.18}$$

where Θ is the output of a randomized algorithm on D, θ^* is the minimizer of $\mathcal{L}_D$, and the expectation is computed over the randomness of the algorithm.

We provide the *optimal* excess empirical risk (utility) of Noisy-GD algorithm under (q, ρ)-Rényi DP constraint. The notion of *optimality* for utility is defined as the smallest upper-bound for excess empirical risk that can be guaranteed under (q, ρ)-Rényi DP constraint by tuning the algorithm's hyperparameters (such as the noise variance $\hat{\sigma}^2$ and the number of iterations K). We focus here on smooth and strongly convex loss functions with a bounded gradient sensitivity.

Lemma 11 (Excess empirical risk of Noisy-GD)**.** *For λ-strongly convex and β-smooth loss function $\ell(\theta; \mathbf{x})$, step-size $\eta \leq \frac{\lambda}{\beta^2}$, and any start parameter $\theta_0 \in \mathbb{R}^d$, the excess empirical risk of Noisy-GD Algorithm 2 on any dataset $D \in \mathcal{X}^n$ is bounded by*

$$\mathbb{E}\left[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)\right] \leq \frac{\beta}{2} \left\|\theta_0 - \theta^*\right\|_2^2 \cdot e^{-\lambda \eta K} + \frac{2\beta d \hat{\sigma}^2}{\lambda n^2}, \tag{3.19}$$

where θ_K is the output of Noisy-GD on D.

We provide a proof of Lemma 11 in § A.6. This lemma shows decreasing excess empirical risk for Noisy-GD algorithm under smooth strongly convex loss function as the number of iterations K increases. The utility is determined by K and the noise variance $\hat{\sigma}^2$, which are constrained under (q, ρ)-RDP. Using our tight Rényi-DP bound presented in Corollary 1, we prove the following optimal utility guarantee for Noisy-GD under a privacy budget.

Theorem 17 (Utility under DP for Noisy GD)**.** *For β-smooth λ-strongly convex loss function $\ell(\theta; \mathbf{x})$ with gradient sensitivity S_ℓ, and datasets $D \in \mathcal{X}^n$ of size n, if the step-size $\eta = \frac{\lambda}{\beta^2}$ and any initial parameter $\theta_0 \in \mathbb{R}^d$, then, for any $q > 1$ and*

	Method	**Utility**	**Gradient Complexity**
Bassily et al. [12]	Noisy SGD	$O\left(\frac{d\log^2(n)}{\lambda n^2\epsilon^2}\right)$	n^2
Wang et al. [112]	DP-SVRG	$O\left(\frac{d\log(n)}{\lambda n^2\epsilon^2}\right)$	$O\left((n+\frac{\beta}{\lambda})\log\left(\frac{\lambda n^2\epsilon^2}{d}\right)\right)$
Zhang et al. [120]	Output Perturbation	$O\left(\frac{\beta d}{\lambda^2 n^2\epsilon^2}\right)$	$O\left(\frac{\beta}{\lambda}n\log\left(\frac{n^2\epsilon^2}{d}\right)\right)$
Kifer et al. [58]	Objective Perturbation	$O\left(\frac{d}{\lambda n^2\epsilon^2}\right)$	NA
Ours	Noisy GD	$O\left(\frac{\beta d}{\lambda^2 n^2\epsilon^2}\right)$	$O\left(\frac{\beta^2}{\lambda^2}n\log\left(\frac{\lambda^2 n^2\epsilon^2}{d}\right)\right)$

Table 3.1: Comparison with the prior (ε,δ)-DP ERM algorithms in terms of utility and gradient complexity. We assume 1-Lipschitz, β-smooth and λ-strongly convex loss. Size of input dataset is n, and dimension of parameter vector θ is d. For objective perturbation, we assume $\varepsilon \geq \frac{\beta}{2\lambda}$, and loss is twice differentiable. For our result, we assume $\varepsilon \leq 2\log(1/\delta)$. The lower bound is $\Omega\left(\min\left\{1, \frac{d}{\varepsilon^2 n^2}\right\}\right)$ [12]. We ignore numerical constants and dependence on $\log(1/\delta)$.

$\rho > 0$, the Noisy-GD Algorithm 2 is (q,ρ)-Rényi differentially private, and achieves an excess empirical risk of

$$\mathbb{E}\left[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)\right] = O\left(\frac{q\beta d S_\ell^2}{\rho\lambda^2 n^2}\right), \tag{3.20}$$

when noise variance $\hat{\sigma}^2 = \frac{qS_\ell^2}{\lambda\rho}$, and number of updates $K^* = \frac{\beta^2}{\lambda^2}\log(\frac{\lambda^2\|\theta_0-\theta^*\|_2^2}{4S_\ell^2}\cdot\frac{\rho n^2}{qd})$. Equivalently, for $\varepsilon \leq 2\log(1/\delta)$ and $\delta > 0$, Noisy-GD is (ε,δ)-differentially private, and satisfies

$$\mathbb{E}\left[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)\right] = O\left(\frac{\beta d S_\ell^2\log(1/\delta)}{\varepsilon^2\lambda^2 n^2}\right), \tag{3.21}$$

by setting noise variance $\hat{\sigma}^2 = \frac{2S_\ell^2(\varepsilon+2\log(1/\delta))}{\lambda\varepsilon^2}$, and the number of updates as $K^* = \frac{\beta^2}{\lambda^2}\log\left(\frac{\lambda^2\|\theta_0-\theta^*\|_2^2}{16S_\ell^2}\cdot\frac{\varepsilon^2 n^2}{d\log(1/\delta)}\right)$.

See § A.6 for the proof of Theorem 17.

Our algorithm achieves this utility guarantee with $O(n\log\left(\frac{\varepsilon^2 n^2}{d}\right))$ gradient computations of $\nabla\ell(\theta;\mathbf{x})$, which is faster than Noisy-SGD algorithm [12] with a factor of n. However, we additionally assume smoothness for the loss function. Our gradient complexity also matches that of other efficient gradient perturbation and output perturbation methods [112, 120], as shown in Table 3.1.

This utility matches the following lower bound by Bassily et al. [12] for the best attainable utility of (ε, δ)-differentially private algorithms on Lipschitz smooth strongly convex loss functions.

Theorem 2 (Lower bound for (ε, δ)-DP algorithms [12])**.** *Let $n, d \in \mathbb{N}$, $\delta = o(\frac{1}{n})$, and $\mathcal{X} = \mathcal{Y} = \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$. For every (ε, δ)-DP algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$, there is a dataset $D \in \mathcal{X}^n$ such that for some 1-Lipschitz and 1-strongly convex loss function $\ell(\theta; \mathbf{x})$, we must have*

$$\mathop{\mathbb{E}}_{\Theta \sim \mathcal{A}(D)} [\mathcal{L}_D(\Theta) - \mathcal{L}_D(\theta^*)] = \Omega\left(\min\left(1, \frac{d}{\varepsilon^2 n^2}\right)\right). \tag{3.22}$$

Our utility matches this lower bound upto the constant factor $\log(1/\delta)$, when assuming $\frac{\beta}{\lambda^2} = O(1)$. This improves upon the previous gradient perturbation methods [12, 113] by a factor of $\log(n)$, and matches the utility of previously know optimal ERM algorithm for Lipschitz smooth strongly convex loss functions, such as objective perturbation [26, 58] and output perturbation [120].

Utility gain from tight privacy guarantee. As shown in Table 3.1, our utility guarantee for Noisy-GD is logarithmically better than that for Noisy SGD in Bassily et al. [12], although the two algorithms are extremely similar. This is because we use our tight RDP guarantee, while Bassily et al. [12] use a composition-based privacy bound. More specifically, Noisy SGD needs n^2 iterations to achieve the optimal utility, as shown in Table 3.1. This number of iterations is large enough for the composition-based privacy bound to grow above our RDP guarantee, thus leaving room for improving privacy utility trade-off. This concludes that our tight privacy guarantee enables providing a superior privacy-utility trade-off, for Lipschitz, strongly convex, and smooth loss functions.

Our algorithm also has significantly smaller gradient complexity than Noisy-SGD [12], for strongly convex loss functions, by a factor of $n/\log n$. We use a (moderately large) constant step-size, thus achieving fast convergence to optimal utility. However, Noisy SGD [12] uses a decreasing step-size, thus requiring more iterations to reach optimal utility.

3.5 Discussions and Future Directions

We have developed a novel theoretical framework for analyzing the dynamics of privacy loss for noisy gradient descent algorithms. Our theoretical results show that by hiding the internal state of the training algorithm (over many iterations over the whole data), we can tightly analyze the rate of information leakage throughout training, and derive a bound that is significantly tighter than that of composition-based approaches. With this tight privacy analysis, we show that Noisy-GD can solve the differentially-private empirical risk minimization problem with an optimal excess empirical risk with a gradient-complexity that is almost linear in the size of the dataset.

Future Work. Our main result is a tight privacy guarantee for Noisy GD on smooth and strongly-convex loss functions. Extending these results to non-convex settings is an important direction. In later chapters (specifically, § 4.5.3), we make progress towards privacy analysis in non-convex settings. Another direction to explore is extending our analysis for Noisy-GD to the Noisy-SGD algorithm. Unfortunately, stochastic selection of mini-batch comes with some technical complications that make it difficult to derive tight privacy bounds. While some recent progress has been made, tight sub-sampling amplification that is also compatible with the convergent privacy analysis is still elusive.

Chapter 4

Deletion in Machine Learning

In this chapter, we reason about the concept of data-deletion from a machine learning model using randomized algorithms. The need for data-deletion is motivated to confer the "*Right to be Forgotten*" to data subjects—if a person wants to revoke access to their data, an organization must comply by erasing all information about them without undue delay. This chapter reviews *machine unlearning* guarantees, which are the pervasive statistical definitions for certifying data-deletion capabilities of randomized algorithms. We scrutinize these notions to highlight several fundamental flaws that betray its ultimate goal of empowering *self-determinism of personal information* and propose a provably robust certification for data deletion. Thereafter, we study Noisy Gradient Descent as a data-deletion algorithm and provide a complete characterization of its four way trade-offs: differential privacy, deletion privacy, excess empirical risk, and gradient complexity. In other words, our work characterizes

...data-deletion property of Noisy-GD under computation constraints.

Related works. The idea of machine unlearning was first studied by Cao and Yang [25] and many following works focused on *exact* unlearning where the output distribution of unlearned model is identical to a freshly trained model [19, 104, 119, 117, 90, 116]. Ginart et al. [46], Guo et al. [50] independently provided the first formal definition of *approximate unlearning*, taking inspiration from the notion of (ε, δ)-DP. Gupta et al. [51] initiated the study of *adaptive machine unlearning* that deals with the additional challenges in unlearning in an online setting where the deletion requests are influenced by the previous models' behavior.

Neel et al. [72] provide unlearning guarantees for perturbed gradient descent, a close variant of Noisy-GD that only adds noise to the last iterate model. Ullah et al. [105] provide unlearning guarantees for a variant of Noisy-SGD with momentum and checkpointing as an unlearning algorithm. Additionally, many other Noisy-GD variants for unlearning have been proposed that do not come with a formal unlearning guarantee [75, 111, 109, 93, 89].

4.1 Problem Setting: Machine Unlearning

Let $\mathcal{X}$ be the data domain. A dataset D is an ordered set of n records from $\mathcal{X}$. We use $\mathcal{Y}$ to denote the space of models. A *learning algorithm* $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ inputs a dataset $D \in \mathcal{X}^n$ and returns a model in $\mathcal{Y}$. Suppose a dataset D can be modified by a replacement *edit request*[1] as follows (visualized in Figure 4.1).

Definition 19 (Edit request)**.** A *replacement operation* $\langle \text{ind}, \mathbf{y} \rangle \in [n] \times \mathcal{X}$ on a dataset $D = (\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathcal{X}^n$ performs the following modification:

$$D \circ \langle \text{ind}, \mathbf{y} \rangle = (\mathbf{x}_1, \cdots, \mathbf{x}_{\text{ind}-1}, \mathbf{y}, \mathbf{x}_{\text{ind}+1}, \cdots, \mathbf{x}_n). \tag{4.1}$$

Let $r \leq n$ and $\mathcal{U}^r = [n]^r \times \mathcal{X}^r$. An *edit request* $u = \{\langle \text{ind}_1, \mathbf{y}_1 \rangle, \cdots, \langle \text{ind}_r, \mathbf{y}_r \rangle\} \in \mathcal{U}^r$ on D is defined as a set of r replacement operations modifying distinct indices atomically, i.e.

$$D \circ u = D \circ \langle \text{ind}_1, \mathbf{y}_1 \rangle \circ \cdots \circ \langle \text{ind}_r, \mathbf{y}_r \rangle, \tag{4.2}$$

where $\text{ind}_i \neq \text{ind}_j$ for all $i \neq j$.

Similar to Ginart et al. [46], we define a *deletion* or an *unlearning algorithm* as a (possibly stochastic) mapping $\bar{\mathcal{A}} : \mathcal{X}^n \times \mathcal{U}^r \times \mathcal{Y} \to \mathcal{Y}$. This algorithm takes a dataset $D \in \mathcal{X}^n$, an edit request $u \in \mathcal{U}^r$ and the current model in $\mathcal{Y}$, and produces an updated model in $\mathcal{Y}$. Since edit requests needs to be processed monthly under RTBF guidelines, we adopt the online setting introduced by Neel et al. [72] in which a stream of edit requests $(u_i)_{i \geq 1} \stackrel{\text{def}}{=} (u_1, u_2, \cdots)$, with $u_i \in \mathcal{U}^r$, arrives sequentially. As per this formulation, the *data curator*, characterized by algorithms $(\mathcal{A}, \bar{\mathcal{A}}, f_{\text{pub}})$,

[1] Over a period of time, some individuals may request additions, while others may request deletions. We model this using batched replacement edits and handle any disparity in the number of additions and deletions by inserting or replacing dummy records $\perp$.

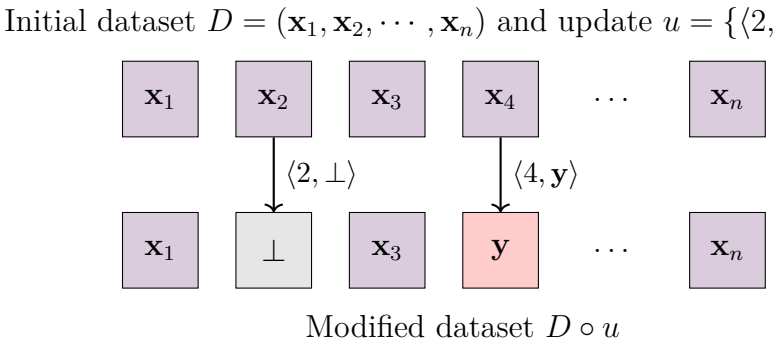


Figure 4.1: Depiction of an *edit request* modifying a dataset.

executes the learning algorithm $\mathcal{A}$ on the initial dataset $D_0 \in \mathcal{X}^n$ during the setup stage before arrival of the first edit request to generate the initial model $\hat{\Theta}_0 \in \mathcal{Y}$, i.e., $\hat{\Theta}_0 = \mathcal{A}(D_0)$. Thereafter, at any edit step $i \geq 1$, to reflect an incoming edit request $u_i \in \mathcal{U}^r$ that transforms $D_{i-1} \circ u_i \to D_i$, the curator executes the unlearning algorithm $\bar{\mathcal{A}}$ on current dataset D_{i-1}, the edit request u_i, and the current model $\hat{\Theta}_{i-1}$ for generating the next model $\hat{\Theta}_i \in \mathcal{Y}$, i.e., $\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) = \hat{\Theta}_i$. Additionally, the curator keeps the sequence $(\hat{\Theta}_i)_{i\geq 0} = (\hat{\Theta}_0, \hat{\Theta}_1, \cdots)$ of learned/unlearned models secret and only releases publishable objects $\phi_i = f_{\text{pub}}(\hat{\Theta}_i)$ for all $i \geq 0$. The publish function $f_{\text{pub}} : \mathcal{Y} \to \Phi$ generates these publishable objects in some space Φ, which can represent any downstream usage such as making predictions.

Ginart et al. [46] note that in real world, deletion requests could often be *adaptive*, i.e., may depend on the prior published objects. For instance, security researchers may demonstrate privacy attacks targeting a minority subpopulation on publicly available models, causing people in that subpopulation to request deletion of their information from training data. Gupta et al. [51] model such an interactive environment through an *adaptive update requester*. We provide the following generalized definition of Gupta et al. [51]'s update requester and describe its interaction with a data curator in Algorithm 3.

Definition 20 (Update requester [51])**.** The edit sequence $(u_i)_{i\geq 1}$ is generated by an update requester $\mathbb{Q}$ that inputs a subset of interaction history between herself and the curator $(\mathcal{A}, \bar{\mathcal{A}}, f_{\text{pub}})$, and outputs a new edit request for the current round. We quantify the strength of $\mathbb{Q}$ with two integers (p, r). Here p is the maximum number of prior published objects that the requester $\mathbb{Q}$ has access to for generating

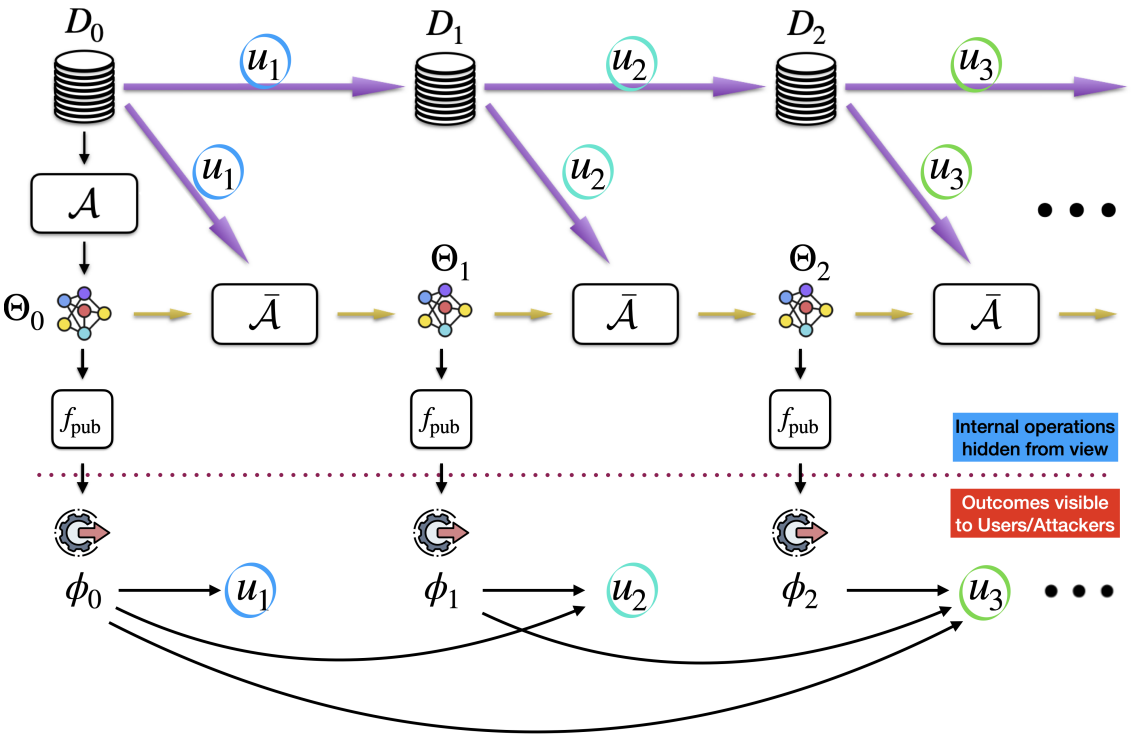


Figure 4.2: Detailed depiction of the problem setup.

the subsequent request and r is the number of records that can be edited per request. More formally, a *p-adaptive r-requester* is a mapping $\mathbb{Q} : \Phi^{\leq p} \times \mathcal{U}^{r*} \to \mathcal{U}^r$. Given a sorted list of indices $\vec{s} = (s_1, \cdots, s_p) \in \mathbb{N}^p$ that the requester $\mathbb{Q}$ can observe, the i^{th} edit request u_i generated by $\mathbb{Q}$ on interaction with $(\mathcal{A}, \bar{\mathcal{A}})$ is defined as

$$u_i = \mathbb{Q}(\underbrace{\phi_{s_1}, \phi_{s_2}, \cdots, \phi_{s_j}}_{\stackrel{\text{def}}{=} \phi_{\vec{s}<i}}; \underbrace{u_1, u_2, \cdots, u_{i-1}}_{\stackrel{\text{def}}{=} u_{<i}}), \tag{4.3}$$

where s_j is the largest index in $\vec{s}$ that is less than i.

Figure 4.3 depicts an example of an adaptive requester $\mathbb{Q}$ modelling the interactive nature of the real-world edit requests inserting, deleting, or replacing entries in response to past outcomes released by the curator $(\mathcal{A}, \bar{\mathcal{A}}, f_{\text{pub}})$.

We consider a 0-adaptive requester as non-adaptive and by ∞-adaptivity or fully-adaptive, we mean requesters that have access to the entire transcript history $(\phi_{<i}; u_{<i})$ at step i.

We present the definitions of *unlearning* and *adaptive unlearning*[2] proposed by

[2]Definition 21 of adaptive unlearning is stronger than Gupta et al. [51]'s since theirs require only one-sided indistinguishability with $(1 - \gamma)$ probability over generated edit requests $u_{\leq i}$.

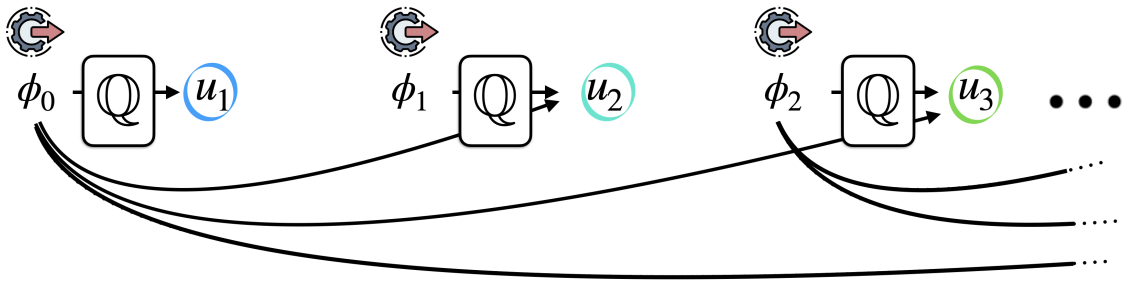


Figure 4.3: Visualization of an *adaptive edit requester* $\mathbb{Q}$ that looks at a (subset of) past outcome releases to (stochastically) determine the next edit request to issue.

Algorithm 3 Interacting curator $(\mathcal{A}, \bar{\mathcal{A}}, f_{\text{pub}})$ & requester $\mathbb{Q}$

Input: dataset $D_0 \in \mathcal{X}^n$, observable indices $\vec{s} \in \mathbb{N}^p$.

1: Initialize $\hat{\Theta}_0 \leftarrow \mathcal{A}(D_0)$
2: Publish $\phi_0 \leftarrow f_{\text{pub}}(\hat{\Theta}_0)$
3: **for** $i = 1, 2, \cdots$ **do**
4: Get next request $u_i \leftarrow \mathbb{Q}(\phi_{\vec{s}<i}; u_{<i})$
5: Update model $\hat{\Theta}_i \leftarrow \bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1})$
6: Publish $\phi_i \leftarrow f_{\text{pub}}(\hat{\Theta}_i)$
7: Update dataset $D_i \leftarrow D_{i-1} \circ u_i$

Neel et al. [72] and Gupta et al. [51] respectively (along with other widely-used data deletion definitions provided in Appendix § B.1.1).

Definition 21 ((Adaptive) Machine unlearning [72, 51])**.** We say that $\bar{\mathcal{A}}$ is an (ε, δ)-*unlearning* algorithm for $\mathcal{A}$ under a publish function f_{pub}, if for all initial datasets $D_0 \in \mathcal{X}^n$ and all non-adaptive 1-requesters $\mathbb{Q}$, the following condition holds. For every edit step $i \geq 1$, and for all generated edit sequences $u_{\leq i} \stackrel{\text{def}}{=} (u_1, \cdots, u_i)$,

$$f_{\text{pub}}(\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}))\Big|_{u_{\leq i}} \stackrel{\varepsilon,\delta}{\approx} f_{\text{pub}}(\mathcal{A}(D_i)), \tag{4.4}$$

where $X \stackrel{\varepsilon,\delta}{\approx} Y$, for any two random variables X, Y, denotes (ε, δ)-DP like indistinguishability between them. If Equation 4.4 holds for all ∞-adaptive 1-requesters $\mathbb{Q}$, we say that $\bar{\mathcal{A}}$ is an (ε, δ)-*adaptive-unlearning* algorithm for $\mathcal{A}$.

4.2 Flaws in Existing Unlearning Definitions

> *The controller shall have the obligation to erase personal data without undue delay where ... the data subject withdraws consent ...*
>
> Right to be Forgotten, Article 17(1)(b), GDPR.

In this section we expose some critical limitations of prior unlearning certifications when it comes to upholding the "Right to be Forgotten". These limitations manifest under a threat model that is often actualized in reality. We highlight multiple reasons why both adaptive and non-adaptive machine unlearning, as described in Definition 21, along with other similar established definitions detailed in Appendix § B.1.1, are inadequate in addressing this threat model.

A real-world threat model for RTBF-compliance. The RTBF guidelines in GDPR and CCPA require *permanent deletion* of personal information, *regardless of its form*, without *undue delay* after receipt of a legitimate deletion request from the user. Considering that data curators are given a grace period to process deletion requests, we assume in our threat model that an attacker targeting a record deleted at the i^{th} step can only observe releases by the curator *after deletion*[3]. In other words, the attacker has access to the entire published sequence post-deletion, which we denote as $\phi_{\geq i} \stackrel{\text{def}}{=} (\phi_i, \phi_{i+1}, \cdots)$.

Furthermore, we assume that users may *interact adaptively* with the curator. Although the attacker cannot see the interaction history until the i^{th} step, we assume that the attacker may be aware of any *dependency relationship* between the published objects $\phi_{<i} \stackrel{\text{def}}{=} (\phi_0, \cdots, \phi_{i-1})$ and the corresponding edit requests $u_{<i} \stackrel{\text{def}}{=} (u_1, \cdots, u_{i-1})$. The assumption is based on the fact that real-world users often exhibit predictable behavior. However, it is crucial to ensure that an attacker cannot extract deleted information from unlearned models, regardless of their understanding of general human behavior patterns. To account for the worst-case scenario, we model the attacker as having *complete knowledge about the adaptive requester* $\mathbb{Q}$ (as defined in Definition 20), which represents any form of dependence

[3]Our threat model doesn't include attacks which involve comparing releases before and after deletion (such as Chen et al. [27]'s). Succeeding in such attacks legally do not violate RTBF as they need information published before a request arrives.

relationships between published outcomes and subsequent requests. However, the attacker does not observe the interaction transcript $(\phi_{<i}; u_{<i})$ of $\mathbb{Q}$.

4.2.1 Unsoundness due to Adaptivity

We highlight the problem with certifying data deletion based on indistinguishability from retraining under adaptive requests with a simple real-life example (which we briefly discussed in § 1.2.2).

Example 1. Imagine publicly posted images on Instagram as a massive dataset, with new uploads acting as addition requests and *copyright infringement notices* as deletion requests. Consider the case of Sam Yang, a renowned Canadian digital illustrator with hundreds of artworks posted on Instagram and a follower count exceeding 2.1 million. Figure 4.4a showcases an original artwork by Sam titled "Night scene with Kara."

Recently, a Redditor utilized Stability AI, a deep-learning platform for creating custom generative models, to train a stable-diffusion model [121]—denoted as $\hat{\Theta}_0$—using a dataset D_0 exclusively comprised of Sam's artworks. The Redditor then posted examples of generated images online (denoted as ϕ_0), such as the one shown in Figure 4.4b. Soon after, internet users began generating and posting stylistic copies on Reddit and Instagram.

Infuriated by this, Sam Yang, along with several other illustrators, filed a class-action lawsuit against Stability AI and Midjourney for copyright infringement [67]. Since the uploads of stylistic copies and the deletion of original artworks occurred almost simultaneously, let's refer to these actions collectively as a *single batched edit request* u_1. In response to the lawsuit, an anonymous Redditor curated a dataset $D_1 = D_0 \circ u_1$ that contained only the stylistic copies, excluding Sam's originals, to circumvent copyright infringement [106]. Figure 4.4c depicts a second-generation stylistic copy ϕ_1 produced by a model $\hat{\Theta}_1 = \mathcal{A}(D_1)$ trained solely on synthetic images. It is evident that Sam's distinctive art style persists, even after retraining from scratch without his original works!

□

To understand the cause of this problem under adaptivity more analytically, let's consider a simplified variant of this problem.

(a) Original "Night scene with Kara" by Sam Yang.

(b) A stylistic copy via the SamDoesArt-V2 model.

(c) A second generation stylistic copy.

Figure 4.4: A real-life example where even retraining does not ensure total deletion.

Example 2. For a data domain $\mathcal{X} = \{-2, -1, 1, 2\}$, consider the following learning and unlearning algorithms $(\mathcal{A}, \bar{\mathcal{A}})$. For any dataset $D \subset \mathcal{X}$ and any subset $S \subset D$ of records to be deleted,

$$\mathcal{A}(D) = \sum_{\mathbf{x} \in D} \mathbf{x}, \quad \text{and} \quad \bar{\mathcal{A}}(D, S, \mathcal{A}(D)) = \sum_{\mathbf{x} \in D \setminus S} \mathbf{x}, \tag{4.5}$$

Note that the unlearning algorithm $\bar{\mathcal{A}}$ perfectly imitates the learning algorithm $\mathcal{A}$ as for any deletion request[4] $u \subset D$, we have $\bar{\mathcal{A}}(D, u, \mathcal{A}(D)) = \mathcal{A}(D \setminus u)$. Now consider two neighboring datasets $D_{-1} = \{-2, -1, 2\}$, $D_1 = \{-2, 1, 2\}$ and the following dependence between the learned model $\mathcal{A}(D)$ and deletion request u:

$$u = \begin{cases} \{\mathbf{x} \in \mathcal{X} | \mathbf{x} < 0\} & \text{if } \mathcal{A}(D) < 0, \\ \{\mathbf{x} \in \mathcal{X} | \mathbf{x} \geq 0\} & \text{otherwise.} \end{cases} \tag{4.6}$$

Knowing this dependence, an attacker can distinguish whether D is D_{-1} or D_1 by looking solely at $\bar{\mathcal{A}}(D, \mathbf{x}, \mathcal{A}(D))$. This is because if $D = D_{-1}$, then the output after deletion is positive, and if $D = D_1$ the output is negative. Note that even though $\bar{\mathcal{A}}$ perfectly imitates retraining via $\mathcal{A}$ and the attacker does not observe either the model $\mathcal{A}(D)$ or the request u, she can still ascertain the identity (-1 or 1) of a deleted record. This example demonstrates two things: **A)** adaptive requests can cause the curator's dataset to have patterns specific to the identity of a target record being deleted, and **B)** an attacker knowing the relationship between unobserved

[4] An edit request containing only deletion operations, i.e., replacement with $\perp$, can be equivalently denoted as a set $u \subset D$ consisting of the records at indices stated in the edit request.

releases and deletion requests can infer the identity of the target record by observing only the unlearned model, even if the curator did full retraining.

□

Given the adaptive nature of the real-world interactions, several data deletion definitions in the literature, such as those proposed by Ginart et al. [46], Guo et al. [50], and Sekhari et al. [92], that quantify data-deletion based on indistinguishability from retraining do not provide reliable certifications for the "Right to be Forgotten" (further detailed in § B.1.1). Even the Adaptive unlearning Definition 21 by Gupta et al. [51], which is specifically designed to ensure RTBF under adaptive deletion requests, fails catastrophically under adaptivity! This is highlighted in the following theorem.

Theorem 18. *There exists an algorithm pair* $(\mathcal{A}, \bar{\mathcal{A}})$ *satisfying* (ε, δ)*-adaptive-unlearning with* $\varepsilon = \delta = 0$ *under publish function* $f_{\text{pub}}(\theta) = \theta$ *such that by designing a* 1*-adaptive* 1*-requester* $\mathbb{Q}$*, an attacker can infer the identity of a record deleted by edit* u_i*, at any arbitrary step* $i > 3$*, with probability at-least* $1 - (1/2)^{i-3}$ *from a single post-edit release* ϕ_i*, even with no access to* $\mathbb{Q}$*'s transcript* $(\phi_{<i}; u_{<i})$*.*

See § B.1 for proof of Theorem 18. This theorem shows that even with perfect indistinguishability between the output of unlearning and retraining algorithm, an adversary who knows the functioning of the update requester $\mathbb{Q}$, i.e. has a *general understanding* of how real-world interactions are like, can decipher the deleted records by only looking for patterns in the post-deletion releases.

4.2.2 Unsoundness due to Secret States

Both adaptive and non-adaptive unlearning guarantees in Definition 21 are bounds on information leakage about a deleted record through a *single released output*. However, our adversary can observe multiple (potentially infinite) releases after deletion. We identify a yet another reason for violation of RTBF under Definition 21, even when edit requests are non-adaptive. This vulnerability arises because Definition 21 permits the curator to store secret models while requiring indistinguishability only over the output of a publishing function f_{pub}. These secret models may propagate encoded information about records even after their deletion from the dataset. So, every subsequent release by an unlearning algorithm can reveal

new information about a record that was purportedly erased multiple edits earlier. We demonstrate in the following theorem that a certified unlearning algorithm can reveal a limited amount of information about a deleted record per release so as not to break the unlearning certification, yet eventually reveal everything about the record to an adversary that observes enough future releases. Figure 4.5 visualizes this vulnerability.

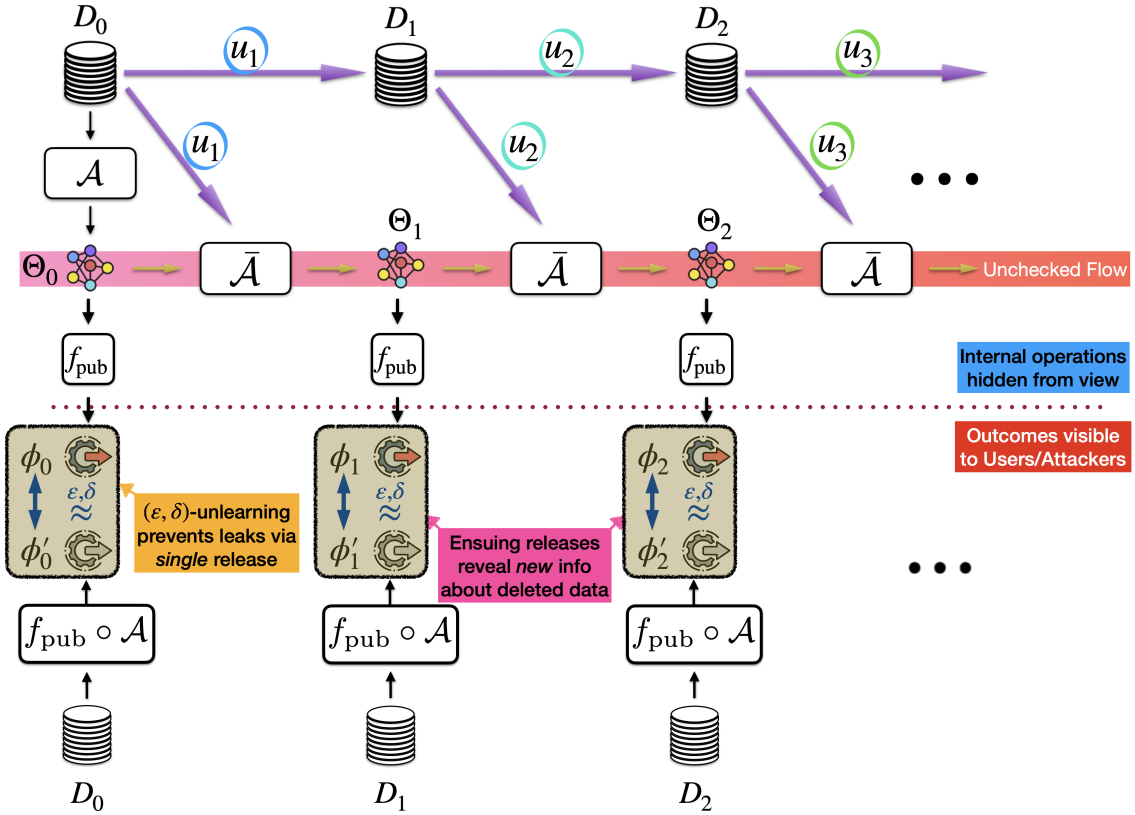


Figure 4.5: Demonstration of the problem with secret states in (ε, δ)-unlearning as described in Definition 21. Consider a data-point deleted by first edit request u_1. Secret states $\hat{\Theta}_1, \hat{\Theta}_2, \cdots$ can carry unbounded information about this deleted record while f_{pub} ensures each individual release carries no more than (ε, δ) amount of information about deleted records. Since each future release can reveal this much *new* information, the total leakage composes without any limits.

Theorem 19. *For every $\varepsilon > 0$, there exists a pair $(\mathcal{A}, \bar{\mathcal{A}})$ of algorithms that satisfy $(\varepsilon, 0)$-unlearning under some publish function f_{pub} such that for all non-adaptive 1-requesters $\mathbb{Q}$, their exists an attacker that can correctly infer the identity of a record deleted at any arbitrary edit step $i \geq 1$ by observing only future releases $\phi_{\geq i}$.*

We provide a proof of Theorem 19 in § B.1. This theorem shows that it is possible to design $(\mathcal{A}, \bar{\mathcal{A}}, f_{\text{pub}})$ such that the (ε, δ)-indistinguishability is preserved per release when passed through the f_{pub} function. However, the internal states retain complete information in a manner that every subsequent release through the f_{pub} leak new information in a way that cumulatively, there is no privacy at all in the long run.

Several unlearning definitions, such as those by Ginart et al. [46], Guo et al. [50] and Sekhari et al. [92] (detailed in Appendix § B.1.1), directly measure the indistinguishability for unlearned models rather than through a publish function f_{pub}. However, the aforementioned vulnerability can still arise if unlearning algorithms satisfying these definitions rely on hidden states that depend on deleted records. It is worth noting that Ginart et al. [46], in their online formulation, advocates a deletion operation to maintain "arbitrary metadata like data structures or partial computations that can be leveraged to help with subsequent deletions", making it susceptible to the vulnerability we have described.

4.2.3 Incompleteness

Existing unlearning definitions measure the effectiveness of a deletion algorithm by assessing how closely its output resembles that of retraining. We emphasize that resembling the output of retraining is *not necessary* for erasing deleted information. For example, let's consider a (useless) mechanism $\bar{\mathcal{A}}$ that outputs a fixed predetermined model $\theta \in \mathcal{Y}$ regardless of its inputs. It is evident that $\bar{\mathcal{A}}$ is a valid deletion algorithm for any learning algorithm, as $\bar{\mathcal{A}}(\cdot, \cdot, \cdot)$ does not rely on the input dataset or the learned model (nor does $f_{\text{pub}}(\bar{\mathcal{A}}(\cdot, \cdot, \cdot))$ for any f_{pub}). Yet $\bar{\mathcal{A}}$ fails to satisfy (ε, δ)-unlearning of Definition 21 for any sensible learning algorithm $\mathcal{A}$. Similarly, many perfectly valid deletion algorithms do not meet existing definitions of unlearning. This makes the (ε, δ)-unlearning an *incomplete* definition for characterizing the data-deletion capability of unlearning algorithms. Several other definitions, such as those by Ginart et al. [46], Guo et al. [50] and Sekhari et al. [92] can also be shown to be incomplete (refer to Appendix § B.1.1).

4.3 Redefining Deletion in Machine Learning

In this section, we introduce a new data-deletion certification for machine unlearning algorithms to address the issues demonstrated in Section § 4.2. We prove its trustworthiness and contend that it offers a more accurate measure of a mechanism's data-deletion abilities. We also study its connections to differential privacy and redefine the data-deletion problem in Machine Learning based on our results.

In Theorem 19, we noted that using secret states that depend on deleted records to speed up unlearning can violate RTBF. Firstly, to prevent such violations, we advocate for designing unlearning algorithms that do not rely on any internal states affected by deleted records and to directly quantify deletion privacy for an unlearned model rather than after applying any publish function, i.e., setting $f_{\text{pub}}(\theta) = \theta$ in Algorithm 3. By doing so, we ensure that the only source of any information about a deleted record that can ever be released by the curator in the future is accounted for. In essence, future unlearning steps become *post-processing operations* for the deleted records, meaning that a valid certification of deletion privacy for the immediate unlearned model applies to all future releases.

Secondly, we propose a new definition of *deletion privacy*. As demonstrated previously, adaptive requests can encode patterns specific to a target record which persists in the dataset even after deletion of the target record, making indistinguishable-from-retraining based deletion certifications unreliable. Our following definition accounts for the worst-case influence adaptive requests by measuring the indistinguishability of an unlearning mechanism's output from that of some mechanism that is not allowed to see the deleted record or edit requests influenced by it.

Definition 22 ((ε, δ)-deletion privacy under p-adaptive r-requesters)**.** Let $\delta \in [0, 1]$, $\varepsilon \geq 0$, and $p, r \in \mathbb{N}$. We say that an algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies (ε, δ)-*deletion privacy under p-adaptive r-requesters* if the following condition holds for all p-adaptive r-requester $\mathbb{Q}$. For every step $i \geq 1$, there exists a randomized mapping $\pi_i^{\mathbb{Q}} : \mathcal{X}^n \to \mathcal{Y}$ such that for all initial datasets $D_0 \in \mathcal{X}^n$,

$$\Pr\left[\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y} \rangle) \in S\right] + \delta, \tag{4.7}$$

for all $u_i \in \mathcal{U}^r$ and all $\langle \text{ind}, \mathbf{y} \rangle \in u_i$.

Our (ε, δ)-deletion privacy is different from existing notions of (ε, δ)-unlearning in the following ways:

1. **Tracking original presence.** We measure the amount of deleted information still residing in an unlearned model based on how indistinguishable the distribution of the unlearned model is from a random variable *free of any influence of the deleted data-point due to its original presence in the dataset in the past*. The construction of an imaginary mechanism $\pi_i^{\mathbb{Q}}$ and its operation on dataset $D_0 \circ \langle \text{ind}, \mathbf{y} \rangle$, where the original record $D_{i-1}[\text{ind}]$ never existed, serves to enforce this independence.

2. **Detached from learning mechanism $\mathcal{A}$.** Certifying (ε, δ)-deletion privacy does not hinge on showing indistinguishabilty of unlearned model from that of freshly retrained model using $\mathcal{A}$, unlike all prior notions of machine unlearning [46, 50, 92]. Therefore, our notion detaches measure of deletion from indistinguishability-to-retraining. We again emphasize that the latter is *not necessary* to achieve the former.

3. **Secure under adaptivity.** Our (ε, δ)-deletion privacy definition allows the construction of $\pi_i^{\mathbb{Q}}$ to take into account *any possible reactive nature* of real-world interactions. That is because we tie the deletion certification to an *arbitrary* adaptive model $\mathbb{Q}$ of the world, where the indistinguishability in Equation 4.7 ensures that even under this model of adaptivity, the original presence of the deleted record cannot be detected.

4. **Quantifying deletion per data-point.** Our notion measures indistinguishability *separately for each deleted data-point*. This makes our definition much more suitable for batched deletions as increasing the batch size does not worsen the certification due to grouping effects, unlike existing unlearning definitions such as Definition 21 that become worse similar to group-privacy guarantees in differential privacy.

5. **Construction $\pi_i^{\mathbb{Q}}$ can be different across steps.** Note that $\pi_i^{\mathbb{Q}}$ can differ across steps $i \in \mathbb{N}$. This freedom enables the analysis to capture any abnormalities due to prolonged processing of really long deletion request sequences.

6. **One-sided indistinguishability.** Our definition requires only a one-sided indistinguishability. For the purpose of ensuring RTBF, one-sided indistinguishability is enough as we only want to prevent events that make an attacker confident that some deleted information was leaked; we don't care if attacker becomes confident that nothing was leaked.

7. **Generalizes (ε, δ)-unlearning under non-adaptivity.** A non-adaptive requester $\mathbb{Q}$ is equivalent to fixing the request sequence $(u_i)_{i\geq 1}$ a-priori, even before the interaction happens. Note that for any operation $\langle \text{ind}, \mathbf{y} \rangle \in u_i$, the associativity of edit operations implies that $D_i = (D_0 \circ \langle \text{ind}, \mathbf{y} \rangle) \circ u_1 \circ \cdots \circ u_i$; that is, that dataset D_i is a function of $D_0 \circ \langle \text{ind}, \mathbf{y} \rangle$ when $\mathbb{Q}$ is non-adaptive. Therefore, for non-adaptive $\mathbb{Q}$, we can set $\pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y} \rangle) = \pi(D_i)$ in Equation 4.7 for any randomized map $\pi : \mathcal{X}^n \to \mathcal{Y}$, including the learning algorithm $\mathcal{A}$. That means, under non-adaptivity, (ε, δ)-unlearning under a publish function $f_{\text{pub}}(\theta) = \theta$ implies (ε, δ)-deletion privacy.

Remark 5. *Note that a construction $\pi_i^{\mathbb{Q}}$ need not be an implementable algorithm. It only serves as an analytical gadget to quantify how well an algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ erases information about deleted records. Later in the chapter, we will present an* easy construction *for $\pi_i^{\mathbb{Q}}$ that is almost always sufficient.*

Soundness of (ε, δ)-deletion privacy. Definition 22 reliably safeguards the "Right to be Forgotten" as the random variable $\pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y} \rangle)$ stays independent of the deleted record $D_0[\text{ind}]$ by design, even when the update requester $\mathbb{Q}$ is ∞-adaptive. When an attacker aims to identify a record at index 'ind' in D_0 that is being replaced by record $\mathbf{y} \in \mathcal{X}$ through one of the replacement operations in edit request $u_i \in \mathcal{U}^r$, the bound in Equation 4.7 ensures that any observer of the unlearned model $\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1})$ cannot be overly certain that the observation was *not* $\pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y} \rangle)$ instead. Consequently, the unlearned model itself must possess minimal information regarding the deleted record $D_0[\text{ind}]$. This argument allows us to establish the following guarantee of soundness.

Theorem 20 (Definition 22 of (ε, δ)-deletion privacy safeguards RTBF)**.** *If the algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies (ε, δ)-deletion privacy guarantee under all p-adaptive*

r-requesters, then even with complete knowledge of a p-adaptive r-requester $\mathbb{Q}$ *that interacts with the curator before a target record* $D_0[\text{ind}]$ *in the initial dataset* D_0 *is deleted at step* $i \geq 1$ *by request* u_i*, any attacker* $MI : \mathcal{Y}^* \to \{0, 1\}$ *observing only the post-deletion models* $\hat{\Theta}_{\geq i} = (\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots)$ *has an advantage*

$$Adv(\text{MI}) \stackrel{def}{=} \Pr\left[\text{MI}(\hat{\Theta}_{\geq i}) = 1 \middle| \mathbf{x}\right] - \Pr\left[\text{MI}(\hat{\Theta}_{\geq i}) = 1 \middle| \mathbf{x}'\right] \tag{4.8}$$

for disambiguating between two possible values $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ *of the deleted record* $D_0[\text{ind}]$ *bounded as follows.*

$$Adv(\text{MI}) \leq e^{\varepsilon} - 1 + 2\delta. \tag{4.9}$$

The proof of Theorem 20 appears in § B.2. As $\varepsilon, \delta \to 0$, r.h.s. of Equation 4.9 approaches 0, implying that no adversary can successfully infer what the deleted entry was. This shows that our definition of (ε, δ)-deletion privacy is *provably sound*.

Completeness of (ε, δ)-deletion privacy. Our Definition 22 allows for the certification of fixed-output unlearning mechanisms described in Section § 4.2, which prior unlearning definitions based on indistinguishability to retraining cannot accomplish. This flexibility stems from the freedom of choice for mechanism $\pi_i^{\mathbb{Q}}$, which can be set to consistently produce the same fixed output.

However, (ε, δ)-deletion privacy may not be a complete guarantee for certifying RTBF. This is because it explicitly prohibits unlearning algorithms from relying on hidden states influenced by previously deleted records. We acknowledge the potential for designing RTBF-compliant unlearning algorithms that utilize partial computations affected by deleted records. However, certifying such algorithms presents greater challenges, as it necessitates establishing bounds on information leakage across all future releases. As shown in Theorem 19, stateful unlearning algorithms of this nature can unveil new information to observers with each subsequent release.

4.4 Deletion Privacy vs. Differential Privacy

A differential privacy guarantee on $\mathcal{A}$ and $\bar{\mathcal{A}}$ sets a limit on the information present in an unlearned model regarding individual records that *remain in the dataset*. However, our concept of deletion privacy specifically restricts the information concerning *only the deleted records*. These two notions are *orthogonal*—an algorithm

pair $(\mathcal{A}, \bar{\mathcal{A}})$ can satisfy $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy without providing $(\varepsilon_{\text{dp}}, \delta)$-differential privacy for any $\infty > \varepsilon_{\text{dp}} > 0$—and *compatible*—an algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ can simultaneously satisfy $(\varepsilon_{\text{dp}}, \delta)$-differential privacy and $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy with $\varepsilon_{\text{dd}} \ll \varepsilon_{\text{dp}}$ for non-adaptive update requesters $\mathbb{Q}$.

In this section, we formalize the connections between these two types of privacy certifications and how they interact when the edit requests are adaptively chosen. We show that $(\varepsilon_{\text{dp}}, \delta)$-differential privacy is both a *necessary* and a *sufficient* condition to ensure that a non-adaptive $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy guarantee for $(\mathcal{A}, \bar{\mathcal{A}})$ extends to adaptive settings with a *graceful degradation* in deletion certification.

For the *sufficient* part, note that when $\mathcal{A}$ and $\bar{\mathcal{A}}$ are both differentially private, they prevent an adaptive requester from establishing dependencies between records in the curator's dataset. By leveraging this property in the following theorem, we establish a reduction from adaptive to non-adaptive deletion privacy only assuming that $\mathcal{A}$ and $\bar{\mathcal{A}}$ also satisfy $(\varepsilon_{\text{dp}}, \delta)$-differential privacy.

Theorem 21 (From adaptive to non-adaptive deletion)**.** *If an algorithm pair* $(\mathcal{A}, \bar{\mathcal{A}})$ *satisfies* $(\varepsilon_{\text{dd}}, \delta)$*-deletion privacy under all non-adaptive* r*-requesters and is also* $(\varepsilon_{\text{dp}}, \delta)$*-differentially private, then pair* $(\mathcal{A}, \bar{\mathcal{A}})$ *also satisfies* $(\varepsilon_{\text{dd}}', (p+2)\delta)$*-deletion privacy under all* p*-adaptive* r*-requesters, for*

$$\varepsilon_{\text{dd}}{}' = \varepsilon_{\text{dd}} + p\varepsilon_{\text{dp}}(e^{\varepsilon_{\text{dp}}} - 1) + \varepsilon_{\text{dp}}\sqrt{2p\log(1/\delta)}. \tag{4.10}$$

We provide proof of Theorem 21 in § B.3. Theorem 21 is essentially the same as composition of the p-fold $(\varepsilon_{\text{dp}}, \delta)$-differential privacy guarantees (as stated in Theorem 9) and a singular $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy guarantee. While we make use of the advanced composition [40] for simplicity, this theorem can be improved with any off-the-shelf composition, including Kairouz et al. [56]'s optimal composition.

When $\varepsilon_{\text{dd}} \ll \varepsilon_{\text{dp}}$, note that the adaptive deletion privacy guarantee is barely larger than just the DP bound on the total information revealed about the deleted record through the past releases before the deletion request was issued. That is to say, when $\mathcal{A}, \bar{\mathcal{A}}$ are differentially private, any deletion privacy guarantee under non-adaptivity *gracefully reduces to deletion-privacy under non-adaptivity* with the worst-case degradation being equivalent to the composition of the two bounds.

Additionally, Theorem 21 simplifies the certification process for unlearning algorithms to ensure RTBF compliance. Under the assumption that deletion requests

are independent of previous releases, i.e. non-adaptive, showing that the algorithm is both differentially private and provides deletion privacy is sufficient.

Not only is differential-privacy a sufficient condition to handle adaptivity, we show that in order to guarantee deletion privacy for erased records in the real-world setting, it is *necessary* for the unlearning algorithm to uphold the privacy of records that remain undeleted. This is because the only effective means of preventing the adaptive world from reacting to the presence of a target record before deletion is by ensuring it never becomes aware of its existence.

Theorem 22 (Membership Privacy is necessary for adaptive deletion privacy)**.** *Let* $Test : \mathcal{Y} \to \{0, 1\}$ *be a membership inference test for* $\mathcal{A}$ *to distinguish between neighboring datasets* $D, D' \in \mathcal{X}^n$*. Similarly, let* $\overline{Test} : \mathcal{Y} \to \{0, 1\}$ *be a membership inference test for* $\bar{\mathcal{A}}$ *to distinguish between* $\bar{D}, \bar{D}' \in \mathcal{X}^n$ *that are neighboring after applying edit* $\bar{u} \in \mathcal{U}^1$*. If* $Adv(Test) > \Delta_{\mathcal{A}}$ *and* $Adv(\overline{Test}) > \Delta_{\bar{\mathcal{A}}}$*, then the pair* $(\mathcal{A}, \bar{\mathcal{A}})$ cannot *satisfy* (ε, δ)*-deletion privacy under* 1*-adaptive* 1*-requester for any*

$$\varepsilon < \log(1 + \Delta_{\mathcal{A}} \cdot \Delta_{\bar{\mathcal{A}}} - 2\delta). \tag{4.11}$$

See § B.3 for a proof of Theorem 22. This theorem shows that without membership privacy, which is ensured by differential privacy, it isn't possible to provide any non-trivial deletion privacy guarantees under adaptivity, no matter what algorithms $(\mathcal{A}, \bar{\mathcal{A}})$ we use. Therefore, for building reliable unlearning algorithms, we should look towards differentially-private algorithms.

Clarifying the need for Differential Privacy. It is important to emphasize that we are not advocating for unlearning solely through differentially private mechanisms, as they uniformly limit the information content of all records, whether deleted or not. Instead, an effective unlearning algorithm should offer two distinct information reattainment bounds: one for the records currently present in the dataset, provided by a differential privacy guarantee, and a significantly smaller bound for the records previously deleted, ensured through a non-adaptive deletion privacy guarantee. Together, these two bounds ensure privacy of deleted records under adaptivity, thanks to Theorem 21.

4.5 ERM with Differential and Deletion Privacy

In the preceding sections, we introduced a *provably reliable* notion for quantifying the data-deletion capabilities of an unlearning algorithm. In this section, we contextualize it for the *empirical risk minimization* problem as discussed in § 2.2, and in the subsequent sections, we study unlearning using Noisy-GD algorithm.

Let space of model parameters be $\mathcal{Y} = \mathbb{R}^d$ and $\boldsymbol{\ell}(\theta; \mathbf{x}) : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ be a loss function of a parameter $\theta \in \mathbb{R}^d$ for a record $\mathbf{x} \in \mathcal{X}$. We consider the problem of *empirical risk minimization* (ERM) of the average $\boldsymbol{\ell}(\theta; \mathbf{x})$ over records in the dataset $D \in \mathcal{X}^n$ under $L2$ regularization, that is, the minimization objective is

$$\mathcal{L}_D(\theta) = \frac{1}{n} \sum_{\mathbf{x} \in D} \boldsymbol{\ell}(\theta; \mathbf{x}) + \mathbf{r}(\theta), \text{ with } \mathbf{r}(\theta) = \frac{\lambda \|\theta\|_2^2}{2}. \tag{4.12}$$

Recall that the *excess empirical risk* of a (random) model Θ on D is defined as

$$\operatorname{err}(\Theta; D) = \mathbb{E}\left[\mathcal{L}_D(\Theta) - \mathcal{L}_D(\theta_D^*)\right], \tag{4.13}$$

where $\theta_D^* = \arg\min_{\theta \in \mathbb{R}^d} \mathcal{L}_D(\theta)$, and expectation is over Θ.

Problem Definition. Let constants $0 < \varepsilon_{\text{dd}} \leq \varepsilon_{\text{dp}}$ and $\delta \in (0, 1]$. Our goal in this paper is to design a learning mechanism $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ and an unlearning mechanism $\bar{\mathcal{A}} : \mathcal{X}^n \times \mathcal{U}^r \times \mathbb{R}^d \to \mathbb{R}^d$ for ERM such that

(1.) both $\mathcal{A}$ and $\bar{\mathcal{A}}$ satisfy $(\varepsilon_{\text{dp}}, \delta)$-differential privacy with respect to individual records in the input dataset,

(2.) pair $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy guarantee for all non-adaptive r-requesters $\mathbb{Q}$,

(3.) and, all models $(\hat{\Theta}_i)_{i \geq 0}$ produced by $(\mathcal{A}, \bar{\mathcal{A}}, \mathbb{Q})$ on any $D_0 \in \mathcal{X}^n$ have $\operatorname{err}(\hat{\Theta}_i; D_i) \leq f(\varepsilon_{\text{dp}}, \varepsilon_{\text{dd}}, \delta, n, d)$ for some excess risk bound[5] $f(\cdots)$.

Objectives **(1.)** and **(2.)** together ensure that $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies deletion-privacy for adaptive requests as well, and objective **(3.)** ensures (un)learned models are useful.

[5]The bound $f(\cdots)$ in **(3.)** should be close to the optimal excess risk attainable by ERM on D_i under $(\varepsilon_{\text{dp}}, \delta)$-DP.

A deletion algorithm $\bar{\mathcal{A}}$ is only useful if it is computationally cheaper than retraining with $\mathcal{A}$. We judge the benefit of $\bar{\mathcal{A}}$ over $\mathcal{A}$ for i^{th} request u_i by the difference in retraining $\text{Cost}(\mathcal{A}; D_{i-1} \circ u_i)$ and deletion $\text{Cost}(\bar{\mathcal{A}}; D_{i-1}, u_i, \hat{\Theta}_{i-1})$.

4.5.1 Deletion using Noisy Gradient Descent

This section proposes a simple and effective data-deletion solution based on Noisy-GD as introduced in Chapter 3, described again in Algorithm 4 below.

Algorithm 4 Noisy-GD: Noisy Gradient Descent [12]

Input: Dataset $D \in \mathcal{X}^n$, loss function $\ell : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$, learning rate η, noise variance $\hat{\sigma}^2$, initial parameter vector r.v. $\Theta_0 \in \mathbb{R}^d$, number of iterations $K \in \mathbb{N}$

1: **for** $k = 0, 1, \cdots, K-1$ **do**

2: $\quad \nabla \mathcal{L}_D(\Theta_{\eta k}) = \frac{1}{n} \sum_{\mathbf{x} \in D} \nabla \boldsymbol{\ell}(\Theta_{\eta k}; \mathbf{x}) + \nabla \mathbf{r}(\Theta_{\eta k})$

3: $\quad$ Update model $\Theta_{\eta(k+1)} \leftarrow \Theta_{\eta k} - \eta \nabla \mathcal{L}_D(\Theta_{\eta k}) + \sqrt{2\eta \left(\frac{\hat{\sigma}}{n}\right)^2} \mathcal{N}(0, I_d)$

4: Output $\Theta_{\eta K}$

Remark 6. *For simplicity, we've denoted $\hat{\sigma}/n = \sigma$ as we did for our analysis in Chapter 3 (cf. Remark 4).*

Our proposed approach falls under the Descent-to-Delete framework proposed by Neel et al. [72], wherein, after each deletion request u_i, we run Noisy-GD starting from the previous model $\hat{\Theta}_{i-1}$ and perform fine-tuning using Noisy-GD over records in the modified dataset $D_i = D_{i-1} \circ u_i$; sufficient to erase information regarding deleted records in the subsequent model $\hat{\Theta}_i$. Our algorithms $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ is defined as follows.

Definition 23 (Noisy-GD based data-deletion solution)**.** Let $K_{\mathcal{A}}, K_{\bar{\mathcal{A}}} \in \mathbb{N}$. For any $D \in \mathcal{X}^n$, our learning algorithm $\mathcal{A}_{\text{Noisy-GD}} : \mathcal{X}^n \to \mathbb{R}^d$ is defined as

$$\mathcal{A}_{\text{Noisy-GD}}(D) = \text{Noisy-GD}(D, \Theta, K_{\mathcal{A}}), \tag{4.14}$$

where $\Theta \in \mathbb{R}^d$ is a randomly initialized model from a distribution ρ. And, for any edit request $u \in \mathcal{U}^r$ on dataset $D \in \mathcal{X}^n$ and any model denoted with its random variable $\Theta \in \mathbb{R}^d$, our unlearning algorithm $\bar{\mathcal{A}}_{\text{Noisy-GD}} : \mathcal{X}^n \times \mathcal{U}^r \times \mathbb{R}^d \to \mathbb{R}^d$ is

$$\bar{\mathcal{A}}_{\text{Noisy-GD}}(D, u, \Theta) = \text{Noisy-GD}(D \circ u_i, \Theta, K_{\bar{\mathcal{A}}}). \tag{4.15}$$

Our curator $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ with any initial dataset $D_0 \in \mathcal{X}^n$ interacts with any update requester $\mathbb{Q}$ as described in Algorithm 3 with publish function $f_{\text{pub}}(\theta) = \theta$.

For this setup, our objective is to provide conditions under which the algorithm pair $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ satisfies objectives **(1.)**, **(2.)**, and **(3.)** as stated in the problem definition § 4.5 and analyze the computational savings of using $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ over $\mathcal{A}_{\text{Noisy-GD}}$ in terms of gradient complexity.

4.5.2 Analysis for Convex Loss Functions

In this subsection, we analyze Noisy-GD based unlearning for convex loss functions $\boldsymbol{\ell}(\theta; \mathbf{x})$ under $L2$ regularization $\mathbf{r}(\theta)$.

Our first goal in this subsection is to prove $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy guarantees on our proposed algorithm pair $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ (in Definition 23) under any non-adaptive r-requester $\mathbb{Q}$. That is, if $(\hat{\Theta}_i)_{i\geq 0}$ is the sequence of models produced by the interaction between $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}}, \mathbb{Q})$ on a $D_0 \in \mathcal{X}^n$, we need to show that their exists a mapping $\pi_i^{\mathbb{Q}}$ such that for all $i \geq 1$ and any $u_i \in \mathcal{U}^r$,

$$\Pr\left[\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) \in S\right] \leq e^{\varepsilon_{\text{dd}}} \cdot \Pr\left[\pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y}\rangle)\right] + \delta, \tag{4.16}$$

for all $S \in \mathcal{Y}$ and $\langle \text{ind}, \mathbf{y}\rangle \in u_i$. Our approach would build on our results on the *dynamics of privacy erosion* introduced in Chapter 3. For convenience, we use the following Rényi variant of our $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy Definition 22.

Definition 24 ((q, ρ_{dd})-deletion privacy)**.** We say that algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies (q, ρ_{dd})-*Rényi deletion privacy under p-adaptive r-requesters* if the following condition holds for all p-adaptive r-requester $\mathbb{Q}$. For every step $i \geq 1$, there exists a randomized mapping $\pi_i^{\mathbb{Q}} : \mathcal{X}^n \to \mathcal{Y}$ such that for all initial datasets $D_0 \in \mathcal{X}^n$,

$$\mathrm{R}_q\left(\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) \middle\| \pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y}\rangle)\right) \leq \rho_{\text{dd}} \quad \text{for all } \langle \text{ind}, \mathbf{y}\rangle \in u_i, \tag{4.17}$$

for all $u_i \in \mathcal{U}^r$ and all $\langle \text{ind}, \mathbf{y}\rangle \in u_i$.

Our approach will be to show that $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ satisfies (q, ρ_{dd})-deletion privacy and then convert the bound in Equation 4.17 to the desired inequality in Equation 4.16 using Theorem 5.

For an arbitrary replacement operation $\langle \text{ind}, \mathbf{y}\rangle$ in u_i, we define a map $\pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y}\rangle) = \hat{\Theta}'_i$, where the model sequence $(\hat{\Theta}'_i)_{i\geq 0}$ is produced by the interaction of

between the same algorithms $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}}, \mathbb{Q})$ but on $D_0 \circ \langle \text{ind}, \mathbf{y} \rangle$. Since non-adaptive requester $\mathbb{Q}$ is equivalent to fixing the edit sequence $(u_i)_{i \geq 1}$ a-priori, note that showing the deletion-privacy guarantee reduces to proving the following Rényi DP bound

$$\mathrm{R}_q \left(\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) \middle\| \bar{\mathcal{A}}(D'_{i-1}, u_i, \hat{\Theta}'_{i-1}) \right) \leq \varepsilon_{\text{dd}}, \tag{4.18}$$

for all $u_{\leq i}$ and for all neighboring datasets D_0, D'_0 s.t. $D'_0 = D_0 \circ \langle \text{ind}, \mathbf{y} \rangle$ with $\langle \text{ind}, \mathbf{y} \rangle \in u_i$.

Note from our Definition 23 that the sequence of models $(\hat{\Theta}_0, \cdots, \hat{\Theta}_i)$ can be seen as being generated from a continuous run of Noisy-GD, where:

1. for iterations $0 \leq k < K_{\mathcal{A}}$, the loss function is $\mathcal{L}_{D_0}$,
2. for the iterations $K_{\mathcal{A}} + (j-1)K_{\bar{\mathcal{A}}} \leq k < K_{\mathcal{A}} + jK_{\bar{\mathcal{A}}}$ on any $1 \leq j \leq i-1$, the loss function is $\mathcal{L}_{D_j}$, and
3. for the iterations $K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}} \leq k < K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}}$, the loss function is $\mathcal{L}_{D_{i-1} \circ u_i}$.

Let $(\Theta_{\eta k})_{0 \leq k \leq K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}}}$ be the sequence representing the intermediate parameters of this extended Noisy-GD run. Similarly, let $(\Theta'_{\eta k})_{k \geq 0}$ be the parameter sequence corresponding to the extended run on the neighbouing dataset D'_0. Since $\langle \text{ind}, \mathbf{y} \rangle \in u_i$, note from the construction that $D'_{i-1} \circ u_i = D_{i-1} \circ u_i$, meaning that the loss functions while processing request u_i is identical for the two processes, i.e. $\mathcal{L}_{D_{i-1} \circ u_i} = \mathcal{L}_{D'_{i-1} \circ u_i}$. For brevity, we refer to the dataset seen in iteration k of the two respective extended runs as $D(k)$ and $D'(k)$ respectively. In short, these two discrete processes induced by Noisy-GD follow the following update rule for any $0 \leq k < K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}}$:

$$\begin{cases} \Theta_{\eta(k+1)} = \Theta_{\eta k} - \eta \nabla \mathcal{L}_{D(k)}(\Theta_{\eta k}) + \sqrt{2\eta\sigma^2}\mathcal{N}(0, I_d), \\ \Theta'_{\eta(k+1)} = \Theta'_{\eta k} - \eta \nabla \mathcal{L}_{D'(k)}(\Theta'_{\eta k}) + \sqrt{2\eta\sigma^2}\mathcal{N}(0, I_d), \end{cases} \tag{4.19}$$

where Θ_0 and Θ'_0 are sampled from same the weight initialization distribution ρ.

To prove the bound in Equation 4.18, we follow the same approach as in Chapter 3, which is interpolating the two discrete stochastic process of Noisy-GD with two piecewise-continuous tracing diffusions Θ_t and Θ'_t in the duration $\eta k < t \leq \eta(k+1)$,

defined as follows.

$$\begin{cases} \Theta_t = \mathbf{T}_k(\Theta_{\eta k}) - (t - \eta k)\mathbf{U}_k(\Theta_{\eta k}) + \sqrt{2\sigma^2}(W_t - W_{\eta k}), \\ \Theta'_t = \mathbf{T}_k(\Theta'_{\eta k}) + (t - \eta k)\mathbf{U}_k(\Theta'_{\eta k}) + \sqrt{2\sigma^2}(W'_t - W'_{\eta k}), \end{cases} \tag{4.20}$$

where W_t, W'_t are two independent Wiener processes, and $\mathbf{T}_k, \mathbf{U}_k$ are two vector maps on $\mathbb{R}^n$ defined as

$$\mathbf{T}_k(\theta) = \theta - \frac{\eta}{2}\left(\nabla \mathcal{L}_{D(k)}(\theta) + \nabla \mathcal{L}_{D'(k)}(\theta)\right), \tag{4.21}$$

and

$$\mathbf{U}_k(\theta) = \frac{1}{2}\left(\nabla \mathcal{L}_{D(k)}(\theta) - \nabla \mathcal{L}_{D'(k)}(\theta)\right) = \frac{1}{2n}\left(\nabla \boldsymbol{\ell}(\theta; D(k)[\text{ind}]) - \nabla \boldsymbol{\ell}(\theta; D'(k)[\text{ind}])\right). \tag{4.22}$$

Recall that Equation 4.20 is identical to Equation 4.19 when $t = \eta(k+1)$, and can be expressed by the following stochastic differential equations (SDEs):

$$\begin{cases} \mathrm{d}\Theta_t = -\mathbf{U}_k(\Theta_{\eta k})\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}W_t \\ \mathrm{d}\Theta'_t = +\mathbf{U}_k(\Theta'_{\eta k})\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}W'_t, \end{cases} \tag{4.23}$$

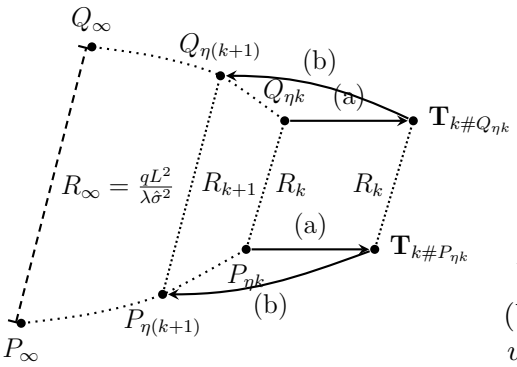
and initial condition $\lim_{t\to\eta k^+}\Theta_t = \mathbf{T}_k(\Theta_{\eta k})$, $\lim_{t\to\eta k^+}\Theta'_t = \mathbf{T}_k(\Theta'_{\eta k})$. These two SDEs can be equivalently described by the following pair of Fokker-Planck equations (cf. Lemma 7).

Lemma 12 (Fokker-Planck equation for SDE Equation 4.23)**.** *Fokker-Planck equation for SDE in Equation 4.23 at time $\eta k < t \le \eta(k+1)$, is*

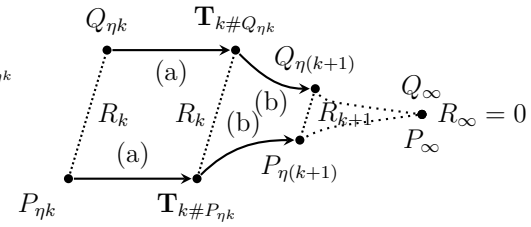
$$\begin{cases} \partial_t P_t(\theta) &= \nabla \cdot \left(P_t(\theta)\mathbb{E}\left[\mathbf{U}_k(\Theta_{\eta k})|\Theta_t = \theta\right]\right) + \sigma^2 \Delta P_t(\theta), \\ \partial_t Q_t(\theta) &= \nabla \cdot \left(Q_t(\theta)\mathbb{E}\left[-\mathbf{U}_k(\Theta'_{\eta k})\middle|\Theta'_t = \theta\right]\right) + \sigma^2 \Delta Q_t(\theta), \end{cases} \tag{4.24}$$

where P_t and Q_t are the densities of Θ_t and Θ'_t respectively.

We provide an overview of how we bound Equation 4.18 in Figure 4.6. Basically, our analysis has two phases; in phase (I) we provide a bound on $\mathrm{R}_q\left(\hat{\Theta}_{i-1}\middle\|\hat{\Theta}'_{i-1}\right)$ that holds for any choice of number of iterations $K_{\mathcal{A}}$ and $K_{\bar{\mathcal{A}}}$, and in phase (II) we prove an exponential contraction in divergence $\mathrm{R}_q\left(\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1})\middle\|\bar{\mathcal{A}}(D'_{i-1}, u_i, \hat{\Theta}'_{i-1})\right)$ with number of iterations $K_{\bar{\mathcal{A}}}$.



(a) Phase I: Processing requests $u_{<i}$, i.e. $0 \le k < K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}}$.



(b) Phase II: Processing request u_i, i.e. $K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}} \le k < K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}}$.

Figure 4.6: Diagram illustrating the technical overview of Theorem 23. Here $P_{\eta k}$ and $Q_{\eta k}$ represent the kth iteration parameter distribution of $\Theta_{\eta k}$ and $\Theta'_{\eta k}$ respectively. We interpolate the two discrete processes in two steps: (a) an identical transformation $\mathbf{T}_k$ (as defined in Equation 4.21, and (b) a diffusion process. If divergence before descent step is $R_k = \mathrm{R}_q\left(P_{\eta k} \| Q_{\eta k}\right)$, the stochastic mapping $\mathbf{T}_k$ in (a) doesn't increase the divergence, while the diffusion (b) either increases it upto an asymptotic constant in phase I or decreases it exponentially to 0 in phase II.

We first introduce a few lemmas that will be used in both phases. The first set of following lemmas show that the transformation $\Theta_{\eta k}, \Theta'_{\eta k} \to \mathbf{T}_k(\Theta_{\eta k}), \mathbf{T}_k(\Theta'_{\eta k})$ preserves the Rényi divergence. To prove this property, we show that $\mathbf{T}_k$ is a differentiable bijective map in Lemma 14 and apply the following Lemma from Vempala and Wibisono [110].

Lemma 13 (Vempala and Wibisono [110, Lemma 15])**.** *If $\mathbf{T} : \mathbb{R}^d \to \mathbb{R}^d$ is a differentiable bijective map, then for any random variables $\Theta, \Theta' \in \mathbb{R}^d$, and for all $q > 0$,*

$$\mathrm{R}_q\left(\mathbf{T}(\Theta) \| \mathbf{T}(\Theta')\right) = \mathrm{R}_q\left(\Theta \| \Theta\right). \tag{4.25}$$

Lemma 14. *If $\boldsymbol{\ell}(\theta; \mathbf{x})$ is a twice continuously differentiable, convex, and β-smooth loss function and regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, then the map $\mathbf{T}_k$ defined in Equation 4.21 is:*

1. *a differentiable bijection for any $\eta < \frac{1}{\lambda+\beta}$, and*
2. *$(1-\eta\lambda)$-Lipschitz for any $\eta \le \frac{2}{2\lambda+\beta}$.*

See § B.3.2 for a proof of Lemma 14. Since $\mathbf{T}_k$ is a differentiable bijective map for small enough learning rate η, it preserves the Rényi divergence as per Lemma 13. However, even if we choose a larger step size, by virtue of the *data-processing inequality* for Rényi divergence, we always have that $\mathrm{R}_q\left(\mathbf{T}_k(\hat{\Theta}_{\eta k})\middle\|\mathbf{T}_k(\hat{\Theta}'_{\eta k})\right) \leq \mathrm{R}_q\left(\hat{\Theta}_{\eta k}\middle\|\hat{\Theta}'_{\eta k}\right)$. The benefit of bijectivity is that it makes the differential inequalities in our proofs a bit neater to solve.

The second set of lemmas presented below describe how $\mathrm{R}_q\left(\Theta_t\|\Theta_t\right)$ evolves with time in both phases I and II. Central to our analysis is the following Lemma 8 from Chapter 3 which bounds the rate of change of Rényi divergence for any pair of diffusion process characterized by their Fokker-Planck equations.

Lemma 15 (Rate of change of Rényi divergence)**.** *Let $\mathbf{V}_t, \mathbf{V}'_t : \mathbb{R}^d \to \mathbb{R}^d$ be two time dependent vector field such that $\max_{\theta\in\mathbb{R}^d} \|\mathbf{V}_t(\theta) - \mathbf{V}'_t(\theta)\|_2 \leq L$ for all $\theta \in \mathbb{R}^d$ and $t \geq 0$. For a diffusion process $(\Theta_t)_{t\geq 0}$ and $(\Theta'_t)_{t\geq 0}$ defined by the Fokker-Planck equations*

$$\begin{cases} \partial_t P_t(\theta) = \nabla \cdot (P_t(\theta)\mathbf{V}_t(\theta)) + \sigma^2 \Delta P_t(\theta) \quad \textit{and} \\ \partial_t Q_t(\theta) = \nabla \cdot (Q_t(\theta)\mathbf{V}'_t(\theta)) + \sigma^2 \Delta Q_t(\theta), \end{cases} \tag{4.26}$$

respectively, where P_t and Q_t are the densities of Θ_t and Θ'_t, the rate of change of Rényi divergence between the two at any $t \geq 0$ is upper bounded as

$$\partial_t \mathrm{R}_q\left(P_t\|Q_t\right) \leq \frac{qL^2}{2\sigma^2} - \frac{q\sigma^2}{2}\frac{\mathrm{I}_q\left(P_t\|Q_t\right)}{\mathrm{E}_q\left(P_t\|Q_t\right)}. \tag{4.27}$$

See § 3.2.1 and § 3.2.2 in Chapter 3 to recap how we derive this rate of change. We apply the above lemma to the Fokker-Planck Equation 4.24 of our pair of tracing diffusion SDE Equation 4.20 and solve the resulting differential inequality to prove the bound in Equation 4.18. To assist our proof, we rely on the following lemma showing that our two tracing diffusions satisfy the Log-Sobolev inequality described in Definition 16.

Lemma 16. *If loss $\boldsymbol{\ell}(\theta; \mathbf{x})$ is convex and β-smooth, regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, and learning rate $\eta \leq \frac{2}{2\lambda+\beta}$, then the tracing diffusion $(\Theta_t)_{0\leq t\leq \eta(K_{\mathcal{A}}+iK_{\bar{\mathcal{A}}})}$ and $(\Theta'_t)_{0\leq t\leq \eta(K_{\mathcal{A}}+iK_{\bar{\mathcal{A}}})}$ defined in Equation 4.20 with $\Theta_0, \Theta'_0 \sim \rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda(1-\eta\lambda/2)} I_d\right)$ satisfy* LSI *inequality with constant $\lambda(1-\eta\lambda/2)$.*

We prove this lemma via induction to show that the alternating application of a contractive map $\mathbf{T}_k$ and Gaussian noise addition preserves the LSI inequality. See the proof of Lemma 16 in § B.3.2 for details.

Remark 7. *Lemma 16 offers a tighter analysis of the constant in the log-Sobolev inequality as compared to Lemma 10 used in Chapter 3, thanks to a tighter analysis of the Lipschitzness in Lemma 14.*

Because the tracing diffusions satisfy the *Log-Sobolev* inequality, we can control the $\mathrm{I}_q/\mathrm{E}_q$ term in Equation 4.27 using the following property.

Lemma 6 (LSI inequality in terms of Rényi Divergence [110])**.** *The distribution Q_t satisfies* LSI(c) *inequality if and only if for all distributions P_t on $\mathbb{R}^d$ such that $\frac{P_t}{Q_t} \in \mathbb{L}^2(Q_t)$ with continuous derivatives $\nabla \frac{P_t}{Q_t}$, and any $q > 1$,*

$$\mathrm{R}_q\left(P_t \| Q_t\right) + q(q-1)\partial_q \mathrm{R}_q\left(P_t \| Q_t\right) \leq \frac{q^2\sigma^2}{2c} \frac{\mathrm{I}_q\left(P_t \| Q_t\right)}{\mathrm{E}_q\left(P_t \| Q_t\right)}. \tag{4.28}$$

The proof of Lemma 6 was covered in § 2.3.1 of Chapter 2.

We are now ready to prove the deletion-privacy bound in Equation 4.18. We do this by solving the privacy erosion dynamics arising on plugging Equation 4.28 in Equation 4.27, where the LSI constant is $c = \lambda(1 - \eta\lambda/2)$ from Lemma 16. On solving this dynamics, we get the following theorem.

Theorem 23 (Deletion guarantee on $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ under convexity)**.** *Let the start distribution be $\rho = \mathcal{N}\left(0, \frac{\hat{\sigma}^2}{\lambda n^2(1-\eta\lambda/2)} I_d\right)$, the loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ be convex, β-smooth, and L-Lipschitz, the regularizer be $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, and learning rate be $\eta < \frac{1}{\lambda+\beta}$. Then algorithm pair $(\mathcal{A}_{\textit{Noisy-GD}}, \bar{\mathcal{A}}_{\textit{Noisy-GD}})$, as defined in Definition 23, satisfies (q, ρ_{dd})-Rényi deletion privacy under all non-adaptive r-requesters for any noise variance $\hat{\sigma}^2 > 0$ and $K_{\mathcal{A}} \geq 0$ if*

$$K_{\bar{\mathcal{A}}} \geq \frac{2}{\eta\lambda} \log\left(\frac{4qL^2}{\lambda\rho_{\mathrm{dd}}\hat{\sigma}^2}\right). \tag{4.29}$$

See the proof of Theorem 23 in § B.3.2 for the details. From (q, ρ_{dd})-Rényi deletion privacy, we can get $(\varepsilon_{\mathrm{dd}}, \delta)$-deletion privacy by simply applying the Rényi divergence to Hockey-stick divergence conversion Theorem 5.

Corollary 3. *Let constants $0 \leq \varepsilon_{\text{dd}} \leq 2\log(1/\delta)$ where $\delta \in (0, 1]$. Under the same conditions as in Theorem 23, the algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy if*

$$K_{\bar{\mathcal{A}}} \geq \frac{2}{\eta\lambda} \log\left(\frac{32L^2\log(1/\delta)}{\lambda{\varepsilon_{\text{dd}}}^2\hat{\sigma}^2}\right). \tag{4.30}$$

Proof. From Mironov [69]'s conversion Theorem 5, note that (q, ρ_{dd})-Rényi deletion privacy where order $q = 1 + \frac{2}{\varepsilon_{\text{dd}}}\log(1/\delta)$ and $\rho_{\text{dd}} = \frac{\varepsilon_{\text{dd}}}{2}$ implies $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy. Since we assume that $\varepsilon_{\text{dd}} \leq 2\log(1/\delta)$, we have that

$$\frac{q}{\rho_{\text{dd}}} = \frac{1 + \frac{2}{\varepsilon_{\text{dd}}}\log(1/\delta)}{\varepsilon_{\text{dd}}/2} = \frac{2(\varepsilon_{\text{dd}} + 2\log(1/\delta))}{{\varepsilon_{\text{dd}}}^2} \leq \frac{8\log(1/\delta)}{{\varepsilon_{\text{dd}}}^2}.$$

So from Theorem 23, it follows that unlearning with $K_{\mathcal{A}} \geq \frac{2}{\eta\lambda}\log\left(\frac{4L^2}{\lambda\hat{\sigma}^2} \cdot \frac{8\log(1/\delta)}{{\varepsilon_{\text{dd}}}^2}\right)$ provides $(1 + \frac{2}{\varepsilon_{\text{dd}}}\log(1/\delta), \frac{\varepsilon_{\text{dd}}}{2})$-Rényi deletion privacy, which implies $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy. □

Utility. Our next goal in this section is to provide utility guarantees for the algorithm pair $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ in form of excess empirical risk bounds. For that, we introduce some additional auxiliary results first. The following Lemma 17 shows that excess empirical risks does not increase too much on replacing r records in a dataset, and the subsequent Lemma 18 provides a convergence guarantee on the excess empirical risk of Noisy-GD algorithm under convexity.

Lemma 17. *Suppose the loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is convex, L-Lipschitz, and β-smooth, and the regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$. Then, the excess empirical risk of any randomly distributed parameter Θ for any dataset $D \in \mathcal{X}^n$ after applying any edit request $u \in \mathcal{U}^r$ that modifies no more than r records is bounded as*

$$\text{err}(\Theta; D \circ u) \leq \left(1 + \frac{\beta}{\lambda}\right)\left[2\ \text{err}(\Theta; D) + \frac{16r^2L^2}{\lambda n^2}\right]. \tag{4.31}$$

Lemma 18 (Accuracy of Noisy-GD)**.** *For convex, L-Lipschitz, and, β-smooth loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ and regularizer $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, if learning rate $\eta < \frac{1}{\lambda+\beta}$, the excess empirical risk of $\Theta_{\eta K} = \text{Noisy-GD}(D, \Theta_0, K) \in \mathbb{R}^d$ for any $D \in \mathcal{X}^n$ is bounded as*

$$\text{err}(\Theta_{\eta K}; D) \leq \text{err}(\Theta_0; D)e^{-\lambda\eta K/2} + \left(1 + \frac{\beta}{\lambda}\right)\frac{d\hat{\sigma}^2}{n^2}. \tag{4.32}$$

We provide proof of Lemma 17 and proof of Lemma 18 in § B.3.2.

Finally, we are ready to combine all the pieces into Theorem 24, showing that the algorithm pair $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ solves the data-deletion problem as described in § 4.5. We basically combine the DP guarantee on Noisy-GD in Chapter 3, non-adaptive deletion-privacy guarantee in Theorem 23, and prove excess empirical risk bound using Lemma 18 and Lemma 17.

Theorem 24 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda, \beta, L > 0$, constant $q > 1$, and constants $\rho_{\text{dp}} \geq \rho_{\text{dd}} > 0$. Define constant $\nu = \frac{\lambda+\beta}{\lambda}$. Let the loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ be twice differentiable, convex, L-Lipschitz, and β-smooth, and let the regularizer be $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$. If the learning rate is $\eta = \frac{1}{2(\lambda+\beta)}$, the gradient noise variance is $\hat{\sigma}^2 = \frac{4qL^2}{\lambda\rho_{\text{dp}}}$, and the weight initialization distribution is $\rho = \mathcal{N}\left(0, \frac{\hat{\sigma}^2}{\lambda n^2(1-\eta\lambda/2)} I_d\right)$, then*

(1.) *both $\mathcal{A}_{\text{Noisy-GD}}$ and $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ are (q, ρ_{dp})-Rényi DP for any $K_{\mathcal{A}}, K_{\bar{\mathcal{A}}} \geq 0$,*

(2.) *pair $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ satisfies (q, ρ_{dd})-deletion-privacy all non-adaptive r-requesters*

$$\text{if} \quad K_{\bar{\mathcal{A}}} \geq 4\nu \log \frac{\rho_{\text{dp}}}{\rho_{\text{dd}}}, \tag{4.33}$$

(3.) *and all models in sequence $(\hat{\Theta}_i)_{i\geq 0}$ produced by the interactions between $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ and $\mathbb{Q}$ on any $D_0 \in \mathcal{X}^n$, where $\mathbb{Q}$ is any r-requester, have an excess empirical risk $\text{err}(\hat{\Theta}_i; D_i) = O\left(\frac{qd}{\rho_{\text{dp}} n^2}\right)$*

$$\text{if} \quad K_{\mathcal{A}} \geq 4\nu \log\left(\frac{\rho_{\text{dp}} n^2}{4qd}\right), \quad \text{and} \quad K_{\bar{\mathcal{A}}} \geq 4\nu \log \max\left\{5\nu, \frac{8\rho_{\text{dp}} r^2}{qd}\right\}. \tag{4.34}$$

The proof of Theorem 24 is also provided in § B.3.2. We remark that, the weight initialization distribution ρ in this theorem can be simply set to *any point-distribution* (i.e., $\Theta_0 = \theta_0$ for some fixed $\theta_0 \in \mathbb{R}^d$) without affecting the guarantees of Theorem 24. More precisely, any start distribution $\rho = \mathcal{N}(\mu, \sigma_0^2 I_d)$ with an arbitrary mean $\mu \in \mathbb{R}^d$ and variance $\sigma_0^2 \leq \frac{\hat{\sigma}^2}{\lambda n^2(1-\eta\lambda/2)}$ suffices. For discussions, we provide the following corollary that expresses Theorem 24 in terms of $(\varepsilon_{\text{dp}}, \delta)$-differential and $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy guarantees.

Corollary 4. *Let constants $\lambda, \beta, L > 0$, constants $0 \leq \varepsilon_{\text{dd}} \leq \varepsilon_{\text{dp}} \leq 2\log(1/\delta)$, and loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ and learing rate η satisfy the conditions of Theorem 24. If the gradient noise variance is $\hat{\sigma}^2 = \frac{16L^2 \log(1/\delta)}{\lambda {\varepsilon_{\text{dp}}}^2}$, then*

(1.) *both $\mathcal{A}_{\textit{Noisy-GD}}$ and $\bar{\mathcal{A}}_{\textit{Noisy-GD}}$ are $(\varepsilon_{\text{dp}}, \delta)$-Rényi DP for any $K_{\mathcal{A}}, K_{\bar{\mathcal{A}}} \geq 0$,*

(2.) *pair $(\mathcal{A}_{\textit{Noisy-GD}}, \bar{\mathcal{A}}_{\textit{Noisy-GD}})$ satisfies $(\varepsilon_{\text{dd}}, \delta)$-deletion-privacy all non-adaptive r-requesters*

$$\textit{if} \quad K_{\bar{\mathcal{A}}} \geq 8\nu \log \frac{\varepsilon_{\text{dp}}}{\varepsilon_{\text{dd}}}, \tag{4.35}$$

(3.) *and all models in sequence $(\hat{\Theta}_i)_{i \geq 0}$ produced by the interactions between $(\mathcal{A}_{\textit{Noisy-GD}}, \bar{\mathcal{A}}_{\textit{Noisy-GD}})$ and $\mathbb{Q}$ on any $D_0 \in \mathcal{X}^n$, where $\mathbb{Q}$ is any r-requester, have an excess empirical risk $\operatorname{err}(\hat{\Theta}_i; D_i) = O\left(\frac{d\log(1/\delta)}{{\varepsilon_{\text{dp}}}^2 n^2}\right)$*

$$\textit{if} \quad K_{\mathcal{A}} \geq 4\nu \log\left(\frac{{\varepsilon_{\text{dp}}}^2 n^2}{16 d \log(1/\delta)}\right), \quad \textit{and} \quad K_{\bar{\mathcal{A}}} \geq 4\nu \log \max\left\{5\nu, \frac{2{\varepsilon_{\text{dp}}}^2 r^2}{d\log(1/\delta)}\right\}. \tag{4.36}$$

Proof. We will use a fact that for any two distributions P, Q, if $\mathrm{R}_q(P \| Q) \leq q \cdot \kappa$ for all $q > 1$, then for all S, we have $P(S) \leq e^{\varepsilon} Q(S) + \delta$ where $\varepsilon = 2\sqrt{\kappa \log(1/\delta)}$. We prove this result formally later in Chapter 5, but for now let's just use it.

Let $0 < \kappa_{\text{dd}} < \kappa_{\text{dp}}$ be two arbitrary constants. If we set $\rho_{\text{dp}} = q \cdot \kappa_{\text{dp}}$ and $\rho_{\text{dd}} = q \cdot \kappa_{\text{dd}}$ in Theorem 23, then all they hyper-paramters, including the bounds on $K_{\mathcal{A}}$ and $K_{\bar{\mathcal{A}}}$ become a constant in terms of the q. In other words, we get $(q, q \cdot \kappa_{\text{dp}})$-Rényi DP and $(q, q \cdot \kappa_{\text{dd}})$-deletion privacy *for all* $q > 1$ if $\hat{\sigma}^2 = \frac{4L^2}{\lambda \kappa_{\text{dp}}}$ and

$$K_{\mathcal{A}} \geq 4\nu \log\left(\max\left\{\frac{\kappa_{\text{dp}}}{\kappa_{\text{dd}}}, \frac{\kappa_{\text{dp}} n^2}{4d}\right\}\right), \quad \text{and} \quad K_{\bar{\mathcal{A}}} \geq 4\nu \log\left(\max\left\{5\nu, \frac{8\kappa_{\text{dp}} r^2}{d}\right\}\right).$$

Since we want $(\varepsilon_{\text{dp}}, \delta)$-DP and $(\varepsilon_{\text{dd}}, \delta)$-Rényi deletion privacy, we can set $\kappa_{\text{dp}} = \frac{{\varepsilon_{\text{dp}}}^2}{4\log(1/\delta)}$ and $\kappa_{\text{dd}} = \frac{{\varepsilon_{\text{dd}}}^2}{4\log(1/\delta)}$ and use the fact stated in the start. Plugging in these values results in the bounds in theorem statement. □

Our utility upper bound in Theorem 24 matches the theoretical lower bound of $\Omega(\min\left\{1, \frac{d}{{\varepsilon_{\text{dp}}}^2 n^2}\right\})$ by Bassily et al. [12] on the optimal empirical risk attainable by any (ε, δ)-DP algorithms on Lipschitz, smooth, strongly-convex loss functions (cf. Theorem 2). Consequently, our unlearning algorithm $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ maintains the optimal privacy-utility trade-offs of retraining with $\mathcal{A}_{\text{Noisy-GD}}$, while being a lot

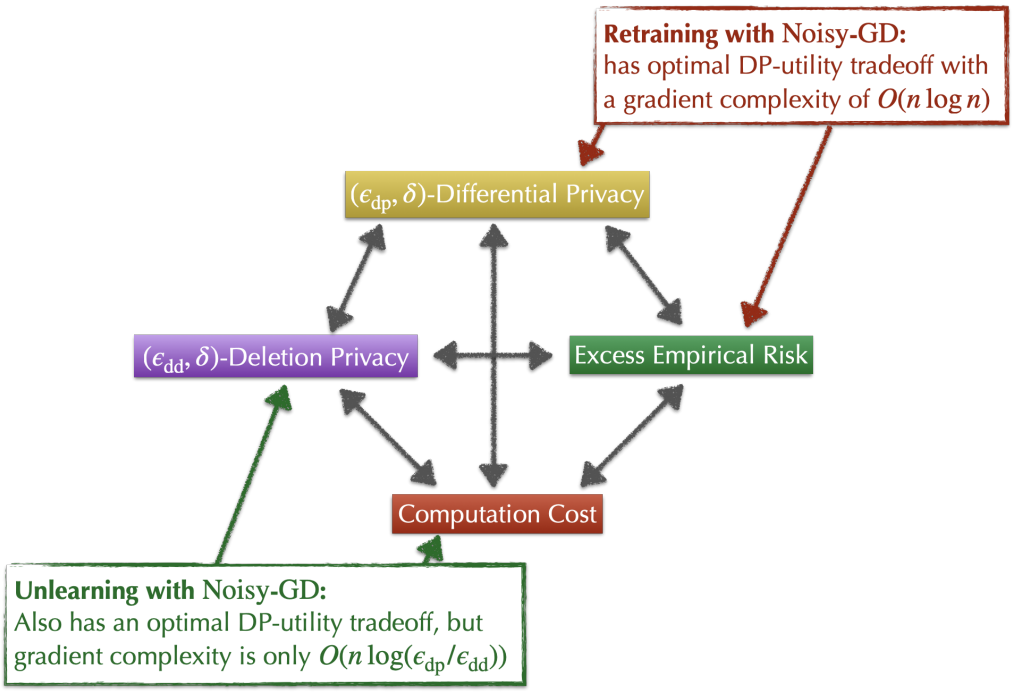


Figure 4.7: Fine-tuning (or unlearning) using Noisy-GD in the setting of streaming edit requests provides optimal DP-utility trade-off, certified deletion guarantee and only takes a fraction of computation cost of full-retraining with Noisy-GD for reaching the same DP & utility level.

cheaper. Figure 4.7 captures this *four-way trade-off* between differential privacy, deletion privacy, utility and computation complexity. It is worth noting that in order to satisfy $(\varepsilon_{\text{dp}}, \delta)$-differential privacy and $(\varepsilon_{\text{dd}}, \delta)$-deletion privacy for non-adaptive r-requesters, the necessary number of iterations $K_{\bar{\mathcal{A}}}$ remains constant regardless of the size, r, of the deletion batch, depending solely on the ratio $\frac{\varepsilon_{\text{dd}}}{\varepsilon_{\text{dp}}}$. However, the number of iterations required to ensure optimal utility under DP increases as r grows. Importantly, when deletion batches are sufficiently small, specifically when $r \leq \frac{\sqrt{d \log(1/\delta)}}{2\varepsilon_{\text{dd}}}$, performing an adequate number of unlearning iterations to satisfy the deletion privacy guarantee is also sufficient to ensure optimal utility of the unlearned model.

In Table 4.1, we provide a comparison of our Noisy-GD based deletion solution with some prior approaches for unlearning. More specifically, we compare the compute savings in terms of gradient complexity of the learning algorithm $\mathcal{A}$ minus that of the respective unlearning algorithm $\bar{\mathcal{A}}$ under identical (ε, δ)-unlearning (Definition 21) constraints. To make the comparison fair, we make the same assumptions

Unlearning Algorithm	**Requires secret states?**	**Compute savings for i^{th} edit**
Noisy-m-A-SGD [105]	No	$\Omega\left(\sqrt{d}\left(1-\frac{\sqrt{d}}{n}\right)\right)$
Perturbed-GD [72]	Yes	$\Omega\left(n\log\left(\frac{\varepsilon n}{\sqrt{d}}\right)\right)$
Perturbed-GD [72]	No	$\Omega\left(n\log\left(\frac{\varepsilon n}{\log^2(id)\sqrt{d}}\right)\right)$
Noisy-GD [Thm. 4, Ours]	No	$\Omega\left(n\log\min\left\{n, \frac{\varepsilon n}{\sqrt{d}}\right\}\right)$

Table 4.1: Comparisons of the computation savings in gradient complexity per edit request along with requirement of secret states with prior unlearning algorithms. Edit requests are non-adaptive and modify $r = 1$ record in n-sized datasets. We assume the loss $\boldsymbol{\ell}(\theta; \mathbf{x})$ of models in $\mathbb{R}^d$ to be convex, 1-Lipschitz, and $O(1)$-smooth, and $L2$ regularization constant to be $O(1)$. For a fair comparison, we require that each of them satisfy (ε, δ)-unlearning guarantee and have the same excess empirical risk $\text{err}(\Theta_i; D_i) = O(1)$ for all unlearning steps $i \geq 0$.

of $O(1)$-smoothness, $O(1)$-convexity and $O(1)$-Lipschitzness on the loss functions.

4.5.3 Analysis for Non-Convex Loss Functions

In this subsection, we study deletion using Noisy-GD when the loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is non-convex (but bounded and with $L2$ regularization). Suppose $D_0 \in \mathcal{X}^n$ is an arbitrary dataset, $\mathbb{Q}$ is any non-adaptive r-requester, and $(\hat{\Theta}_i)_{i\geq 0}$ is the model sequence generated by the interaction of $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}}, \mathbb{Q})$. Our first goal will be to prove (q, ρ_{dd})- Rényi deletion privacy guarantee on $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ and we will later use it for arguing utility as well. Recall from Definition 24 that to prove (q, ρ_{dd})-deletion privacy, we need to construct a map $\pi_i^{\mathbb{Q}} : \mathcal{X}^n \to \mathcal{Y}$ such that for all $i \geq 1$ and any $u_i \in \mathcal{U}^r$,

$$\mathrm{R}_q\left(\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) \middle\| \pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y}\rangle)\right) \leq \rho_{\text{dd}} \quad \text{for all } \langle \text{ind}, \mathbf{y}\rangle \in u_i. \tag{4.37}$$

Our construction of $\pi_i^{\mathbb{Q}}$ for this proof is completely different from the one described in preceeding subsection § 4.5.2. As discussed in point (7.) in § 4.3, since $\mathbb{Q}$ is non-adaptive, it suffices to show that there exists a map $\pi : \mathcal{X}^n \to \mathcal{Y}$ such that for all $i \geq 1$,

$$\mathrm{R}_q\left(\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) \middle\| \pi(D_i)\right) \leq \rho_{\text{dd}}, \tag{4.38}$$

for all $D_0 \in \mathcal{X}^n$ and all edit sequences $(u_i)_{i\geq 1}$ from $\mathcal{U}^r$.

Our mapping of choice for the purpose is the Gibbs distribution with the following density:

$$\pi(D)(\theta) \propto e^{-\mathcal{L}_D(\theta)/\sigma^2}. \tag{4.39}$$

The high-level intuition for this construction is that Noisy-GD can be interpreted as Unadjusted Langevin Algorithm (ULA) [81], which is a discretization of the Langevin diffusion (described in Equation 2.31) that eventually converges to this Gibbs distribution (see § 2.3 for a quick refresher). However, showing a convergence for ULA (in indistinguishability notions like Rényi divergence) to this Gibbs distribution, especially in form of non-asymptotic bounds on the mixing time and discretization error has been a long-standing open problem. Recent breakthrough results by Vempala and Wibisono [110] followed by Chewi et al. [29] resolved this problem with an elegant argument, relying solely on isoperimetric assumptions over Equation 4.39 that hold for non-convex losses. Our deletion-privacy argument leverages this rapid convergence result to basically show that once Noisy-GD reaches near-indistinguishability to its Gibbs mixing distribution, maintaining indistinguishability to subsequent Gibbs distribution corresponding to database edits require much fewer Noisy-GD iterations than fresh retraining (i.e. data deletion is faster than retraining).

We start by presenting Chewi et al. [29]'s convergence argument adapted to our Noisy-GD formulation, with a *tighter analysis that results in a* $\log(q)$ *improvement in the discretization error* over the original. Consider the discrete stochastic process $(\Theta_{\eta k})_{0\leq k\leq K}$ induced by parameter update step in Noisy-GD Algorithm 4 when run for K iterations on a database D with an arbitrary start distribution $\Theta_0 \sim P_0$. We interpolate each discrete update from $\Theta_{\eta k}$ to $\Theta_{\eta(k+1)}$ via a diffusion process Θ_t defined over time $\eta k \leq t \leq \eta(k+1)$ as

$$\Theta_t = \Theta_{\eta k} - (t - \eta k)\nabla\mathcal{L}_D(\Theta_{\eta k}) + \sqrt{2\sigma^2}(W_t - W_{\eta k}), \tag{4.40}$$

where $(W_t)_{t\geq 0}$ is a Wiener process and, as before, $\sigma = \hat{\sigma}/n$ (cf. Remark 4). Note that if $\Theta_{\eta k}$ models the parameter distribution after the k^{th} update, then $\Theta_{\eta(k+1)}$ models the parameter distribution after the $k + 1^{th}$ update. On repeating this construction for all $k = 0, \cdots, K$, we get a *tracing diffusion* $\{\Theta_t\}_{t\geq 0}$ for Noisy-GD (which is different from Equation 4.20). We denote the distribution of random variable Θ_t

with P_t. The tracing diffusion during the duration $\eta k \leq t \leq \eta(k+1)$ is characterized by the following Fokker-Planck equation.

Lemma 19 (Proposition 14 [29]). *For tracing diffusion Θ_t defined in Equation 4.40, the equivalent Fokker-Planck equation in the interval $\eta k \leq t \leq \eta(k+1)$ is*

$$\partial_t P_t(\theta) = \nabla \cdot \left(\left\{ \mathbb{E}\left[\nabla \mathcal{L}_D(\Theta_{\eta k}) - \nabla \mathcal{L}_D(\Theta_t) | \Theta_t = \theta\right] + \sigma^2 \nabla \log \frac{P_t(\theta)}{\pi(D)(\theta)} \right\} P_t(\theta) \right), \tag{4.41}$$

where $\pi(D)$ is the Gibbs distribution defined in Equation 4.39.

The proof of Lemma 19 is provided in § B.3.3. This expression of the Fokker-Planck equation allows us to directly track the *effect of discretization* in the dynamics of convergence. The following lemma provides a partial differential inequality that bounds the rate of change in Rényi divergence $\mathrm{R}_q\left(P_t \| \pi(D)\right)$ using this Fokker-Planck equation.

Lemma 20 ([29, Proposition 15]). *Let $Q_t \overset{def}{=} P_t/\pi(D)$ where $\pi(D)$ is the Gibbs distribution defined in Equation 4.39 and $\psi_t \overset{def}{=} Q_t^{q-1}/\mathrm{E}_q\left(Q_t \| \pi(D)\right)$. The rate of change in $\mathrm{R}_q\left(P_t \| \pi(D)\right)$ along racing diffusion in time $\eta k \leq t \leq \eta(k+1)$ is bounded as*

$$\partial_t \mathrm{R}_q\left(P_t \| \pi(D)\right) \leq -\frac{3q\sigma^2}{4} \frac{\mathrm{I}_q\left(P_t \| \pi(D)\right)}{\mathrm{E}_q\left(P_t \| \pi(D)\right)} + \frac{q}{\sigma^2} \mathbb{E}\left[\psi_t(\Theta_t) \left\| \nabla \mathcal{L}_D(\Theta_{\eta k}) - \nabla \mathcal{L}_D(\Theta_t) \right\|_2^2\right]. \tag{4.42}$$

The proof of Lemma 20 is provided in § B.3.3. The next step is to solve the PDI Equation 4.42 to get a convergence bound for Noisy-GD. To help in that, we first introduce the change of measure inequalities shown in Chewi et al. [29].

Lemma 21 (Change of measure inequality [29]). *If $\boldsymbol{\ell}(\theta; \mathbf{x})$ is β-smooth, and regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, then for any probability density P on $\mathbb{R}^d$,*

$$\mathop{\mathbb{E}}_{P}\left[\|\nabla \mathcal{L}_D\|_2^2\right] \leq 4\sigma^4 \mathop{\mathbb{E}}_{\pi(D)}\left[\left\|\nabla \sqrt{\frac{P}{\pi(D)}}\right\|_2^2\right] + 2d\sigma^2(\beta + \lambda), \tag{4.43}$$

where $\pi(D)$ is the Gibbs distribution defined in Equation 4.39.

For the sake of completeness, we provide proof of Lemma 21 in § B.3.3. Another change in measure inequality needed for the proof is the following well-known Donsker-Varadhan variational principle.

Lemma 22 (Donsker-Varadhan Variational principle [35])**.** *If P and Q are two distributions on $\mathbb{R}^d$ such that $P \ll Q$, then for all functions $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathop{\mathbb{E}}_{\Theta\sim P}[f(\Theta)] \leq \mathrm{KL}\left(P\|Q\right) + \log \mathop{\mathbb{E}}_{\Theta'\sim Q}\left[\exp(f(\Theta'))\right]. \tag{4.44}$$

We are now ready to prove the rate of convergence guarantee for Noisy-GD following Chewi et al. [29]'s method, but with a more refined analysis that leads to a improvement of $\log q$ factor in the discretization error (compared to the original [29, Theorem 4]).

Theorem 25 (Convergence of Noisy-GD in Rényi divergence)**.** *Let constants $\beta, \lambda, \sigma^2 > 0$ and $q, B > 1$. Suppose the loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is $(\sigma^2 \log(B)/4)$-bounded and β-smooth, and regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$. If step size is $\eta \leq \frac{\lambda}{64Bq^2(\beta+\lambda)^2}$, then for any database $D \in \mathcal{X}^n$ and any weight initialization distribution P_0 for Θ_0, the Rényi divergence of distribution $P_{\eta K}$ of output model $\Theta_{\eta K} = \text{Noisy-GD}(D, \Theta_0, K)$ with respect to the Gibbs distribution $\pi(D)$ defined in Equation 4.39 shrinks as follows:*

$$\mathrm{R}_q\left(P_{\eta K}\|\pi(D)\right) \leq q \exp\left(-\frac{\lambda\eta K}{2B}\right) \mathrm{R}_q\left(P_0\|\pi(D)\right) + \frac{32d\eta qB(\beta+\lambda)^2}{\lambda}. \tag{4.45}$$

See Appendix § B.3.3 for the proof of Theorem 25. We will use Theorem 25 for proving the deletion-privacy and utility guarantee on the pair $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$. For that, we need the following additional result that shows that Gibbs distributions enjoy strong indistinguishability on bounded perturbations to its potential function (which is basically why the exponential mechanism satisfies $(\varepsilon, 0)$-DP [114, 41]).

Lemma 23 (Indistinguishability under bounded perturbations)**.** *For two potential functions $\mathcal{L}, \mathcal{L}' : \mathbb{R}^d \to \mathbb{R}$ and some constant σ^2, let $P \propto e^{-\mathcal{L}/\sigma^2}$ and $Q \propto e^{-\mathcal{L}'/\sigma^2}$ be the respective Gibbs distributions. If $|\mathcal{L}(\theta) - \mathcal{L}'(\theta)| \leq c$ for all $\theta \in \mathbb{R}^d$, then $\mathrm{R}_q\left(P\|Q\right) \leq \frac{2c}{\sigma^2}$ for all $q > 1$.*

See proof of Lemma 23 for details. In the Theorem 27 that follows next, we show that $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ solves the data-deletion problem described in Section § 4.5 even for non-convex losses. Our proof uses the convergence Theorem 25 and indistinguishability for bounded perturbation Lemma 23 to show that the unlearning algorithm $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ can consistently produce models indistinguishable to

the corresponding Gibbs distribution Equation 4.39 in the online setting at a fraction of computation cost of $\mathcal{A}_{\text{Noisy-GD}}$. As discussed in pt. (7) in Section § 4.3, such an indistinguishability is sufficient for ensuring deletion-privacy for non-adaptive requests. As for adaptive requests, the well-known Rényi DP guarantee in Theorem 39 combined with our reduction Theorem 21 offers a deletion-privacy guarantee for $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ under adaptivity.

Our proof of accuracy for the unlearned models leverages the fact that Gibbs distribution Equation 4.39 is an almost excess risk minimizer as shown in the following Theorem 26. Since our deletion-privacy guarantee is based on near-indistinguishability to Equation 4.39, this property also ensures near-optimal excess risk of unlearned models.

Theorem 26 (Near optimality of Gibbs sampling)**.** *If the loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is $\sigma^2 \log(B)/4$-bounded and β-smooth, the regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2} \|\theta\|_2^2$, then the excess empirical risk for a model $\bar{\Theta} \in \mathbb{R}^d$ sampled from the Gibbs distribution $\pi(D) \propto e^{-\mathcal{L}_D/\sigma^2}$ is*

$$\operatorname{err}(\bar{\Theta}; D) = \mathbb{E}\left[\mathcal{L}_D(\bar{\Theta}) - \mathcal{L}_D(\theta_D^*)\right] \leq \frac{d\sigma^2}{2}\left(\log\frac{\beta+\lambda}{\lambda} + \sqrt{B}\right). \tag{4.46}$$

See proof of Theorem 26 in § B.3.3. Following is the main theorem that combines all the ingredients in this section together.

Theorem 27 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda, \beta, L, \sigma^2, \eta > 0$, constants $q, B > 1$, and constants $d > \rho_{\text{dp}} \geq \rho_{\text{dd}} > 0$. Let the loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ be $\frac{\sigma^2 \log(B)}{4}$-bounded, L-Lipschitz and β-smooth, the regularizer be $\mathbf{r}(\theta) = \frac{\lambda}{2} \|\theta\|_2^2$, and the weight initialization distribution be $\rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda} I_d\right)$. Then,*

(1.) *both $\mathcal{A}_{\text{Noisy-GD}}$ and $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ are (q, ρ_{dp})-Rényi DP for any $\eta \geq 0$ and any $K_{\mathcal{A}}, K_{\bar{\mathcal{A}}} \geq 0$ if*

$$\sigma^2 \geq \frac{qL^2}{\rho_{\text{dp}} n^2} \cdot \eta \max\{K_{\mathcal{A}}, K_{\bar{\mathcal{A}}}\}, \tag{4.47}$$

(2.) *pair $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ satisfies (q, ρ_{dd})-Rényi deletion privacy under all non-adaptive r-requesters for any $\sigma^2 > 0$, if learning rate is $\eta \leq \frac{\lambda \rho_{\text{dd}}}{64 dq B(\beta+\lambda)^2}$*

and the number of iterations satisfy

$$\begin{aligned} K_{\mathcal{A}} &\geq \frac{2B}{\lambda\eta}\log\left(\frac{q\log(B)}{\rho_{\mathrm{dd}}}\right), \text{ and} \\ K_{\bar{\mathcal{A}}} &\geq K_{\mathcal{A}} - \frac{2B}{\lambda\eta}\log\left(\frac{\log(B)}{2\left(\rho_{\mathrm{dd}} + \frac{r}{n}\log(B)\right)}\right), \end{aligned} \tag{4.48}$$

(3.) *and all models in the sequence* $(\hat{\Theta}_i)_{i\geq 0}$ *produced by interactions between* $(\mathcal{A}_{\textit{Noisy-GD}}, \bar{\mathcal{A}}_{\textit{Noisy-GD}})$ *and* $\mathbb{Q}$ *on any* $D_0 \in \mathcal{X}^n$*, where* $\mathbb{Q}$ *is an* r*-requester, have an excess empirical risk* $\mathrm{err}(\hat{\Theta}_i; D_i) = \tilde{O}\left(\frac{dq}{\rho_{\mathrm{dp}} n^2} + \frac{1}{n}\sqrt{\frac{q\rho_{\mathrm{dd}}}{\rho_{\mathrm{dp}}}}\right)$ *when inequalities in Equation 4.48 and Equation 4.47 are equalities.*

The proof of Theorem 27 is also provided in § B.3.3. Previous studies on unlearning under non-convexity mainly focused on empirical analysis for utility. To the best of our knowledge, we are the first to offer utility guarantees in this context. Furthermore, our non-convex utility bound only exceeds the optimal privacy-preserving utility under convexity by approximately $\tilde{O}\left(\frac{1}{n}\sqrt{\frac{q\rho_{\mathrm{dd}}}{\rho_{\mathrm{dp}}}}\right)$. This term becomes negligible when dealing with large databases or a small deletion privacy to differential privacy budget ratio.

Our results show a significant computational advantage on unlearning with $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ in scenarios where the proportion of edited records in a single edit request satisfies $\frac{r}{n} \leq \frac{1}{2} - \frac{\rho_{\mathrm{dd}}}{\log B}$. For instance, in the deletion regime where we desire $\rho_{\mathrm{dd}} = \log(B)/4$, employing $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ instead of retraining with $\mathcal{A}_{\text{Noisy-GD}}$ requires $\Omega(dn\log\frac{n}{r})$ fewer gradient steps.

Remark 8. *Both Theorem 24 and Theorem 27 also hold when gradients* $\nabla\boldsymbol{\ell}(\theta; \mathbf{x})$ *are clipped to* L *instead of assuming* L*-Lipschitzness. Appendix § B.4 discusses how gradient clipping is compatible with other assumptions we make.*

4.6 Discussions and Future Directions

We showed that existing unlearning certifications in literature are unreliable in real-world scenarios, mainly due to their failure in handling adaptive deletion requests and because they permit unchecked propagation of deleted information through internal states as they only measure indistinguishability for outputs. To

mitigate these, we proposed a new deletion guarantee that safeguards the "Right to be Forgotten" in a provably secure way. We also showed the importance of protecting the privacy of existing records in order to ensure privacy of deleted records under adaptive deletions, and established connections between deletion privacy and differential privacy. Our results on a Noisy-GD based unlearning algorithm show a significant computation saving compared to retraining at no loss in utility, for both convex and non-convex losses.

Future work. Our guarantees on Noisy-GD relies on assumption of smoothness which we believe can be eliminated by not relying on isoperimetry based analysis, but using the techniques developed in Bok et al. [18]. Additionally, evaluating our bounds experimentally as well as testing the performance of Noisy-GD for deep learning can be an opportunity for future research. We also leave extending the analysis for Noisy-SGD, which is a non-trivial and challenging problem, as part of future work.

Chapter 5

Laplace Interpretation of Privacy

In this chapter we propose a set of useful expressions of differential privacy notions in terms of Laplace transformations. Its bare form expression appears in several related works on analyzing DP, either as an integral or an expectation. We show that recognizing the expression as a Laplace transform unlocks a new way to reason about differential privacy properties by exploiting the duality between time and frequency domains. Leveraging our interpretation, we connect the Rényi divergence $\mathrm{R}_q\left(P\|Q\right)$ as a function of order q and the (ε,δ)-DP profile $\delta_{P|Q}(\varepsilon)$ as a function of ε for any two distributions P,Q as being the Laplace and inverse-Laplace transforms of one another. This connection shows that not only is Rényi divergence is well-defined for complex orders $q=\gamma+i\omega$, complex orders play a pivotal role in functional equivalence between privacy profile $\delta_{P|Q}(\varepsilon)$ and Rényi divergence curve $\mathrm{R}_q\left(P\|Q\right)$. Using our Laplace transform-based analysis, we prove an *exactly tight* (i.e., matching even in the constants) adaptive composition theorem for (ε,δ)-DP point guarantees as well as $(\varepsilon,\delta_{P|Q}(\varepsilon))$-DP curve guarantees. Additionally, we highlight that the notion of f-DP has a symmetry problem that prevents it from being a tight characterizing of DP. The same problem arises when considering dominating privacy profiles. We fix the issue by showing that all four functional notions of DP divergences, namely privacy profile, Rényi divergence profile, privacy loss distribution, and trade-off curve, are equivalent to one another and also enjoy reversal properties, which eliminates the need to enforce symmetry.

To summarize, we propose

> *...a novel interpretation of differential privacy that improves our understanding of it and offers multiple analytical benefits.*

5.1 Motivation

Differential privacy (DP) [37] has become a widely adopted standard for quantifying privacy of algorithms that process statistical data. In simple terms, differential privacy bounds the influence a single data-point may have on the outcome probabilities. Being a statistical property, the design of differentially private algorithms involves a *pen-and-paper analysis* of any randomness internal to the processing that obscures the influence a data-point might have on its output. A clear understanding of the nature of differential privacy notions is therefore tantamount to study and design of privacy-preserving algorithms.

Throughout its exploration, various functional interpretations of the concept of differential privacy have emerged over the years. These include the privacy-profile curve $\delta(\epsilon)$ [10] that traces the (ϵ, δ)-DP point guarantees, the f-DP [34] view of worst-case *trade-off curve* between type I and type II errors for hypothesis testing membership [56, 11], the Rényi DP [69] function of order q that admits a natural analytical composition [1, 69], the view of the *privacy loss distribution (PLD)* [96] that allows for approximate numerical composition [59, 47], and the recent *characteristic function formulation* of the dominating privacy loss random variables Zhu et al. [122]. Each of these formalisms have their own properties and use-cases, and none of them seem to be superior in all aspects.

Regardless of their differences, they all have some shared difficulties—certain types of manipulations on them are harder to perform in the time-domain, but considerably simpler to do in the frequency-domain. For instance, Koskela et al. [59] noted that composing PLDs of two mechanisms involve convolving their probability densities, which can be numerically approximated efficiently by multiplying their Discrete Fast-Fourier Transformations (DFFT) and then inverting it back to get the convolved density using Inverse-DFFT. Such maneuvers are also frequently performed for analytical reasons while proving properties of differential privacy, often without even realizing this detour through the frequency-domain. A notable example of this is the analysis of Moments' accountant by Abadi et al. [1], where the authors bound higher-order moments of subsampled Gaussian distributions, compose the moments through multiplication, and then derive the (ε, δ)-DP bound on the DP-SGD mechanism. Their analysis goes the through frequency space,

as the moment generating function of a random variable corresponds to the two-sided Laplace transform of its probability density function [68]. Often times when dealing with a functional notion of DP, expressing components like expectations or cumulative densities their integral form ends up being a Fourier or a Laplace transform. Realizing them as such can be tremendously useful in analysis.

In this chapter, we formalize these time-frequency domain dualisms enjoyed by the functional representations into a new interpretation of differential privacy. In addition to augmenting existing perspectives on DP, this interpretation provides a flexible analytical toolkit that greatly extends our cognitive reach in reasoning about DP and its underpinnings. This interpretation is based on recognizing that the privacy-profile $\delta_{P|Q}(\varepsilon) := \sup_S P(S) - e^{\varepsilon} \cdot Q(S)$ and the Rényi-divergence $\mathrm{R}_q\left(P\|Q\right) := \frac{1}{q-1}\int_\Omega P^q Q^{1-q}\mathrm{d}\theta$ between any two distributions P, Q on the same space Ω can be seen as a Laplace transform[1] of the privacy loss distribution $\mathrm{PLD}(P\|Q)$, the distribution of privacy loss random variable $Z = L_{P|Q}(\Theta)$ where $\Theta \sim P$:

$$\forall \varepsilon \in \mathbb{R} \;:\; \delta_{P|Q}(\varepsilon) = \mathop{\mathbb{E}}_{Z\leftarrow \mathrm{PLD}(P\|Q)}\left[\max\{0, 1 - e^{\varepsilon - Z}\}\right] = \mathcal{L}\left\{1 - F_Z(t+\varepsilon)\right\}(1), \tag{5.1}$$

$$\forall q \in \mathbb{C} \;:\; e^{(q-1)\cdot \mathrm{R}_q(P\|Q)} = \mathop{\mathbb{E}}_{Z\leftarrow \mathrm{PLD}(P\|Q)}\left[e^{(q-1)\cdot Z}\right] = \mathcal{B}\left\{f_Z(t)\right\}(1-q), \tag{5.2}$$

where $F_Z(t) = \Pr[Z < t]$ is the cumulative distribution function and $f_Z(t) = \int_{\{\theta\in\Omega : L_{P|Q}(\theta) = z\}} P\mathrm{d}\theta$ is the (generalized) density function of the privacy loss random variable Z. The first equality in Equation 5.1 is a widely used way to represent the $(\varepsilon, \delta(\varepsilon))$-DP curve in literature [96, 11, 10, 59, 47, 99, 24]. Similarly, the first equality in Equation 5.2 represents the well-known moment-generating function of privacy loss [69, 1, 11]. The second equalities above are part of a set of Laplace expressions presented in this chapter. Together, these expressions unlock a formal approach to perform a wide-variety of manipulations and transformations on them using the fundamental properties of the Laplace functional (see Table C.1). Using them, we show that the privacy-profile and Rényi divergence between any two distributions have the following equivalence.

$$\forall q \in \mathbb{C} \;:\; e^{(q-1)\cdot \mathrm{R}_q(P\|Q)} = q(q-1)\cdot \mathcal{B}\left\{\delta_{P|Q}(t)\right\}(1-q), \tag{5.3}$$

[1] Laplace transform maps a time-domain function $g(t)$ with $t \in \mathbb{R}$ to a function $\mathcal{L}\left\{g\right\}(s) := \int_0^\infty e^{-st} g(t)\mathrm{d}t$ with $s \in \mathbb{C}$ in the complex space. Similarly, bilateral Laplace transform of $g(t)$ is defined as $\mathcal{B}\left\{g\right\}(s) := \int_{-\infty}^{\infty} e^{-st} g(t)\mathrm{d}t$.

which again is a Laplace transform expression. Furthermore, Zhu et al. [122]'s characteristic function of the privacy loss $\phi_{P|Q}(q) := \mathop{\mathbb{E}}_{P}\left[e^{iq\log(P/Q)}\right]$ also turns out to be a Fourier transform, which is a special case of the bilateral Laplace transform:

$$\forall q \in \mathbb{R} \ : \ \phi_{P|Q}(q) = \mathop{\mathbb{E}}_{Z\leftarrow \mathrm{PLD}(P\|Q)}\left[e^{iqZ}\right] = \mathcal{B}\left\{f_Z\right\}(-iq). \tag{5.4}$$

These expressions can take advantage of the relationship between their time-domain and complex frequency-domain representations, as certain manipulations are more straightforward in one domain as compared to the other. Using the Laplace transform interpretations extensively, this chapter presents the following findings.

1. We note that the Laplace transform expression of Rényi divergence permits the order q to be a complex number in $\mathbb{C}$. Based on this observation, we revisit the discussion on equivalence and interconversion between (q, ρ)-Rényi DP and (ε, δ)-DP in literature [11, 122, 5, 24]. We show that the privacy-profile curve $\delta_{P|Q}(\varepsilon)$ and the Rényi divergence $\mathrm{R}_q\left(P\|Q\right)$ as a function of q are equivalent as long as either P is absolutely continuous w.r.t. Q (denote as $P \ll Q$) or $Q \ll P$; absolute continuity in both directions is not necessary. Moreover, we establish that while $\delta_{P_1|Q_1}(\varepsilon) \leq \delta_{P_2|Q_2}(\varepsilon)$ for all ε implies that $\mathrm{R}_q\left(P_1\|Q_1\right) \leq \mathrm{R}_q\left(P_2\|Q_2\right)$ for all $q > 1$, the converse does not hold. This is due to the fact that the dominance relationship between privacy profiles $\delta_{P_1|Q_1}(\varepsilon)$ and $\delta_{P_2|Q_2}(\varepsilon)$ depends on how the Rényi divergence curves $\mathrm{R}_q\left(P_1\|Q_1\right)$ and $\mathrm{R}_q\left(P_2\|Q_2\right)$ behave along the complex line $\{q \in \mathbb{C} : \mathfrak{Re}(q) = c\}$ at any $c \in \mathbb{R}$ for which the two divergences exist; not along $\{q \in \mathbb{R} : q > 1\}$ that prior works [11, 122, 5, 24] focus on.

2. Among all functional notions of DP, *exactly tight adaptive* composition theorem is only known for Rényi DP in an explicit form[2][69, 20]. And, for the PLD formalism, only non-adaptive composition theorems are known that are *exactly tight*[3] [47, 96, 59]. In this thesis, we establish an exactly tight theorem for

[2]Dong et al. [34] examine tight composition under the f-DP framework by defining an abstract composition operation, denoted $f_1 \otimes f_2$. However, they do not provide an explicit form for this operator for a general trade-off function f, offering it only for the specific case of the Gaussian trade-off function G_μ.

[3]Unlike under non-adaptivity, composing two privacy loss random variables Z_1 and Z_2 does not amount to convolving their privacy loss distributions (PLDs) $f_{Z_1} \circledast f_{Z_2}$ when the mechanisms are adaptive because Z_1, Z_2 become dependent. We note that Gopi et al. [47] seem to incorrectly assert their Theorem 5.5 to be valid under adaptivity.

composing any two privacy profiles, $\delta_{P_1|Q_1}(\varepsilon)$ and $\delta_{P_2|Q_2}(\varepsilon)$, leveraging time-frequency dualities with their Rényi divergence curves. Our composition method also extends to adaptive scenarios, provided that the conditional distributions P_2^θ and Q_2^θ, given an observation θ from the first distribution pair, are dominated by a privacy profile for all θ, which is a standard assumption for adaptive composition guarantees.

3. We apply our composition theorem for privacy profiles to derive an optimal composition theorem for (ϵ_i, δ_i)-point guarantees of differential privacy. Our approach begins by determining the worst-case privacy profile $\delta_i(\epsilon)$ that *any* (ϵ_i, δ_i)-DP mechanism must satisfy. We then use our composition theorem to derive the combined privacy profile $\delta^{\otimes}(\epsilon)$. This provides the most precise composition guarantee possible when given only that a sequence of mechanisms each satisfies an (ϵ_i, δ_i)-DP point guarantee. Our bound surpasses the optimal composition result in Kairouz et al. [56, Theorem 3.3] because, whereas their result only provides a discrete set of (ϵ, δ) values met by the composed curve, ours forms a continuous curve. This continuity enables us to determine the tightest ϵ value for any given δ budget. We also show that our results align with the bounds generated by numerical accountants such as Google's PLDAccountant [36] and Microsoft's PRVAccountant [47].

4. The concept of f-DP introduced by Dong et al. [34] provides a functional perspective on the indistinguishability between two distributions P and Q through hypothesis testing. The function $f : [0, 1] \to [0, 1]$ represents a bound on the trade-off $T(P, Q) : [0, 1] \to [0, 1]$ between Type-I and Type-II errors for any test aimed at determining whether a sample θ originates from P or Q. Unlike other functional notions of differential privacy, f-DP is unique in being *not connected* to the rest via a Laplace transform. Instead, Dong et al. [34] establish that the privacy profile $\delta(\varepsilon)$ and the trade-off curve f of a mechanism exhibit a convex-conjugate relationship, also known as Fenchel duality. However, Dong et al. [34, Proposition 2.12] confirm this functional equivalence only when f is *symmetric*. With Poisson subsampling at probability p, the resulting amplified curve $f_p(x) = pf(x) + (1-p) \cdot x$ becomes asymmetric. To ensure symmetry, the subsampling result [34, Theorem 4.2] applies a p-

sampling operator $C_p(f) \stackrel{\text{def}}{=} \min\{f_p, f_p^{-1}\}^{**}$ that overestimates the f_p-curve. We show that this symmetrization step disrupts the equivalence between f-DP and privacy profile formalisms (and thereby with other functional notions). To address this, we propose maintaining the natural asymmetry in f-DP and avoid the need for this symmetrization step by adopting a convention on the direction of skew. This completes the equivalences across all functional notions of DP.

Related work: Our work builds an interpretation of differential privacy by leveraging several works that appeared before. This includes, but not limited to the works on various interpretations of privacy by Dong et al. [34], Dwork and Rothblum [39], Dwork et al. [41], Mironov [69], Bun and Steinke [20], Sommer et al. [96], Gopi et al. [47], Koskela et al. [59], Zhu et al. [122]. In particular, the work of Zhu et al. [122] shares the most similarity with ours, as they were the first to observe that many functional notions of differential privacy appear to be linked via Laplace or Fourier transforms. However, their work centers on using the characteristic function of privacy loss (in Equation 5.4) as an intermediate functional representation connecting various DP notions. In contrast, we examine the nature of these connections themselves to harness the perspective of Laplace transformations as an analytical toolkit for differential privacy.

Relevant studies on composition theorems for differential privacy include Dwork et al. [40], Kairouz et al. [56], Murtagh and Vadhan [71], Bun and Steinke [20], Mironov [69], along with numerical accounting methods such as those by Gopi et al. [47], Koskela et al. [59], Doroshenko et al. [36].

5.2 Notations and Preliminaries

For a data universe $\mathcal{X}$, we consider datasets D of size n: $D = (x_0, \ldots, x_n) \in \mathcal{X}^n$ and algorithms $\mathcal{A} : \mathcal{X}^n \to \Omega$ that return a random output in space Ω. We assume that $\mathcal{X}$ includes a sentinel element $\perp$, representing an empty entry, to simulate 'add' and 'remove' adjacency within 'replace' adjacency. Two datasets D and D' in $\mathcal{X}^n$ are considered *neighboring* (denoted by $D \simeq D'$) if they differ by a single record replacement. Throughout the chapter, we denote the distributions of the output

random variables $\mathcal{A}(D)$ and $\mathcal{A}(D')$ as P and Q, respectively, and focus our study on the indistinguishability behavior of these two distributions. For simplicity, we use the same symbols P and Q to refer to their probability mass or density functions.

We recall the definitions of hockey-stick and (ε, δ)-differential privacy.

Definition 1 ((ε, δ)-Differential Privacy [41])**.** *Let $\varepsilon \geq 0$ and $0 \leq \delta \leq 1$. A randomized algorithm $\mathcal{A}$ is* (ε, δ)-differentially private *(hereon (ε, δ)-DP) if for all $D \simeq D'$ and for all $S \subset \mathcal{Y}$, we have*

$$\Pr\left[\mathcal{A}(D) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\mathcal{A}(D') \in S\right] + \delta. \tag{5.5}$$

Equivalently, $\mathcal{A}$ is (ε, δ)-DP if the hockey-stick divergence (see Definition 4 below) between output distributions P and Q satisfies $\mathrm{H}_{e^{\varepsilon}}(P \| Q) \leq \delta$.

Definition 4 (Hockey-stick divergence [87])**.** *The* hockey-stick divergence *of P w.r.t. Q for any $\gamma > 0$ is defined as*

$$\mathrm{H}_{\gamma}(P \| Q) \overset{def}{=} \sup_{S \subset \mathcal{Y}} P(S) - \gamma \cdot Q(S) = \mathop{\mathbb{E}}_{\Theta \sim P}\left[1 - \gamma \frac{P(\Theta)}{Q(\Theta)}\right]_{+}. \tag{5.6}$$

Following Balle et al. [10], for any two distributions P, Q on the same space, we will refer to

$$\delta_{P|Q}(\varepsilon) \overset{\text{def}}{=} \mathrm{H}_{e^{\varepsilon}}(P \| Q) \tag{5.7}$$

as the *privacy-profile* of P with respect to Q.

Remark 9. *Our definition of a privacy profile permits $\varepsilon < 0$. This differs from the original definition by Balle et al. [10], which only considers non-negative ε. Although allowing $\varepsilon < 0$ might seem counterintuitive, it is both accurate and efficient because negative values of ε yield the privacy profile with P and Q switched:*

$$\begin{aligned}
\delta_{Q|P}(\varepsilon) &= \sup_{S \subset \mathcal{Y}} Q(S) - e^{\varepsilon} P(S) \\
&= 1 - e^{\varepsilon} + e^{\varepsilon}[\sup_{S' \subset \mathcal{Y}} P(S') - e^{-\varepsilon} Q(S')] \\
&= 1 - e^{\varepsilon} + e^{\varepsilon} \delta_{P|Q}(-\varepsilon).
\end{aligned} \tag{5.8}$$

Another popular notion that we recall is the Rényi divergence. We provide the following extended definition for it.

Definition 5 (Rényi divergence [80]). *For two distributions P, Q over the same space, we define the Rényi divergence $\mathrm{R}_q(P\|Q)$ of any order $q \in \mathrm{ROC}_{P,Q}$ as*

$$\mathrm{R}_q(P\|Q) \stackrel{def}{=} \frac{1}{q-1}\log \mathrm{E}_q(P\|Q), \qquad \text{where} \qquad \mathrm{E}_q(P\|Q) \stackrel{def}{=} \int_{\theta\in\Omega} P(\theta)^q Q(\theta)^{1-q} d\theta, \tag{5.9}$$

and $\mathrm{ROC}_{P,Q}$ is the region consisting of all orders $q \in \mathbb{C}$ where the integral is conditionally convergent*.*

Definition 11 (Rényi Differential privacy [69]). *A randomized algorithm $\mathcal{A}$ is said to be* (q,ρ)-Rényi differentially private *(hereon (q,ρ)-Rényi DP) if, for all neighboring datasets $D \simeq D' \in \mathcal{X}^n$, we have $\mathrm{R}_q(P\|Q) \le \rho$.*

Remark 10. *It is known that Rényi divergence converges to Kullback-Leibler divergence as order q tends to one.*[4] *Among real orders $q \in \mathbb{R}$, privacy researchers typically restrict themselves to $q > 1$ without thinking much about values smaller than 1. Just like privacy profile $\delta_{P|Q}(\varepsilon)$, Rényi divergence for orders $q < 1$ yield the Rényi divergence with P and Q reversed.*

$$\begin{aligned}(q-1)\cdot \mathrm{R}_q(P\|Q) &= \log \mathrm{E}_q(P\|Q)\\ &= \log \int_{\theta\in\Omega} P(\theta)^q Q(\theta)^{1-q} d\theta\\ &= \log \mathrm{E}_{1-q}(Q\|P) = -q\cdot \mathrm{R}_{1-q}(Q\|P).\end{aligned} \tag{5.10}$$

Yet another notion of differential privacy is the f-DP notion.

Definition 25 (f-Differential Privacy [34]). Let $f : [0,1] \to [0,1]$ be a convex, continuous, non-increasing function such that $f(x) \le 1-x$ for all $x \in [0,1]$. A randomized algorithm $\mathcal{A}$ is *f-differentially private* (henceforth f-DP) if

$$\forall \alpha \in [0,1] : f_{\mathcal{A}(D)|\mathcal{A}(D')}(\alpha) \ge f(\alpha), \tag{5.11}$$

where for any distributions P, Q on Ω, the $f_{P|Q} : [0,1] \to [0,1]$ is the trade-off function defined as

$$\forall \alpha \in [0,1] : f_{P|Q}(\alpha) \stackrel{\text{def}}{=} \inf_{\phi:\Omega\to[0,1]} \{\beta_\phi : \alpha_\phi \le \alpha\}, \tag{5.12}$$

where $\alpha_\phi \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{P}[\phi]$, and $\beta_\phi \stackrel{\text{def}}{=} 1 - \mathop{\mathbb{E}}_{Q}[\phi]$.

[4] It follows from the *replica trick*: $\mathbb{E}[\log X] = \lim_{n\to 0} \frac{1}{n}\log \mathbb{E}[X^n]$, when $\log X$ is the privacy loss rv $Z \leftarrow \mathrm{PLD}(P\|Q)$.

Remark 11. *For a hypothesis test ϕ on a sample θ originating from either P or Q, we follow the convention that $\theta \sim Q$ represents the positive event, and $\phi(\theta) = 1$ denotes a positive prediction. With this convention, α_ϕ and β_ϕ represent the* false positive rate *(Type I error) and* false negative rate *(Type II error), respectively. Note that the left-continuous inverse of the trade-off function $f^{-1}_{P|Q}(\beta)$ gives the trade-off curve with P and Q reversed.*

$$\begin{aligned} f^{-1}_{P|Q}(\beta) &:= \{\inf \alpha \in [0,1] : f_{P|Q}(\alpha) \leq \beta\} \\ &= \inf_{\phi:\Omega\to[0,1]} \{\alpha_\phi : \beta_\phi \leq \beta\} \\ &= f_{Q|P}(\beta). \end{aligned} \tag{5.13}$$

We also recall the privacy quantification of Gaussian mechanism.

Theorem 28 (Privacy of Gaussian Mechanism [41, 7, 69, 34])**.** *If $P = \mathcal{N}(\mu, \sigma^2 I_d)$ and $Q = \mathcal{N}(\mu', \sigma^2 I_d)$ are two multivariate Gaussian distributions on $\mathbb{R}^d$ such that $\kappa = \|\mu - \mu'\|_2^2/(2\sigma^2)$, then the privacy loss distribution is $\mathrm{PLD}(P\|Q) = \mathcal{N}(\kappa, 2\kappa)$, the Rényi divergence is $\mathrm{R}_q(P\|Q) = \kappa q$ for all $q > 1$ and for all $\varepsilon \in \mathbb{R}$,*[5]

$$\delta_{P|Q}(\varepsilon) = \Phi\left(\frac{-\varepsilon + \kappa}{\sqrt{2\kappa}}\right) - e^{\varepsilon}\Phi\left(\frac{-\varepsilon - \kappa}{\sqrt{2\kappa}}\right) = O(e^{-\varepsilon^2/4\kappa}), \tag{5.14}$$

where $\Phi(t) = \Pr_{G\sim\mathcal{N}(0,1)}[G \leq t]$ is the standard normal CDF, and

$$f_{P|Q}(\alpha) = \Phi(\Phi^{-1}(1-\alpha) - \sqrt{2\kappa}). \tag{5.15}$$

5.3 Laplace Transform of Privacy Loss

Differential privacy bounds the maximum divergence in the output distribution caused by including or omitting a data-point from the dataset. This principle is mirrored in the notion of *privacy loss* which expresses how much an algorithm's output reveals about the inclusion of a specific data-point.

Definition 26 (Privacy Loss Distribution [96])**.** The *privacy loss* of an observation $\theta \in \mathcal{Y}$ from an algorithm $\mathcal{A}$, when comparing datasets $D \simeq D'$, is defined as

[5] While the original result by Balle and Wang [7] was stated only for non-negative ε, the expression remains identical for $\varepsilon < 0$ on invoking Equation 5.8, thanks to the symmetry property $\Phi(-x) = 1 - \Phi(x)$ of normal distribution.

$L_{P|Q}(\theta) \stackrel{\text{def}}{=} \log(P(\theta)/Q(\theta))$, where P and Q are the probability mass or probability density functions of $\mathcal{A}(D)$ and $\mathcal{A}(D')$, respectively. The *privacy loss distribution* $\text{PLD}(P\|Q)$ is the distribution of $L_{P|Q}(\Theta)$ when $\Theta \sim P$.

Like other notions of differential privacy, the privacy loss distribution formalism also enjoys a reversal property.

Remark 12. *Similar to other functional privacy notions, the PLD formalism possesses a reversal property [47, Definition 3.1]. Let f_Z and $f_{Z'}$ represent the* generalized density functions[6] *of the random variables $Z \leftarrow \text{PLD}(P\|Q)$ and $Z' \leftarrow \text{PLD}(Q\|P)$, respectively. Then for all $z \in \mathbb{R}$,*

$$\begin{aligned} f_Z(z) &= \int_{\{\theta\in\Omega: L_{P|Q}(\theta)=z\}} P(\theta)\mathrm{d}\theta \\ &= \int_{\{\theta\in\Omega: L_{P|Q}(\theta)=z\}} e^{L_{P|Q}(\theta)} \cdot Q(\theta)\mathrm{d}\theta \\ &= e^z \cdot \int_{\{\theta\in\Omega: L_{Q|P}(\theta)=-z\}} Q(\theta)\mathrm{d}\theta = e^z \cdot f_{Z'}(-z). \end{aligned} \tag{5.16}$$

The $\text{PLD}(P\|Q)$ describes how outputs arising from D increase an observer's confidence that they did not come from D'. Several divergence notions can be described in terms of $\text{PLD}(P\|Q)$. The following Table 5.1 describes summarizes a few of them and the following Theorem 29 how (ε, δ)-DP and Rényi DP constraints the $\text{PLD}(P\|Q)$ to have an exponentially decaying tail.

Theorem 29. *Let P and Q be distributions on the same space denoting the output distribution of some algorithm $\mathcal{A}$ on neighbouring datasets $D \simeq D'$ respectively. Additionally, let $Z \leftarrow \text{PLD}(P\|Q)$ denote the* privacy loss distribution*. Then,*

$$\text{R}_q(P\|Q) \le \rho \implies \forall h > 0 \ : \ \Pr_{Z\leftarrow\text{PLD}(P\|Q)}[Z \ge h + \rho] \le e^{-h(q-1)}, \tag{5.17}$$

$$\text{and } \delta_{P|Q}(\varepsilon) \le \delta \implies \forall h > 0 \ : \ \Pr_{Z\leftarrow\text{PLD}(P\|Q)}[Z \ge h + \varepsilon] \le \frac{\delta}{1 - e^{-h}}. \tag{5.18}$$

[6] We define density as $f_X(t)dt = \lim_{a\to 0^+} \int_{t-a}^{t+a} F_X(u)du$ to handle cases where F_X isn't differentiable everywhere, such as when PLD is a discrete distribution. This density is expressible with *Dirac delta* $\Delta(t)$ as $f_X(t) = \begin{cases} \dot{F}_X(t) & \text{if derivative } \dot{F}_X \text{ exists at } t \\ [F_X(t^+) - F_X(t^-)]\Delta(t) & \text{otherwise} \end{cases}$, and satisfies $F_X(t) = \int_{-\infty}^{t^+} f_X(u)du$.

Divergence notion of P w.r.t Q	**In terms of $Z \sim \mathrm{PLD}(P\|Q)$**
Total-Variation distance	$\mathrm{TV}(P,Q) = \mathbb{E}\left[\max\{0, 1-e^{-Z}\}\right]$
Hockey-stick divergence	$\mathrm{H}_\gamma(P\|Q) = \mathbb{E}\left[\max\{0, 1-\gamma\cdot e^{-Z}\}\right]$
Rényi divergence	$\mathrm{R}_q(P\|Q) = \frac{1}{q-1}\log \mathbb{E}\left[e^{(q-1)Z}\right]$
Kullback-Leibler divergence	$\mathrm{KL}(P\|Q) = \mathbb{E}[Z]$

Table 5.1: Some popular divergence notions in terms of the privacy loss distribution.

Proof. Let $s = 1-q < 0$ and $h' \geq \rho$.

$$\begin{aligned}
\Pr_{Z\leftarrow\mathrm{PLD}(P\|Q)}[Z \geq h'] &= \Pr_{Z\leftarrow\mathrm{PLD}(P\|Q)}\left[e^{-sZ} \geq e^{-sh'}\right] \\
&\leq e^{sh'} \mathop{\mathbb{E}}_{Z\leftarrow\mathrm{PLD}(P\|Q)}\left[e^{-sZ}\right] && \text{(From Markov inequality)} \\
&= e^{sh'}\int_{-\infty}^{\infty} e^{-st} f_Z(t)\mathrm{d}t \\
&= e^{sh'}\int_{\Omega}\left(\frac{P(\theta)}{Q(\theta)}\right)^{-s} P(\theta)\mathrm{d}\theta \\
&= e^{sh'}\cdot \mathrm{E}_{1-s}(P\|Q) \\
&= e^{-(q-1)h'}\cdot e^{(q-1)\mathrm{R}_q(P\|Q)} && (s \to 1-q) \\
&\leq e^{-(q-1)(h'-\rho)} && \text{(Since } \mathrm{R}_q(P\|Q) \leq \rho\text{)}
\end{aligned}$$

On substituting $h' = h + \rho$ proves the first part.

For the second part, let $h > 0$. From Theorem 30 (that we introduce shortly), we have

$$\begin{aligned}
\delta_{P|Q}(\varepsilon) &= \mathcal{L}\{1 - F_Z(t+\varepsilon)\}(1) \\
&= \int_0^{\infty} e^{-t}[1 - F_Z(t+\varepsilon)]\mathrm{d}t \\
&\geq \int_0^{h} e^{-t}[1 - F_Z(h+\varepsilon)]\mathrm{d}t && \text{(Since } 1-F_Z(\cdot) \text{ a decreasing function)} \\
&= \Pr_{Z\leftarrow\mathrm{PLD}(P\|Q)}[Z \geq h+\varepsilon]\cdot\int_0^{h} e^{-t}\mathrm{d}t \\
&= \Pr_{Z\leftarrow\mathrm{PLD}(P\|Q)}[Z \geq h+\varepsilon]\cdot(1-e^{-h}).
\end{aligned}$$

Plugging in $\delta_{P|Q}(\varepsilon) \leq \delta$ and rearranging concludes the second part. □

While being able to express the divergence notions in terms of privacy loss distribution is great, often they are not directly useful in arguing about their properties. In the following theorem, we present a more dynamic version of these relationships by expressing them as a set of *Laplace transforms* of the privacy loss distributions $\mathrm{PLD}(P\|Q)$ and $\mathrm{PLD}(Q\|P)$.

Theorem 30. *For a random variable X, let $F_X(t) \stackrel{def}{=} \Pr[X \leq t]$ denote its* cumulative distribution function *and $f_X(t)$ denote its* generalized probability density function. *Let P and Q be probability distributions and $Z \sim \mathrm{PLD}(P\|Q)$ and $Z' \sim \mathrm{PLD}(Q\|P)$ denote their privacy loss random variables. If $Z \sim \mathrm{PLD}(P\|Q)$ and $Z' \sim \mathrm{PLD}(Q\|P)$, then for all $\varepsilon \in \mathbb{R}$,*

$$\begin{aligned}
\delta_{P|Q}(\varepsilon) &= \mathcal{L}\left\{1 - F_Z(t+\varepsilon)\right\}(1) && (5.19)\\
&= e^{\varepsilon} \cdot \mathcal{L}\left\{F_{Z'}(-t-\varepsilon)\right\}(-1) && (5.20)\\
&= e^{\varepsilon} \cdot \mathcal{L}\left\{f_{Z'}(-t-\varepsilon)\right\}(-1) - \mathcal{L}\left\{f_Z(t+\varepsilon)\right\}(1) && (5.21)\\
&= \mathcal{L}\left\{f_Z(t+\varepsilon)\right\}(0) - e^{\varepsilon} \cdot \mathcal{L}\left\{f_{Z'}(-t-\varepsilon)\right\}(0). && (5.22)
\end{aligned}$$

And, for all $q \in \mathrm{ROC}_{\mathcal{B}}\{f_{Z'}\}$ (or equivalently, $1-q \in \mathrm{ROC}_{\mathcal{B}}\{f_Z\}$),

$$e^{(q-1)\cdot \mathrm{R}_q(P\|Q)} = \mathrm{E}_q\left(P\|Q\right) = \mathcal{B}\{f_Z(t)\}(1-q) = \mathcal{B}\{f_{Z'}(t)\}(q). \tag{5.23}$$

We provide a proof of Theorem 30 in § C.2. Laplace expressions in Theorem 30 often arise in their explicit integral forms within several proofs in related works on differential privacy, for instance [24, Lemma 9], [99, Proposition 7], and [7, Theorem 5]. In their integral forms, they frequently undergo manipulations like integration-by-parts or change-of-variables which can quickly get complicated. Our Theorem 30 offers a way to simplify the complexity of such manipulations as one can express the concerned terms in their Laplace expressions and invoke its properties from Table C.1, like Equation C.2 and Equation C.4 for shifting or scaling variable of integration and Equation C.6 and Equation C.7 for integrating-by-parts. Examples in this thesis will illustrate that reasoning about privacy this way through its Laplace transform interpretation could be quite effective.

In the following theorem, we show that the Rényi divergence $\mathrm{R}_q\left(P\|Q\right)$ and the privacy profile $\delta_{P|Q}(\varepsilon)$ are connected through a Laplace transform as well. We can show this using only the expressions in Theorem 30.

Theorem 31 (Rényi DP from privacy profile)**.** *Let $q > 1$. For any two distributions P and Q,*

$$e^{(q-1)\cdot \mathrm{R}_q(P\|Q)} = \mathrm{E}_q\left(P\|Q\right) = q(q-1)\cdot \mathcal{B}\left\{\delta_{P|Q}(t)\right\}(1-q), \tag{5.24}$$

for all orders q such that $1-q \in \mathrm{ROC}_{\mathcal{B}}\{\delta_{P|Q}\}$.

Proof. Let $Z \sim \mathrm{PLD}(P\|Q)$ and $Z' \sim \mathrm{PLD}(Q\|P)$. From Equation 5.22,

$$\begin{aligned}
\delta_{P|Q}(\varepsilon) &= \mathcal{L}\left\{f_Z(t+\varepsilon)\right\}(0) - e^{\varepsilon}\cdot \mathcal{L}\left\{f_{Z'}(-t-\varepsilon)\right\}(0) && (5.25)\\
&= \int_{0^+}^{\infty} e^0 \cdot f_Z(t+\varepsilon)\mathrm{d}t - e^{\varepsilon}\cdot \int_{0^+}^{\infty} e^0 \cdot f_{Z'}(-t-\varepsilon)\mathrm{d}t && (5.26)\\
&= 1 - F_Z(\varepsilon) - e^{\varepsilon}\cdot F_{Z'}(-\varepsilon). && (5.27)
\end{aligned}$$

We apply the linearity, time-shifting, and reversal properties of two-sided Laplace transforms to simplify the Laplace transform of privacy profile as follows

$$\begin{aligned}
\mathcal{B}\left\{\delta_{P|Q}(t)\right\}(1-q) &= \mathcal{B}\left\{1 - F_Z(t) - e^t\cdot F_{Z'}(-t)\right\}(1-q) && (5.28)\\
&\overset{\textit{Equation C.1}}{=} \mathcal{B}\left\{1-F_Z(t)\right\}(1-q) - \mathcal{B}\left\{e^t\cdot F_{Z'}(-t)\right\}(1-q) && (5.29)\\
&\overset{\textit{Equation C.3}}{=} \mathcal{B}\left\{1-F_Z(t)\right\}(1-q) - \mathcal{B}\left\{F_{Z'}(-t)\right\}(-q) && (5.30)\\
&\overset{\textit{Equation C.5}}{=} \mathcal{B}\left\{1-F_Z(t)\right\}(1-q) - \mathcal{B}\left\{F_{Z'}(t)\right\}(q). && (5.31)
\end{aligned}$$

Next, we apply the derivative property of Laplace transforms to get

$$\mathcal{B}\left\{1-F_Z(t)\right\}(1-q) \overset{\textit{Equation C.6}}{=} \frac{\mathcal{B}\left\{f_Z(t)\right\}(1-q)}{q-1}, \tag{5.32}$$

and

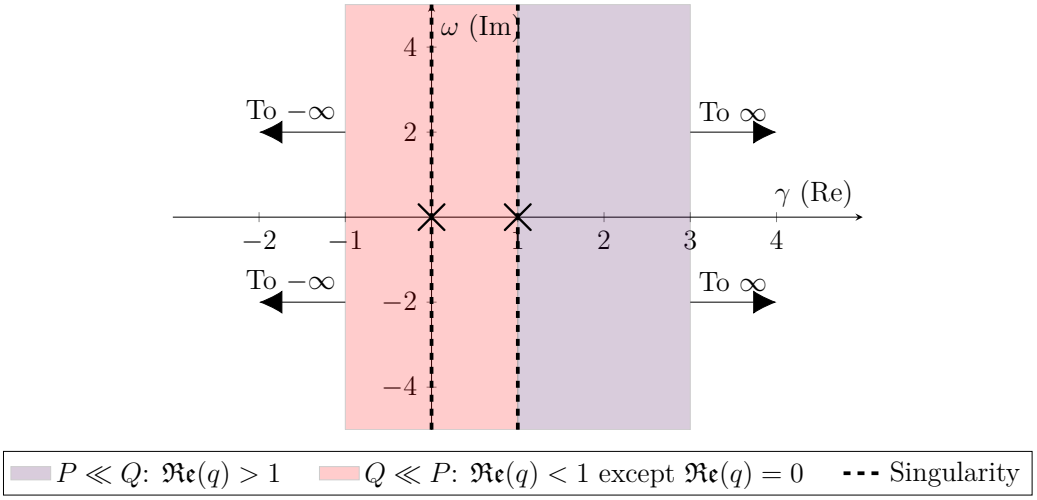
$$\mathcal{B}\left\{F_{Z'}(t)\right\}(q) \overset{\textit{Equation C.6}}{=} \frac{\mathcal{B}\left\{f_{Z'}\right\}(q)}{q}. \tag{5.33}$$

Finally, noting from Theorem 30 and Equation 5.9 that $\mathcal{B}\left\{f_Z\right\}(1-q) = \mathcal{B}\left\{f_{Z'}\right\}(q) = \mathrm{E}_q\left(P\|Q\right)$, we have

$$\mathcal{B}\left\{\delta_{P|Q}(t)\right\}(1-q) = \mathrm{E}_q\left(P\|Q\right)\left[\frac{1}{q-1} - \frac{1}{q}\right] = \frac{e^{(q-1)\mathrm{R}_q(P\|Q)}}{q(q-1)}. \tag{5.34}$$

□

Region of Convergence of $\mathcal{B}\left\{\delta_{P|Q}\right\}(1-q)$ for $P \ll Q$ and $Q \ll P$ at Rényi order $q = \gamma + i\omega$



The privacy profile $\delta_{P|Q}$, for any pair of distributions P, Q, is a continuous function with bounded range $[0, 1]$.[7] Therefore, by Lerch's theorem (cf. discussion on uniqueness in § 2.3.2), *any two privacy profiles share the same Rényi divergence curve if and only if they are identical.* In other words, a privacy profile $\delta_{P|Q}$ is equivalent to its Rényi divergence curve $\mathrm{R}_q(P\|Q)$, provided the Laplace transform $\mathcal{B}\left\{\delta_{P|Q}\right\}(1-q)$ exists at some $q \in \mathbb{C}$, i.e., $\mathrm{ROC}_{\mathcal{B}}\{\delta_{P|Q}\} \neq \emptyset$.

Note that Theorem 31 establishes a connection between Rényi divergence and privacy profile that applies to *all distributions P and Q*, whether P is absolutely continuous[8] (denoted as $P \ll Q$) with respect to Q or not. The choice of P, Q however influences the region of convergence $\mathrm{ROC}_{\mathcal{B}}\{\delta_{P|Q}\}$ where the Rényi divergence can be defined. One can verify that if $P \ll Q$, the Laplace transform $\mathcal{B}\left\{\delta_{P|Q}\right\}(1-q)$ converges for all real $q > 1$; and if $Q \ll P$, it converges for all $q < 1$, except for $q = 0$. At $q = 0$ or $q = 1$, the transform does not converge as the expression for $\delta_{P|Q}(\varepsilon)$ has singularities at these points because numerator $\mathrm{E}_q(P\|Q) = \int P^q Q^{1-q} \mathrm{d}\theta = 1$ when $q = 0$ or 1 while denominator becomes zero. Since region of convergence is always a strip in the complex plane $\mathbb{C}$ parallel to the imaginary line, the Rényi divergence exists not just for real orders q, but also for imaginary orders: $\{q \in \mathbb{C} \ : \ \mathfrak{Re}(q) > 1\}$ when

[7] We can write $\delta_{P|Q}(\varepsilon) = \sup_S \psi_S(\varepsilon)$ where $\psi_S(\varepsilon) := P(S) - e^{\varepsilon} \cdot Q(S)$ is a continuous decreasing function. Therefore, the supremum of ϕ_S over all S is also a continuous decreasing function.

[8] P is absolutely continuous with respect to Q if for all measurable subsets $S \subset \Omega$, $P(S) > 0 \implies Q(S) > 0$. We say that distribution pair (P, Q) is absolutely continuous (w.r.t. each other) if $P \ll Q$ and $Q \ll P$.

$P \ll Q$ and $\{q \in \mathbb{C} \ : \ \mathfrak{Re}(q) < 1 \text{ and } \mathfrak{Re}(q) \neq 0\}$ when $Q \ll P$. Therefore, as long as either $P \ll Q$ *or* $Q \ll P$, the region $\mathrm{ROC}_{\mathcal{B}}\{\delta_{P|Q}\}$ is not empty and a characterizing curve $\mathrm{R}_q\left(P\|Q\right)$ exists for the profile $\delta_{P|Q}(\varepsilon)$. That is to say, Theorem 31 can be used to derive the exact privacy profile $\delta_{P|Q}$ using the Rényi divergence $\mathrm{R}_q\left(P\|Q\right)$ function. We can do this by substituting $1 - q = s$ in Equation 5.24, rearranging, and applying the *inverse Laplace transform* Equation 2.51, to get the following explicit form:

$$\delta_{P|Q}(\varepsilon) = \mathcal{L}^{-1}\left\{\frac{e^{-s\mathrm{R}_{1-s}(P\|Q)}}{s(s-1)}\right\}(\varepsilon) = \frac{1}{2\pi i}\lim_{\omega\to\infty}\int_{\gamma-i\omega}^{\gamma+i\omega} e^{s\varepsilon}\cdot\frac{e^{-s\mathrm{R}_{1-s}(P\|Q)}}{s(s-1)}ds, \tag{5.35}$$

where $\gamma \in \mathbb{R}$ can be *any* real point in $\mathrm{ROC}_{\mathcal{B}}\{\delta_{P|Q}\}$.

Case study. The following example demonstrates an application of the Laplace transform identity in Equation 31. We first describe the privacy profile of the randomized response mechanism [56] and then use Equation 5.24 on it to reason about its Rényi divergence characteristics.

Theorem 32 (Privacy profile of randomized response)**.** *Fix $\varepsilon > 0$ and $\delta \in [0, 1]$. Let $\mathcal{A}^{\varepsilon,\delta}_{\mathrm{RR}} : \{0, 1\} \to \{0, 1\} \times \{\bot, \top\}$ be the randomized response mechanism, which has the following output probabilities.*

$$\mathcal{A}^{\varepsilon,\delta}_{\mathrm{RR}}(0) = \begin{cases} (0,\bot) & \textit{with probability } \delta, \\ (0,\top) & \textit{with probability } \frac{(1-\delta)e^{\varepsilon}}{e^{\varepsilon}+1}, \\ (1,\top) & \textit{with probability } \frac{(1-\delta)}{e^{\varepsilon}+1}, \\ (1,\bot) & \textit{with probability } 0, \end{cases} \quad \mathcal{A}^{\varepsilon,\delta}_{\mathrm{RR}}(1) = \begin{cases} (0,\bot) & \textit{with probability } 0, \\ (0,\top) & \textit{with probability } \frac{(1-\delta)}{e^{\varepsilon}+1}, \\ (1,\top) & \textit{with probability } \frac{(1-\delta)e^{\varepsilon}}{e^{\varepsilon}+1}, \\ (1,\bot) & \textit{with probability } \delta. \end{cases}$$

For $P = \mathcal{A}^{\varepsilon,\delta}_{\mathrm{RR}}(0)$ and $Q = \mathcal{A}^{\varepsilon,\delta}_{\mathrm{RR}}(1)$, the privacy profiles are

$$\forall t \in \mathbb{R} \ : \ \delta_{\mathrm{RR}} := \delta_{P|Q}(t) = \delta_{Q|P}(t) = \begin{cases} \delta & \textit{if } \varepsilon < t, \\ 1 - \frac{e^t+1}{e^{\varepsilon}+1}(1-\delta) & \textit{if } -\varepsilon < t \leq \varepsilon, \\ 1 - e^t(1-\delta) & \textit{if } t \leq -\varepsilon. \end{cases} \tag{5.36}$$

Theorem 32 extends Balle et al. [10, Theorem 2], which provides the privacy profile for randomized response only in the case where $\delta = 0$ and for values $t \geq 0$. From Equation 32, we can see that neither $P \ll Q$ nor $Q \ll P$ holds for the output

distributions of randomized response mechanism when $\delta > 0$. As such, the Laplace transform $\mathcal{B}\left\{\delta_{P|Q}\right\}(1-q)$ must not converge for any $q \in \mathbb{R}$. One can check that when $q \geq 1$, the transform has a term $\int_{\varepsilon}^{\infty} e^{(q-1)t} \cdot \delta \mathrm{d}t$ which blows up to ∞, and when $q < 1$, the transform has a term $\int_{-\infty}^{-\varepsilon} e^{(q-1)t} \cdot (1 - e^t(1-\delta))\mathrm{d}t$ that blows up to $-\infty$. Hence, the Rényi divergence of the privacy profile Equation 5.36 cannot be defined for any order $q \in \mathbb{R}$ when $\delta > 0$.

When $\delta = 0$, note that both $P \ll Q$ and $Q \ll P$. Therefore, the Laplace transform $\delta_{P|Q}(\varepsilon)$ must exist for all $q \in \mathbb{R} \setminus \{0, 1\}$. The following theorem shows the resulting Rényi divergence curve, derived by computing the Laplace transform.

Theorem 33 (Rényi DP of $(\varepsilon, 0)$-Randomized Response)**.** *For any $\varepsilon > 0$ and $\delta = 0$, the output distributions of randomized response mechanism in Theorem 32 exhibit the following Rényi divergence for all orders $q \in \mathbb{C}$ such that $\mathfrak{Re}(q) \notin \{0, 1\}$:*

$$\mathrm{R}_q\left(P \| Q\right) = \frac{1}{q-1} \log\left(\frac{e^{\varepsilon}}{1+e^{\varepsilon}} e^{-q\varepsilon} + \frac{1}{1+e^{\varepsilon}} e^{q\varepsilon}\right). \tag{5.37}$$

Theorem 33 generalizes Mironov [69, Proposition 5], which gives the Rényi divergence of randomized response only for real orders $q > 1$. In the following section, we elaborate on the significance of complex orders in Rényi divergence. Figure 5.1 visualizes the privacy profile $\delta_{P|Q}$, its Laplace transform $\mathcal{B}\left\{\delta_{P|Q}\right\}$, and the corresponding Rényi divergence $\mathrm{R}_q\left(P \| Q\right)$ of this randomized response mechanism, and compares it with that of Gaussian mechanism (cf. Theorem 28).

5.3.1 Dominance: Rényi Divergence vs. Privacy Profile

Differential privacy is a study of distributional divergence between output distributions P and Q not just for a pair of neighboring inputs D, D' but across *all* neighboring inputs. When considering functional notions of DP, comparing indistinguishability characteristics of two output-distribution pairs, say (P_1, Q_1) and (P_2, Q_2), requires a notion of *dominance*. Zhu et al. [122] define a *dominating pair of distribution* specific to a mechanism $\mathcal{A}$ as a pair of distributions P, Q such that

$$\forall \varepsilon \in \mathbb{R} \ : \ \sup_{D \simeq D'} \delta_{\mathcal{A}(D)|\mathcal{A}(D')}(\varepsilon) \leq \delta_{P|Q}(\varepsilon). \tag{5.38}$$

[9] The tightest conversion by Asoodeh et al. [5] lacks a closed-form expression and is challenging to approximate numerically in a stable way. Therefore, we compare with Canonne et al. [24, Corollary 13], which yields similar values.

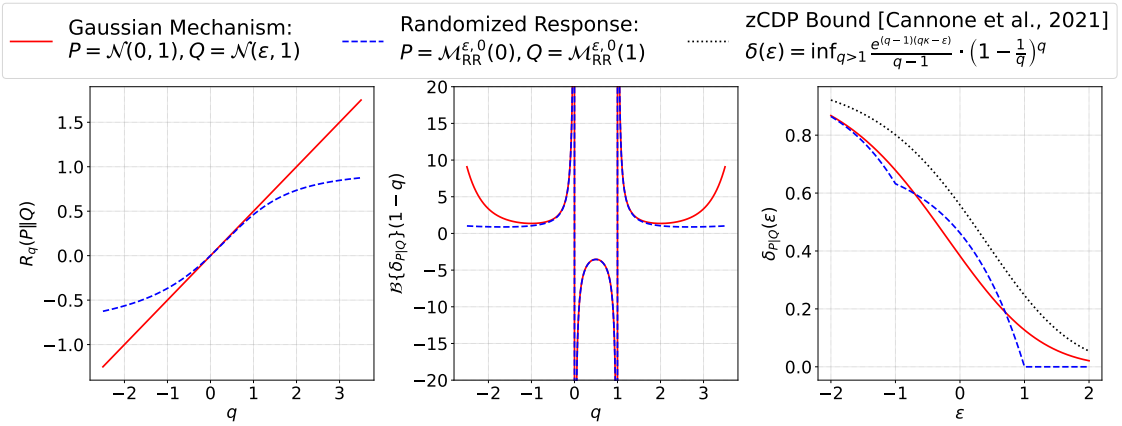


Figure 5.1: Comparison between the indistinguishability characteristic of Gaussian mechanism of Theorem 28 ($\kappa = \|\mu - \mu'\|^2/2\sigma^2 = \varepsilon^2/2$) and randomized response mechanism defined in Theorem 32 (with $\delta = 0$). This figure visualizes the singularities at $q = 0$ and $q = 1$ that exists for the Laplace transform $\mathcal{B}\left\{\delta_{P|Q}\right\}(1-q)$ disappears for Rényi divergence $\mathrm{R}_q(P\|Q)$, which is an effect of the replica trick unfolding. This figure also demonstrates that neither the dominance $\mathrm{R}_q(P_1\|Q_1) \le \mathrm{R}_q(P_2\|Q_2)$ for all $q > 1$, nor the dominance $\mathcal{B}\left\{\delta_{P_1|Q_1}\right\}(1-q) \le \mathcal{B}\left\{\delta_{P_2|Q_2}\right\}(1-q)$ for all $q \in \mathbb{R}\setminus\{0,1\}$ is enough to bound $\delta_{P_1|Q_1}(\varepsilon) \le \delta_{P_2|Q_2}(\varepsilon)$ at all $\varepsilon \in \mathbb{R}$. Additionally, the black dotted line in the rightmost plot shows that even the tightest[9]conversion on the Rényi curve considering only real orders $q > 1$ fails to characterize its own privacy profile.

Following the definition by Zhu et al. [122], we define the following notions of dominance.

Definition 27 (Dominance for Distribution Pairs)**.** We say that the distribution pair (P_2, Q_2) dominates (P_1, Q_1) in privacy profile (denote as $(P_1, Q_1) \preceq_\delta (P_2, Q_2)$) if

$$\forall \varepsilon \in \mathbb{R} \ : \ \delta_{P_1|Q_1}(\varepsilon) \le \delta_{P_2|Q_2}(\varepsilon). \tag{5.39}$$

And, we say (P_2, Q_2) dominates (P_1, Q_1) in Rényi divergence (denote as $(P_1, Q_1) \preceq_R (P_2, Q_2)$) if

$$\forall q > 1 \ : \mathrm{R}_q(P_1\|Q_1) \le \mathrm{R}_q(P_2\|Q_2)\,. \tag{5.40}$$

The following theorem reveals the surprising fact that, while the Rényi divergence curve and privacy profile curve are equivalent for all absolutely continuous distribution pairs, this equivalence does not imply an identical dominance ordering across absolutely continuous distribution pairs.

Theorem 34. *Consider two absolutely continuous distribution pairs* (P_1, Q_1) *and* (P_2, Q_2)*. Then*

$$(P_1, Q_1) \preceq_\delta (P_2, Q_2) \implies (P_1, Q_1) \preceq_R (P_2, Q_2). \tag{5.41}$$

However, the opposite direction does not hold:

$$(P_1, Q_1) \preceq_R (P_2, Q_2) \not\Longrightarrow (P_1, Q_1) \preceq_\delta (P_2, Q_2). \tag{5.42}$$

Proof. First part directly follows from Equation 5.23 by noting that

$$\mathcal{B}\left\{\delta_{P_1|Q_1}\right\}(1-q) = \int_{-\infty}^{\infty} e^{-(1-q)t} \cdot \delta_{P_1|Q_1}(t)\mathrm{d}t \tag{5.43}$$

$$\leq \int_{-\infty}^{\infty} e^{-(1-q)t} \cdot \delta_{P_2|Q_2}(t)\mathrm{d}t = \mathcal{B}\left\{\delta_{P_2|Q_2}\right\}(1-q) \tag{5.44}$$

and that for $q > 1$, $q(q-1) > 0$.

The proof of the second statement is based on the example by Zhu et al. [122] comparing the Rényi DP curve and privacy profile of Gaussian mechanism described in Theorem 28 (denote with (P_2, Q_2)) with that of randomized response mechanism, defined in Theorem 32 (denote with (P_1, Q_1)). For any $\varepsilon > 0$, we set $\kappa = \varepsilon^2/2$ in the Gaussian mechanism where $(\varepsilon, 0)$ is the parameter of the randomized response and compare them in Figure 5.1. From the leftmost plot, we can see that $(P_1, Q_1) \preceq_R (P_2, Q_2)$ holds.[10] However, note that in the rightmost plot that their privacy profiles cross one another. Hence, the dominance $(P_1, Q_1) \preceq_\delta (P_2, Q_2)$ does not hold. □

Zhu et al. [122] point out that it is troubling that identifying a dominating pair of distributions in Rényi divergence does not guarantee a corresponding dominance in privacy profiles. Although the Rényi divergence curve $\mathrm{R}_q(P\|Q)$ is an equivalent representation of the privacy profile $\delta_{P|Q}(\varepsilon)$, converting the $\mathrm{R}_q(P\|Q)$ curve into a tight upper bound for privacy profile will *always* introduce a gap. This gap can be substantial, as shown in Figure 5.1, where we compare the (nearly) tight upper bound (black dotted line) derived for the privacy profile from the Rényi divergence curve (red line on the left) of the Gaussian mechanism with the actual privacy profile of this Gaussian mechanism.

[10] Dong et al. [34, Proposition B.7 (a)] show that $\mathrm{R}_q\left(\mathcal{M}_{\mathrm{RR}}^{\varepsilon,0}(0)\middle\|\mathcal{M}_{\mathrm{RR}}^{\varepsilon,\delta}(1)\right) \leq \mathrm{R}_q\left(\mathcal{N}(0,1)\|\mathcal{N}(\varepsilon,1)\right)$ for all $q > 1$, but the bound actually holds for all $q > 0$.

Role of complex orders. While recognizing the fundamental gap between the tightest achievable bound on the privacy profile derived from a bound on the Rényi divergence curve and the privacy profile itself, no such tight bound has yet been established. The best available bounds rely on taking an infimum over pointwise conversions from Rényi divergence at a single order $q > 1$ to (ε, δ)-DP [24, 20]. However, since pointwise guarantees offer a lossy characterization of indistinguishability, taking an infimum over privacy profiles implied by these guarantees is unlikely to achieve tight upper bounds. Notably, from the Inverse Laplace Transform expression Equation 5.35 for privacy profiles, we observe that the value of $\delta_{P|Q}$ at any ϵ depends on the behavior of the Rényi divergence $\mathrm{R}_q(P\|Q)$ along the complex line $\mathfrak{Re}(1-q) = \gamma$ for any $\gamma \in \mathbb{R} \setminus \{0, 1\}$, and not along the real line. In fact, the choice of γ on $\mathbb{R}$ does not matter at all, as long as it lies in the $\mathrm{ROC}_{\mathcal{B}}\{\delta_{P|Q}\}$. Thus, we believe that establishing a tight functional conversion will require examining how the Rényi divergence *curls as we move the order q along the complex line $\gamma + i\omega$.*

5.4 Exactly-Tight Composition Theorems

Unlike functional DP guarantees, point guarantees like (ε, δ)-DP or (q, ρ)-Rényi DP are a lossy characterization of the indistinguishability between two distributions P and Q. Despite this, using them for reporting or certifying an algorithm's worst-case privacy is generally acceptable as the privacy protection they guarantee, although conservative, is adequate. The main issue arises when attempting to compose point DP guarantees, as the quantification loss often compounds drastically, resulting in a significant overestimation of the actual privacy protection provided by an algorithm. Consider the k-fold (non-adaptive) composition of a one-dimensional Gaussian mechanism with L_2 sensitivity 1 and noise variance $\sigma^2 = 1$. For individual 1D Gaussian distributions $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(1, 1)$, the privacy profile is of the order $\delta_{P|Q}(\varepsilon) = O(e^{-\varepsilon^2/2})$ (cf. Theorem 28), indicating that the mechanism satisfies a point guarantee of $(O(\sqrt{\log(1/\delta)}), \delta)$-DP for any $\delta \in (0, 1]$. When we extend this to k-fold non-adaptive self-composition, the resulting output distribution is a k-dimensional Gaussian, specifically $P^{\otimes k} = \mathcal{N}(\vec{0}, I_k)$ and $Q^{\otimes k} = \mathcal{N}(\vec{1}, I_k)$. The privacy profile of this composition is of the asymptotic order $\delta_{P^{\otimes k}|Q^{\otimes k}}(\varepsilon) = O(e^{-\varepsilon^2/2k})$, yield-

ing a point guarantee of $(O(\sqrt{k\log(1/\delta)}), \delta)$-DP. However, if we attempt to compose the individual $(O(\sqrt{\log(1/\delta)}), \delta)$-DP point guarantees for each Gaussian mechanism, even with the optimal composition theorem for (ε, δ)-DP point guarantees from Kairouz et al. [56], the best achievable guarantee is $(O(\sqrt{k}\log(1/\delta)), (k+1)\delta)$-DP. This result is significantly more pessimistic than the true privacy profile—a factor of $O(\sqrt{\log(k/\delta)})$ in the ε for the same δ.

Functional notions of DP effectively address this problem by capturing the indistinguishability between any two distributions P and Q accurately [34, 59]. However, Rényi DP (as a function of order q) remains the only functional DP notion with an *exactly tight* composition theorem (i.e., matching even the constants) that can accommodate *adaptively chosen heterogeneous mechanisms*.[11]

Remark 13. *The previous statement requires some justifications. We note that the PLD formalism for composition is limited to non-adaptive mechanisms [96, 59, 47]. On the other hand for f-DP, there is no general composition formula for arbitrary trade-off curves $f_{P_i|Q_i}$. Instead, Dong et al. [34] provides an explicit composition operator expression applicable only to Gaussian trade-off functions (Corollary 3.3). Furthermore, Zhu et al. [122]'s characteristic function formalism Equation 5.4 appears equivalent to Rényi divergence, as we have learned that the order q in Equation 5.23 can assume imaginary values.*

In the following theorem, we provide an *exactly tight* composition result that applies to arbitrary privacy profiles $\delta_{P_1|Q_1}$ and $\delta_{P_2|Q_2}$. Later we also extend this theorem to handle adaptivity.

Theorem 35 (Exactly Tight Composition of Privacy Profiles)**.** *If $P = P_1 \times P_2$ and $Q = Q_1 \times Q_2$ are two product distributions on $\Omega_1 \times \Omega_2$ such that (P_1, Q_1) and (P_2, Q_2) are absolutely continuous at least in one direction, then*

$$\delta_{P|Q}(\varepsilon) = \left(\delta_{P_1|Q_1} \circledast \left(\ddot{\delta}_{P_2|Q_2} - \dot{\delta}_{P_2|Q_2}\right)\right)(\varepsilon) \tag{5.45}$$

$$= \int_{-\infty}^{\infty} \delta_{P_1|Q_1}(\varepsilon - \tau) \cdot \left(\ddot{\delta}_{P_2|Q_2}(\tau) - \dot{\delta}_{P_2|Q_2}(\tau)\right) \mathrm{d}\tau, \tag{5.46}$$

where $\dot{\delta}_{P_2|Q_2}$ and $\ddot{\delta}_{P_2|Q_2}$ are the first and second order gradient functions of $\delta_{P_2|Q_2}$.

[11] Adaptive means each mechanism's output may depend on previous outputs; heterogeneous means the mechanisms need not be identical.

Proof. Let's consider the random variables $Z \sim \mathrm{PLD}(P\|Q)$, $Z_1 \sim \mathrm{PLD}(P_1\|Q_1)$, and $Z_2 \sim \mathrm{PLD}(P_2\|Q_2)$. For a pair $(\Theta_1, \Theta_2) \sim P$, the privacy loss random variable Z is given by $L_{P|Q}(\Theta_1, \Theta_2)$, which simplifies to:

$$\log \frac{P_1(\Theta_1)P_2(\Theta_2)}{Q_1(\Theta_1)Q_2(\Theta_2)} = L_{P_1|Q_1}(\Theta_1) + L_{P_2|Q_2}(\Theta_2). \tag{5.47}$$

This decomposition implies that Z can be expressed as the sum of Z_1 and Z_2, i.e., $Z = Z_1 + Z_2$. Consequently, the probability density of Z, f_Z, is the convolution of f_{Z_1} and f_{Z_2}:

$$f_Z(t) = \int_{-\infty}^{\infty} f_{Z_1}(\tau) f_{Z_2}(t-\tau) d\tau = (f_{Z_1} \circledast f_{Z_2})(t). \tag{5.48}$$

Invoking Theorem 30, we then obtain the Laplace transform of f_Z at $(1-q)$:

$$\mathcal{B}\left\{f_Z(t)\right\}(1-q) \overset{\text{Equation C.8}}{=} \mathcal{B}\left\{f_{Z_1}(t)\right\}(1-q) \cdot \mathcal{B}\left\{f_{Z_2}(t)\right\}(1-q). \tag{5.49}$$

From Definition 11, this directly implies that the Rényi divergence of order q for the pair (P, Q) is the sum of the Rényi divergences for the pairs (P_1, Q_1) and (P_2, Q_2) as

$$\mathrm{R}_q\left(P\|Q\right) = \mathrm{R}_q\left(P_1\|Q_1\right) + \mathrm{R}_q\left(P_2\|Q_2\right). \tag{5.50}$$

Now suppose $s = 1 - q < 0$. Thanks to absolute continuity, we can express the Rényi divergences in terms of the Laplace transform using Theorem 31 as follows, cancelling out the common terms:

$$s(s-1)\mathcal{B}\left\{\delta_{P|Q}(t)\right\}(s) = s(s-1)\mathcal{B}\left\{\delta_{P_1|Q_1}(t)\right\}(s) \cdot s(s-1)\mathcal{B}\left\{\delta_{P_2|Q_2}(t)\right\}(s). \tag{5.51}$$

Since $q = 1 - s$ cannot be 0 or 1, we can divide by $s(s-1)$ on both sides:

$$\begin{aligned}
\mathcal{B}\left\{\delta_{P|Q}(t)\right\}(s) &= \mathcal{B}\left\{\delta_{P_1|Q_1}(t)\right\}(s) \cdot s(s-1)\mathcal{B}\left\{\delta_{P_2|Q_2}(t)\right\}(s) && (5.52)\\
&= \mathcal{B}\left\{\delta_{P_1|Q_1}(t)\right\}(s) \cdot \left(s^2\mathcal{B}\left\{\delta_{P_2|Q_2}(t)\right\}(s) - s\mathcal{B}\left\{\delta_{P_2|Q_2}(t)\right\}(s)\right) && (5.53)\\
&\overset{\text{Equation C.6}}{=} \mathcal{B}\left\{\delta_{P_1|Q_1}(t)\right\}(s) \cdot \left(\mathcal{B}\left\{\ddot{\delta}_{P_2|Q_2}(t)\right\}(s) - \mathcal{B}\left\{\dot{\delta}_{P_2|Q_2}(t)\right\}(s)\right) && (5.54)\\
&\overset{\text{Equation C.1}}{=} \mathcal{B}\left\{\delta_{P_1|Q_1}(t)\right\}(s) \cdot \mathcal{B}\left\{\ddot{\delta}_{P_2|Q_2}(t) - \dot{\delta}_{P_2|Q_2}(t)\right\}(s) && (5.55)\\
&\overset{\text{Equation C.8}}{=} \mathcal{B}\left\{\left(\delta_{P_1|Q_1} \circledast \left(\ddot{\delta}_{P_2|Q_2} - \dot{\delta}_{P_2|Q_2}\right)\right)(t)\right\}(s). && (5.56)
\end{aligned}$$

Hence, from the uniqueness of Laplace transform, we get

$$\forall \varepsilon \in \mathbb{R}, \quad \delta_{P|Q}(\varepsilon) = \left(\delta_{P_1|Q_1} \circledast \left(\ddot{\delta}_{P_2|Q_2} - \dot{\delta}_{P_2|Q_2}\right)\right)(\varepsilon). \tag{5.57}$$

□

Theorem 35 provides an *exactly tight* composition theorem—not only because the terms in Equation 5.45 are equal, but also because the privacy profiles $\delta_{P_1|Q_1}$ and $\delta_{P_2|Q_2}$ precisely capture the indistinguishability of their respective distributions. This theorem mirrors the composition property of Rényi divergence but works in the time domain ε instead of the frequency domain q. It assumes absolute continuity in at least one direction ($P_i \ll Q_i$ or $Q_i \ll P_i$) for both $i = 1$ and 2.

Interestingly however, our result in Equation 5.45 appears to hold even when absolute continuity fails in either direction. We will see an example of this in the next section where we apply Equation 5.45 to compose the privacy profile of randomized mechanisms $\delta_{\mathrm{RR}}^{\varepsilon_1,\delta_1} \otimes \delta_{\mathrm{RR}}^{\varepsilon_2,\delta_2}$ and get a tight expression for $\delta_1, \delta_2 > 0$ without running into a singularity. This happens because even when the Laplace transform is undefined everywhere, the frequency-domain manipulations performed on it still correspond to valid manipulation steps in the time domain. Since our main interest lies in the time domain function—the composed privacy profile—taking advantage of this flexibility proves beneficial. To emphasize the significance of this, Theorem 35 allows us to tightly derive composed privacy profiles with the same exact-tightness as Rényi DP even for mechanisms that do not satisfy Rényi-DP for any order $q \in \mathbb{R} \setminus \{0, 1\}$, opening exciting possibilities beyond the limitations of Rényi DP.

Remark 14. *Formally proving why dropping the absolute continuity assumption in Theorem 35 does not compromise the validity of Equation 5.45 appears to be a challenging yet intriguing problem. Our efforts to resolve this suggests that a mathematical understanding of the number* 0^i *is necessary.*

Adaptive Composition. For two mechanisms $\mathcal{A}_1$ and $\mathcal{A}_2$, if mechanism $\mathcal{A}_2$ sees the output from $\mathcal{A}_1$, then the output distribution of $\mathcal{A}_1$ and $\mathcal{A}_2$ are no longer independent. Theorem 35 can still be applied as long as their exists a distribution pair P_2, Q_2 that *dominates the privacy profile* of output distribution pairs $\mathcal{A}_2(D, \theta)$ and $\mathcal{A}_2(D', \theta)$ across all θ for a given dataset pair $D \simeq D'$, which is a reasonable assumption for adaptively compositing functional guarantees for DP.[12] This is

[12] Adaptive composition for point DP guarantees requires that for fixed $D \simeq D'$, conditioned on any observation from $\mathcal{A}_1$, the point DP guarantee holds for $\mathcal{A}_2$. Adaptively composing DP curves needs a stronger assumption that conditioned on any output of $\mathcal{A}_1$, the privacy profile of $\mathcal{A}_2$ lies below a worst-case DP curve.

possible due to the following result.

Lemma 24 (Zhu et al. [122, Theorem 27])**.** *Let $P(x,y) = P_1(x) \cdot P_2^x(y)$ and $Q(x,y) = Q_1(x) \cdot Q_2^x(y)$ be two joint distributions on $\Omega_1 \times \Omega_2$. Then for any distributions P_2 and Q_2 on Ω_2 such that $\delta_{P_2^x|Q_2^x}(\varepsilon) \leq \delta_{P_2|Q_2}(\varepsilon)$ for all $\varepsilon \in \mathbb{R}$ and $x \in \Omega_1$, we have $\delta_{P|Q}(\varepsilon) \leq \delta_{P_1 \times P_2|Q_1 \times Q_2}(\varepsilon)$.*

Our main composition result in the following theorem follows from applying Lemma 24 to Theorem 35.

Theorem 4. *If algorithms $\mathcal{A}_1 : \mathcal{X}^n \to \mathcal{Y}_1$ and $\mathcal{A}_2 : \mathcal{Y}_1 \times \mathcal{X}^n \to \mathcal{Y}_2$ are such that their exists distribution pairs (P_1, Q_1) on $\mathcal{Y}_1$ and (P_2, Q_2) on $\mathcal{Y}_2$ such that for all neighboring datasets $D \simeq D'$,*

$$\forall \varepsilon \in \mathbb{R} \ : \ \delta_{\mathcal{A}_1(D)|\mathcal{A}_1(D')}(\varepsilon) \leq \delta_{P_1|Q_1}(\varepsilon), \tag{5.58}$$

and

$$\forall \varepsilon \in \mathbb{R}, \ \forall X \in \mathcal{Y}_1 \ : \ \delta_{A_2(X,D)|A_2(X,D')}(\varepsilon) \leq \delta_{P_2|Q_2}(\varepsilon), \tag{5.59}$$

then the composed *mechanism defined as $\mathcal{A}(D) \stackrel{def}{=} (X, Y)$ where $X \sim \mathcal{A}_1(D)$ and $Y \sim \mathcal{A}_2(X, D)$ satisfies $(\varepsilon, \delta(\varepsilon))$-DP for all $\varepsilon \in \mathbb{R}$ with*

$$\delta(\varepsilon) \leq \left(\delta_{P_1|Q_1} \circledast \left(\ddot{\delta}_{P_2|Q_2} - \dot{\delta}_{P_2|Q_2}\right)\right)(\varepsilon) \tag{5.60}$$

where $\dot{\delta}_{P_2|Q_2}(\varepsilon)$ and $\ddot{\delta}_{P_2|Q_2}(\varepsilon)$ are the first and second order gradients of $\delta_{P_2|Q_2}(\varepsilon)$ respectively.

Proof. For any neighbouring datasets $D \simeq D'$, if we denote the distributions of $\mathcal{A}(D)$ and $\mathcal{A}(D')$ as P and Q respectively, we have from Lemma 24 and Theorem 35 that for all $\varepsilon \in \mathbb{R}$,

$$\delta_{P|Q}(\varepsilon) \leq \delta_{P_1 \times P_2|Q_1 \times Q_2}(\varepsilon) = \left(\delta_{P_1|Q_1} \circledast \left(\ddot{\delta}_{P_2|Q_2} - \dot{\delta}_{P_2|Q_2}\right)\right)(\varepsilon). \tag{5.61}$$

□

5.4.1 Tight Composition for (ε, δ)-DP

In this section, we use Theorem 35 to prove an exactly-tight composition theorem for adaptive composition of a sequence of $(\varepsilon_i, \delta_i)$-DP mechanisms. We begin with

a variant of Kairouz et al. [56]'s result that the privacy profile $\delta_{P|Q}(\varepsilon)$ under an (ε, δ)-DP point guarantee is dominated by the privacy profile of randomized response mechanism.

Theorem 36 (Dominating Privacy Profile under (ε, δ)-DP [56])**.** *Fix $\varepsilon \geq 0$ and $\delta \in [0, 1]$. Suppose distributions P and Q over Ω satisfy (ε, δ)-differential privacy. Then,*

$$\forall \varepsilon \in \mathbb{R} \ : \ \delta_{P|Q}(\varepsilon) \leq \delta_{\mathrm{RR}}(\varepsilon) \quad \textit{and} \quad \delta_{Q|P}(\varepsilon) \leq \delta_{\mathrm{RR}}(\varepsilon), \tag{5.62}$$

where $\delta_{\mathrm{RR}}(t)$ is the privacy profile of the randomized response mechanism $\mathcal{A}_{\mathrm{RR}}^{\varepsilon,\delta}$.

Kairouz et al. [56] do not express their notion of *dominance* in the same way as we do in Definition 27—they say distribution pair (P_1, Q_1) dominates (P_2, Q_2) if their trade-off curves satisfy

$$\forall \alpha \in [0, 1] \ : \ f_{P_1|Q_1}(\alpha) \geq f_{P_2|Q_2}(\alpha). \tag{5.63}$$

Nonetheless, our notion of dominance and their notion is equivalent, which was established by [34] by showing that the privacy profile $\delta_{P_i|Q_i}$ and the corresponding trade-off curve $f_{P_i|Q_i}$ are *primal* and *dual* with respect to Frenchel duality. We also provide a direct proof of Theorem 36 in the Appendix C.3.

As a side note, observe that combining Theorem 36 with Theorem 33 gives a tight Rényi DP guarantee for a pure ε-DP mechanism, which has recently attracted interest [79].

$$\delta_{P|Q}(\varepsilon) = \delta_{Q|P}(\varepsilon) = 0 \implies \forall q > 1 \ : \ \mathrm{R}_q\left(P \| Q\right) \leq \frac{1}{q-1} \log\left(\frac{e^{(1-q)\cdot\varepsilon}}{e^{\varepsilon}+1} + \frac{e^{q\cdot\varepsilon}}{e^{\varepsilon}+1}\right). \tag{5.64}$$

Following the objective of this section, we use our Theorem 35 on the above worst-case privacy profile under (ε, δ)-DP point guarantees, resulting in an exactly-tight composition guarantee as stated below.

Theorem 37 (Tight Composition for (ε, δ)-DP)**.** *For any $\varepsilon_i \geq 0$, $\delta_i \in [0, 1]$ for $i \in \{1, \cdots, k\}$, the k-fold composition of $(\varepsilon_i, \delta_i)$-differentially private mechanisms satisfies $(\varepsilon, \delta^{\otimes k}(\varepsilon))$-DP for all ε, defined recursively as*

$$\forall t \in \mathbb{R} \ : \ \delta^{\otimes l}(t) = \delta_l + \frac{(1-\delta_l)}{e^{\varepsilon_l}+1}\left[e^{\varepsilon_l} \cdot \delta^{\otimes l-1}(t-\varepsilon_l) + \delta^{\otimes l-1}(t+\varepsilon_l)\right], \tag{5.65}$$

with $\delta^{\otimes 0}(t) = [1 - e^t]_+$.

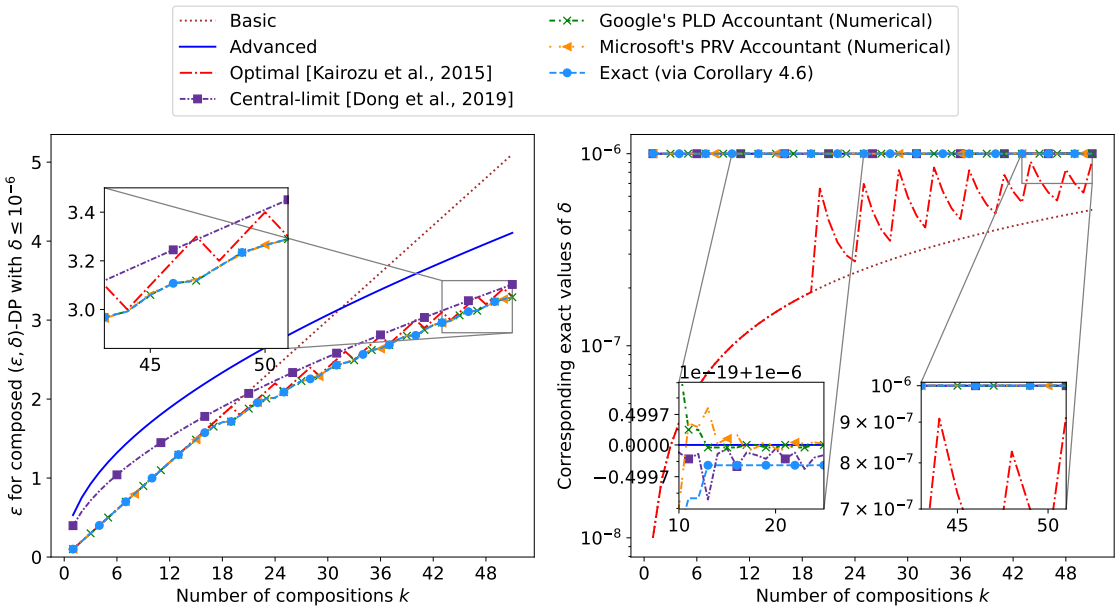


Figure 5.2: Comparison of (ε, δ)-DP bounds from various composition theorems for k-fold composition of a $(0.1, 10^{-8})$-DP point guarantee, with the budget constraint $\delta < 10^{-6}$. The spikes in the right plot, showing exact $\delta < 10^{-6}$ values from Kairouz et al. [56, Theorem 3.3], occur because, out of a set of $\lfloor k/2 \rfloor$ DP point guarantees by their result, we select the smallest ε corresponding to the largest $\delta < 10^{-6}$ in the set, which fluctuates as k increases.

Theorem 37 introduces a convenient recursive method for computing compositions of heterogeneous DP guarantees. While this bound matches each of the $\lfloor k \rfloor$ discrete (ε, δ)-DP values given by the optimal composition theorem from Kairouz et al. [56, Theorem 3.3], our result offers a continuous curve over all $\varepsilon \in \mathbb{R}$. Consequently, for a given budget on δ, our bound provides a tighter limit on ε than that of [56], as shown in Figure 5.2. Furthermore, the recursion reduces to the following exact expression when composing homogeneous DP guarantees.

Corollary 5. *For any $\varepsilon \geq 0$, $\delta \in [0, 1]$, the k-fold composition of (ε, δ)-DP mechanisms satisfies $(\varepsilon, \delta^{\otimes k}(\varepsilon))$-DP for all ε, where*

$$\forall t \in \mathbb{R} \;:\; \delta^{\otimes k}(t) = 1 - (1-\delta)^k \left(1 - \mathop{\mathbb{E}}_{Y \leftarrow \mathrm{Binomial}\left(k, \frac{e^\varepsilon}{1+e^\varepsilon}\right)} \left[1 - e^{t-\varepsilon\cdot(2Y-k)}\right]_+\right). \quad (5.66)$$

Figure 5.2 illustrates the enhanced privacy quantification achieved by our bound for k-fold composition of (ε, δ)-DP guarantees. We also compare our results with Google's PLDAccountant [36] and Microsoft's PRVAccountant [47], which utilize

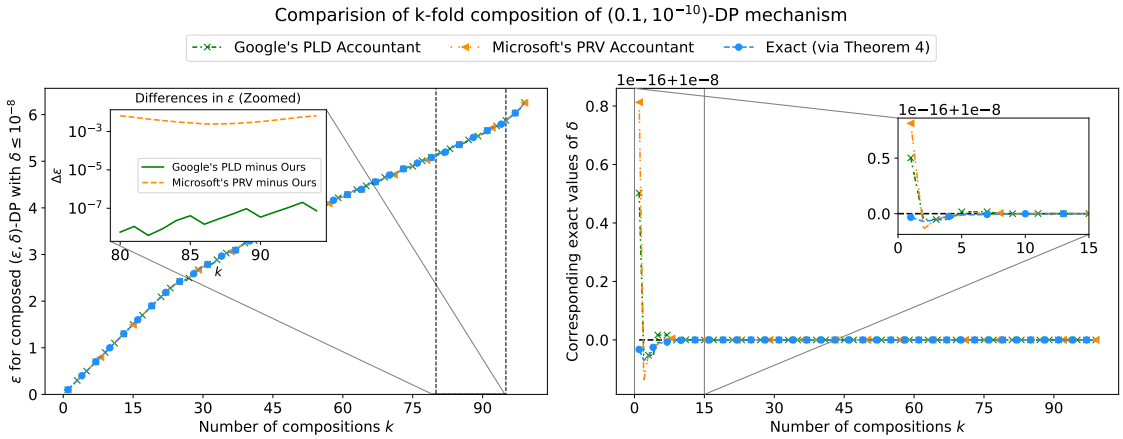


Figure 5.3: Comparison of (ε, δ)-DP bounds between numerical accountants and our Corollary 5 for 100-fold composition of a $(0.1, 10^{-10})$-DP point guarantee, with the budget constraint $\delta < 10^{-8}$. Note that at $k = 100$, the constraint on δ cannot be satisfied and so the corresponding $\varepsilon = \infty$ at that value. We note that at smaller values of k, numerical accountants can sometimes over exceed the budget constraints on δ. Additionally, the gap for ε between our exact bound and those approximated by numerical accountant tend to be of the order $\approx 10^{-7}$ for Google's PLDAccountant and $\approx 10^{-3}$ for Microsoft's PRVAccountant.

the Discrete Fast Fourier Transform. This comparison shows that our analytical approach closely matches the values approximated by these numerical methods. Additionally, through Figure 5.3, we show that these numerical methods can be unstable at edge case values and yield non-negligible gaps in their approximation.

5.5 Asymmetry and DP Notion Equivalences

An important characteristic of functional notions of DP is that they can be *asymmetric*, in the sense that switching $P \leftrightarrow Q$ may yield a different curve $\delta_{Q|P}$ than the original $\delta_{P|Q}$. In context of a mechanism $\mathcal{A}$, such an asymmetry between its output distribution $P = \mathcal{A}(D)$ and $Q = \mathcal{A}(D')$ means that a sample $\Theta \sim P$ might reveal more (or less) information that it came from D than a sample $\Theta' \sim Q$ reveals about coming from D'. Since D and D' are neighboring datasets differing by the presence or absence of a single record, this asymmetry in indistinguishability means that an attacker might have an easier time trying to detect the presence of a record from the output of $\mathcal{M}$ than to detect its absence. In other words, optimal hypothesis tests would experience a skew in the trade-off between their false positive

and false negative rates.

Such skewness often arises due to *subsampling*, which is a heavily used technique to boost an algorithm's intrinsic privacy properties [10, 12, 1, 114, 8]. For instance, Poisson subsampling in the context of the *add or remove* relationship between neighboring datasets, which is commonly used in DP-SGD [1] algorithm, skews the privacy profile of a base mechanism, as illustrated in the following example.

Effect of Poisson Subsampling on $\delta_{P|Q}$. Without loss of generality, assume datasets $D \simeq D'$ are such that the record at index i is present in D but empty in D', i.e., $D[i] \neq D'[i] = \perp$. If we randomly filter the records using an iid selection mask $U \sim \text{Bernoulli}(\lambda)^{\otimes n}$, the subsampled datasets D_U and D'_U are defined as follows for all $i \in [n]$:

$$D_U[i] \stackrel{\text{def}}{=} \begin{cases} D[i] & \text{if } U_i = 1 \\ \perp & \text{otherwise} \end{cases} \quad \text{and} \quad D'_U[i] \stackrel{\text{def}}{=} \begin{cases} D'[i] & \text{if } U_i = 1 \\ \perp & \text{otherwise} \end{cases}. \tag{5.67}$$

The distributions P and Q of the outputs $\mathcal{A}(D_U)$ and $\mathcal{A}(D'_U)$ for any algorithm $\mathcal{A}$ will be identical with probability $1 - \lambda$, which amplifies privacy considerably. More precisely, let P_{IN} and Q_{IN} be the distributions of $\mathcal{A}(D_U)$ and $\mathcal{A}(D'_U)$ conditioned on $i \in U$, and P_{OUT} and Q_{OUT} be the distributions conditioned on $i \notin U$. For any event $S \subseteq \Omega$, we have:

$$\begin{aligned} P(S) &= \Pr[i \notin U] \cdot P_{\text{OUT}}(S) + \Pr[i \in U] \cdot P_{\text{IN}}(S) \\ &= (1 - \lambda) \cdot Q_{\text{OUT}}(S) + \lambda \cdot P_{\text{IN}}(S) \qquad (5.68) \\ &= (1 - \lambda) \cdot Q_{\text{IN}}(S) + \lambda \cdot P_{\text{IN}}(S). \qquad (5.69) \end{aligned}$$

Equation Equation 5.68 holds because if $i \notin U$, then $D_U = D'_U$, and equation Equation 5.69 holds because the ith record in D' is empty, so conditioning on $i \in U$ or $i \notin U$ does not affect the output distribution, i.e., $Q_{\text{IN}} = Q_{\text{OUT}} = Q$. Using this fact, we the following theorem shows the *exact effect* Poisson subsampling has on the privacy profile.

Theorem 38 (Poisson subsampling on add/remove neighbours)**.** *Let $0 < \lambda \leq 1$. For any distributions P, Q, P_{IN} and Q_{IN} such that $P = \lambda P_{\text{IN}} + (1 - \lambda) Q_{\text{IN}}$ and*

$Q = Q_{\text{IN}}$,

$$\delta_{P|Q}(\varepsilon) = \begin{cases} \lambda \delta_{P_{\text{IN}}|Q_{\text{IN}}}(\log(1 + (e^{\varepsilon} - 1)/\lambda)) & \text{if } \varepsilon > \log(1 - \lambda), \\ 1 - e^{\varepsilon} & \text{otherwise.} \end{cases} \tag{5.70}$$

This privacy amplification result was first provided by Li et al. [64], and since has appeared in several works [10, 8, 1, 99]. But unlike other works, we present this amplification effect as a *single curve* that exactly captures the impact of subsampling on both directions. This effect is visualized in the top-left plot in Figure 5.4 (solid orange curve vs. dashed orange curve).

Imprecise Handling of Asymmetry. We observe that several works on privacy handle asymmetric notions of DP in a somewhat imprecise manner, which can introduce significant slack in the analysis or numerical bounds. For example, Dong et al. [34, Theorem 4.2] uses the biconjugation operation $\min\{f_{P|Q}, f_{Q|P}\}^{**}$ to quantify amplification for Poisson subsampling on the trade-off curve $f_{P|Q}$ (corresponding to the subsampled profile $\delta_{P|Q}$ in Theorem 38), leading to an overestimation of the actual trade-off curve $f_{P|Q}$ (see the bottom-right plot in Figure 5.4 for a comparison of the overestimated trade-off after symmetrization in the red line versus the actual trade-off $f_{P|Q}$ in the orange dashed line). Similarly, Dong et al. [34, Proposition 30] defines a dominating profile under subsampling by taking the max of $\delta_{P|Q}$ and $\delta_{Q|P}$ (see the top-left plot in Figure 5.4 to compare the overestimated privacy profile in red with the actual profiles).

The reasoning behind these operations is that a DP guarantee function should hold in both directions, requiring consideration of the worst-case aspects from each direction. However, these symmetrization steps can introduce small gaps that may compound significantly when several privacy profiles are composed. Moreover, these operations disrupt the equivalence across different privacy notions, as after symmetrization, converting to another notion, such as the privacy profile, no longer aligns with the actual privacy profile.

Proposed Solution. We note that all these functional notions of privacy possess a reversal property (see Remarks 9, 10, 11, and 12), which allows us to bypass the need for symmetrization operations. The key idea is to retain the chosen characterizing function *in only one direction*, without attempting to control its asymmetry. Additionally, while composing functional notions, we refrain from

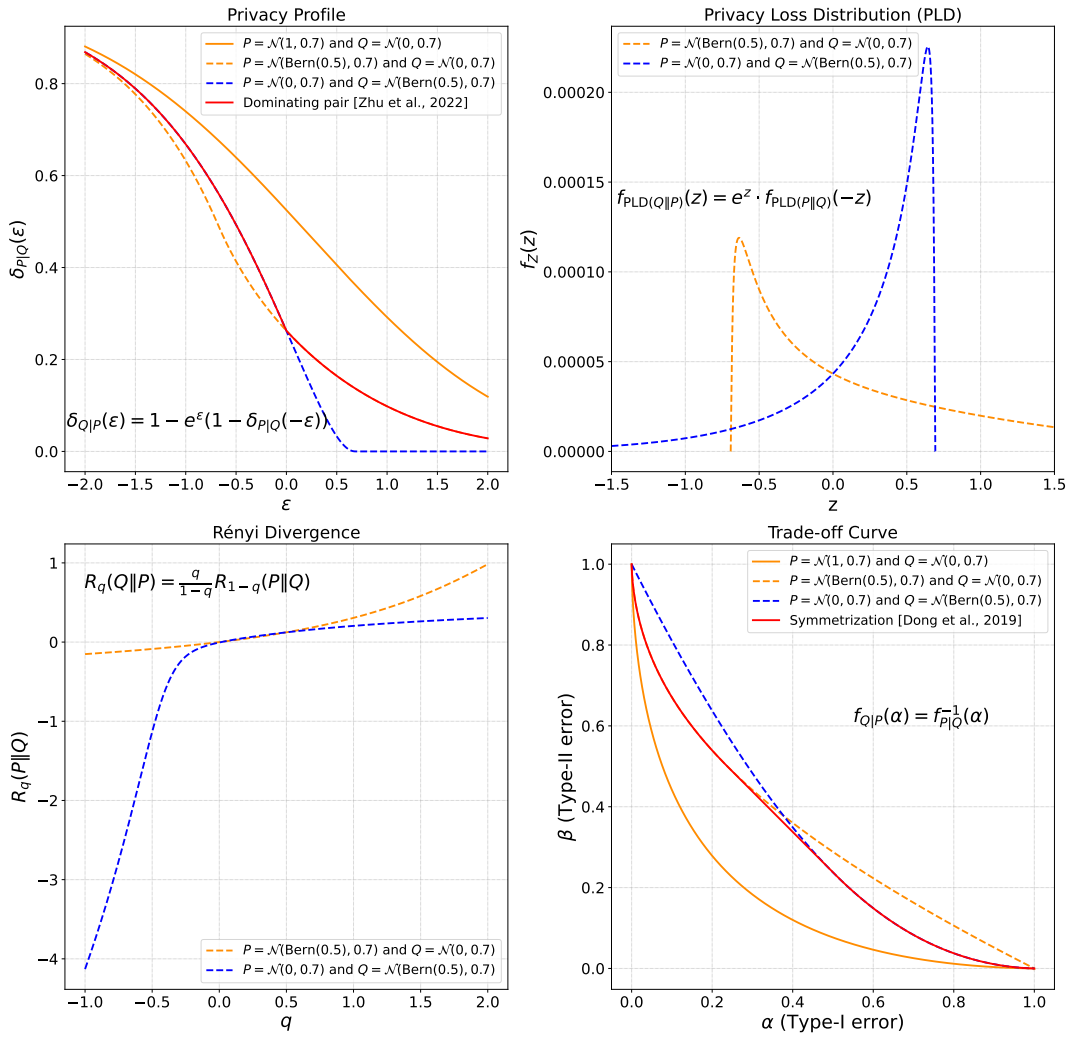


Figure 5.4: Visualization of the four functional notions of DP, namely privacy profile $\delta_{P|Q}(\varepsilon)$ as a function of ε, the generalized density function of privacy loss distribution $\text{PLD}(P\|Q)$, the Rényi divergence $\text{R}_q(P\|Q)$ as a function of order q, and the trade-off function $f_{P|Q}(\alpha)$ for hypothesis testing between P and Q. We also provide the reversal theorems for each of the plots.

changing the notion of adjacency (or its direction). Doing this enables lossless composition operations, thanks to Theorem 35 or Rényi DP composition [69], without running the risk of accidentally underestimating the privacy. When a pointwise DP guarantee is required, we can leverage the reversal properties to query the curve at a specific budget constraint in both directions, providing the max of the two, which results in an exactly-tight pointwise DP guarantee. Additionally, we also avoid the overhead of maintaining the functional representation, such as the PLD, in both directions as currently done in Google's PLDAccountant.

5.6 Discussions and Future Directions

This chapter presents a new interpretation of differential privacy by framing well-known notions—such as the privacy-profile curve and Rényi divergence—in terms of Laplace transforms of the privacy loss distribution. This perspective unites various functional representations of DP, offering a powerful analytical toolkit for understanding and manipulating privacy guarantees. By expressing key DP notions as Laplace transforms, we provide exact conections between them and derive tight composition theorems. The resulting framework not only broadens our theoretical understanding of DP but also paves the way for more nuanced applications and improvements in privacy accounting.

Future work. There are several promising directions where our Laplace transform interpretation could be applied. First, deriving a closed-form, tight subsampling guarantee for Rényi DP remains an unresolved issue [115, 99]. Similarly, obtaining a tight conversion from Concentrated Differential Privacy [20, 39] to (ε, δ)-DP is an ongoing challenge that our framework may help address. Moreover, using the recursive integral expression for privacy profiles in Theorem 35 together with the subsampled Gaussian profile from Theorem 38 could potentially yield an exact, closed-form composition result for Noisy-SGD. Such a result could eliminate the need for numerical accountants like Google's PLDAccountant [36] or Microsoft's PRVAccountant [47].

Chapter 6

Conclusion

This thesis studies the learning and unlearning dynamics of *Noisy Gradient Descent* algorithm from an information-theoretic lens. The notion of information we study is *differential privacy*—the effect a single data-point can impart on a trained model when the point is included in the dataset. We provide results that capture the *dynamics of differential privacy* for Noisy-GD that describes the behavior of information flow from the training dataset into the machine learning model when trained with Noisy-GD. Our study refines the bounds on (ε, δ)-differential privacy guarantees for Noisy-GD, tightening the bound on ε under strong convexity and smoothness from $O(\sqrt{K})$ dependence on number of iteration to only $O(1)$. We also show that this refined bound implies for strongly-convex and smooth losses, Noisy-GD achieves *optimal privacy-utility* trade-off achievable under an (ε, δ)-DP constraint, requiring only $O(n \log n)$ gradient computations, where n is the training dataset size. In contrast, other works in literature either do not achieve optimal privacy-utility trade-off [112] or requires $\Omega(n^2)$ gradient computations [12, 13].

This thesis also demonstrates that Noisy-GD can facilitate *machine unlearning*. By analyzing Noisy-GD's privacy dynamics, we find that simply fine-tuning a pre-trained model on a dataset from which the chosen records have been removed can selectively erase information pertaining to those records. Computationally, unlearning this way using Noisy-GD under strongly convex and smooth losses is very effective, requiring only $O(n)$ gradient computations (i.e., $O(1)$ number of passes over the dataset) while maintaining (ε, δ)-DP guarantee for remaining records as well as optimal utility attainable under this DP constraint.

Additionally, we show that unlearning algorithms must provide (ε, δ)-DP guarantees *in addition to* their (ε_u, δ)-unlearning guarantees to ensure reliable data deletion

in realistic settings where deletion requests can depend on the model's current state. Without a DP guarantee, the released model may retain unbounded information from the original training set. Consequently, even as users add or modify data after the model is published, previously deleted information can be reintroduced into the dataset and future models, undermining the intended unlearning process. We further demonstrate that (ε_u, δ)-unlearning algorithms that also achieve (ε, δ)-DP guarantees provide $(\varepsilon_u + \varepsilon, 2\delta)$-unlearning guarantees under adaptive deletion requests—an elegant way to extend non-adaptive unlearning guarantees to adaptive ones.

In the last chapter, this thesis presents a novel interpretation of differential privacy by leveraging time-frequency dualities across various functional representations, including privacy profiles, Rényi divergence, and privacy loss distributions. By framing these within the context of Laplace transforms, we develop a versatile analytical toolkit for DP, enhancing both theoretical understanding and practical composition methods. Our approach addresses limitations in existing adaptive composition bounds, provides continuous guarantees for composed privacy profiles, and bridges gaps between different DP frameworks without needing approximations or symmetrizations. Together, these results push forward the capabilities of differential privacy research, setting a foundation for more nuanced and robust privacy-preserving algorithms.

Bibliography

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Kenneth Adelman and Gabrielle Adelman. California coastal records project, 2002.

[3] Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially private identity and closeness testing of discrete distributions. *arXiv preprint arXiv:1707.05497*, 2017.

[4] Jason Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *Advances in Neural Information Processing Systems*, 35:3788–3800, 2022.

[5] Shahab Asoodeh, Jiachun Liao, Flavio P Calmon, Oliver Kosut, and Lalitha Sankar. A better bound gives a hundred rounds: Enhanced privacy guarantees via f-divergences. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 920–925. IEEE, 2020.

[6] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

[7] Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403. PMLR, 2018.

[8] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31, 2018.

[9] Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. In *Advances in Cryptology–CRYPTO 2019: 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2019, Proceedings, Part II 39*, pages 638–667. Springer, 2019.

[10] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1), 2020.

[11] Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2496–2506. PMLR, 2020.

[12] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE, 2014.

[13] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

[14] Ernst Benda. The protection of human dignity (article 1 of the basic law). *SMUL Rev.*, 53:443, 2000.

[15] Jerry Berman and Paula Bruening. Is privacy still possible in the twenty-first century? *Social Research*, pages 306–318, 2001.

[16] Lillian R BeVier. Information about individuals in the hands of government: Some reflections on mechanisms for privacy protection. *Wm. & Mary Bill Rts. J.*, 4:455, 1995.

[17] Sergey G Bobkov. On isoperimetric constants for log-concave probability distributions. In *Geometric aspects of functional analysis*, pages 81–88. Springer, 2007.

[18] Jinho Bok, Weijie Su, and Jason M Altschuler. Shifted interpolation for differential privacy. *arXiv preprint arXiv:2403.00278*, 2024.

[19] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

[20] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[21] Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 634–649. IEEE, 2015.

[22] Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv'it: Private and sample efficient identity testing. In *International Conference on Machine Learning*, pages 635–644. PMLR, 2017.

[23] Clément L Canonne, Gautam Kamath, Audra McMillan, Jonathan Ullman, and Lydia Zakynthinou. Private identity testing for high-dimensional distributions. *Advances in neural information processing systems*, 33:10099–10111, 2020.

[24] Clément L Canonne, Gautam Kamath, and Thomas Steinke. The discrete gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33:15676–15688, 2020.

[25] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.

[26] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

[27] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 896–911, 2021.

[28] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Advances in Cryptology–EUROCRYPT 2019: 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19–23, 2019, Proceedings, Part I 38*, pages 375–403. Springer, 2019.

[29] Sinho Chewi, Murat A Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of langevin monte carlo from poincar\'e to log-sobolev. *arXiv preprint arXiv:2112.12662*, 2021.

[30] Alan M Cohen. *Numerical methods for Laplace transform inversion*, volume 5. Springer Science & Business Media, 2007.

[31] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[32] Bernhard Debatin. Ethics, privacy, and self-restraint in social networking. In *Privacy online: Perspectives on privacy and self-disclosure in the social web*, pages 47–60. Springer, 2011.

[33] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[34] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

[35] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212, 1983.

[36] Vadym Doroshenko, Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. Connect the dots: Tighter discrete approximations of privacy loss distributions. *arXiv preprint arXiv:2207.04380*, 2022.

[37] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

[38] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

[39] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[40] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

[41] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407, 2014.

[42] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014.

[43] Phil PG Dyke and PP Dyke. *An introduction to Laplace transforms and Fourier series*. Springer, 2001.

[44] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.

[45] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[46] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[47] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34:11631–11642, 2021.

[48] Dimitris Gritzalis, Miltiadis Kandias, Vasilis Stavrou, and Lilian Mitrou. History of information: the case of privacy and security in social media. In *Proc. of the History of Information Conference*, pages 283–310, 2014.

[49] Leonard Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[50] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

[51] Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34, 2021.

[52] R Holley and D Stroock. Logarithmic sobolev inequalities and stochastic ising models. *J. Stat. Phys.;(United States)*, 46(5/6), 1987.

[53] Jan Holvast. History of privacy. In *The history of information security*, pages 737–769. Elsevier, 2007.

[54] Yiyang Huang and Clément L Canonne. Tight bounds for machine unlearning via differential privacy. *arXiv preprint arXiv:2309.00886*, 2023.

[55] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pages 2008–2016. PMLR, 2021.

[56] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[57] Gautam Kamath and Jonathan Ullman. A primer on private statistics. *arXiv preprint arXiv:2005.00010*, 2020.

[58] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

[59] Antti Koskela, Joonas Jälkö, and Antti Honkela. Computing tight differential privacy guarantees using fft. In *International Conference on Artificial Intelligence and Statistics*, pages 2560–2569. PMLR, 2020.

[60] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[61] Michel Ledoux. The geometry of markov diffusion generators. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 305–366, 2000.

[62] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

[63] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE, 2006.

[64] Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, pages 32–33, 2012.

[65] Qinbin Li, Zhaomin Wu, Zeyi Wen, and Bingsheng He. Privacy-preserving gradient boosting decision trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 784–791, 2020.

[66] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1):3–es, 2007.

[67] Shanti Escalante-De Mattei. Midjourney and other ai art image generators face legal challenges, 2023. URL https://www.artnews.com/art-in-america/features/midjourney-ai-art-image-generators-lawsuit-1234665579/. Accessed: 2024-06-18.

[68] James F Miller. Moment-generating functions and laplace transforms. *Journal of the Arkansas Academy of Science*, 4(1):97–100, 1951.

[69] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[70] Ilya Mironov, Kunal Talwar, and Li Zhang. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

[71] Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pages 157–175. Springer, 2015.

[72] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pages 931–962. PMLR, 2021.

[73] Aleš Nekvinda and Luděk Zajíček. A simple proof of the rademacher theorem. *Časopis pro pěstování matematiky*, 113(4):337–341, 1988.

[74] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[75] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

[76] Alan V Oppenhiem, Alan S Willsky, and S Hamid Nawab. Signals and systems, 1996.

[77] Jeremy Orloff. Uniqueness of laplace transform, 2015.

[78] Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[79] Differential Privacy. Tight RDP & zCDP Bounds from Pure DP — differentialprivacy.org. https://differentialprivacy.org/pdp-to-zcdp/, 2024. [Accessed 12-11-2024].

[80] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1. Berkeley, California, USA, 1961.

[81] Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

[82] Beate Roessler and Judith DeCew. Privacy. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

[83] Antoinette Rouvroy and Yves Poullet. The right to informational self-determination and the value of self-development: Reassessing the importance of privacy for democracy. In *Reinventing data protection?*, pages 45–76. Springer, 2009.

[84] r/StableDiffusion. Reddit search results for "samdoesarts" in r/stablediffusion, 2024. URL https://www.reddit.com/r/StableDiffusion/search/?q=samdoesarts. Accessed: 2024-06-18.

[85] Jed Rubenfeld. The right of privacy. *Harvard Law Review*, pages 737–807, 1989.

[86] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *arXiv preprint arXiv:0911.5708*, 2009.

[87] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

[88] Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning*, pages 982–990. PMLR, 2014.

[89] Sebastian Schelter. Amnesia-a selection of machine learning models that can forget user data very fast. *suicide*, 8364(44035):46992, 2020.

[90] Sebastian Schelter, Mozhdeh Ariannezhad, and Maarten de Rijke. Forget me now: Fast and exact unlearning in neighborhood-based recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2011–2015, 2023.

[91] Ferdinand David Schoeman. *Philosophical dimensions of privacy: An anthology*. Cambridge University Press, 1984.

[92] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18075–18086. Curran Associates, Inc., 2021.

[93] Saurabh Shintre, Kevin A Roundy, and Jasjeet Dhaliwal. Making machine learning forget. In *Privacy Technologies and Policy: 7th Annual Privacy Forum, APF 2019, Rome, Italy, June 13–14, 2019, Proceedings 7*, pages 72–83. Springer, 2019.

[94] Daniel J Solove. Conceptualizing privacy. *Calif. L. Rev.*, 90:1087, 2002.

[95] Daniel J Solove. A brief history of information privacy law. *Proskauer on privacy, PLI*, 2016.

[96] David Sommer, Sebastian Meiser, and Esfandiar Mohammadi. Privacy loss classes: The central limit theorem in differential privacy. *Cryptology ePrint Archive*, 2018.

[97] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pages 245–248. IEEE, 2013.

[98] G. Spence. *Code Napoleon: Or, the French Civil Code*. William Benning Law, 1827. URL https://books.google.com.sg/books?id=NZoB0AEACAAJ.

[99] Thomas Steinke. Composition of differential privacy & privacy amplification by subsampling. *arXiv preprint arXiv:2210.00597*, 2022.

[100] Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997.

[101] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

[102] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

[103] Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pages 819–850. PMLR, 2013.

[104] Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022.

[105] Enayat Ullah, Tung Mai, Anup Rao, Ryan A Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pages 4126–4142. PMLR, 2021.

[106] u/Sandro Halpo. A much less antagonizing post about a free samdoesart model that will probably trigger the man anyways just because it exists..., 2022. URL https://www.reddit.com/r/StableDiffusion/comments/yrjtwu/a_much_less_antagonizing_post_about_a_free/. Accessed: 2024-06-18.

[107] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[108] Leonid Nisonovich Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

[109] Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376 (2133):20180083, 2018.

[110] Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

[111] Eduard Fosch Villaronga, Peter Kieseberg, and Tiffany Li. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law & Security Review*, 34(2):304–313, 2018.

[112] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, pages 2722–2731, 2017.

[113] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *arXiv preprint arXiv:1802.05251*, 2018.

[114] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502. PMLR, 2015.

[115] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd international conference on artificial intelligence and statistics*, pages 1226–1235. PMLR, 2019.

[116] Zuobin Xiong, Wei Li, Yingshu Li, and Zhipeng Cai. Exact-fun: An exact and efficient federated unlearning approach. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 1439–1444. IEEE, 2023.

[117] Haonan Yan, Xiaoguang Li, Ziyao Guo, Hui Li, Fenghua Li, and Xiaodong Lin. Arcane: An efficient architecture for exact machine unlearning. In *IJCAI*, volume 6, page 19, 2022.

[118] Jiayuan Ye and Reza Shokri. Differentially private learning needs hidden state (or much faster convergence). *Advances in Neural Information Processing Systems*, 35:703–715, 2022.

[119] Sihao Yu, Fei Sun, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. Legonet: A fast and exact unlearning architecture. *arXiv preprint arXiv:2210.16023*, 2022.

[120] Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private erm for smooth objectives. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3922–3928, 2017.

[121] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[122] Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, pages 4782–4817. PMLR, 2022.

Appendix A

Supplement for Chapter 3

A.1 Deferred Proofs for § 3.2.1

Lemma 7. *For coupled tracing diffusion processes Equation 3.4 in time $\eta k < t \leq \eta(k+1)$, the equivalent Fokker-Planck equations are*

$$\begin{cases} \partial_t P_t(\theta) = \nabla \cdot (P_t(\theta)\mathbf{V}_t(\theta)) + \sigma^2 \Delta P_t(\theta), \\ \partial_t Q_t(\theta) = \nabla \cdot (Q_t(\theta)\mathbf{V}'_t(\theta)) + \sigma^2 \Delta Q_t(\theta), \end{cases} \tag{A.1}$$

where $\mathbf{V}_t(\theta) = \mathop{\mathbb{E}}_{\Theta_{\eta k} \sim P_{\eta k|t}} [\mathbf{U}(\Theta_{\eta k}) | \Theta_t = \theta]$ *and* $\mathbf{V}'_t(\theta) = \mathop{\mathbb{E}}_{\Theta'_{\eta k} \sim Q_{\eta k|t}} \left[-\mathbf{U}(\Theta'_{\eta k}) | \Theta'_t = \theta\right]$ *are time-dependent maps, and* $\mathbf{U}(\theta) = \frac{1}{2}(\nabla \mathcal{L}_D(\theta) - \nabla \mathcal{L}_{D'}(\theta))$.

Proof. We only prove $\partial_t P_t(\theta) = \nabla \cdot (P_t(\theta)\mathbf{V}_t(\theta)) + \sigma^2 \Delta P_t(\theta)$. The proof for the other Fokker-Planck equation is similar. Recall that conditionals of joint distribution $P_{\eta k,t}$ is

$$P_{\eta k,t}(\theta_k, \theta) = P_{\eta k}(\theta_k) P_{t|\eta k}(\theta|\theta_k) = P_t(\theta) P_{\eta k|t}(\theta_k|\theta). \tag{A.2}$$

By marginalizing away θ_k in Equation A.2, and taking partial derivative w.r.t. t on both sides, we obtain the following:

$$\begin{aligned} \partial_t P_t(\theta) &= \int_{\mathbb{R}^d} \partial_t P_{t|\eta k}(\theta|\theta_k) P_{\eta k}(\theta_k) \mathrm{d}\theta_k \\ &= \int_{\mathbb{R}^d} \left(\nabla \cdot (P_{\eta k,t}(\theta_k, \theta)\mathbf{U}(\theta_k)) + \sigma^2 \Delta P_{\eta k,t}(\theta_k, \theta)\right) \mathrm{d}\theta_k \quad \text{(By Equation 3.5)} \\ &= \nabla \cdot \left(P_t(\theta) \int_{\mathbb{R}^d} P_{\eta k|t}(\theta_k|\theta)\mathbf{U}(\theta_k)\mathrm{d}\theta_k\right) + \sigma^2 \Delta P_t(\theta) \\ &= \nabla \cdot \left(P_t(\theta) \mathop{\mathbb{E}}_{\Theta_{\eta k} \sim P_{\eta k|t}} [\mathbf{U}(\Theta_{\eta k}) | \Theta_t = \theta]\right) + \sigma^2 \Delta P_t(\theta) \\ &= \nabla \cdot (P_t(\theta) \cdot \mathbf{V}_t(\theta)) + \sigma^2 \Delta P_t(\theta). \\ &\qquad \text{(where } \mathbf{V}_t(\theta) = \mathop{\mathbb{E}}_{\Theta_{\eta k} \sim P_{\eta k|t}} [\mathbf{U}(\Theta_{\eta k}) | \Theta_t = \theta]) \end{aligned}$$

□

A.2 Deferred Proofs for § 3.2.2

Lemma 25 (Leibniz integral rule)**.** *Suppose $f_t : \mathbb{R}^d \to \mathbb{R}$ is Lebesgue-integrable for each $t \geq 0$. If for almost all $\theta \in \mathbb{R}^d$, the derivative $\frac{\mathrm{d}f_t}{\mathrm{d}t}$ exists and there exists an integrable function $g : \mathbb{R}^d \to \mathbb{R}$ such that $\left|\frac{\partial f_t}{\partial t}(\theta)\right| \leq g(\theta)$ for all $t \geq 0$ and almost every $\theta \in \mathbb{R}^d$, then*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} f_t(\theta)\mathrm{d}\theta = \int_{\mathbb{R}^d} \frac{\partial f_t}{\partial t}(\theta)\mathrm{d}\theta, \quad \text{for all } t \geq 0. \tag{A.3}$$

Lemma 8 (Rate of Rényi privacy loss)**.** *Let $\mathbf{V}_t$ and $\mathbf{V}'_t$ be two vector fields on $\mathbb{R}^d$ corresponding to a pair of arbitrarily chosen neighboring datasets D and D' with $\max_{\theta\in\mathbb{R}^d} \|\mathbf{V}_t(\theta) - \mathbf{V}'_t(\theta)\|_2 \leq S_v$ for all $t \geq 0$. Then, for corresponding coupled diffusions $\{P_t\}_{t\geq 0}$ and $\{Q_t\}_{t\geq 0}$ under vector fields $\mathbf{V}_t$ and $\mathbf{V}'_t$ and noise variance σ^2, the Rényi privacy loss rate at any $t \geq 0$ is upper bounded by*

$$\partial_t \mathrm{R}_q\left(P_t \| Q_t\right) \leq \frac{1}{\gamma}\frac{qS_v^2}{4\sigma^2} - (1-\gamma)\sigma^2 q \frac{\mathrm{I}_q\left(P_t\|Q_t\right)}{\mathrm{E}_q\left(P_t\|Q_t\right)}, \tag{A.4}$$

where $\gamma > 0$ is a tuning parameter that we can fix arbitrarily according to our need.

Proof. For brevity, let the functions $R(q,t) = \mathrm{R}_q\left(P_t\|Q_t\right)$, $E(q,t) = \mathrm{E}_q\left(P_t\|Q_t\right)$, and $I(q,t) = \mathrm{I}_q\left(P_t\|Q_t\right)$. Under the stated assumptions $\frac{\partial E(q,t)}{\partial t}$ is bounded as follows.

$$\frac{\partial E(q,t)}{\partial t} = \frac{\partial}{\partial t}\int_{\mathbb{R}^d} \frac{P_t^q}{Q_t^{q-1}}\mathrm{d}\theta \tag{A.5}$$

By Leibniz integral rule (Lemma 25), we exchange order of derivative and integration in Equation A.5. The necessary conditions are satisfied because of the following properties about P_t and Q_t:

1. P_t is absolutely continuous with respect to Q_t.
2. The distributions of coupled tracing diffusions $\{P_t\}_{\eta k<t\leq\eta(k+1)}$ and $\{Q_t\}_{\eta k<t\leq\eta(k+1)}$ have full support and smooth densities due to convolution with Gaussian noise.
3. The evolutions of probability densities P_t and Q_t with regard to time t satisfy the Fokker-Planck equations Equation 3.5.

Therefore, we obtain:

$$\frac{\partial E(q,t)}{\partial t} = q\int_{\mathbb{R}^d} \frac{\partial P_t}{\partial t}\left(\frac{P_t}{Q_t}\right)^{q-1} \mathrm{d}\theta - (q-1)\int_{\mathbb{R}^d} \frac{\partial Q_t}{\partial t}\left(\frac{P_t}{Q_t}\right)^{q} \mathrm{d}\theta \qquad \text{(By Lemma 25)}$$

$$= q\int_{\mathbb{R}^d}\left(\sigma^2 \Delta P_t + \nabla\cdot(P_t \mathbf{V}_t)\right)\left(\frac{P_t}{Q_t}\right)^{q-1}\mathrm{d}\theta$$

$$- (q-1)\int_{\mathbb{R}^d}\left(\sigma^2 \Delta Q_t + \nabla\cdot(Q_t \mathbf{V}'_t)\right)\left(\frac{P_t}{Q_t}\right)^{q}\mathrm{d}\theta \qquad \text{(From Equation 3.6)}$$

$$= \underbrace{\sigma^2 q\int_{\mathbb{R}^d}\left(\frac{P_t}{Q_t}\right)^{q-1}\Delta P_t \mathrm{d}\theta - \sigma^2(q-1)\int_{\mathbb{R}^d}\left(\frac{P_t}{Q_t}\right)^{q}\Delta Q_t \mathrm{d}\theta}_{\stackrel{\text{def}}{=}F_1}$$

$$+ \underbrace{q\int_{\mathbb{R}^d}\left(\frac{P_t}{Q_t}\right)^{q-1}\nabla\cdot((P_t\mathbf{V}_t))\,\mathrm{d}\theta - (q-1)\int_{\mathbb{R}^d}\left(\frac{P_t}{Q_t}\right)^{q}\nabla\cdot((Q_t\mathbf{V}'_t))\,\mathrm{d}\theta}_{\stackrel{\text{def}}{=}F_2}$$

We simplify F_1 as following:

$$F_1 = \sigma^2(q-1)\int_{\mathbb{R}^d}\left\langle \nabla\left(\frac{P_t}{Q_t}\right)^{q}, \nabla Q_t\right\rangle \mathrm{d}\theta - \sigma^2 q\int_{\mathbb{R}^d}\left\langle \nabla\left(\frac{P_t}{Q_t}\right)^{q-1}, \nabla P_t\right\rangle \mathrm{d}\theta$$

(From Equation 2.24)

$$= \sigma^2 q(q-1)\int_{\mathbb{R}^d}\left(\frac{P_t}{Q_t}\right)^{q-2}\left\langle \nabla\frac{P_t}{Q_t}, \frac{P_t}{Q_t^2}\nabla Q_t - \frac{\nabla P_t}{Q_t}\right\rangle Q_t \mathrm{d}\theta$$

$$= -\sigma^2 q(q-1)\mathop{\mathbb{E}}_{Q_t}\left[\left(\frac{P_t}{Q_t}\right)^{q-2}\left\|\nabla\frac{P_t}{Q_t}\right\|_2^2\right] \qquad \left(\because \nabla\tfrac{P_t}{Q_t} = \tfrac{\nabla P_t}{Q_t} - \tfrac{P_t}{Q_t^2}\nabla Q_t\right)$$

$$= -\sigma^2 q(q-1)\mathop{\mathbb{E}}_{Q_t}\left[\left\|\nabla\left[\left(\frac{P_t}{Q_t}\right)^{\frac{q}{2}}\right]\right\|_2^2\right]$$

$$= -\sigma^2 q(q-1) I(q,t) \qquad \text{(From Equation 2.10)}$$

We upper bound F_2 as following:

$$F_2 = -q \int_{\mathbb{R}^d} \left\langle \nabla \left(\frac{P_t}{Q_t}\right)^{q-1}, P_t \mathbf{V}_t \right\rangle \mathrm{d}\theta \; + (q-1) \int_{\mathbb{R}^d} \left\langle \nabla \left(\frac{P_t}{Q_t}\right)^{q}, Q_t \mathbf{V}_t' \right\rangle \mathrm{d}\theta$$

(From Equation 2.23)

$$= q(q-1) \int_{\mathbb{R}^d} \left(\frac{P_t}{Q_t}\right)^{q-2} \left\langle \nabla \frac{P_t}{Q_t}, \frac{P_t}{Q_t} \left(\mathbf{V}_t' - \mathbf{V}_t\right) \right\rangle Q_t \mathrm{d}\theta$$

$$\leq \gamma q(q-1)\sigma^2 \int_{\mathbb{R}^d} \left(\frac{P_t}{Q_t}\right)^{q-2} \left\| \nabla \frac{P_t}{Q_t} \right\|_2^2 Q_t \mathrm{d}\theta$$

(From Equation 2.25 with $b = 2\gamma\sigma^2$)

$$+ \frac{q(q-1)S_v^2}{4\gamma\sigma^2} \int_{\mathbb{R}^d} \left(\frac{P_t}{Q_t}\right)^{q-2} \times \left(\frac{P_t}{Q_t}\right)^{2} Q_t \mathrm{d}\theta \quad (\because \max_{\theta \in \mathbb{R}^d} \|\mathbf{V}_t(\theta) - \mathbf{V}_t'(\theta)\|_2 \leq S_v)$$

$$= \gamma\sigma^2 q(q-1) I(q,t) + \frac{1}{\gamma} \frac{q(q-1)S_v^2}{4\sigma^2} E(q,t)$$

(From Equation 2.2 & Equation 2.10)

Therefore, we get the following bound on the rate of Renyi divergence:

$$\begin{aligned} \frac{\partial R(q,t)}{\partial t} &= \frac{1}{q-1} \times \frac{1}{E(q,t)} \times \frac{\partial E(q,t)}{\partial t} \\ &\leq -(1-\gamma)\sigma^2 q \frac{I(q,t)}{E(q,t)} + \frac{1}{\gamma} \frac{qS_v^2}{4\sigma^2} \end{aligned}$$

□

Discussions about the terms in Lemma 8. Lemma 8 bounds the rate of privacy loss with various terms. Generally speaking, the term $\frac{qS_v^2}{4\sigma^2}$ bounds the worst-case privacy loss growth due to noisy gradient update when $S_v = S_\ell$, while the term $\frac{\mathrm{I}_q(P_t \| Q_t)}{\mathrm{E}_q(P_t \| Q_t)}$ amplifies our bound for the rate of privacy loss, as the Rényi privacy loss accumulates during the process. We offer more explanations as the following.

1. $\frac{qS_v^2}{4\sigma^2}$: This is the first term in the right hand side of Equation 3.7. It quantifies the worst-case privacy loss due of one noisy gradient update in Noisy-GD Algorithm 2. The constant S_ℓ is the sensitivity of loss gradient $\ell(\theta; \mathbf{x})$ over two data-points $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. The larger S_ℓ is, the further apart the parameters θ and θ' after the gradient descent updates on two neighboring dataset $D \simeq D'$ could be, where $\theta = \theta_0 - \eta \nabla \mathcal{L}_D(\theta_0)$ and $\theta' = \theta_0 - \eta \nabla \mathcal{L}_{D'}(\theta_0)$. The term σ^2 is the variance of Gaussian noise. Because additive noise shrink the expected

trajectory difference between θ and θ' in Noisy-GD updates, the larger σ^2 is, the more indistinguishable the distributions of sum of θ, θ' and Gaussian noise will be, therefore the smaller the privacy loss (Rényi divergence between end distributions) will be.

2. $\frac{\mathrm{I}_q(P_t\|Q_t)}{\mathrm{E}_q(P_t\|Q_t)}$: This term is the second term in the right-hand side of Equation 3.7, which originates from the partial derivative of P_t, Q_t with regard to time t. To obtain the expression I_q/E_q, we are using the Fokker-Planck equation to substitute the terms related to $\partial_t P_t, \partial_t Q_t$ with terms determined by the gradient and Laplacian of P_t, Q_t over θ.

   The term $\mathrm{I}_q\left(P_t\|Q_t\right)$ is the *Rényi information* defined in Definition 10, which equals $\mathop{\mathbb{E}}_{\Theta\sim Q_t}\left[\left\|\nabla \log \frac{P_t(\Theta)}{Q_t(\Theta)}\right\|_2^2 \left(\frac{P_t(\Theta)}{Q_t(\Theta)}\right)^q\right]$. The term $\mathrm{E}_q\left(P_t\|Q_t\right)$ is the *moment of likelihood ratio* defined in Definition 5, which equals $\mathop{\mathbb{E}}_{\Theta\sim Q_t}\left[\left(\frac{P_t(\Theta)}{Q_t(\Theta)}\right)^q\right]$. These two terms differ by a *multiplicative ratio* $\left\|\nabla \frac{P_t(\Theta)}{Q_t(\Theta)}\right\|_2^2$ for their quantities inside expectation. This ratio characterizes the variation of log likelihood ratio function across Θ, where Θ is taken from distribution Q_t. This is intuitive in the one dimensional case, because $\int_{\theta_1}^{\theta_2} \nabla \log \frac{P_t(\theta)}{Q_t(\theta)} \mathrm{d}\theta = \log \frac{P_t(\theta_2)}{Q_t(\theta_2)} - \log \frac{P_t(\theta_1)}{Q_t(\theta_2)}$. Meanwhile, since $P_t(\theta), Q_t(\theta)$ are continuous and $\int P_t(\theta)\mathrm{d}\theta = \int Q_t(\theta)\mathrm{d}\theta = 1$, by mean value theorem, there exists $\tilde{\theta} \in \mathbb{R}^d$ such that the log likelihood ratio $\log \frac{P_t(\tilde{\theta})}{Q_t(\tilde{\theta})}$ is zero. Therefore, the variation of log likelihood ratio across θ implicitly increases the largest log likelihood ratio $\max_{\theta\in\mathbb{R}^d}\left[\log(\frac{P_t(\theta)}{Q_t(\theta)}) - \log(\frac{P_t(\tilde{\theta})}{Q_t\tilde{\theta})})\right] = \max_{\theta\in\mathbb{R}^d}\left[\log(\frac{P_t(\theta)}{Q_t(\theta)})\right]$ across θ, which reflects the Rńyi privacy loss R_q.

   As a result, intuitively, under some conditions, the larger the Rényi privacy loss R_q is, the larger the variation of log likelihood ratio across θ will be, and therefore the larger the term $\frac{\mathrm{I}_q(P_t\|Q_t)}{\mathrm{E}_q(P_t\|Q_t)}$ will be. Therefore, when the Rényi privacy loss R_q is large, the bound for the rate of privacy loss in Equation 3.7 in Lemma 8 will also be smaller (under $(1-\gamma) > 0$).

3. γ is a tuning constant to balance the privacy growth rate estimated using the above two terms, thus helping us tune the privacy loss accumulation.

A.3 Naïve DP guarantees for Noisy-GD

In this section, we derive naíve Rényi differential privacy bounds for Noisy-GD (Algorithm 2) algorithm and compare it to the (ε, δ)-DP bound by Abadi et al. [1].

Theorem 39 (Rényi DP guarantee for Noisy-GD Algorithm 2)**.** *If $\ell(\theta; \mathbf{x})$ has gradient sensitivity S_ℓ, then Noisy-GD satisfies (q, ρ)-Rényi DP with $\rho = \frac{qS_\ell^2}{4\hat{\sigma}^2} \cdot \eta K$.*

Proof. The L_2 sensitivity of gradient $\nabla \mathcal{L}_D(\theta) = \frac{1}{n} \sum_{\mathbf{x} \in D} \nabla \ell(\theta; \mathbf{x})$ computed in step 2 of Algorithm 2 for neighboring databases in $\mathcal{X}^n$ that differ in a single record is $\frac{S_\ell}{n}$ since $\ell(\theta; \mathbf{x})$ has a gradient sensitivity of S_ℓ.

Conditioned on observing the intermediate model $\Theta_{\eta k} = \theta_k$ at step k, the next model $\Theta_{\eta(k+1)}$ after the noisy gradient update is a Gaussian mechanism with noise variance $2\hat{\sigma}^2/\eta n^2$. So, for neighboring databases $D, D' \in \mathcal{X}^n$, we have from the Rényi DP bound of Gaussian mechanisms proposed by Mironov [69, Proposition 7] that

$$\mathrm{R}_q \left(\Theta_{\eta(k+1)} \mid_{\Theta_{\eta k} = \theta_k} \middle\| \Theta'_{\eta(k+1)} \mid_{\Theta'_{\eta k} = \theta_k} \right) \leq \frac{\eta q S_\ell^2}{4\hat{\sigma}^2}, \tag{A.6}$$

where $(\Theta_{\eta k})_{0 \leq k \leq K}$ and $(\Theta'_{\eta k})_{0 \leq k \leq K}$ are intermediate parameters in Algorithm 2 when run on databases D and D' respectively. Finally, from Rényi composition Mironov [69, Proposition 1], we have

$$\begin{aligned}
\mathrm{R}_q \left(\Theta_{\eta K} \middle\| \Theta'_{\eta K} \right) &\leq \mathrm{R}_q \left((\Theta_0, \Theta_\eta, \cdots, \Theta_{\eta K}) \middle\| (\Theta'_0, \Theta'_\eta, \cdots, \Theta'_{\eta K}) \right) \\
&\leq \sum_{k=0}^{K-1} \mathrm{R}_q \left(\Theta_{\eta(k+1)} \mid_{\Theta_{\eta k} = \theta_k} \middle\| \Theta'_{\eta(k+1)} \mid_{\Theta'_{\eta k} = \theta_k} \right) \\
&\leq \frac{qS_\ell^2}{4\hat{\sigma}^2} \cdot \eta K.
\end{aligned}$$

□

Remark 15. *Different papers discussing Noisy-GD variants adopt different notational conventions for the total noise added to the gradients. The noise in our Algorithm 2 is $\mathcal{N}\left(0, \frac{2\hat{\sigma}^2}{\eta} I_d\right)$; but is $\mathcal{N}\left(0, \hat{\sigma}^2 S_\ell^2 I_d\right)$ in the full-batch setting of DP-SGD by Abadi et al. [1]. To translate the bound in Theorem 39, one can simply rescale $\hat{\sigma}^2$ across different conventions to have the same noise variance.*

Our Theorem 39 is somewhat identical to Abadi et al. [1]'s (ε, δ)-DP bound. To verify this, note from Rényi divergence to (ε, δ)-indistinguishability conversion Theorem 5 that $(1 + \frac{2}{\varepsilon}\log\frac{1}{\delta}, \frac{\varepsilon}{2})$-Rényi DP implies (ε, δ)-DP. So, setting the bound in Theorem 39 to be smaller than $\frac{\varepsilon}{2}$ and substituting $q = 1 + \frac{2}{\varepsilon}\log\frac{1}{\delta}$, we get

$$\left(\frac{\varepsilon + 2\log\frac{1}{\delta}}{\varepsilon}\right)\frac{S_\ell^2}{4\hat{\sigma}^2}\cdot\eta K \le \frac{\varepsilon}{2} \iff \frac{S_\ell\sqrt{K(\frac{\varepsilon}{2} + \log\frac{1}{\delta})}}{\varepsilon} \le \hat{\sigma}\sqrt{\frac{2}{\eta}}. \tag{A.7}$$

The r.h.s. term is the *effective noise multiplier* used in full-batch DP-SGD of Abadi et al. [1]. For $\varepsilon \le 2\log\frac{1}{\delta}$ we get the same noise bound as in Abadi et al. [1, Theorem 1].

A.4 Deferred Proofs for § 3.2.4

Theorem 15 (Rényi-DP for Noisy-GD under LSI(c))**.** *Let $\{P_t\}_{t\ge0}$ and $\{Q'_t\}_{t\ge0}$ be the tracing diffusion for Noisy-GD on neighboring datasets $D \simeq D' \in \mathcal{X}^n$, under noise variance $\hat{\sigma}^2$ and loss function $\ell(\theta;\mathbf{x})$. Let $\ell(\theta;\mathbf{x})$ be a loss function on $\mathcal{Y} = \mathbb{R}^d$, with a finite gradient sensitivity S_ℓ. If for any neighboring datasets D and D', the corresponding coupled tracing diffusions P_t and Q_t satisfy* LSI(c) *throughout $0 \le t \le \eta K$, then for all $q > 1$, Noisy-GD satisfies (q, ρ) Rényi Differential Privacy for*

$$\rho = \frac{qS_\ell^2}{2c\hat{\sigma}^2}(1 - e^{-c\eta K}). \tag{A.8}$$

Proof. Recall that the Rényi-DP evolution Equation 3.13 holds for the tracing diffusion for Noisy-GD in the time duration $\eta k < t \le \eta(k+1)$, which is restated as follows.

$$\partial_t R(q,t) + 2c(1-\gamma)\left(\frac{R(q,t)}{q} + (q-1)\partial_q R(q,t)\right) \le \frac{1}{\gamma}\frac{qS_\ell^2}{4\hat{\sigma}^2}.$$

We fix the constant $\gamma = 1/2$ to get the following simplification:

$$\partial_t R(q,t) + c\left(\frac{R(q,t)}{q} + (q-1)\partial_q R(q,t)\right) \le \frac{qS_\ell^2}{2\hat{\sigma}^2}.$$

For brevity, suppose constant $\Delta_s = \frac{S_\ell^2}{2\hat{\sigma}^2}$. We define $u(q,t) = \frac{R(q,t)}{q}$. Then,

$$\partial_t R(q,t) + c\left(\frac{R(q,t)}{q} + (q-1)\partial_q R(q,t)\right) \le \Delta_s q$$
$$\implies \partial_t u(q,t) + cu(q,t) + c(q-1)\partial_q u(q,t) \le \Delta_s.$$

For any arbitrary constant $\bar{q} > 1$, define a curve $(q(s), t(s))$ that follows $q(s) = (\bar{q} - 1)\exp(c(s - \eta K)) + 1$ and $t(s) = s$. Note that $\frac{\mathrm{d}q(s)}{\mathrm{d}s} = c(q(s) - 1)$ and $\frac{\mathrm{d}t(s)}{\mathrm{d}s} = 1$. Therefore, for any $\eta k < s \leq \eta(k+1)$, the differential inequality followed along the path $u(s) = u(q(s), t(s))$ is

$$\frac{\mathrm{d}u(s)}{\mathrm{d}s} + cu(s) \leq \Delta_s \implies \frac{\mathrm{d}}{\mathrm{d}s}(e^{cs}u(s)) \leq \Delta_s e^{cs}.$$

Additionally, the discontinuities at times $t = \eta k$ for integral k does not increase the Rényi divergence. This is due to the post-processing property of Rényi divergence because of which $\mathrm{R}_q\left(\mathbf{T}_{\#P_{\eta k}} \middle\| \mathbf{T}_{\#Q_{\eta k}}\right) \leq \mathrm{R}_q\left(P_{\eta k} \| Q_{\eta k}\right)$ for any step k. Therefore, we can directly integrate in the duration $0 \leq t \leq \eta K$ to get

$$\begin{aligned}
&[e^{cs}u(s)]_0^{\eta K} \leq \int_0^{\eta K} \Delta_s e^{cs} \mathrm{d}s \\
\implies & e^{c\eta K}u(\eta K) - u(0) \leq \frac{\Delta_s}{c}\left(e^{c\eta K} - 1\right) \\
\implies & u(\eta K) \leq \frac{\Delta_s}{c}\left(1 - e^{-c\eta K}\right) = \frac{S_\ell^2}{2c\hat{\sigma}^2}\left(1 - e^{-c\eta K}\right). \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Since } u(0) = R(q(0), 0)/q(0) = 0.\text{)}
\end{aligned}$$

Noting that $q(0) > 1$, on reverting the substitution, we get

$$\mathrm{R}_{\bar{q}}\left(P_{\eta K} \| Q_{\eta K}\right) \leq \bar{q} \cdot \frac{S_\ell^2}{2c\hat{\sigma}^2}\left(1 - e^{-c\eta K}\right).$$

□

Isoperimetric constants for Noisy-GD. To prove that LSI holds for the tracing diffusion for noisy GD, we first note that the diffusion process Equation 3.3 can be written as composition of Lipschitz mapping and Gaussian noise for any $\eta k < t \leq \eta(k+1)$. Then, we rely on Lemma 2 and Lemma 1 introduced in Chapter 2 that show Lipschitz transformation and Gaussian perturbation of a probability distribution preserve its LSI property.

Lemma 10 (LSI for Noisy-GD)**.** *If loss function $\ell(\theta; \mathbf{x})$ is λ-strongly convex and β-smooth over $\mathbb{R}^d$, the step-size is $\eta < \frac{1}{\beta}$, and for any initla $\theta_0 \in \mathbb{R}^d$, the coupled tracing diffusion processes $\{P_t\}_{t\geq 0}$ and $\{Q_t\}_{t\geq 0}$ for Noisy-GD on any neighboring datasets D and D' satisfy* LSI(c) *for any $t \geq 0$ with $c = \frac{\lambda}{2}$.*

Proof. We only prove LSI(c) for the tracing diffusion process $\{P_t\}_{t\geq 0}$ on dataset D. The proof for $\{Q_t\}_{t\geq 0}$ is similar.

For any $D \in \mathcal{X}^n$, and any $0 < s \leq \eta$, recall that the update step in tracing diffusion Equation 3.3 equals the following random mapping:

$$\Theta_{\eta k+s} = \mathbf{T}_s(\Theta_{\eta k}) + \sqrt{2s\sigma^2}\mathcal{N}(0, I_d), \tag{A.9}$$

where the mapping $\mathbf{T}_s(\theta) = \theta - \eta \cdot \frac{1}{2}(\nabla\mathcal{L}_D(\theta) + \nabla\mathcal{L}_{D'}(\theta)) - s\cdot\frac{1}{2}(\nabla\mathcal{L}_D(\theta) - \nabla\mathcal{L}_{D'}(\theta))$. We first show that $\mathbf{T}_s(\theta)$ is $(1-\eta\lambda)$-Lipschitz. For any $w, v \in \mathbb{R}^d$, we have

$$\begin{aligned}
\mathbf{T}_s(w) - \mathbf{T}_s(v) &= w - v - \frac{\eta+s}{2}[\nabla\mathcal{L}_D(w) - \nabla\mathcal{L}_D(v)] - \frac{\eta-s}{2}[\nabla\mathcal{L}_{D'}(w) - \nabla\mathcal{L}_{D'}(v)] \\
&= w - v - \left[\frac{\eta+s}{2}\nabla^2\mathcal{L}_D(z) + \frac{\eta-s}{2}\nabla^2\mathcal{L}_{D'}(z')\right](w-v) \\
&\qquad\qquad \text{(for some } z, z' \in \mathbb{R}^d \text{ by the mid-value theorems)} \\
&= \left(I - \left[\frac{\eta+s}{2}\nabla^2\mathcal{L}_D(z) + \frac{\eta-s}{2}\nabla^2\mathcal{L}_{D'}(z')\right]\right)(w-v)
\end{aligned}$$

By λ-strong convexity and β-smoothness of loss function $\ell(\theta;\mathbf{x})$ on $\mathbb{R}^d$, we know that $\nabla^2\mathcal{L}_D(z)$ and $\nabla^2\mathcal{L}_{D'}(z')$ both have eigenvalues in the range $[\lambda, \beta]$. Since $s < \eta < \frac{1}{\beta}$, all eigenvalues of $I - \left[\frac{\eta+s}{2}\nabla^2\mathcal{L}_D(z) + \frac{\eta-s}{2}\nabla^2\mathcal{L}_{D'}(z')\right]$ is in $(0, 1-\eta\lambda]$. So, $\mathbf{T}_s$ is $(1-\eta\lambda)$-Lipshitz.

Now, using induction we prove P_t statisfies c-LSI for $c = \frac{\lambda}{2}$ for any $t \geq 0$.

Base step: Being a point distribution with the entire probability mass at θ_0, note that P_0 satisfies c-LSI for any $c > 0$, including $c = \frac{\lambda}{2}$.

Induction step: Suppose $P_{\eta k}$ satisfies c-LSI with the constant $c = \frac{\lambda}{2}$ for some $k \in \mathbb{N}$. Distribution P_t for $\eta k < t \leq \eta(k+1)$ is same as $\mathbf{T}_s$ pushover distribution plus Gaussian noise distribution, i.e. $P_t = \mathbf{T}_{s\#P_{\eta k}} \circledast \mathcal{N}(0, 2s\sigma^2 I_d)$ for $s = t - \eta k$. By using Lemma 2 and Lemma 1, we get $\left(\frac{c}{(1-\eta\lambda)^2+2sc}\right)$-LSI for P_t. Since $s < \eta < \frac{1}{\lambda}$, we have

$$(1-\eta\lambda)^2 + 2sc < 1 - \eta\lambda + 2sc = 1 - \eta\lambda + s\lambda \leq 1.$$

Hence, for $\eta k < t \leq \eta(k+1)$, P_t satisfies c'-LSI with constant $c' > c$, which means it also satisfies c-LSI by definition. □

A.5 Deferred Proofs for § 3.3

Theorem 16 (Lower bound on Rényi DP of Noisy-GD)**.** *There exist neighboring datasets $D \simeq D' \in \mathcal{X}^n$, a start parameter $\theta_0 \in \mathbb{R}^d$, and a smooth loss function $\ell(\theta; \mathbf{x})$ on $\mathbb{R}^d$, with a gradient sensitivity S_ℓ, such that for any step-size $\eta < 1$, noise variance $\hat{\sigma}^2 > 0$, and iteration $K \in \mathbb{N}$, the Rényi divergence of the output distributions, $P_{\eta K}$ and $Q_{\eta K}$, of Noisy-GD Algorithm 2 on respective datasets D, D' is lower-bounded by*

$$\mathrm{R}_q\left(P_{\eta K} \| Q_{\eta K}\right) \geq \frac{q S_\ell^2}{4\hat{\sigma}^2}\left(1 - e^{-\eta K}\right). \tag{A.10}$$

Proof. We consider the following L_2-norm squared loss function with bounded data universe:

$$\ell(\theta; \mathbf{x}) = \frac{1}{2}\|\theta - \mathbf{x}\|_2^2, \text{where } \theta \in \mathbb{R}^d \text{ and } \mathbf{x} \in \mathcal{X} = \left\{\mathbf{x}' \in \mathbb{R}^d \middle| \|\mathbf{x}'\|_2 \leq \frac{S_\ell}{2}\right\}. \tag{A.11}$$

For any dataset $D = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, and any $\theta \in \mathbb{R}^d$, the loss is

$$\mathcal{L}_D(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\|\theta - \mathbf{x}_i\|_2^2.$$

It is easy to verify that $\mathcal{L}_D(\theta)$ is 1-smooth. The gradient sensitivity of $\ell(\theta; \mathbf{x})$ is

$$\max_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \max_{\theta \in \mathbb{R}^d} \|\nabla \ell(\theta; \mathbf{x}) - \nabla \ell(\theta; \mathbf{x}')\|_2 = \max_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}} \max_{\theta \in \mathbb{R}^d} \| \ (\theta - \mathbf{x}) - (\theta - \mathbf{x}')\|_2 = S_\ell.$$

We construct the two neighboring datasets $D, D' \in \mathcal{X}^n$ such that $D = (\mathbf{x}_1, 0^d, \cdots, 0^d)$ and $D' = (\mathbf{x}_1', 0^d, \cdots, 0^d)$, where $\mathbf{x}_1, \mathbf{x}_1' \in \mathcal{X}$ are two records that are S_ℓ distance apart (i.e. $\|\mathbf{x}_1 - \mathbf{x}_1'\|_2 = S_\ell$).

Under dataset D, we can express the random variable $\Theta_{\eta K}$ at the K'th iteration of Noisy-GD using the following recursion with starting parameter $\Theta_0 = 0^d$.

$$\begin{aligned}
\Theta_{\eta K} =& (1-\eta)\Theta_{\eta(K-1)} + \eta\frac{\mathbf{x}_1}{n} + \sqrt{2\eta\left(\frac{\hat{\sigma}}{n}\right)^2} \cdot W_{K-1} \\
=& (1-\eta)^K \Theta_0 + \eta \sum_{i=0}^{K-1} (1-\eta)^i \frac{\mathbf{x}_1}{n} + \sqrt{2\eta\left(\frac{\hat{\sigma}}{n}\right)^2} \sum_{i=0}^{K-1} (1-\eta)^{K-1-i} W_i \\
=& \frac{\eta \mathbf{x}_1}{n} \sum_{i=0}^{K-1} (1-\eta)^i + \sqrt{2\eta\left(\frac{\hat{\sigma}}{n}\right)^2 \sum_{i=0}^{K-1} (1-\eta)^{2i}} \cdot W \quad (\text{where } W_i, W \sim \mathcal{N}(0, I_d))
\end{aligned}$$

A similar recursion can be used for $\Theta'_{\eta K}$ in Noisy-GD under dataset D'. Both $\Theta_{\eta K}$ and $\Theta'_{\eta K}$ are Gaussian random variables with variance $\frac{2\eta\hat{\sigma}^2}{n^2}\sum_{i=0}^{K-1}(1-\eta)^{2i}$ in

each dimension. Thus, we can calculate their exact divergence.

$$\begin{aligned}
\mathrm{R}_q\left(\Theta_{\eta K}\middle\|\Theta'_{\eta K}\right) &= \frac{q\cdot\left\|\eta(\mathbf{x}_1-\mathbf{x}'_1)\sum_{i=0}^{K-1}(1-\eta)^i\right\|_2^2}{2\cdot 2\eta\hat{\sigma}^2\sum_{i=0}^{K-1}(1-\eta)^{2i}} \\
&= \frac{q\eta^2 S_\ell^2}{4\eta\hat{\sigma}^2}\cdot\frac{\left(1-(1-\eta)^K\right)^2/\eta^2}{(1-(1-\eta)^{2K})/\left(\eta(2-\eta)\right)} \\
&= \frac{qS_\ell^2}{4\hat{\sigma}^2}\cdot\frac{2-\eta}{1+(1-\eta)^K}\left(1-(1-\eta)^K\right) \\
&\geq \frac{qS_\ell^2}{4\hat{\sigma}^2}(1-e^{-\eta K})
\end{aligned}$$

This inequality concludes the proof. □

Corollary 2 (Rényi-DP guarantee of Noisy-GD on L_2-norm squared loss)**.** *For any two neighboring datasets $D \simeq D' \in \mathcal{X}^n$, start parameter $\theta_0 \in \mathbb{R}^d$, step-size $\eta \leq 1$, noise variance $\hat{\sigma}^2$, and $K \in \mathbb{N}$, if the loss function is L_2-norm squared loss $\ell(\theta;\mathbf{x}) = \frac{1}{2}\|\theta-\mathbf{x}\|_2^2$ on $\mathbb{R}^d$, with a gradient sensitivity S_ℓ, the Rényi divergence of Noisy-GD's outputs on D, D' is upper-bounded by*

$$\mathrm{R}_q\left(P_{\eta K}\|Q_{\eta K}\right) \leq \frac{qS_\ell^2}{(2-\eta)\hat{\sigma}^2}(1-e^{-\frac{2-\eta}{2}\eta K}). \tag{A.12}$$

Proof. To use Theorem 15, we still need to verify c-LSI for the tracing diffusion on L_2-norm squared loss.

We use the explicit expression for tracing diffusion proved in Theorem 16 to prove c-LSI. We utilize the fact that $P_{\eta K}$, the tracing diffusion for L_2-norm squared loss at discrete update time ηK, is Gaussian with bounded variance $\frac{2\eta\hat{\sigma}^2}{n^2}\sum_{i=0}^{K-1}(1-\eta)^{2i} \leq \frac{2\hat{\sigma}^2}{2-\eta}$ in each dimension. Therefore, based on Lemma 1, which shows the LSI properties of Gaussian distributions, $P_{\eta K}$ satisfies c-LSI with $c = \frac{2-\eta}{2}$. Similarly, by computing the explicit expression for tracing diffusion at time $\eta k < t < \eta(k+1)$, one can verify P_t satisfies c-LSI.

Now, we can directly use Theorem 15 to derive an upper-bound for RDP for Noisy GD under L_2-squared norm loss. □

Discussion about tightness results. Figure 3.3 shows the gap between lower bound in Theorem 16 and upper bound in Corollary 2, under small step-size $\eta = 0.02$. The upper bound is roughly two times larger than the lower bound, which shows

tightness of our privacy guarantee. Figure 3.3 compares these bounds with the composition-based result in Theorem 39, which grows as fast as the lower bound in early iterations, but keeps growing linearly and eventually exceeding our Rényi-DP guarantee around $K \approx \Omega(\frac{1}{\eta}) \approx 100$. Moreover, the larger the RDP order q is, the smaller the required number of iterations K is for our Rényi DP guarantee to surpass the composition-based privacy bound.

Convergence rate difference in upper and lower bounds. There is a gap between the exponent and constant of our privacy upper bound Corollary 2 and the lower bound Theorem Theorem 16. We analyze the gap as follows.

1. **The gap in exponent:** There is a $\frac{2-\eta}{2}$ multiplicative gap between the exponent of our privacy upper bound and the lower bound. This is because discretized Noisy GD-converges to a biased stationary distribution. Therefore, our LSI constant bound $c = \frac{2-\eta}{2\sigma^2}$ depends on the discretization bias caused by step-size η, thus causing the exponent gap in our privacy bound.

2. **The gap in constant:** Our upper bound is larger than the lower bound by roughly a multiplicative constant of two. This is due to the **balancing ratio** $\gamma > 0$ in Lemma 8 for bounding the rate of privacy loss.

    a) **At the start of Noisy-GD:** setting $\gamma = 1$ in Equation 3.7 results in a smaller privacy loss rate bound. This is because, at the start of Noisy-GD, the accumulated privacy loss $\mathrm{R}_q\left(P_t \| Q_t\right)$ is small, thus leading to a small second term I_q/E_q in Equation 3.7, by Lemma 6. Setting $\gamma = 1$ reduces the coefficient $\frac{1}{\gamma}$ for the dominating first term of Equation 3.7, at a small cost of increasing the coefficient for the smaller second term I_q/E_q. This facilitates a smaller privacy loss rate bound, and is reflected in the similar growth of composition bound (equivalent to setting $\gamma = 1$) and our lower bound in Figure Figure 3.3.

    b) **As Noisy-GD converges:** setting $\gamma \to 0$ in Equation 3.7 results in a smaller privacy loss rate bound. This is because, at convergence, the accumulated privacy loss $\mathrm{R}_q\left(P_t \| Q_t\right)$ is larger, thus leading to more significant second term I_q/E_q in Equation 3.7. Setting $\gamma \to 0$ in Equation 3.7

reduces the coefficient $-(1-\gamma)$ for the dominating second term I_q/E_q, thus facilitate a smaller bound for the privacy loss rate.

3. In our proof for Theorem 15, we set $\gamma = \frac{1}{2}$ to *balance* privacy loss rate estimates at the start and convergence of Noisy-GD, thus obtaining the smallest privacy bound at convergence, as shown in the proof.

A.6 Deferred Proofs for § 3.4

Lemma 11 (Excess empirical risk of Noisy-GD)**.** *For λ-strongly convex and β-smooth loss function $\ell(\theta;\mathbf{x})$, step-size $\eta \leq \frac{\lambda}{\beta^2}$, and any start parameter $\theta_0 \in \mathbb{R}^d$, the excess empirical risk of Noisy-GD Algorithm 2 on any dataset $D \in \mathcal{X}^n$ is bounded by*

$$\mathbb{E}\left[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)\right] \leq \frac{\beta}{2}\left\|\theta_0 - \theta^*\right\|_2^2 \cdot e^{-\lambda\eta K} + \frac{2\beta d\hat{\sigma}^2}{\lambda n^2}, \tag{A.13}$$

where θ_K is the output of Noisy-GD on D.

Proof. The update step of the Noisy-GD performs the following operation:

$$\theta_{k+1} = \theta_k - \eta\nabla\mathcal{L}_D(\theta_k) + \sqrt{2\eta\left(\frac{\hat{\sigma}}{n}\right)^2}\mathcal{N}\left(0, I_d\right). \tag{A.14}$$

For brevity, let's denote $\sigma = \hat{\sigma}/n$. Since θ^* is the empirical risk minimizer, $\nabla\mathcal{L}_D(\theta^*) = 0$. Therefore, $\theta^* = \theta^* - \eta\nabla\mathcal{L}_D(\theta^*)$. So, we can expand the following as

$$\begin{aligned}
\left\|\theta_{k+1} - \theta^*\right\|_2^2 &= \left\|\theta_k - \eta\nabla\mathcal{L}_D(\theta_k) + \sqrt{2\eta\sigma^2}\mathcal{N}\left(0, I_d\right) - \left(\theta^* - \eta\nabla\mathcal{L}_D(\theta^*)\right)\right\|_2^2 \\
&= \left\|\theta_k - \theta^*\right\|_2^2 + \eta^2\left\|\nabla\mathcal{L}_D(\theta_k) - \nabla\mathcal{L}_D(\theta^*)\right\|_2^2 + 2\eta\sigma^2\left\|\mathcal{N}\left(0, I_d\right)\right\|_2^2 \\
&\quad + 2\left\langle\theta_k - \theta^*, \sqrt{2\eta\sigma^2}\mathcal{N}\left(0, I_d\right)\right\rangle - 2\eta\left\langle\nabla\mathcal{L}_D(\theta_k) - \nabla\mathcal{L}_D(\theta^*), \sqrt{2\eta\sigma^2}\mathcal{N}\left(0, I_d\right)\right\rangle \\
&\quad - 2\eta\left\langle\theta_k - \theta^*, \nabla\mathcal{L}_D(\theta_k) - \nabla\mathcal{L}_D(\theta^*)\right\rangle.
\end{aligned}$$

By β-smoothness of $\mathcal{L}_D$ and $\eta \leq \frac{\lambda}{\beta^2}$, we have

$$\eta^2\left\|\nabla\mathcal{L}_D(\theta_k) - \nabla\mathcal{L}_D(\theta^*)\right\|_2^2 \leq \eta\lambda\left\|\theta_k - \theta^*\right\|_2^2. \tag{A.15}$$

By strong convexity of $\mathcal{L}_D$, we have

$$\begin{aligned}
\mathbb{E}\left[\langle\nabla\mathcal{L}_D(\theta_k), \theta_k - \theta^*\rangle\right] &\geq \mathbb{E}\left[\mathcal{L}_D(\theta_k) - \mathcal{L}_D(\theta^*)\right] + \frac{\lambda}{2}\mathbb{E}\left[\left\|\theta_k - \theta^*\right\|_2^2\right] \\
&\geq \frac{\lambda}{2}\mathbb{E}\left[\left\|\theta_k - \theta^*\right\|_2^2\right] + \frac{\lambda}{2}\mathbb{E}\left[\left\|\theta_k - \theta^*\right\|_2^2\right] \\
&\geq \lambda\mathbb{E}\left[\left\|\theta_k - \theta^*\right\|_2^2\right].
\end{aligned}$$

By taking expectations on the controlling inequality, and plugging the above results, we get

$$\mathbb{E}\left[\|\theta_{k+1} - \theta^*\|_2^2\right] \leq (1 - \eta\lambda)\mathbb{E}\left[\|\theta_k - \theta^*\|_2^2\right] + 2\eta\sigma^2 d. \tag{A.16}$$

By β-smoothness, and optimality of θ^*,

$$\mathcal{L}_D(\theta_k) - \mathcal{L}_D(\theta^*) \leq \langle \nabla\mathcal{L}_D(\theta^*), \theta_k - \theta^*\rangle + \frac{\beta}{2}\|\theta_k - \theta^*\|_2^2 = \frac{\beta}{2}\|\theta_k - \theta^*\|_2^2.$$

On taking expectation over $\Theta_{\eta K}$ which is the distribution of θ_k, we have

$$\mathbb{E}\left[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)\right] \leq \frac{\beta}{2}\mathbb{E}\left[\|\theta_K - \theta^*\|_2^2\right].$$

On unrolling the recursion in Equation A.16, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)\right] \leq& \frac{\beta}{2}(1-\eta\lambda)^K\mathbb{E}\left[\|\theta_0 - \theta^*\|_2^2\right] + 2\beta d\sigma^2\sum_{k=0}^{K-1}(1-\eta\lambda)^k \\ \leq& \frac{\beta}{2}\mathbb{E}\left[\|\theta_0 - \theta^*\|_2^2\right]e^{-\lambda\eta K} + \frac{2\beta d\hat{\sigma}^2}{\lambda n^2}.\end{aligned}$$

Since θ_0 is a point distribution, $\mathbb{E}\left[\|\theta_0 - \theta^*\|_2^2\right] = \|\theta_0 - \theta^*\|_2^2$. This completes the proof. □

Theorem 17 (Utility under DP for Noisy GD)**.** *For β-smooth λ-strongly convex loss function $\ell(\theta; \mathbf{x})$ with gradient sensitivity S_ℓ, and datasets $D \in \mathcal{X}^n$ of size n, if the step-size $\eta = \frac{\lambda}{\beta^2}$ and any initial parameter $\theta_0 \in \mathbb{R}^d$, then, for any $q > 1$ and $\rho > 0$, the Noisy-GD Algorithm 2 is (q, ρ)-Rényi differentially private, and achieves an excess empirical risk of*

$$\mathbb{E}\left[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)\right] \leq \frac{4q\beta d S_\ell^2}{\rho\lambda^2 n^2}, \tag{A.17}$$

when noise variance $\hat{\sigma}^2 = \frac{qS_\ell^2}{\lambda\rho}$, and number of updates $K^* = \frac{\beta^2}{\lambda^2}\log(\frac{\lambda^2\|\theta_0 - \theta^*\|_2^2}{4S_\ell^2} \cdot \frac{\rho n^2}{qd})$. Equivalently, for $\varepsilon \leq 2\log(1/\delta)$ and $\delta > 0$, Noisy-GD is (ε, δ)-differentially private, and satisfies

$$\mathbb{E}\left[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)\right] \leq \frac{32\beta d S_\ell^2\log(1/\delta)}{\varepsilon^2\lambda^2 n^2}, \tag{A.18}$$

by setting noise variance $\hat{\sigma}^2 = \frac{2S_\ell^2(\varepsilon + 2\log(1/\delta))}{\lambda\varepsilon^2}$, and the number of updates as $K^* = \frac{\beta^2}{\lambda^2}\log\left(\frac{\lambda^2\|\theta_0 - \theta^*\|_2^2}{16S_\ell^2} \cdot \frac{\varepsilon^2 n^2}{d\log(1/\delta)}\right)$.

Proof. From Lemma 11, we have

$$\mathbb{E}\left[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)\right] \leq \frac{\beta}{2}\left\|\theta_0 - \theta^*\right\|_2^2 \cdot e^{-\lambda\eta K} + \frac{2\beta d\hat{\sigma}^2}{\lambda n^2} \tag{A.19}$$

Since $\eta = \frac{\lambda}{\beta^2} \leq \frac{1}{\beta}$, by Corollary 1, the Noisy-GD with K iterations will be (q,ρ)-Rényi DP as long as $\hat{\sigma}^2 \geq \frac{qS_\ell^2}{\lambda\rho}(1 - e^{-\lambda\eta K/2})$. Therefore, if we set $\hat{\sigma}^2 = \frac{qS_\ell^2}{\lambda\rho}$, Noisy-GD is (q,ρ)-Rényi DP for any K. On substituting this noise variance in Equation A.19, we get

$$\mathbb{E}\left[\mathcal{L}_D(\theta_K) - \mathcal{L}_D(\theta^*)\right] \leq \frac{\beta}{2}\left\|\theta_0 - \theta^*\right\|_2^2 \cdot e^{-\lambda\eta K} + \frac{2q\beta d S_\ell^2}{\rho\lambda^2 n^2}. \tag{A.20}$$

By setting $K^* = \frac{1}{\lambda\eta}\log\left(\frac{\rho\lambda^2 n^2\|\theta_0-\theta^*\|_2^2}{4qdS_\ell^2}\right) = \frac{\beta^2}{\lambda^2}\log\left(\frac{\rho\lambda^2 n^2\|\theta_0-\theta^*\|_2^2}{4qdS_\ell^2}\right)$, we can control the excess empirical risk to be

$$\mathbb{E}\left[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)\right] \leq \frac{4q\beta d S_\ell^2}{\rho\lambda^2 n^2}. \tag{A.21}$$

Now, we convert the optimal excess risk guarantee under an (q,ρ) Rényi-DP constraint to an optimal excess risk guarantee under (ε,δ)-DP constraint. Let $\varepsilon > 0$ and $0 < \delta < 1$ be two constants such that $\varepsilon \leq 2\log(1/\delta)$. As per Rényi divergence to hockey-stick divergence conversion Theorem 5, (q,ρ)-RDP implies (ε,δ)-DP for $q = 1 + \frac{2}{\varepsilon}\log(1/\delta)$ and $\rho = \frac{\varepsilon}{2}$. By using this conversion, we bound Equation A.21 in terms of DP parameters as

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_D(\theta_{K^*}) - \mathcal{L}_D(\theta^*)\right] &\leq \frac{4\beta d S_\ell^2}{\lambda^2 n^2} \cdot \frac{q}{\rho} \\
&= \frac{4\beta d S_\ell^2}{\lambda^2 n^2} \cdot \frac{1 + \frac{2}{\varepsilon}\log(1/\delta)}{\frac{\varepsilon}{2}} \\
\because \varepsilon \leq 2\log(1/\delta) \quad &\leq \frac{4\beta d S_\ell^2}{\lambda^2 n^2} \cdot \frac{8\log(1/\delta)}{\varepsilon^2}.
\end{aligned}$$

The amount of noise needed in terms of DP parameters is

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{S_\ell^2}{\lambda} \cdot \frac{q}{\varepsilon'} \\
&= \frac{S_\ell^2}{\lambda} \cdot \frac{1 + \frac{2}{\varepsilon}\log(1/\delta)}{\frac{\varepsilon}{2}}
\end{aligned}$$

The optimal number of updates K^* in terms of DP parameters is bounded as

$$\begin{aligned} K^* &= \frac{\beta^2}{\lambda^2} \log\left(\frac{\lambda^2 \left\|\theta_0 - \theta^*\right\|_2^2}{4S_\ell^2} \cdot \frac{n^2}{d} \cdot \frac{\rho}{q}\right) \\ &= \frac{\beta^2}{\lambda^2} \log\left(\frac{\lambda^2 \left\|\theta_0 - \theta^*\right\|_2^2}{4S_\ell^2} \cdot \frac{n^2}{d} \cdot \frac{1 + \frac{2}{\varepsilon}}{\frac{\varepsilon}{2}}\right) \\ &\leq \frac{\beta^2}{\lambda^2} \log\left(\frac{\lambda^2 \left\|\theta_0 - \theta^*\right\|_2^2}{4S_\ell^2} \cdot \frac{n^2}{d} \cdot \frac{\varepsilon^2}{4\log(1/\delta)}\right). \end{aligned}$$

□

Appendix B

Supplement for Chapter 4

B.1 Deferred Proofs for § 4.2

Theorem 18. *There exists an algorithm pair* $(\mathcal{A}, \bar{\mathcal{A}})$ *satisfying* (ε, δ)*-adaptive-unlearning with* $\varepsilon = \delta = 0$ *under publish function* $f_{\text{pub}}(\theta) = \theta$ *such that by designing a* 1*-adaptive* 1*-requester* $\mathbb{Q}$*, an attacker can infer the identity of a record deleted by edit* u_i*, at any arbitrary step* $i > 3$*, with probability at-least* $1 - (1/2)^{i-3}$ *from a single post-edit release* ϕ_i*, even with no access to* $\mathbb{Q}$*'s transcript* $(\phi_{<i}; u_{<i})$*.*

Proof. Let data universe $\mathcal{X}$, the internal model space $\mathcal{Y}$, as well as publishable outcome space Φ be $\mathbb{R}$. Consider the task of releasing a sequence of medians using function $\text{med} : \mathbb{R}^* \to \mathbb{R}$ in the online setting when the initial database $D_0 \in \mathcal{X}^n$ is being modified by some adaptive requester $\mathbb{Q}$. Given a database $D \in \mathcal{X}^n$, our learning algorithm is defined as $\mathcal{A}(D) = \text{med}(D)$. For an arbitrary edit request $u \in \mathcal{U}^r$, our unlearning algorithm is defined as $\bar{\mathcal{A}}(D, u, \bullet) = \text{med}(D \circ u)$ for any $\bullet \in \mathcal{Y}$. Let the publish function $f_{\text{pub}} : \mathcal{Y} \to \Phi$ be an identity function, i.e. $f_{\text{pub}}(\theta) = \theta$.

For any initial database $D_0 \in \mathcal{X}^n$ and an adaptive sequence $(u_i)_{i \geq 1}$ generated by any ∞-adaptive 1-requester $\mathbb{Q}$, note that

$$f_{\text{pub}}(\bar{\mathcal{A}}(D_{i-1}, u_i, \bullet)) = f_{\text{pub}}(\mathcal{A}(D_i)), \quad \text{for all } i \geq 1 \text{ and any } \bullet \in \mathcal{Y}. \tag{B.1}$$

Therefore, $\bar{\mathcal{A}}$ is a $(0, 0)$-adaptive unlearning algorithm for $\mathcal{A}$ under f_{pub}.

Now suppose that n is odd and D_0 consists of unique entries. W.L.O.G assume that the median record $\text{med}(D_0)$ is at index ind^m and its owner will be deleting it at step i by sending a non-adaptive edit request $u_i = \{\langle \text{ind}^m, \mathbf{y} \rangle\}$ such that $\mathbf{y} \neq \text{med}(D_0)$. We design the following 1-adaptive 1-requester $\mathbb{Q}$ that sends edit

requests in the first $i-1$ steps to ensure with high probability that the published outcome at step i remains the deleted record, i.e., $\text{med}(D_i) = \text{med}(D_0)$:

$$\mathbb{Q}(\phi_0, u_1, u_2, \cdots, u_{j-1}) = \{\langle \text{ind}_j, \phi_0 \rangle\} \quad \forall\ 1 \le j < i, \tag{B.2}$$

where ind_j is randomly sampled from $[n] \setminus \{\text{ind}_1, \cdots, \text{ind}_{j-1}\}$ without replacement. Note that by the end of interaction, $\mathbb{Q}$ replaces at-least $i-2$ unique records in D_0 with $\phi_0 = \text{med}(D_0)$. If one of those original records was larger than $\text{med}(D_0)$ and another was smaller than $\text{med}(D_0)$, then it is guaranteed that $\text{med}(D_i) = \text{med}(D_0)$. Therefore, $\Pr[\text{med}(D_i) = \text{med}(D_0)]$ is at-least

$$\begin{aligned}
\Pr\left[\exists \text{ind}^l, \text{ind}^u \in \{\text{ind}_1, \cdots, \text{ind}_{i-1}\} \text{ s.t. } D_0[\text{ind}^l] < D_0[\text{ind}^m] < D_0[\text{ind}^u]\right] \\
&\ge 1 - 2 \times \binom{\lfloor n \rfloor / 2}{i-2} \Big/ \binom{n}{i-2} \\
&\ge 1 - \left(\frac{1}{2}\right)^{i-3}.
\end{aligned}$$

In other words, a copy of the record deleted at i^{th} step will be blatantly revealed after processing the deletion request with probability at least $1 - (1/2)^{i-3}$, despite using a $(0,0)$-adaptive-unlearning mechanism $\bar{\mathcal{A}}$ for deletion.

□

Theorem 19. *For every $\varepsilon > 0$, there exists a pair $(\mathcal{A}, \bar{\mathcal{A}})$ of algorithms that satisfy $(\varepsilon, 0)$-unlearning under some publish function f_{pub} such that for all non-adaptive 1-requesters $\mathbb{Q}$, their exists an attacker that can correctly infer the identity of a record deleted at any arbitrary edit step $i \ge 1$ by observing only future releases $\phi_{\ge i}$.*

Proof. For a query $h : \mathcal{X} \to \{0, 1\}$, consider the task of learning the count over a database that is being edited online by a non-adaptive 1-requester $\mathbb{Q}$. Since $\mathbb{Q}$ is non-adaptive by assumption, it is equivalent to the entire edit sequence $\{u_i\}_{i \ge 1}$ being fixed before interaction. We design an algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ for this task with secret model space being $\mathcal{Y} = \mathbb{N}^3$ and published outcome space being $\Phi = \mathbb{R}$, with the publish function being $f_{\text{pub}}(\langle a, b, c \rangle) = a + b/c + \text{Lap}\left(\frac{1}{\varepsilon}\right)$ (with the convention that $b/c = 0$ if $b = c = 0$). At any step $i \ge 0$, our internal model $\hat{\Theta}_i = \langle \text{cnt}_i, \text{del}_i, i \rangle$ encodes the current count of h on database D_i, the count of h on records previously deleted by $u_{\le i}$, and the current step index i. Our learning algorithm initializes

the secret model as $\hat{\Theta}_0 = \mathcal{A}(D_0) = \langle \sum_{\mathbf{x}\in D_0} h(\mathbf{x}), 0, 0\rangle$, and, for an edit request $u_i = \{\langle \mathrm{ind}_i, \mathbf{y}_i\rangle\}$, our algorithm $\bar{\mathcal{A}}$ updates the secret model $\hat{\Theta}_{i-1} \to \hat{\Theta}_i$ following the rule

$$\hat{\Theta}_i = \bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) = \langle \mathrm{cnt}_i, \mathrm{del}_i, i\rangle \text{ where } \begin{cases} \mathrm{cnt}_i = \mathrm{cnt}_{i-1} + h(\mathbf{y}_i) - h(D_{i-1}[\mathrm{ind}_i]), \\ \mathrm{del}_i = \mathrm{del}_{i-1} + h(D_{i-1}[\mathrm{ind}_i]). \end{cases}$$

Note that $\forall i \geq 1$, $\Delta_i \stackrel{\text{def}}{=} \mathrm{del}_i/i \in [0, 1]$. Therefore, from properties of Laplace mechanism [41], it is straightforward to see that for all $i \geq 1$,

$$\begin{aligned} f_{\text{pub}}(\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}))\Big| u_{\leq i} &= \sum_{\mathbf{x}\in D_i} h(\mathbf{x}) + \Delta_i + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right) \\ &\stackrel{\varepsilon,0}{\approx} \sum_{\mathbf{x}\in D_i} h(\mathbf{x}) + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right) = f_{\text{pub}}(\mathcal{A}(D_i)). \end{aligned}$$

Hence, $\bar{\mathcal{A}}$ is an $(\varepsilon, 0)$-unlearning algorithm for $\mathcal{A}$ under f_{pub}.

To show that an adversary can still infer the identity of record deleted by edit request $u_i = \{\langle \mathrm{ind}_i, \bullet\rangle\}$, consider a database D'_{i-1} that differs from D_{i-1} only at index ind_i such that $h(D'_{i-1}[\mathrm{ind}_i]) \neq h(D_{i-1}[\mathrm{ind}_i])$. Let random variable sequences $\phi_{\geq i}$ and $\phi'_{\geq i}$ denote the releases by $\bar{\mathcal{A}}$ in the scenarios that the $(i-1)^{\text{th}}$ database was D_{i-1} and D'_{i-1} respectively. The divergence between these two random variable sequences reflect the capacity of any adversary to infer the record deleted by u_i. Since, we have identical databases after u_i, i.e. $D_{j-1} \circ u_j = D'_{j-1} \circ u_j$ for all $j \geq i$, note that both ϕ_j and ϕ'_j are independent Laplace distributions with a shift of exactly $\frac{1}{j}$ units. Therefore,

$$\max_{S\subset\Phi^*} \log \frac{\Pr[\phi_{\geq i} \in S]}{\Pr\left[\phi'_{\geq i} \in S\right]} = \sum_{j=i}^{\infty} \max_{S_j\subset\mathbb{R}} \log \frac{\Pr[\phi_j \in S_j]}{\Pr\left[\phi'_j \in S_j\right]} = \sum_{j=i}^{\infty} \log e^{\varepsilon/j} = \sum_{j=1}^{\infty} \frac{\varepsilon}{j} = \infty.$$

□

B.1.1 Flaws in Related Unlearning Definitions

In this subsection, we show that our criticisms on trustworthiness of unlearning notions under adaptive requests in § 4.2 also apply to the other popular deletion privacy notions like Ginart et al. [46], Guo et al. [50] and Sekhari et al. [92].

Definition 28 (Data deletion operation [46])**.** Algorithm $\bar{\mathcal{A}}$ is a *data deletion operation for a learning algorithm* $\mathcal{A}$ if $\bar{\mathcal{A}}(D, S, \mathcal{A}(D)) \stackrel{0}{\approx} 0\mathcal{A}(D \setminus S)$ for all $D \subset \mathcal{X}$ and all subsets $S \subset D$ that is *selected independently* of $\mathcal{A}(D)$.

Definition 29 ((ε, δ)-certified removal [50])**.** A removal mechanism $\bar{\mathcal{A}}$ performs (ε, δ)-*certified removal* for learning algorithm $\mathcal{A}$ if for all databases $D \subset \mathcal{X}$ and deletion subsets $S \subset D$,

$$\bar{\mathcal{A}}(D, S, \mathcal{A}(D)) \overset{\varepsilon,\delta}{\approx} \mathcal{A}(D \setminus S). \tag{B.3}$$

Definition 30 ((ε, δ)-unlearning [92])**.** For all $D \subset \mathcal{X}$ of size n and deletion subsets $S \subset D$ such that $|S| \leq m$, a learning algorithm $\mathcal{A}$ and an unlearning algorithm $\bar{\mathcal{A}}$ is (ε, δ)-*unlearning* if

$$\bar{\mathcal{A}}(T(D), S, \mathcal{A}(D)) \overset{\varepsilon,\delta}{\approx} \bar{\mathcal{A}}(T(D \setminus S), \varnothing, \mathcal{A}(D \setminus S)), \tag{B.4}$$

where $\varnothing$ denotes the empty set and $T(D)$ denotes the data statistics available to $\bar{\mathcal{A}}$ about D.

Unsoundness. Definition 28 explicitly assumes that there is no influence of the learned model $\mathcal{A}(D)$ on the selection of deletion subset S. However, such a dependence is common in the real world; for example, people sometimes delete their information if they don't like what a model $\mathcal{A}(D)$ reveals about them. Therefore, Definition 28's certification in these settings is trivially void. Unlike Definition 28 however, Definition 29 and Definition 30 make no assumptions about dependence between the deletion request S and the learned model $\mathcal{A}(D)$. So, request S can depend on $\mathcal{A}(D)$ under these two certifications. We recall the example we provide in Section § 4.2 to show that Definition 29 and Definition 30, are also unsound under adaptivity.

For the universe of records $\mathcal{X} = \{-2, -1, 1, 2\}$, consider the following learning and unlearning algorithms:

$$\mathcal{A}(D) = \sum_{\mathbf{x} \in D} \mathbf{x}, \quad \text{and} \quad \bar{\mathcal{A}}(D, S, \mathcal{A}(D)) = \sum_{\mathbf{x} \in D \setminus S} \mathbf{x}. \tag{B.5}$$

Note that for any $D \subset \mathcal{X}$ and any $S \subset D$, the above algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies Definition 28, 29 and 30 for $\varepsilon = \delta = 0$ and $\mathbf{T}(D) = D$. Suppose the adversary is aware that the following dependence holds between the learned model $\mathcal{A}(D)$ and deletion request S:

$$S = \begin{cases} \{\mathbf{x} < 0 : \forall \mathbf{x} \in \mathcal{X}\} & \text{if } \mathcal{A}(D) < 0, \\ \{\mathbf{x} \geq 0 : \forall \mathbf{x} \in \mathcal{X}\} & \text{otherwise.} \end{cases} \tag{B.6}$$

Consider two neighbouring databases $D_{-1} = \{-2, -1, 2\}$ and $D_1 = \{-2, 1, 2\}$. Knowing the above dependence, an adversary can determine whether $D = D_{-1}$ or $D = D_1$ by looking only at $\bar{\mathcal{A}}(D, S, \mathcal{A}(D))$. This is because if $D = D_{-1}$, then the observation after unlearning is 2, and if $D = D_1$, the observation after unlearning is -2. So, even though $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies the guarantees of Ginart et al. [46], Guo et al. [50] and Sekhari et al. [92], it blatantly reveals the identity (-1 or 1) of a deleted record to an adversary observing only the post-deletion release.

Incompleteness. Definition 28, 29 and 30 are also incomplete. Consider an unlearning algorithm $\bar{\mathcal{A}}$ that outputs a fixed output $\mathbf{x}_1 \in \mathcal{X}$ if the deletion request $S = \varnothing$ and outputs another fixed output $\mathbf{x}_2 \in \mathcal{X}$ if the deletion request $S \neq \varnothing$. It is easy to see that $\bar{\mathcal{A}}$ is a valid deletion algorithm as its output does not depend on the input database D or the learned model $\mathcal{A}(D)$. However, note that $\bar{\mathcal{A}}$ does not satisfy the unlearning Definition 30, for any learning algorithm $\mathcal{A}$. And, for a learning algorithm $\mathcal{A}(D) = \sum_{\mathbf{x} \in D} \mathbf{x}$, one can verify that the pair $(\mathcal{A}, \bar{\mathcal{A}})$ does not satisfy Definition 28 and 29 either.

B.2 Deferred Proofs for § 4.3

Theorem 20 (Definition 22 of (ε, δ)-deletion privacy safeguards RTBF)**.** *If the algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies (ε, δ)-deletion-privacy guarantee under all p-adaptive r-requesters, then even with complete knowledge of a p-adaptive r-requester $\mathbb{Q}$ that interacts with the curator before a target record $D_0[\text{ind}]$ in the initial database D_0 is deleted at step $i \geq 1$ by request u_i, any attacker $MI : \mathcal{Y}^* \to \{0, 1\}$ observing only the post-deletion models $\hat{\Theta}_{\geq i} = (\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots)$ has an advantage*

$$Adv(\text{MI}) \stackrel{def}{=} \Pr\left[\text{MI}(\hat{\Theta}_{\geq i}) = 1 \middle| \mathbf{x}\right] - \Pr\left[\text{MI}(\hat{\Theta}_{\geq i}) = 1 \middle| \mathbf{x}'\right] \tag{B.7}$$

for disambiguating between two possible values $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ of the deleted record $D_0[\text{ind}]$ bounded as follows.

$$Adv(\text{MI}) \leq e^{\varepsilon} - 1 + 2\delta. \tag{B.8}$$

Proof. For an arbitrary step $i \geq 1$, suppose one of the replacement operations in the edit request $u_i \in \mathcal{U}^r$ replaces a record at index 'ind' from the database D_{i-1} with '$\mathbf{y}$'. In the worst case, this record $D_{i-1}[\text{ind}]$ might have been there from the start,

i.e. $D_0[\text{ind}] = D_0[\text{ind}]$, and influenced all the decisions of the adaptive requester $\mathbb{Q}$ in the edit steps $1, \cdots, i-1$. To prove soundness, we need to show that if $(\mathcal{A}, \bar{\mathcal{A}})$ satisfies (ε, δ)-deletion-privacy, then even in this worst-case scenario, no adaptive adversary can design a membership inference test $\text{MI}(\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots) \in \{0, 1\}$ that can distinguish with high probability the null hypothesis $H_0 = \{D_0[\text{ind}] = \mathbf{x}\}$ from the alternate hypothesis $H_1 = \{D_0[\text{ind}] = \mathbf{x}'\}$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. That is, the advantage of any test MI, defined as

$$\text{Adv}(\text{MI}) \stackrel{\text{def}}{=} \Pr\left[\text{MI}(\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots) = 1 | H_0\right] - \Pr\left[\text{MI}(\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots) = 1 | H_1\right], \quad \text{(B.9)}$$

must be small. Since after processing edit request u_i, the databases $D_i, D_{i+1}, \cdots$ no longer contain the deleted record $D_{i-1}[\text{ind}]$, the data-processing inequality implies that future models $\hat{\Theta}_{i+1}, \hat{\Theta}_{i+2}, \cdots$ cannot have more information about $D_{i-1}[\text{ind}]$ that what is present in $\hat{\Theta}_i$. Therefore, any test $\text{MI}(\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots)$ has a smaller advantage than the optimal test $\text{MI}^*(\hat{\Theta}_i) \in \{0, 1\}$ that only uses $\hat{\Theta}_i$.

Also, since $(\mathcal{A}, \bar{\mathcal{A}})$ satisfy (ε, δ)-deletion-privacy for any p-adaptive r-requester $\mathbb{Q}$, we know from Definition 22 that there exists a mapping $\pi_i^{\mathbb{Q}}$ such that for all $D_0 \in \mathcal{X}^n$, the model $\hat{\Theta}_i$ generated by the interaction between $(\mathcal{A}, \bar{\mathcal{A}}, \mathbb{Q})$ on D_0 after i^{th} edit satisfies the inequality $\Pr\left[\hat{\Theta}_i \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\bar{\Theta} \in S\right] + \delta$ for all S, where $\bar{\Theta} = \pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y} \rangle)$. As the database $D_0 \circ \langle \text{ind}, \mathbf{y} \rangle$ is identical under both hypothesis H_0 and H_1, we have

$$\Pr\left[\text{MI}^*(\hat{\Theta}_i) = 1 | H_0\right] \leq e^{\varepsilon} \Pr\left[\text{MI}^*(\bar{\Theta}) = 1\right] + \delta, \text{ and} \quad \text{(B.10)}$$

$$\Pr\left[\text{MI}^*(\hat{\Theta}_i) = 0 | H_1\right] \leq e^{\varepsilon} \Pr\left[\text{MI}^*(\bar{\Theta}) = 0\right] + \delta. \quad \text{(B.11)}$$

On adding the two inequalities, we get:

$$\begin{aligned} \text{Adv}(\text{MI}) \leq \text{Adv}(\text{MI}^*) &= \Pr\left[\text{MI}^*(\hat{\Theta}_i) = 1 | H_0\right] - \Pr\left[\text{MI}^*(\hat{\Theta}_i) = 1 | H_1\right] \\ &= \Pr\left[\text{MI}^*(\hat{\Theta}_i) = 1 | H_0\right] + \Pr\left[\text{MI}^*(\hat{\Theta}_i) = 0 | H_1\right] - 1 \\ &\leq e^{\varepsilon} - 1 + 2\delta. \end{aligned}$$

□

B.3 Deferred Proofs for § 4.4

Theorem 21 (From adaptive to non-adaptive deletion)**.** *If an algorithm pair* $(\mathcal{A}, \bar{\mathcal{A}})$ *satisfies* $(\varepsilon_{\text{dd}}, \delta)$*-deletion privacy under all non-adaptive* r*-requesters and is also* $(\varepsilon_{\text{dp}}, \delta)$*-differentially private, then pair* $(\mathcal{A}, \bar{\mathcal{A}})$ *also satisfies* $(\varepsilon_{\text{dd}}{}', (p+2)\delta)$*-deletion privacy under all* p*-adaptive* r*-requesters, for*

$$\varepsilon_{\text{dd}}{}' = \varepsilon_{\text{dd}} + p\varepsilon_{\text{dp}}(e^{\varepsilon_{\text{dp}}} - 1) + \varepsilon_{\text{dp}}\sqrt{2p\log(1/\delta)}. \tag{B.12}$$

Proof. To prove this theorem, we need to show that for any p-adaptive r-requester $\mathbb{Q}$, there exists a construction for a map $\pi_i^{\mathbb{Q}} : \mathcal{X}^n \to \mathcal{Y}$ such that for all $D_0 \in \mathcal{X}^n$, the sequence of model $(\hat{\Theta}_i)_{i\geq 0}$ generated by the interaction between $(\mathbb{Q}, \mathcal{A}, \bar{\mathcal{A}})$ on D_0 satisfies the following inequality for all $i \geq 1$:

$$\Pr\left[\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) \in S\right] \leq e^{\varepsilon'} \cdot \Pr\left[\pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y}\rangle) \in S\right] + \delta'. \tag{B.13}$$

Fix a database $D_0 \in \mathcal{X}^n$ and an edit request $u_i \in \mathcal{U}^r$. Let $D_0' \in \mathcal{X}^n$ be a neighbouring database defined to be $D_0' = D_0 \circ \langle \text{ind}, \mathbf{y}\rangle$ for an arbitrary replacement operation $\langle \text{ind}, \mathbf{y}\rangle \in u_i$. Given any p-adaptive ρ-requester $\mathbb{Q}$, let $(\hat{\Theta}_i)_{i\geq 0}$ and $(U_i)_{i\geq 1}$ be the sequence of released model and edit request random variables generated on $\mathbb{Q}$'s interaction with $(\mathcal{A}, \bar{\mathcal{A}})$ with initial database as D_0. Similarly, let $(\hat{\Theta}_i')_{i\geq 0}$ and $(U_i')_{i\geq 1}$ be the corresponding sequences generated due to the interaction among $(\mathbb{Q}, \mathcal{A}, \bar{\mathcal{A}})$ on D_0'.

Since $(\mathcal{A}, \bar{\mathcal{A}})$ is assumed to satisfy $(\varepsilon_{\text{dd}}, \delta)$-deletion-privacy guarantee under non-adaptive r-requesters, recall from pt. 7 in Section § 4.3 that there exists a mapping $\pi : \mathcal{X}^n \to \mathcal{Y}$ such that for any fixed edit sequence $u_{\leq i} \stackrel{\text{def}}{=} (u_1, u_2, \cdots, u_i)$,

$$\begin{aligned}
&\Pr\left[\hat{\Theta}_i \in S \middle| U_{\leq i} = u_{\leq i}\right] \leq e^{\varepsilon_{\text{dd}}} \cdot \Pr\left[\pi(D_0 \circ u_{\leq i}) \in S\right] + \delta \\
\implies &\Pr\left[\bar{\mathcal{A}}(D_0 \circ U_{<i}, u_i, \hat{\Theta}_i) \in S \middle| U_{<i} = u_{<i}\right] \\
&\qquad \leq e^{\varepsilon_{\text{dd}}} \cdot \Pr\left[\pi(D_0 \circ U_{<i}' \circ u_i) \in S | U_{<i}' = u_{<i}\right] + \delta.
\end{aligned}$$

Note that since the replacement operation $\langle \text{ind}, \mathbf{y}\rangle$ is part of the edit request u_i, we have $D_0 \circ U_{<i}' \circ u_i = D_0' \circ U_{<i}' \circ u_i$. Moreover, since the sequence $U_{<i}'$ of edit requests is generated by the interaction of $(\mathbb{Q}, \mathcal{A}, \bar{\mathcal{A}})$ on $D_0' = D_0 \circ \langle \text{ind}, u\rangle$ and the i^{th} edit request u_i is fixed beforehand, we can define a valid construction of a map

$\pi_i^{\mathbb{Q}} : \mathcal{X}^n \to \mathcal{Y}$ as per Definition 22 as follows:

$$\pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y} \rangle) = \pi(D_0' \circ U_{<i}' \circ u_i). \tag{B.14}$$

For brevity, let $\hat{\Theta}_u = \bar{\mathcal{A}}(D_0 \circ U_{<i}, u_i, \hat{\Theta}_{i-1})$, and $\hat{\Theta}_u' = \pi_i^{\mathbb{Q}}(D_0 \circ \langle \text{ind}, \mathbf{y} \rangle)$. For this construction, we prove the requisite bound in Equation B.13 as follows using *Hockey-stick* divergence (introduced in Definition 4). □

Theorem 22 (Differential Privacy is necessary for adaptive deletion privacy)**.** *Let* $\text{Test} : \mathcal{Y} \to \{0,1\}$ *be a membership inference test for* $\mathcal{A}$ *to distinguish between neighboring datasets* $D, D' \in \mathcal{X}^n$*. Similarly, let* $\overline{\text{Test}} : \mathcal{Y} \to \{0,1\}$ *be a membership inference test for* $\bar{\mathcal{A}}$ *to distinguish between* $\bar{D}, \bar{D}' \in \mathcal{X}^n$ *that are neighboring after applying edit* $\bar{u} \in \mathcal{U}^1$*. If* $Adv(\text{Test}) > \Delta_{\mathcal{A}}$ *and* $Adv(\overline{\text{Test}}) > \Delta_{\bar{\mathcal{A}}}$*, then the pair* $(\mathcal{A}, \bar{\mathcal{A}})$ cannot *satisfy* (ε, δ)*-deletion privacy under* 1*-adaptive* 1*-requester for any*

$$\varepsilon < \log(1 + \Delta_{\mathcal{A}} \cdot \Delta_{\bar{\mathcal{A}}} - 2\delta). \tag{B.15}$$

Proof. By assumption, we know that there exists tests $\text{Test}, \overline{\text{Test}} : \mathcal{Y} \to \{0,1\}$ such that

$$\text{Adv}(\text{Test}) \stackrel{\text{def}}{=} \Pr\left[\text{Test}(\mathcal{A}(D)) = 1\right] - \Pr\left[\text{Test}(\mathcal{A}(D')) = 1\right] > \Delta_{\mathcal{A}}, \tag{B.16}$$

and for all $\theta \in \mathcal{Y}$,

$$\text{Adv}(\overline{\text{Test}}) \stackrel{\text{def}}{=} \Pr\left[\overline{\text{Test}}(\bar{\mathcal{A}}(\bar{D}, \bar{u}, \theta)) = 1\right] - \Pr\left[\overline{\text{Test}}(\bar{\mathcal{A}}(\bar{D}', \bar{u}, \theta)) = 1\right] > \Delta_{\bar{\mathcal{A}}}. \tag{B.17}$$

Define $S' = \{\theta \in \mathcal{Y} | \text{Test}(\theta) = 1\}$ and $\bar{S}' = \{\theta \in \mathcal{Y} | \overline{\text{Test}}(\theta) = 1\}$. We have that the total variation distance between $\mathcal{A}(D)$ and $\mathcal{A}(D')$ is lower bounded as

$$\begin{aligned}
\text{TV}\left(\mathcal{A}(D); \mathcal{A}(D')\right) &= \sup_{S \subset \mathcal{Y}} \left|\Pr\left[\mathcal{A}(D) \in S\right] - \Pr\left[\mathcal{A}(D') \in S\right]\right| && \text{(B.18)}\\
&> \Pr\left[\mathcal{A}(D) \in S'\right] - \Pr\left[\mathcal{A}(D') \in S'\right] && \text{(B.19)}\\
&= \Pr\left[\text{Test}(\mathcal{A}(D)) = 1\right] - \Pr\left[\text{Test}(\mathcal{A}(D')) = 1\right] > \Delta_{\mathcal{A}}. && \text{(B.20)}
\end{aligned}$$

Similarly, we also have that for all $\theta \in \mathcal{Y}$, the total variation distance between $\bar{\mathcal{A}}(\bar{D}, \bar{u}, \theta)$ and $\bar{\mathcal{A}}(\bar{D}', \bar{u}, \theta)$ is lower bounded as

$$\begin{aligned}
\text{TV}\left(\bar{\mathcal{A}}(\bar{D}, \bar{u}, \theta); \bar{\mathcal{A}}(\bar{D}', \bar{u}, \theta)\right) &= \sup_{S \subset \mathcal{Y}} \left|\Pr\left[\bar{\mathcal{A}}(\bar{D}, \bar{u}, \theta) \in S\right] - \Pr\left[\bar{\mathcal{A}}(\bar{D}', \bar{u}, \theta) \in S\right]\right|\\
&> \Pr\left[\bar{\mathcal{A}}(\bar{D}, \bar{u}, \theta) \in \bar{S}'\right] - \Pr\left[\bar{\mathcal{A}}(\bar{D}', \bar{u}, \theta) \in \bar{S}'\right]\\
&= \Pr\left[\overline{\text{Test}}(\bar{\mathcal{A}}(\bar{D}, \bar{u}, \theta)) = 1\right] - \Pr\left[\overline{\text{Test}}(\bar{\mathcal{A}}(\bar{D}', \bar{u}, \theta)) = 1\right]\\
&> \Delta_{\bar{\mathcal{A}}}.
\end{aligned}$$

Assume W.L.O.G. that $\bar{u}$ replaces at index n and the edited databases $\bar{D} \circ u, \bar{D}' \circ u$ differs only at index 1. Also assume that D, D' differs at index n.

Recall from Definition 22 that satisfying (ε, δ)-deletion privacy under 1-adaptive 1-requesters requires existence of a map $\pi_n^{\mathbb{Q}} : \mathcal{X}^n \to \mathcal{Y}$ for each $\mathbb{Q}$ such that for all $D_0 \in \mathcal{X}^n$,

$$\Pr\left[\bar{\mathcal{A}}(D_{n-1}, u_n, \hat{\Theta}_{n-1}) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\pi_n^{\mathbb{Q}}(D_0 \circ u_n) \in S\right] + \delta. \tag{B.21}$$

To prove the theorem statement, we show that for a starting database $D_0 \in \{D, D'\}$ and an edit request $u_n = \bar{u}$ that deletes the differing record in choices of D_0 at edit step n, there exists a 1-adaptive 1-requester $\mathbb{Q}$ that sends adaptive edit requests $u_1, \cdots, u_{n-1}$ in the first $n-1$ steps such that no map $\pi_n^{\mathbb{Q}}$ exists that satisfies Equation B.21 for both choices of D_0 when ε follows inequality Equation B.15.

Consider the following construction of 1-adaptive 1-requester $\mathbb{Q}$ that only observes the first model $\hat{\Theta}_0 = \mathcal{A}(D_0)$ and generates the edit requests $(u_1, \cdots, u_{n-1})$ as follows:

$$\mathbb{Q}(\hat{\Theta}_0; u_1, u_2, \cdots, u_{i-1}) = \begin{cases} \langle i, \bar{D}[i] \rangle & \text{if } \mathrm{Test}(\hat{\Theta}_0) = 1, \\ \langle i, \bar{D}'[i] \rangle & \text{otherwise.} \end{cases} \tag{B.22}$$

This requester $\mathbb{Q}$ transforms any initial database D_0 to $D_{n-1} = \bar{D}$ if the outcome $\mathrm{Test}(\hat{\Theta}_0) = 1$, otherwise to $D_{n-1} = \bar{D}'$. Consider an adversary that does not observe the interaction transcript $(\hat{\Theta}_{<n}; u_{<n})$, but is interested in identifying whether D_0 was D or D'. The adversary gets to observe only the output $\hat{\Theta}_n = \bar{\mathcal{A}}(D_{n-1}, u_n, \hat{\Theta}_{n-1})$ generated after processing the edit request $u_n = \bar{u}$. On this observation, the adversary runs the membership inference test $\mathrm{MI}(\hat{\Theta}_n) = \overline{\mathrm{Test}}(\hat{\Theta}_n)$. The membership inference advantage of MI is

$$\begin{aligned}
\mathrm{Adv}(\mathrm{MI}; D, D') &\stackrel{\text{def}}{=} \Pr\left[\mathrm{MI}(\hat{\Theta}_n) = 1 | D_0 = D\right] - \Pr\left[\mathrm{MI}(\hat{\Theta}_n) = 1 | D_0 = D'\right] \\
&= \sum_{b \in \{0,1\}} \Pr\left[\overline{\mathrm{Test}}(\hat{\Theta}_n) = 1 | \mathrm{Test}(\hat{\Theta}_0) = b\right] \times \Pr\left[\mathrm{Test}(\hat{\Theta}_0) = b | D_0 = D\right] \\
&\quad - \sum_{b \in \{0,1\}} \Pr\left[\overline{\mathrm{Test}}(\hat{\Theta}_n) = 1 | \mathrm{Test}(\hat{\Theta}_0) = b\right] \times \Pr\left[\mathrm{Test}(\hat{\Theta}_0) = b | D_0 = D'\right] \\
&= \mathrm{Adv}(\mathrm{Test}; D, D') \Big(\Pr\left[\overline{\mathrm{Test}}(\hat{\Theta}_n) = 1 | D_{n-1} = \bar{D}\right] \\
&\qquad - \Pr\left[\overline{\mathrm{Test}}(\hat{\Theta}_n) = 1 | D_{n-1} = \bar{D}'\right] \Big) \\
&= \mathrm{Adv}(\mathrm{Test}; D, D') \times \mathrm{Adv}(\overline{\mathrm{Test}}; \bar{D}, \bar{D}', \bar{u}) \\
&> \Delta_{\mathcal{A}} \times \Delta_{\bar{\mathcal{A}}}.
\end{aligned}$$

So, from the contrapositive of our Theorem 20, we have that $(\mathcal{A}, \bar{\mathcal{A}})$ cannot be an (ε, δ)-deletion private algorithm pair for ε and δ satisfying

$$\Delta_{\mathcal{A}} \cdot \Delta_{\bar{\mathcal{A}}} > e^{\varepsilon} - 1 + 2\delta \tag{B.23}$$

$$\iff \varepsilon < \log(1 + \Delta_{\mathcal{A}} \cdot \Delta_{\bar{\mathcal{A}}} - 2\delta). \tag{B.24}$$

□

B.3.1 Our Reduction Theorem 21 versus Gupta et al. [51]'s

Adaptive unlearning guarantee in [51, Definition 2.3] is designed to ensure that no adaptive requester $\mathbb{Q}$ can force the output distribution of the unlearning algorithm $\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1})$ to diverge substantially from that of retraining algorithm $\mathcal{A}(D_i)$ with high probability. Such an attack is possible in unlearning algorithms that rely on some persistent states that are only randomized once during initialization. For example, Bourtoule et al. [19]'s SISA unlearning algorithm randomly partitions the initial database D_0 during setup and uses the same partitioning for processing edit requests, deleting records from respective shards on request. Gupta et al. [51] show that an adaptive update requester $\mathbb{Q}$ can interactively send deletion requests $u_1, \cdots, u_i$ to SISA so that after some time, the partitioning of remaining records in $D_i = D_0 \circ u_1 \cdots u_i$ follows a pattern that is unlikely to occur on repartitioning of D_i if we execute $\mathcal{A}(D_i)$.

They provide a general reduction [51, Theorem 3.1] from adaptive to non-adaptive unlearning guarantee under differential privacy. Their reduction relies on DP with regards to a change in the description of learning/unlearning algorithm's internal randomness and not with regards to the standard replacement of records. DP with respect to internal description of randomness means that an adversary observing an unlearned model remains uncertain about persistent states like database partitioning in SISA during setup. So from a triangle inequality type argument, Gupta et al. [51] show that with DP with respect to learning/unlearning algorithms' coins along with a non-adaptive unlearning guarantee implies an adaptive unlearning guarantee.

Our work shows that satisfying adaptive unlearning definition of Gupta et al. [51] still does not guarantee deletion privacy as per the RTBF guidelines. In Theorem 18, we demonstrate that there exists an algorithm pair $(\mathcal{A}, \bar{\mathcal{A}})$ satisfying adaptive

unlearning Definition 21 (a strictly stronger version of [51, Definition 2.3]), but still causes blatant non-privacy of deleted records in post-deletion release. The vulnerability we identify occurs because an adaptive requester can learn the identity of any target record before it is deleted and re-encode it back in the curator's database by sending edit requests. Because of this, an adversary (who knows how the adaptive requester works but does not have access to the requester's interaction transcript) can extract the identity of the target record from the model released after processing the deletion request. In our work, we argue that a reliable (and necessary) way to prevent this attack is to make sure that no adaptive requester ever learns the identity of a target record from the pre-deletion model releases it has access to. Consequently, our reduction in Theorem 21 from adaptive to non-adaptive requests relies on differential privacy with respect to the standard replacement of records instead.

B.3.2 Deferred Proofs for § 4.5.2

Lemma 14. *If $\boldsymbol{\ell}(\theta; \mathbf{x})$ is a twice continuously differentiable, convex, and β-smooth loss function and regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, then the map $\mathbf{T}_k$ defined in Equation 4.21 is:*

1. *a differentiable bijection for any $\eta < \frac{1}{\lambda+\beta}$, and*
2. *$(1-\eta\lambda)$-Lipschitz for any $\eta \leq \frac{2}{2\lambda+\beta}$.*

Proof. **Differentiable bijection.** To see that $\mathbf{T}_k$ is injective, assume $\mathbf{T}_k(\theta) = \mathbf{T}_k(\theta')$ for some $\theta, \theta' \in \mathbb{R}^d$. Then, by $(\beta+\lambda)$-smoothness of $\mathcal{L} \stackrel{\text{def}}{=} (\mathcal{L}_{D(k)} + \mathcal{L}_{D'(k)})/2$,

$$\begin{aligned}\|\theta - \theta'\|_2 &= \|\mathbf{T}_k(\theta) + \eta\nabla\mathcal{L}(\theta) - \mathbf{T}_k(\theta') - \eta\nabla\mathcal{L}(\theta')\|_2 \\ &= \eta\,\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2 \\ &\leq \eta(\lambda+\beta)\,\|\theta - \theta'\|_2\,.\end{aligned}$$

Since $\eta < 1/(\lambda+\beta)$, we must have $\|\theta - \theta'\|_2 = 0$. For showing $\mathbf{T}_k$ is surjective, consider the proximal mapping

$$\operatorname{prox}_{\mathcal{L}}(\theta) = \arg\min_{\theta' \in \mathbb{R}^d} \frac{\|\theta' - \theta\|_2^2}{2} - \eta\mathcal{L}(\theta'). \tag{B.25}$$

Note that $\text{prox}_{\mathcal{L}}(\cdot)$ is strongly convex for $\eta < \frac{1}{\lambda+\beta}$. Therefore, from KKT conditions, we have $\theta = \text{prox}_{\mathcal{L}}(\theta) - \eta\nabla\mathcal{L}(\text{prox}_{\mathcal{L}}(\theta)) = \mathbf{T}_k(\text{prox}_{\mathcal{L}}(\theta))$. Differentiability of $\mathbf{T}_k$ follows from the twice continuously differentiable assumption on $\boldsymbol{\ell}(\theta; \mathbf{x})$.

Lipschitzness. Let $\mathcal{L} \stackrel{\text{def}}{=} (\mathcal{L}_{D(k)} + \mathcal{L}_{D'(k)})/2$. For any $\theta, \theta' \in \mathbb{R}^d$,

$$\begin{aligned}\|\mathbf{T}_k(\theta) - \mathbf{T}_k(\theta')\|_2^2 &= \|\theta - \eta\nabla\mathcal{L}(\theta) - \theta' + \eta\nabla\mathcal{L}(\theta')\|_2^2 \\ &= \|\theta - \theta'\|_2^2 + \eta^2\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2^2 - 2\eta\langle\theta - \theta', \nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\rangle.\end{aligned}$$

We define a function $g(\theta) = \mathcal{L}(\theta) - \frac{\lambda}{2}\|\theta\|_2^2$, which is convex and β-smooth. By co-coercivity property of convex and β-smooth functions, we have

$$\begin{aligned}&\langle\theta - \theta', \nabla g(\theta) - \nabla g(\theta')\rangle \geq \frac{1}{\beta}\|\nabla g(\theta) - \nabla g(\theta')\|_2^2 \\ \implies &\langle\theta - \theta', \nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\rangle - \lambda\|\theta - \theta'\|_2^2 \geq \frac{1}{\beta}\Big(\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2^2 + \lambda^2\|\theta - \theta'\|_2^2 \\ &\qquad\qquad - 2\lambda\langle\theta - \theta', \nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\rangle\Big) \\ \implies &\langle\theta - \theta', \nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\rangle \geq \frac{1}{2\lambda+\beta}\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2^2 + \frac{\lambda(\lambda+\beta)}{2\lambda+\beta}\|\theta - \theta'\|_2^2.\end{aligned}$$

Substituting this in the above inequality, and noting that $\eta \leq \frac{2}{2\lambda+\beta}$, we get

$$\begin{aligned}\|\mathbf{T}_k(\theta) - \mathbf{T}_k(\theta')\|_2^2 &\leq \left(1 - \frac{2\eta\lambda(\lambda+\beta)}{2\lambda+\beta}\right)\|\theta - \theta'\|_2^2 + \left(\eta^2 - \frac{2\eta}{\beta+2\lambda}\right)\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2^2 \\ &\leq \left(1 - \frac{2\eta\lambda(\lambda+\beta)}{2\lambda+\beta}\right)\|\theta - \theta'\|_2^2 + \left(\eta^2\lambda^2 - \frac{2\eta\lambda^2}{\beta+2\lambda}\right)\|\theta - \theta'\|_2^2 \\ &= (1 - \eta\lambda)^2\|\theta - \theta'\|_2^2.\end{aligned}$$

□

Lemma 16. *If loss $\boldsymbol{\ell}(\theta; \mathbf{x})$ is convex and β-smooth, regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, and learning rate $\eta \leq \frac{2}{2\lambda+\beta}$, then the tracing diffusion $(\Theta_t)_{0\leq t\leq\eta(K_{\mathcal{A}}+iK_{\bar{\mathcal{A}}})}$ and $(\Theta'_t)_{0\leq t\leq\eta(K_{\mathcal{A}}+iK_{\bar{\mathcal{A}}})}$ defined in Equation 4.20 with $\Theta_0, \Theta'_0 \sim \rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda(1-\eta\lambda/2)}I_d\right)$ satisfy* LSI *inequality with constant $\lambda(1 - \eta\lambda/2)$.*

Proof. For any iteration $0 \leq k < K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}}$ in the extended run of Noisy-GD, and any $0 \leq s \leq \eta$, let's define two functions $\mathcal{L}_s, \mathcal{L}'_s : \mathbb{R}^d \to \mathbb{R}^d$ as follows:

$$\mathcal{L}_s = \frac{1 + s/\eta}{2}\mathcal{L}_{D(k)} + \frac{1 - s/\eta}{2}\mathcal{L}_{D'(k)}, \quad \text{and } \mathcal{L}'_s = \frac{1 - s/\eta}{2}\mathcal{L}_{D(k)} + \frac{1 + s/\eta}{2}\mathcal{L}_{D'(k)}. \tag{B.26}$$

Since $\mathbf{r}(\cdot)$ is the $L2(\lambda)$ regularizer and $\boldsymbol{\ell}(\theta;\mathbf{x})$ is convex and β-smoothness, both $\mathcal{L}_s$ and $\mathcal{L}'_s$ are λ-strongly convex and $(\lambda+\beta)$-smooth for all $0 \le s \le \eta$ and any k. We define maps $\mathbf{T}_s, \mathbf{T}'_s : \mathbb{R}^d \to \mathbb{R}^d$ as

$$\mathbf{T}_s(\theta) = \theta - \eta\nabla\mathcal{L}_s(\theta), \quad \text{and } \mathbf{T}'_s(\theta) = \theta - \nabla\mathcal{L}'_s(\theta). \tag{B.27}$$

From a similar argument as in Lemma 14, both $\mathbf{T}_s$ and $\mathbf{T}'_s$ are $(1-\eta\lambda)$-Lipschitz for learning rate $\eta \le \frac{2}{2\lambda+\beta}$.

Note that the densities of Θ_t and Θ'_t of the tracing diffusion for $t = \eta k + s$ can be respectively expressed as

$$P_t = \mathbf{T}_{s\#}(P_{\eta k}) \circledast \mathcal{N}\left(0, 2s\sigma^2 I_d\right), \quad \text{and } Q_t = \mathbf{T}'_{s\#}(Q_{\eta k}) \circledast \mathcal{N}\left(0, 2s\sigma^2 I_d\right), \tag{B.28}$$

where $P_{\eta k}$ and $Q_{\eta k}$ represent the distributions of $\Theta_{\eta k}$ and $\Theta'_{\eta k}$. We prove the lemma via induction.

Base step: Since Θ_0, Θ'_0 are both Gaussian distributed with variance $\frac{\sigma^2}{\lambda(1-\eta\lambda/2)}$, from Lemma 1 they satisfy LSI inequality with constant $\lambda(1-\eta\lambda/2)$.

Induction step: Suppose $P_{\eta k}$ and $Q_{\eta k}$ satisfy LSI inequality with constant $\lambda(1-\eta\lambda/2)$. Since equation Equation B.28 shows that P_t, Q_t are both Gaussian convolution on a pushforward distribution of $P_{\eta k}, Q_{\eta k}$ respectively over a Lipschitz function, from Lemma 1 and Lemma 2, both P_t, Q_t satisfy LSI inequality with constant

$$\left(\frac{(1-\eta\lambda)^2}{\lambda(1-\eta\lambda/2)} + 2s\right)^{-1} \ge \lambda(1-\eta\lambda/2) \times \underbrace{[(1-\eta\lambda)^2 + \lambda\eta(2-\eta\lambda)]^{-1}}_{=1}, \tag{B.29}$$

for all $\eta k \le t \le \eta(k+1)$. □

Theorem 23 (Deletion guarantee on $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ under convexity)**.** *Let the start distribution be* $\rho = \mathcal{N}\left(0, \frac{\hat{\sigma}^2}{\lambda n^2(1-\eta\lambda/2)} I_d\right)$*, the loss function* $\boldsymbol{\ell}(\theta;\mathbf{x})$ *be convex,* β*-smooth, and* L*-Lipschitz, the regularizer be* $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$*, and learning rate be* $\eta < \frac{1}{\lambda+\beta}$*. Then Algorithm pair* $(\mathcal{A}_{\textit{Noisy-GD}}, \bar{\mathcal{A}}_{\textit{Noisy-GD}})$*, as defined in Definition 23, satisfies* (q, ρ_{dd})*-Rényi deletion privacy under all non-adaptive* r*-requesters for any noise variance* $\hat{\sigma}^2 > 0$ *and* $K_{\mathcal{A}} \ge 0$ *if*

$$K_{\bar{\mathcal{A}}} \ge \frac{2}{\eta\lambda} \log\left(\frac{4qL^2}{\lambda\rho_{\text{dd}}\hat{\sigma}^2}\right). \tag{B.30}$$

Proof. Following the preceding discussion, to prove this theorem, it suffices to show that the inequality Equation 4.18 holds under the stated conditions. Consider the Fokker-Planck equation described in Lemma 12 for the pair of tracing diffusions SDEs in Equation 4.23: at any time t in duration $\eta k < t \leq \eta(k+1)$ for any iteration $0 \leq k < K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}}$,

$$\begin{cases} \partial_t P_t(\theta) & = \nabla \cdot \left(P_t(\theta) \mathbb{E}\left[\mathbf{U}_k(\Theta_{\eta k}) | \Theta_t = \theta \right] \right) + \sigma^2 \Delta P_t(\theta), \\ \partial_t Q_t(\theta) & = \nabla \cdot \left(Q_t(\theta) \mathbb{E}\left[-\mathbf{U}_k(\Theta'_{\eta k}) \middle| \Theta'_t = \theta \right] \right) + \sigma^2 \Delta Q_t(\theta), \end{cases} \tag{B.31}$$

where P_t and Q_t are the distribution of Θ_t and Θ'_t. Since $\boldsymbol{\ell}(\theta; \mathbf{x})$ is L-Lipschitz and for any $K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}} \leq k < K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}}$ we have $D(k)[\text{ind}] = D'(k)[\text{ind}]$, note from the definition of $\mathbf{U}_k(\theta)$ in Equation 4.23 that

$$\left\| \mathbb{E}\left[\mathbf{U}_k(\Theta_{\eta k}) | \Theta_t = \theta \right] - \mathbb{E}\left[-\mathbf{U}_k(\Theta'_{\eta k}) \middle| \Theta'_t = \theta \right] \right\|_2 \leq \begin{cases} \frac{2L}{n} & \text{if } k < K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}} \\ 0 & \text{otherwise} \end{cases}. \tag{B.32}$$

Therefore, applying Lemma 15 to the above pair of Fokker-Planck equations gives that for any t in duration $\eta k < t \leq \eta(k+1)$,

$$\partial_t \mathrm{R}_q\left(P_t \| Q_t\right) \leq \frac{2qL^2}{\sigma^2 n^2} \mathbb{1}\left\{ t \leq \eta(K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}}) \right\} - \frac{q\sigma^2}{2} \frac{\mathrm{I}_q\left(P_t \| Q_t\right)}{\mathrm{E}_q\left(P_t \| Q_t\right)}. \tag{B.33}$$

Equation B.33 suggests a phase change in the dynamics at iteration $k = K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}}$. In the phase I, the divergence bound increases with time due to the effect of the differing record in database pairs $(D_j, D'_j)_{0 \leq j \leq i-1}$. In phase II however, the update request u_i makes $D_{i-1} \circ u_i = D'_{i-1} \circ u_i$, and so doing gradient descent rapidly shrinks the divergence bound. This phase change is illustrated in Figure 4.6.

For brevity, we denote $R(q,t) = \mathrm{R}_q\left(P_t \| Q_t\right)$. Since $\eta < \frac{1}{\lambda+\beta} < \frac{2}{2\lambda+\beta}$, from Lemma 16, the distribution Q_t satisfies LSI inequality with constant $\lambda(1 - \lambda\eta/2)$. So, we can apply Lemma 6 to simplify the above partial differential inequality as follows.

$$\partial_t R(q,t) + \lambda(1 - \lambda\eta/2)\left(\frac{R(q,t)}{q} + (q-1)\partial_q R(q,t) \right) \leq q \begin{cases} \frac{2L^2}{\sigma^2 n^2} & \text{if } t \leq \eta(K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}}) \\ 0 & \text{otherwise} \end{cases}. \tag{B.34}$$

Recall from Remark 4 that $\sigma = \hat{\sigma}/n$. For brevity, let constant $c = \lambda(1 - \lambda\eta/2)$ and constant $\Delta_s = \frac{2L^2}{\hat{\sigma}^2}$. We define $u(q,t) = \frac{R(q,t)}{q}$. Then,

$$\partial_t R(q,t) + c\left(\frac{R(q,t)}{q} + (q-1)\partial_q R(q,t)\right) \leq q \begin{cases} \Delta_s & \text{if } t \leq \eta(K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}}) \\ 0 & \text{otherwise} \end{cases}$$

$$\implies \partial_t u(q,t) + cu(q,t) + c(q-1)\partial_q u(q,t) \leq \begin{cases} \Delta_s & \text{if } t \leq \eta(K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}}) \\ 0 & \text{otherwise} \end{cases}.$$

For some constant $\bar{q} > 1$, let $q(s) = (\bar{q}-1)\exp\left(c\left(s - \eta(K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}})\right)\right)+1$ and $t(s) = s$. Note that $\frac{\mathrm{d}q(s)}{\mathrm{d}s} = c(q(s)-1)$ and $\frac{\mathrm{d}t(s)}{\mathrm{d}s} = 1$. Therefore, for any $\eta k < s \leq \eta(k+1)$, the differential inequality followed along the path $u(s) = u(q(s), t(s))$ is

$$\begin{aligned} &\frac{\mathrm{d}u(s)}{\mathrm{d}s} + cu(s) \leq \Delta_s \times \mathbb{1}\left\{t \leq \eta(K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}})\right\} \\ \implies &\frac{\mathrm{d}}{\mathrm{d}s}\{e^{cs}u(s)\} \leq \Delta_s \times \mathbb{1}\left\{t \leq \eta(K_{\mathcal{A}} + (i-1)K_{\bar{\mathcal{A}}})\right\}. \end{aligned}$$

Since the map $\mathbf{T}_k(\cdot)$ in Equation 4.21 is a differentiable bijection for $\eta < \frac{1}{\lambda+\beta}$ as per Lemma 14, note that Lemma 13 implies that $\lim_{s\to\eta k^+} u(s) = u(\eta k)$. Therefore, we can directly integrate in the duration $0 \leq t \leq \eta(K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}})$ to get

$$\begin{aligned} &[e^{cs}u(s)]_0^{\eta(K_{\mathcal{A}}+iK_{\bar{\mathcal{A}}})} \leq \int_0^{\eta(K_{\mathcal{A}}+(i-1)K_{\bar{\mathcal{A}}})} \Delta_s e^{cs}\mathrm{d}s \\ \implies &e^{c\eta(K_{\mathcal{A}}+iK_{\bar{\mathcal{A}}})}u(\eta(K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}})) - u(0) \leq \frac{\Delta_s}{c}[e^{c\eta(K_{\mathcal{A}}+(i-1)K_{\bar{\mathcal{A}}})} - 1] \\ \implies &u(\eta(K_{\mathcal{A}} + iK_{\bar{\mathcal{A}}})) \leq \frac{\Delta_s}{c}e^{-c\eta K_{\bar{\mathcal{A}}}}. \qquad (\text{Since } u(0) = R(q(0),0)/q(0) = 0.) \end{aligned}$$

Noting that $q(0) \geq 1$, on reverting the substitution, we get

$$\begin{aligned} \mathrm{R}_{\bar{q}}\left(P_{\eta(K_{\mathcal{A}}+iK_{\bar{\mathcal{A}}})} \middle\| Q_{\eta(K_{\mathcal{A}}+iK_{\bar{\mathcal{A}}})}\right) &\leq \frac{2\bar{q}L^2}{\lambda\hat{\sigma}^2(1-\eta\lambda/2)}\exp\left(-\eta\lambda K_{\bar{\mathcal{A}}}(1-\eta\lambda/2)\right) \\ &\leq \frac{4\bar{q}L^2}{\lambda\hat{\sigma}^2}\exp\left(-\frac{\eta\lambda K_{\bar{\mathcal{A}}}}{2}\right) \qquad (\text{Since } \eta < \tfrac{1}{\lambda+\beta}) \end{aligned}$$

Recall from our construction that $P_{\eta(K_{\mathcal{A}}+iK_{\bar{\mathcal{A}}})}$ and $Q_{\eta(K_{\mathcal{A}}+iK_{\bar{\mathcal{A}}})}$ are the distributions of outputs $\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1})$ and $\bar{\mathcal{A}}(D'_{i-1}, u_i, \hat{\Theta}'_{i-1})$ respectively. Therefore, choosing $K_{\bar{\mathcal{A}}}$ as specified in the theorem statement concludes the proof. □

Lemma 17. *Suppose the loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is convex, L-Lipschitz, and β-smooth, and the regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$. Then, the excess empirical risk of any randomly*

distributed parameter Θ *for any database* $D \in \mathcal{X}^n$ *after applying any edit request* $u \in \mathcal{U}^r$ *that modifies no more than* r *records is bounded as*

$$\operatorname{err}(\Theta; D \circ u) \leq \left(1 + \frac{\beta}{\lambda}\right)\left[2 \operatorname{err}(\Theta; D) + \frac{16 r^2 L^2}{\lambda n^2}\right]. \tag{B.35}$$

Proof. Let θ^*_D and $\theta^*_{D\circ u}$ be the minimizers of objectives $\mathcal{L}_D(\cdot)$ and $\mathcal{L}_{D\circ u}(\cdot)$ as defined in Equation 4.12. From λ-strong convexity of the $\mathcal{L}_D$,

$$\mathcal{L}_D(\theta^*_{D\circ u}) - \mathcal{L}_D(\theta^*_D) \geq \frac{\lambda}{2}\left\|\theta^*_{D\circ u} - \theta^*_D\right\|_2^2. \tag{B.36}$$

From optimality of $\theta^*_{D\circ u}$ and L-Lipschitzness of $\boldsymbol{\ell}(\theta; \mathbf{x})$, we have

$$\begin{aligned}
\mathcal{L}_D(\theta^*_{D\circ u}) &= \mathcal{L}_{D\circ u}(\theta^*_{D\circ u}) + \frac{1}{n}\left(\sum_{\mathbf{x}\in D} \boldsymbol{\ell}(\theta^*_{D\circ u}; \mathbf{x}) - \sum_{\mathbf{x}\in D\circ u} \boldsymbol{\ell}(\theta^*_{D\circ u}; \mathbf{x})\right) \\
&\leq \mathcal{L}_{D\circ u}(\theta^*_D) + \frac{1}{n}\left(\sum_{\mathbf{x}\in D} \boldsymbol{\ell}(\theta^*_{D\circ u}; \mathbf{x}) - \sum_{\mathbf{x}\in D\circ u} \boldsymbol{\ell}(\theta^*_{D\circ u}; \mathbf{x})\right) \\
&= \mathcal{L}_D(\theta^*_D) + \frac{1}{n}\sum_{\mathbf{x}\in D}\left(\boldsymbol{\ell}(\theta^*_{D\circ u}; \mathbf{x}) - \boldsymbol{\ell}(\theta^*_D; \mathbf{x})\right) + \frac{1}{n}\sum_{\mathbf{x}\in D\circ u}\left(\boldsymbol{\ell}(\theta^*_D; \mathbf{x}) - \boldsymbol{\ell}(\theta^*_{D\circ u}; \mathbf{x})\right) \\
&\leq \mathcal{L}_D(\theta^*_D) + \frac{2rL}{n}\left\|\theta^*_{D\circ u} - \theta^*_D\right\|_2.
\end{aligned}$$

Combining the two inequalities give

$$\left\|\theta^*_{D\circ u} - \theta^*_D\right\|_2 \leq \frac{4rL}{\lambda n}. \tag{B.37}$$

Therefore, from $(\lambda+\beta)$-smoothness of $\mathcal{L}_{D\circ u}$ and λ-strong convexity of $\mathcal{L}_D$, we have

$$\begin{aligned}
\operatorname{err}(\Theta; D \circ u) &= \mathbb{E}\left[\mathcal{L}_{D\circ u}(\Theta) - \mathcal{L}_{D\circ u}(\theta^*_{D\circ u})\right] \\
&\leq \frac{\lambda+\beta}{2}\mathbb{E}\left[\left\|\Theta - \theta^*_{D\circ u}\right\|_2^2\right] \\
&\leq (\lambda+\beta)\left[\mathbb{E}\left[\left\|\Theta - \theta^*_D\right\|_2^2\right] + \left\|\theta^*_D - \theta^*_{D\circ u}\right\|_2^2\right] \\
&\leq \left(1 + \frac{\beta}{\lambda}\right)\left[2\mathbb{E}\left[\mathcal{L}_D(\Theta) - \mathcal{L}_D(\theta^*_D)\right] + \frac{16 r^2 L^2}{\lambda n^2}\right].
\end{aligned}$$

□

Lemma 18 (Accuracy of Noisy-GD)**.** *For convex, L-Lipschitz, and,* β*-smooth loss function* $\boldsymbol{\ell}(\theta; \mathbf{x})$ *and regularizer* $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$*, if learning rate* $\eta < \frac{1}{\lambda+\beta}$*, the excess empirical risk of* $\Theta_{\eta K} = \textit{Noisy-GD}(D, \Theta_0, K) \in \mathbb{R}^d$ *for any* $D \in \mathcal{X}^n$ *is bounded as*

$$\operatorname{err}(\Theta_{\eta K}; D) \leq \operatorname{err}(\Theta_0; D) e^{-\lambda\eta K/2} + \left(1 + \frac{\beta}{\lambda}\right)\frac{d\hat{\sigma}^2}{n^2} \tag{B.38}$$

Proof. Let $\Theta_{\eta k}$ denote the k^{th} iteration parameter of Noisy-GD run. Recall that $k+1^{\text{th}}$ noisy gradient update step is

$$\Theta_{\eta(k+1)} = \Theta_{\eta k} - \eta \nabla \mathcal{L}_D(\Theta_{\eta k}) + \sqrt{2\eta\sigma^2} W_k, \tag{B.39}$$

where $\sigma = \hat{\sigma}/n$. From $(\beta+\lambda)$-smoothness of $\mathcal{L}_D$, we have

$$\begin{aligned}
\mathcal{L}_D(\Theta_{\eta(k+1)}) &\leq \mathcal{L}_D(\Theta_{\eta k}) + \left\langle \nabla \mathcal{L}_D(\Theta_{\eta k}), \Theta_{\eta(k+1)} - \Theta_{\eta k} \right\rangle + \frac{\beta+\lambda}{2} \left\| \Theta_{\eta(k+1)} - \Theta_{\eta k} \right\|_2^2 \\
&= \mathcal{L}_D(\Theta_{\eta k}) - \eta \left\| \nabla \mathcal{L}_D(\Theta_{\eta k}) \right\|_2^2 + \sqrt{2\eta\sigma^2} \left\langle \nabla \mathcal{L}_D(\Theta_{\eta k}), W_k \right\rangle \\
&\quad + \frac{\eta^2(\beta+\lambda)}{2} \left\| \nabla \mathcal{L}_D(\Theta_{\eta k}) \right\|_2^2 + \eta\sigma^2(\beta+\lambda) \left\| W_k \right\|_2^2 \\
&\quad - \eta\sqrt{2\eta\sigma^2}(\beta+\lambda) \left\langle \nabla \mathcal{L}_D(\Theta_{\eta k}), W_k \right\rangle
\end{aligned}$$

On taking expectation over the joint distribution of $\Theta_{\eta k}, \Theta_{\eta(k+1)}, W_k$, the above simplifies to

$$\mathbb{E}\left[\mathcal{L}_D(\Theta_{\eta(k+1)})\right] \leq \mathbb{E}\left[\mathcal{L}_D(\Theta_{\eta k})\right] - \eta\left(1 - \frac{\eta(\lambda+\beta)}{2}\right) \mathbb{E}\left[\left\| \nabla \mathcal{L}_D(\Theta_{\eta k}) \right\|_2^2\right] + \eta d\sigma^2(\beta+\lambda). \tag{B.40}$$

Let $\theta_D^* = \arg\min_{\theta \in \mathbb{R}^d} \mathcal{L}_D(\theta)$. From λ-strong convexity of $\mathcal{L}_D$, for any $\theta \in \mathbb{R}^d$, we have

$$\left\| \nabla \mathcal{L}_D(\theta) \right\|_2^2 \geq 2\lambda(\mathcal{L}_D(\theta) - \mathcal{L}_D(\theta_D^*)). \tag{B.41}$$

Let $\gamma = \lambda\eta(2 - \eta(\lambda+\beta))$. Plugging this in the above inequality, and subtracting $\mathcal{L}_D(\theta_D^*)$ on both sides, for $\eta < \frac{1}{\lambda+\beta}$, we get

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_D(\Theta_{\eta(k+1)}) - \mathcal{L}_D(\theta_D^*)\right] &\leq (1-\gamma)\mathbb{E}\left[\mathcal{L}_D(\Theta_{\eta k}) - \mathcal{L}_D(\theta_D^*)\right] + \eta d\sigma^2(\beta+\lambda) \\
&\leq (1-\gamma)^{k+1}\mathbb{E}\left[\mathcal{L}_D(\Theta_0) - \mathcal{L}_D(\theta^*)\right] \\
&\quad + \eta d\sigma^2(\beta+\lambda)(1 + \cdots + (1-\gamma)^{k+1}) \\
&\leq e^{-\gamma(k+1)/2}\mathbb{E}\left[\mathcal{L}_D(\Theta_0) - \mathcal{L}_D(\theta_D^*)\right] + \frac{\eta d\sigma^2(\beta+\lambda)}{\gamma}.
\end{aligned}$$

For $\eta < \frac{1}{\lambda+\beta}$, note that $\gamma \geq \lambda\eta$, and so

$$\operatorname{err}(\Theta_{\eta K}; D) \leq \operatorname{err}(\Theta_0; D) e^{-\lambda\eta K/2} + \left(1 + \frac{\beta}{\lambda}\right) d\sigma^2. \tag{B.42}$$

□

Theorem 24 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants* $\lambda, \beta, L > 0$*, constant* $q > 1$*, and constants* $\rho_{\text{dp}} \geq \rho_{\text{dd}} > 0$*. Define constant* $\nu = \frac{\lambda+\beta}{\lambda}$*. Let the loss function* $\boldsymbol{\ell}(\theta; \mathbf{x})$ *be twice differentiable, convex, L-Lipschitz, and* β*-smooth, and let the regularizer be* $\mathbf{r}(\theta) = \frac{\lambda}{2} \|\theta\|_2^2$*. If the learning rate is* $\eta = \frac{1}{2(\lambda+\beta)}$*, the gradient noise variance is* $\hat{\sigma}^2 = \frac{4qL^2}{\lambda\rho_{\text{dp}}}$*, and the weight initialization distribution is* $\rho = \mathcal{N}\left(0, \frac{\hat{\sigma}^2}{\lambda n^2(1-\eta\lambda/2)} I_d\right)$*, then*

(1.) *both* $\mathcal{A}_{\textit{Noisy-GD}}$ *and* $\bar{\mathcal{A}}_{\textit{Noisy-GD}}$ *are* (q, ρ_{dp})*-Rényi DP for any* $K_{\mathcal{A}}, K_{\bar{\mathcal{A}}} \geq 0$*,*

(2.) *pair* $(\mathcal{A}_{\textit{Noisy-GD}}, \bar{\mathcal{A}}_{\textit{Noisy-GD}})$ *satisfies* (q, ρ_{dd})*-deletion-privacy all non-adaptive r-requesters*

$$\textit{if} \quad K_{\bar{\mathcal{A}}} \geq 4\nu \log \frac{\rho_{\text{dp}}}{\rho_{\text{dd}}}, \tag{B.43}$$

(3.) *and all models in sequence* $(\hat{\Theta}_i)_{i\geq 0}$ *produced by the interactions between* $(\mathcal{A}_{\textit{Noisy-GD}}, \bar{\mathcal{A}}_{\textit{Noisy-GD}})$ *and* $\mathbb{Q}$ *on any* $D_0 \in \mathcal{X}^n$*, where* $\mathbb{Q}$ *is any r-requester, have an excess empirical risk* $\text{err}(\hat{\Theta}_i; D_i) = \frac{10\nu q d L^2}{\lambda\rho_{\text{dp}} n^2}$

$$\textit{if} \quad K_{\mathcal{A}} \geq 4\nu \log\left(\frac{\rho_{\text{dp}} n^2}{4qd}\right), \quad \textit{and} \quad K_{\bar{\mathcal{A}}} \geq 4\nu \log \max\left\{5\nu, \frac{8\rho_{\text{dp}} r^2}{qd}\right\}. \tag{B.44}$$

Proof. **(1.) Privacy.** Note than if $\boldsymbol{\ell}(\theta; \mathbf{x})$ is L-Lipschitz, it has a gradient sensitivity $S_\ell = 2L$. By Corollary 1, the Noisy-GD with K iterations will be (q, ρ_{dp})-Rényi DP for the stated choice of loss function, regularizer, and learning rate as long as $\hat{\sigma}^2 \geq \frac{4qL^2}{\lambda\rho_{\text{dp}}}\left(1 - e^{-\lambda\eta K/2}\right)$. Therefore, if we set $\hat{\sigma}^2 = \frac{4qL^2}{\lambda\rho_{\text{dp}}}$, Noisy-GD is (q, ρ_{dp})-Rényi DP for any K. For the same $\hat{\sigma}^2$, both $\mathcal{A}_{\text{Noisy-GD}}$ and $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ are also (q, ρ_{dp})-Rényi DP for any $K_{\mathcal{A}}$ and $K_{\bar{\mathcal{A}}}$ as they run Noisy-GD on respective databases for generating the output.

(2.) Deletion. By Theorem 23, for the stated choice of loss function, regularizer, learning rate, and weight initialization distribution, the algorithm pair $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}})$ satisfies (q, ρ_{dd})-deletion-privacy under all non-adaptive r-requesters $\mathbb{Q}$ if $K_{\bar{\mathcal{A}}} \geq \frac{2}{\eta\lambda} \log\left(\frac{4qL^2}{\lambda\rho_{\text{dd}}\hat{\sigma}^2}\right)$. By plugging in $\hat{\sigma}^2 = \frac{4qL^2}{\lambda\rho_{\text{dp}}}$ and $\eta = \frac{1}{2(\lambda+\beta)}$, this constraint simplifies to $K_{\bar{\mathcal{A}}} \geq 4\nu \log \frac{\rho_{\text{dp}}}{\rho_{\text{dd}}}$.

(3.) Accuracy. We prove the induction hypothesis that under the conditions stated in the theorem, $\text{err}(\hat{\Theta}_i; D_i) \leq \frac{10\nu q d L^2}{\lambda\rho_{\text{dp}} n^2}$ for all $i \geq 0$.

Base case: The minimizer $\theta^*_{D_0}$ of $\mathcal{L}_{D_0}$ satisfies

$$\nabla \mathcal{L}_{D_0}(\theta^*_{D_0}) = \frac{1}{n}\sum_{\mathbf{x}\in D_0} \nabla \boldsymbol{\ell}(\theta^*_{D_0}; \mathbf{x}) - \lambda\theta^*_{D_0} = 0 \implies \left\|\theta^*_{D_0}\right\|_2 \leq \frac{L}{\lambda}. \tag{B.45}$$

As a result, the excess empirical risk of initialization weights $\Theta_0 \sim \rho = \mathcal{N}\left(0, \frac{\hat{\sigma}^2}{\lambda n^2(1-\eta\lambda/2)} I_d\right)$ on $\mathcal{L}_{D_0}$ is bounded as

$$\begin{aligned}
\operatorname{err}(\Theta_0; D_0) &= \mathbb{E}\left[\mathcal{L}_{D_0}(\Theta_0) - \mathcal{L}_{D_0}(\theta^*_{D_0})\right] \\
&\leq \frac{(\lambda+\beta)}{2}\mathbb{E}\left[\left\|\Theta_0 - \theta^*_{D_0}\right\|_2^2\right] \qquad \text{(From } (\lambda+\beta)\text{-smoothness of } \mathcal{L}_{D_0}) \\
&= \frac{(\lambda+\beta)}{2}\left[\left\|\theta^*_{D_0}\right\|_2^2 + \mathbb{E}\left[\|\Theta_0\|_2^2\right] - 2\mathbb{E}\left[\left\langle\theta^*_{D_0}, \Theta_0\right\rangle\right]\right] \\
&\leq \left(1+\frac{\beta}{\lambda}\right)\left[\frac{L^2}{2\lambda} + \frac{\hat{\sigma}^2 d}{n^2(2-\lambda\eta)}\right] \\
&\qquad \text{(From Equation B.45 and } \mathbb{E}\left[\|W\|_2^2\right] = d \text{ if } W \sim \mathcal{N}(0, I_d).) \\
&\leq \nu\left[\frac{L^2}{2\lambda} + \frac{d\hat{\sigma}^2}{n^2}\right].
\end{aligned}$$

Since $\hat{\Theta}_0 = \mathcal{A}_{\text{Noisy-GD}}(D_0) = \text{Noisy-GD}(D_0, \Theta_0, K_{\mathcal{A}})$, by Lemma 18, $K_{\mathcal{A}} \geq 2\nu \log\left(\frac{\rho_{\text{dp}} n^2}{4qd}\right)$ iterations gives

$$\begin{aligned}
\operatorname{err}(\hat{\Theta}_0; D_0) &\leq \operatorname{err}(\Theta_0; D_0) e^{-\lambda\eta K_{\mathcal{A}}/2} + \frac{\nu d\hat{\sigma}^2}{n^2} \\
&\leq \nu\left[\frac{L^2}{2\lambda} + \frac{d\hat{\sigma}^2}{n^2}\right] e^{-\lambda\eta K_{\mathcal{A}}/2} + \frac{\nu d\hat{\sigma}^2}{n^2} \\
&\leq \frac{\nu L^2}{2\lambda} e^{-\lambda\eta K_{\mathcal{A}}/2} + \frac{8\nu d L^2}{\lambda\rho_{\text{dp}} n^2} \qquad \text{(On substituting } \hat{\sigma}^2 = \tfrac{4qL^2}{\lambda\rho_{\text{dp}}}) \\
&\leq \frac{10\nu q d L^2}{\lambda\rho_{\text{dp}} n^2} \qquad \text{(Since } K_{\mathcal{A}} \geq 4\nu\log\left(\tfrac{\rho_{\text{dp}} n^2}{4qd}\right))
\end{aligned}$$

Induction step: Assume that $\operatorname{err}(\hat{\Theta}_{i-1}; D_{i-1}) \leq \frac{10\nu q d L^2}{\lambda\rho_{\text{dp}} n^2}$. Since $\hat{\Theta}_i = \bar{\mathcal{A}}_{\text{Noisy-GD}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) = \text{Noisy-GD}(D_i, \hat{\Theta}_{i-1}, K_{\bar{\mathcal{A}}})$, by Lemma 18 and Lemma 17, running $K_{\bar{\mathcal{A}}} \geq 2\nu \log\max\left\{5\nu, \frac{8r^2}{qd}\right\}$ iterations gives

$$\begin{aligned}
\operatorname{err}(\hat{\Theta}_i; D_i) &\leq \nu\left[2\operatorname{err}(\hat{\Theta}_{i-1}; D_{i-1}) + \frac{16r^2L^2}{\lambda n^2}\right] e^{-\lambda\eta K_{\bar{\mathcal{A}}}/2} + \frac{\nu d\hat{\sigma}^2}{n^2} \\
&\leq \nu\left[\frac{20\nu q d L^2}{\lambda\rho_{\text{dp}} n^2} + \frac{16r^2L^2}{\lambda n^2}\right] e^{-\lambda\eta K_{\bar{\mathcal{A}}}/2} + \frac{4\nu q d L^2}{\lambda\rho_{\text{dp}} n^2} \qquad \text{(Substituting } \hat{\sigma}^2) \\
&\leq \frac{16\nu r^2 L^2}{\lambda n^2} e^{-\lambda\eta K_{\bar{\mathcal{A}}}/2} + \frac{8\nu q d L^2}{\lambda\rho_{\text{dp}} n^2} \qquad \text{(Since } K_{\bar{\mathcal{A}}} \geq 4\nu\log(5\nu)) \\
&\leq \frac{10\nu q d L^2}{\lambda\rho_{\text{dp}} n^2} \qquad \text{(Since } K_{\bar{\mathcal{A}}} \geq 4\nu\log\tfrac{8\rho_{\text{dp}} r^2}{qd})
\end{aligned}$$

□

B.3.3 Deferred Proofs for § 4.5.3

Lemma 19 (Proposition 14 [29]). *For tracing diffusion Θ_t defined in Equation 4.40, the equivalent Fokker-Planck equation in the interval $\eta k \le t \le \eta(k+1)$ is*

$$\partial_t P_t(\theta) = \nabla \cdot \left(\left\{ \mathbb{E}\left[\nabla_D(\Theta_{\eta k}) - \nabla \mathcal{L}_D(\Theta_t) | \Theta_t = \theta\right] + \sigma^2 \nabla \log \frac{P_t(\theta)}{\pi(D)(\theta)} \right\} P_t(\theta) \right), \tag{B.46}$$

where $\pi(D)$ is the Gibbs distribution defined in Equation 2.39.

Proof. Conditioned on observing parameter $\Theta_{\eta k} = \theta_{\eta k}$, the process $(\Theta_t)_{\eta k \le t \le \eta(k+1)}$ is a Langevin diffusion along a constant Vector field $\nabla \mathcal{L}_D(\theta_{\eta k})$. Therefore, the conditional probability density $P_{t|\eta k}(\cdot|\theta_{\eta k})$ of Θ_t given $\theta_{\eta k}$ follows the following Fokker-Planck equation.

$$\partial_t P_{t|\eta k}(\cdot|\theta_{\eta k}) = \sigma^2 \Delta P_{t|\eta k}(\cdot|\theta_{\eta k}) + \nabla \cdot \left(P_{t|\eta k}(\cdot|\theta_{\eta k}) \nabla \mathcal{L}_D(\theta_{\eta k}) \right) \tag{B.47}$$

Taking expectation over $\Theta_{\eta k}$, we have

$$\begin{aligned}
\partial_t P_t(\cdot) &= \int P_{\eta k}(\theta_{\eta k}) \left\{ \sigma^2 \Delta P_{t|\eta k}(\cdot|\theta_{\eta k}) + \nabla \cdot \left(P_{t|\eta k}(\cdot|\theta_{\eta k}) \nabla \mathcal{L}_D(\theta_{\eta k}) \right) \right\} \mathrm{d}\theta_{\eta k} \\
&= \sigma^2 \Delta P_t(\cdot) + \nabla \cdot \left(P_t(\cdot) \nabla \mathcal{L}_D(\cdot) \right) + \nabla \cdot \left(P_t(\cdot) \int \left[\nabla \mathcal{L}_D(\theta_{\eta k}) - \nabla \mathcal{L}_D(\cdot) \right] P_{\eta k|t}(\theta_{\eta k}|\cdot) \mathrm{d}\theta_{\eta k} \right) \\
&= \sigma^2 \nabla \cdot \left(P_t(\cdot) \nabla \log \frac{P_t(\cdot)}{\pi(D)(\cdot)} \right) + \nabla \cdot \left(\mathbb{E}\left[\nabla \mathcal{L}_D(\Theta_{\eta k}) - \nabla \mathcal{L}_D(\cdot) | \Theta_t = \cdot \right] P_t(\cdot) \right),
\end{aligned}$$

where $P_{\eta k|t}$ is the conditional density of $\Theta_{\eta k}$ given Θ_t. For the last equality, we have used the fact that $\nabla \mathcal{L}_D = -\sigma^2 \nabla \log \pi(D)$ from Equation 4.39. □

Lemma 20 ([29, Proposition 15]). *Let $Q_t \overset{def}{=} P_t/\pi(D)$ where $\pi(D)$ is the Gibbs distribution defined in Equation 4.39 and $\psi_t \overset{def}{=} Q_t^{q-1}/\mathrm{E}_q\left(Q_t \| \pi(D)\right)$. The rate of change in $\mathrm{R}_q\left(P_t \| \pi(D)\right)$ along racing diffusion in time $\eta k \le t \le \eta(k+1)$ is bounded as*

$$\partial_t \mathrm{R}_q\left(P_t \| \pi(D)\right) \le -\frac{3q\sigma^2}{4} \frac{\mathrm{I}_q\left(P_t \| \pi(D)\right)}{\mathrm{E}_q\left(P_t \| \pi(D)\right)} + \frac{q}{\sigma^2} \mathbb{E}\left[\psi_t(\Theta_t) \left\| \nabla \mathcal{L}_D(\Theta_{\eta k}) - \nabla \mathcal{L}_D(\Theta_t) \right\|_2^2 \right]. \tag{B.48}$$

Proof. For brevity, let $\Delta_t(\cdot) = \mathbb{E}\left[\nabla\mathcal{L}_D(\Theta_{\eta k}) - \nabla\mathcal{L}_D(\Theta_t)|\Theta_t = \cdot\right]$ in context of this proof. From Lebinz integral rule, we have

$$\begin{aligned}
\partial_t \mathrm{R}_q\left(P_t\|\pi(D)\right) &= \frac{q}{(q-1)\mathrm{E}_q\left(P_t\|\pi(D)\right)} \int \left(\frac{P_t}{\pi(D)}\right)^{q-1} \partial_t P_t \mathrm{d}\theta \\
&= \frac{q}{(q-1)\mathrm{E}_q\left(P_t\|\pi(D)\right)} \int Q_t^{q-1} \nabla \cdot \left(\left\{\Delta_t + \sigma^2 \nabla \log Q_t\right\} P_t\right) \mathrm{d}\theta \\
&\qquad\qquad \text{(From Equation 4.41)} \\
&= -\frac{q}{(q-1)\mathrm{E}_q\left(P_t\|\pi(D)\right)} \int \left\langle \nabla\left(Q_t^{q-1}\right), \Delta_t + \sigma^2 \nabla \log Q_t \right\rangle P_t \mathrm{d}\theta \\
&= -\frac{q}{\mathrm{E}_q\left(P_t\|\pi(D)\right)} \int Q_t^{q-2} \left\langle \nabla Q_t, \Delta_t + \sigma^2 \frac{\nabla Q_t}{Q_t} \right\rangle P_t \mathrm{d}\theta \\
&= -\frac{q}{\mathrm{E}_q\left(P_t\|\pi(D)\right)} \left\{ \sigma^2 \mathrm{I}_q\left(P_t\|\pi(D)\right) + \frac{2}{q} \underbrace{\mathop{\mathbb{E}}_{P_t}\left[Q_t^{q/2-1} \left\langle \nabla\left(Q_t^{q/2}\right), \Delta_t \right\rangle\right]}_{\stackrel{\text{def}}{=} F_1} \right\} \\
&\qquad\qquad \text{(From Equation 2.10)}
\end{aligned}$$

Note that the expectation in $\Delta_t(\cdot)$ is over the conditional distribution $P_{\eta k|t}$ while the expectation in F_1 is over P_t. Therefore, we can combine the two to get an expectation over the unconditional joint distribution over Θ_t and $\Theta_{\eta k}$ as follows.

$$\begin{aligned}
-F_1 &= \mathop{\mathbb{E}}_{\Theta_t \sim P_t}\left[Q_t^{q/2-1}(\Theta_t) \left\langle \nabla\left(Q_t^{q/2}\right)(\Theta_t), \mathop{\mathbb{E}}_{\Theta_{\eta k} \sim P_{\eta k|t}}\left[\nabla\mathcal{L}_D(\Theta_t) - \nabla\mathcal{L}_D(\Theta_{\eta k})\right] \right\rangle\right] \\
&= \mathop{\mathbb{E}}_{P_{\eta k,t}}\left[Q_t^{q/2-1}(\Theta_t) \left\langle \nabla\left(Q_t^{q/2}\right)(\Theta_t), \nabla\mathcal{L}_D(\Theta_t) - \nabla\mathcal{L}_D(\Theta_{\eta k}) \right\rangle\right] \\
&\leq \frac{\sigma^2}{2q}\mathbb{E}\left[Q_t^{-1}(\Theta_t) \left\|\nabla\left(Q_t^{q/2}\right)(\Theta_t)\right\|_2^2\right] + \frac{q}{2\sigma^2}\mathbb{E}\left[Q_t^{q-1}(\Theta_t) \left\|\nabla\mathcal{L}_D(\Theta_t) - \nabla\mathcal{L}_{B_k}(\Theta_{\eta k})\right\|_2^2\right] \\
&= \frac{q\sigma^2}{8}\mathrm{I}_q\left(Q_t\|P\right) + \frac{q}{2\sigma^2}\mathbb{E}\left[Q_t^{q-1}(\Theta_t) \left\|\nabla\mathcal{L}_D(\Theta_t) - \nabla\mathcal{L}_{B_k}(\Theta_{\eta k})\right\|_2^2\right] \\
&\qquad\qquad \text{(From Equation 2.10)}
\end{aligned}$$

Substituting it in the preceding inequality proves the proposition. □

Lemma 21 (Change of measure inequality [29])**.** *If $\boldsymbol{\ell}(\theta; \mathbf{x})$ is β-smooth, and regu-*

larizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$*, then for any probability density* P *on* $\mathbb{R}^d$*,*

$$\mathop{\mathbb{E}}_{P}\left[\|\nabla\mathcal{L}_D\|_2^2\right] \leq 4\sigma^4 \mathop{\mathbb{E}}_{\pi(D)}\left[\left\|\nabla\sqrt{\frac{P}{\pi(D)}}\right\|_2^2\right] + 2d\sigma^2(\beta+\lambda), \tag{B.49}$$

where $\pi(D)$ *is the Gibbs distribution defined in Equation 4.39.*

Proof. Consider the Langevin diffusion Equation 2.31 described in § 2.3 over the potential $\mathcal{L}_D$. The Gibbs distribution $\pi(D)$ is its stationary distribution, and the diffusion's infinitesimal generator $\mathcal{G}$ applied on the $\mathcal{L}_D$ gives

$$\mathcal{G}\mathcal{L}_D = \sigma^2\Delta\mathcal{L}_D - \|\nabla\mathcal{L}_D\|_2^2. \tag{B.50}$$

Therefore,

$$\begin{aligned}
\mathop{\mathbb{E}}_{P}\left[\|\nabla\mathcal{L}_D\|_2^2\right] &= \sigma^2\mathop{\mathbb{E}}_{P}\left[\Delta\mathcal{L}_D\right] - \mathop{\mathbb{E}}_{P}\left[\mathcal{G}\mathcal{L}_D\right] && \text{(From Equation B.50)}\\
&\leq d\sigma^2(\beta+\lambda) - \int \mathcal{G}\mathcal{L}_D\left(\frac{P}{\pi(D)} - 1\right)\pi(D)\mathrm{d}\theta\\
& && \text{(From } \beta\text{-smoothness and Equation 2.40)}\\
&= d\beta\sigma^2(\beta+\lambda) + \int\left[\|\nabla\mathcal{L}_D\|_2^2 - \sigma^2\Delta\mathcal{L}_D\right]\left(\frac{P}{\pi(D)} - 1\right)\pi(D)\mathrm{d}\theta\\
&= d\beta\sigma^2(\beta+\lambda) + \int\|\nabla\mathcal{L}_D\|_2^2\,(P-\pi(D))\mathrm{d}\theta\\
&\quad + \sigma^2\int\left\langle\nabla\mathcal{L}_D, \nabla\left[\left(\frac{P}{\pi(D)} - 1\right)\pi(D)\right]\right\rangle\mathrm{d}\theta && \text{(From Equation 2.24)}\\
&= d\beta\sigma^2(\beta+\lambda) + \int\|\nabla\mathcal{L}_D\|_2^2\,(P-\pi(D))\mathrm{d}\theta\\
&\quad + \sigma^2\int\left\langle\nabla\mathcal{L}_D, -\frac{\nabla\mathcal{L}_D}{\sigma^2}\right\rangle(P-\pi(D))\mathrm{d}\theta\\
&\quad + \sigma^2\int\left\langle\nabla\mathcal{L}_D, \nabla\frac{P}{\pi(D)}\right\rangle\pi(D)\mathrm{d}\theta && \text{(Since } \nabla\pi(D) = -\tfrac{\nabla\mathcal{L}_D}{\sigma^2}\pi(D)\text{)}\\
&= d\beta\sigma^2(\beta+\lambda) + 0 + 2\sigma^2\int\left\langle\sqrt{\frac{P}{\pi(D)}}\nabla\mathcal{L}_D, \nabla\sqrt{\frac{P}{\pi(D)}}\right\rangle\pi(D)\mathrm{d}\theta\\
&\leq d\beta\sigma^2(\beta+\lambda) + \frac{1}{2}\mathop{\mathbb{E}}_{P}\left[\|\nabla\mathcal{L}_D\|_2^2\right] + 2\sigma^4\mathop{\mathbb{E}}_{\pi(D)}\left[\left\|\nabla\sqrt{\frac{P}{\pi(D)}}\right\|_2^2\right]\\
& && \text{(From Equation 2.25 with } a = 2\sigma^2\text{)}
\end{aligned}$$

□

Theorem 25 (Convergence of Noisy-GD in Rényi divergence)**.** *Let constants* $\beta, \lambda, \sigma^2 > 0$ *and* $q, B > 1$*. Suppose the loss function* $\boldsymbol{\ell}(\theta; \mathbf{x})$ *is* $(\sigma^2\log(B)/4)$*-bounded*

and β-smooth, and regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$. If step size is $\eta \leq \frac{\lambda}{64Bq^2(\beta+\lambda)^2}$, then for any database $D \in \mathcal{X}^n$ and any weight initialization distribution P_0 for Θ_0, the Rényi divergence of distribution $P_{\eta K}$ of output model $\Theta_{\eta K} = \text{Noisy-GD}(D, \Theta_0, K)$ with respect to the Gibbs distribution $\pi(D)$ defined in Equation 4.39 shrinks as follows:

$$\mathrm{R}_q\left(P_{\eta K}\|\pi(D)\right) \leq q\exp\left(-\frac{\lambda\eta K}{2B}\right)\mathrm{R}_q\left(P_0\|\pi(D)\right) + \frac{32d\eta qB(\beta+\lambda)^2}{\lambda}. \tag{B.51}$$

Proof. From $(\beta+\lambda)$-smoothness of loss $\mathcal{L}_D$ we have that for any $\eta k \leq t \leq \eta(k+1)$,

$$\begin{aligned}
\|\nabla\mathcal{L}_D(\Theta_{\eta k}) - \nabla\mathcal{L}_D(\Theta_t)\|_2^2 &\leq (\beta+\lambda)^2\|\Theta_{\eta k} - \Theta_t\|_2^2 \\
&= (\beta+\lambda)^2\left\|(t-\eta k)\nabla\mathcal{L}_D(\Theta_{\eta k}) - \sqrt{2(t-\eta k)\sigma^2}W_k\right\|_2^2 \\
&\qquad\qquad \text{(From Equation 4.40)} \\
&\leq 2\eta^2(\beta+\lambda)^2\|\nabla\mathcal{L}_D(\Theta_{\eta k})\|_2^2 + 4\eta\sigma^2(\beta+\lambda)^2\|W_k\|_2^2 \\
&\leq 4\eta^2(\beta+\lambda)^2\|\nabla\mathcal{L}_D(\Theta_{\eta k}) - \nabla\mathcal{L}_D(\Theta_t)\|_2^2 \\
&\quad + 4\eta^2(\beta+\lambda)^2\|\nabla\mathcal{L}_D(\Theta_t)\|_2^2 + 4\eta\sigma^2(\beta+\lambda)^2\|W_k\|_2^2
\end{aligned}$$

Let $Q_t \overset{\text{def}}{=} \frac{P_t}{\pi(D)}$ and $\psi_t \overset{\text{def}}{=} Q_t^{q-1}/\mathrm{E}_q\left(Q_t\|\pi(D)\right)$. If $\eta \leq \frac{1}{2\sqrt{2}(\beta+\lambda)}$, we rearrange to get the following and use it to get the following bound on the discretization error in Equation 4.42:

$$\begin{aligned}
\mathbb{E}\left[\psi_t(\Theta_t)\|\nabla\mathcal{L}_D(\Theta_{\eta k}) - \nabla\mathcal{L}_D(\Theta_t)\|_2^2\right] &\leq 8\eta^2(\beta+\lambda)^2\underbrace{\mathbb{E}\left[\psi_t(\Theta_t)\|\nabla\mathcal{L}_D(\Theta_t)\|_2^2\right]}_{\overset{\text{def}}{=}F_1} \\
&\quad + 32\eta\sigma^2(\beta+\lambda)^2\underbrace{\mathbb{E}\left[\psi_t(\Theta_t)\|W_k\|_2^2/4\right]}_{\overset{\text{def}}{=}F_2}.
\end{aligned}$$

Hence, for solving the PDI Equation 4.42, we have to bound the three expectations F_1 and F_2.

1. **Bounding F_1.** Note that $\mathop{\mathbb{E}}_{\Theta_t\sim P_t}\left[\psi_t(\Theta_t)\right] = \int\psi_t(\theta)P_t(\theta)\mathrm{d}\theta = \frac{1}{\mathrm{E}_q(Q_t\|\pi(D))}\int\frac{P_t^q}{\pi(D)^{q-1}}\mathrm{d}\theta = 1$. So, $\psi_t P_t(\theta) \overset{\text{def}}{=} \psi_t(\theta)P_t(\theta)$ is a probability density function on $\mathbb{R}^d$. On applying

the measure change Lemma 21 on it, we get

$$F_1 = \mathop{\mathbb{E}}_{\psi_t P_t}\left[\|\nabla \mathcal{L}_D\|_2^2\right] \leq 4\sigma^4 \mathop{\mathbb{E}}_{\pi(D)}\left[\left\|\nabla\sqrt{\frac{\psi_t P_t}{\pi(D)}}\right\|_2^2\right] + 2d\sigma^2(\beta+\lambda)$$

(From Equation 4.43)

$$= 4\sigma^4 \mathop{\mathbb{E}}_{\pi(D)}\left[\frac{\left\|\nabla\sqrt{Q_t^q}\right\|_2^2}{\mathrm{E}_q\left(P_t\|\pi(D)\right)}\right] + 2d\sigma^2(\beta+\lambda)$$

$$= \sigma^4 q^2 \frac{\mathrm{I}_q\left(P_t\|\pi(D)\right)}{\mathrm{E}_q\left(P_t\|\pi(D)\right)} + 2d\sigma^2(\beta+\lambda).$$

(From Equation 2.10)

2. **Bounding F_2.** Since $\psi_t P_t$ is a valid density on $\mathbb{R}^d$, the joint density $\psi_t P_{t,z}(\theta, z) \stackrel{\text{def}}{=} \psi_t(\theta) P_{t,z}(\theta, z)$ where $P_{t,z}$ is the joint density of Θ_t and W_k is also a valid density. Note that the F_2 is an expectation on $\|W_k\|_2^2$ taken over the joint density $\psi_t P_{t,z}$. We can perform a measure change operation using Donsker-Varadhan principle to get

$$F_2 = \mathop{\mathbb{E}}_{\psi_t P_{t,z}}\left[\|W_k\|_2^2/4\right] \leq \mathrm{KL}\left(\psi_t P_{t,z}\|P_{t,z}\right) + \log \mathop{\mathbb{E}}_{P_z}\left[\exp(\|W_k\|_2^2/4)\right],$$

where we simplified the second term using the fact that the marginal P_z of $P_{t,z}$ is a standard normal Gaussian. The random variable $\|W_k\|_2^2$ is distributed according to the Chi-squared distribution χ_d^2 with d degrees of freedom. Since the moment generating function of Chi-squared distribution is $\mathrm{M}_{\chi_d^2}(t) = \mathop{\mathbb{E}}_{X\sim\chi_d^2}[\exp(tX)] = (1-2t)^{-d/2}$ for $t < \frac{1}{2}$, we can simplify the second term in F_2 as

$$\log \mathop{\mathbb{E}}_{P_z}\left[\exp(\|W_k\|_2^2/4)\right] = \log \mathrm{M}_{\chi_d^2}\left(\frac{1}{4}\right) = \frac{d\log 2}{2}. \tag{B.52}$$

The KL divergence term can be simplified as follows.

$$\begin{aligned}
\mathrm{KL}\left(\psi_t P_{t,z} \| P_{t,z}\right) &= \int\int \psi_t P_{t,z}(\theta_t, z) \log \psi_t(\theta_t) \mathrm{d}\theta_t \mathrm{d}z \\
&= \int \psi_t P_t \log \frac{Q_t^{q-1}}{\mathrm{E}_q\left(P_t \| \pi(D)\right)} \mathrm{d}\theta_t \qquad \text{(On marginalization of } z\text{)} \\
&= \frac{q-1}{q} \int P_t \psi_t \log \left\{ \frac{Q_t^q}{\mathrm{E}_q\left(P_t \| \pi(D)\right)} - \log \mathrm{E}_q\left(P_t \| \pi(D)\right)^{1/(q-1)} \right\} \mathrm{d}\theta_t \\
&= \frac{q-1}{q} \left\{ \mathrm{KL}\left(P_t \psi_t \| \pi(D)\right) - \mathrm{R}_q\left(P_t \| \pi(D)\right) \right\} \\
&\leq \mathrm{KL}\left(P_t \psi_t \| \pi(D)\right) \qquad \text{(Since } \mathrm{R}_q\left(P_t \| \pi(D)\right) > 0\text{)}
\end{aligned}$$

Note that under the assumptions of the Theorem, $\pi(D)$ satisfies log-Sobolev inequality Equation 2.43 with constant λ/B (i.e. satisfies LSI(λ/B)). To see this, recall from Lemma 1 that the Gaussian distribution $\rho(\theta) = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda} I_d\right)$ satisfies LSI(λ) inequality. Since loss $\boldsymbol{\ell}(\theta; \mathbf{x})$ is $(\sigma^2 \log(B)/4)$-bounded, the density ratio $\frac{\pi(D)(\theta)}{\rho(\theta)} \in \left[\frac{1}{\sqrt{B}}, \sqrt{B}\right]$. The claim therefore follows from Lemma 3. Using this inequality, from Lemma 4 we have

$$\begin{aligned}
\mathrm{KL}\left(P_t \psi_t \| \pi(D)\right) &\leq \frac{\sigma^2 B}{2\lambda} \int P_t \psi_t \left\| \nabla \log \left(\frac{P_t \psi_t}{\pi(D)} \right) \right\|_2^2 \mathrm{d}\theta_t \\
&= \frac{\sigma^2 B}{2\lambda} \int \frac{Q_t^q}{\mathrm{E}_q\left(P_t \| \pi(D)\right)} \left\| \nabla \log(Q_t^q) \right\|_2^2 \pi(D) \mathrm{d}\theta_t \\
&= \frac{2\sigma^2 B}{\lambda} \frac{1}{\mathrm{E}_q\left(P_t \| \pi(D)\right)} \int \left\| \nabla (Q_t^{q/2}) \right\|_2^2 \pi(D) \mathrm{d}\theta_t \\
&= \frac{q^2 \sigma^2 B}{2\lambda} \frac{\mathrm{I}_q\left(P_t \| \pi(D)\right)}{\mathrm{E}_q\left(P_t \| \pi(D)\right)}
\end{aligned}$$

On combining all the two bounds on F_1 and F_2 and rearranging, we have

$$\begin{aligned}
\mathbb{E}\left[\psi_t(\Theta_t) \left\| \nabla \mathcal{L}_D(\Theta_{\eta k}) - \nabla \mathcal{L}_D(\Theta_t) \right\|_2^2 \right] \leq\ & 8\eta q^2 \sigma^4 (\beta + \lambda)^2 \frac{\mathrm{I}_q\left(P_t \| \pi(D)\right)}{\mathrm{E}_q\left(P_t \| \pi(D)\right)} \left(\eta + \frac{2B}{\lambda} \right) \\
& + 16\eta d \sigma^2 (\beta + \lambda)^2 \left(\eta(\beta + \lambda) + \log 2 \right)
\end{aligned}$$

Let step size be $\eta \leq \min\left\{ \frac{2B}{\lambda}, \frac{\lambda}{64 B q^2 (\beta+\lambda)^2} \right\}$. Then, the first term above is bounded as

$$8\eta q^2 \sigma^4 (\beta + \lambda)^2 \frac{\mathrm{I}_q\left(P_t \| \pi(D)\right)}{\mathrm{E}_q\left(P_t \| \pi(D)\right)} \left(\eta + \frac{2B}{\lambda} \right) \leq \frac{\sigma^4}{2} \frac{\mathrm{I}_q\left(P_t \| \pi(D)\right)}{\mathrm{E}_q\left(P_t \| \pi(D)\right)}. \tag{B.53}$$

Let $\eta \leq \frac{1}{4(\beta+\lambda)}$. Then, in the third term, $\left(\eta(\beta + \lambda) + \log 2\right) \leq 1$. Plugging the bound on discretization error back in the PDI Equation 4.42, we get

$$\partial_t \mathrm{R}_q\left(P_t \| \pi(D)\right) \leq -\frac{q\sigma^2}{4} \frac{\mathrm{I}_q\left(P_t \| \pi(D)\right)}{\mathrm{E}_q\left(P_t \| \pi(D)\right)} + 16\eta d q (\beta + \lambda)^2. \tag{B.54}$$

Since $\pi(D)$ satisfies LSI(λ/B) inequality, from Lemma 6 this PDI reduces to

$$\partial_t \mathrm{R}_q\left(P_t \| \pi(D)\right) + \frac{\lambda}{2B}\left(\frac{\mathrm{R}_q\left(P_t \| \pi(D)\right)}{q} + (q-1)\partial_q \mathrm{R}_q\left(P_t \| \pi(D)\right)\right) \leq 16 d\eta q(\beta+\lambda)^2. \tag{B.55}$$

Let $c_1 = \frac{\lambda}{2B}$ and $c_2 = 16d\eta(\beta+\lambda)^2$. Additionally, let $u(q,t) = \frac{\mathrm{R}_q(P_t\|\pi(D))}{q}$. Then,

$$\begin{aligned}
&\partial_t \mathrm{R}_q\left(P_t \| \pi(D)\right) + c_1\left(\frac{\mathrm{R}_q\left(P_t \| \pi(D)\right)}{q} + (q-1)\partial_q \mathrm{R}_q\left(P_t \| \pi(D)\right)\right) \leq c_2 q \\
\implies &\frac{\partial_t \mathrm{R}_q\left(P_t \| \pi(D)\right)}{q} + c_1 \frac{\mathrm{R}_q\left(P_t \| \pi(D)\right)}{q} + c_1(q-1)\left(\frac{\partial_q \mathrm{R}_q\left(P_t \| \pi(D)\right)}{q} - \frac{\mathrm{R}_q\left(P_t \| \pi(D)\right)}{q^2}\right) \leq c_2 \\
\implies &\partial_t u(q,t) + c_1 u(q,t) + c_1(q-1)\partial_q u(q,t) \leq c_2.
\end{aligned}$$

For some constant $\bar{q} \geq 1$, let $q(s) = (\bar{q}-1)\exp(c_1(s-\eta K)) + 1$, and $t(s) = s$. Note that $\frac{\mathrm{d}q(s)}{\mathrm{d}s} = c_1(q(s)-1)$ and $\frac{\mathrm{d}t(s)}{\mathrm{d}s} = 1$. Therefore, for any $0 \leq t \leq \eta K$, the PDI above implies the following differential inequality is followed along the path $u(s) = u(q(s), t(s))$.

$$\begin{aligned}
\frac{\mathrm{d}u(s)}{\mathrm{d}s} + c_1 u(s) \leq c_2 \implies &\frac{\mathrm{d}}{\mathrm{d}s}\{e^{c_1 s} u(s)\} \leq c_2 e^{c_1 s} \\
\implies &[e^{c_1 s} u(s)]_0^{\eta K} \leq \int_0^{\eta K} c_2 e^{c_1 s} \mathrm{d}s \\
\implies &e^{c_1 \eta K} u(\eta K) - u(0) \leq \frac{c_2(e^{c_1 \eta K} - 1)}{c_1} \\
\implies &u(\eta K) \leq e^{-c_1 \eta K} u(0) + \frac{c_2}{c_1}(1 - e^{-c_1 \eta K}).
\end{aligned}$$

On reversing the parameterization of q and t, we get

$$\begin{aligned}
\mathrm{R}_{q(\eta K)}\left(P_{\eta K} \| \pi(D)\right) &\leq \frac{q(\eta K)}{q(0)} e^{-c_1 \eta K} \mathrm{R}_{q(0)}\left(P_0 \| \pi(D)\right) + \frac{c_2}{c_1} q(\eta K) \\
&\leq \frac{q(\eta K)}{q(0)} \exp\left(-\frac{\lambda \eta K}{2B}\right) \mathrm{R}_{q(0)}\left(P_0 \| \pi(D)\right) + \frac{32 d\eta B(\beta+\lambda)^2}{\lambda} q(\eta K).
\end{aligned}$$

Since $q(0) > 1$ and $\bar{q} = q(\eta K) > q(0)$, from monotonicity of Rényi divergence in q, we get

$$\mathrm{R}_{\bar{q}}\left(P_{\eta K} \| \pi(D)\right) \leq \bar{q} \exp\left(-\frac{\lambda \eta K}{2B}\right) \mathrm{R}_{\bar{q}}\left(P_0 \| \pi(D)\right) + \frac{32 d\eta \bar{q} B(\beta+\lambda)^2}{\lambda}. \tag{B.56}$$

Finally, noting that for constants $B, q > 1$ and $\beta, \lambda > 0$,

$$\eta \leq \min\{\frac{1}{2\sqrt{2}(\beta+\lambda)}, \frac{1}{4(\beta+\lambda)}, \frac{2B}{\lambda}, \frac{\lambda}{64Bq^2(\beta+\lambda)^2}\} = \frac{\lambda}{64Bq^2(\beta+\lambda)^2}, \tag{B.57}$$

completes the proof. □

Lemma 23 (Indistinguishability under bounded perturbations). *For two potential functions $\mathcal{L}, \mathcal{L}' : \mathbb{R}^d \to \mathbb{R}$ and some constant σ^2, let $P \propto e^{-\mathcal{L}/\sigma^2}$ and $Q \propto e^{-\mathcal{L}'/\sigma^2}$ be the respective Gibbs distributions. If $|\mathcal{L}(\theta) - \mathcal{L}'(\theta)| \leq c$ for all $\theta \in \mathbb{R}^d$, then $\mathrm{R}_q\left(P \| Q\right) \leq \frac{2c}{\sigma^2}$ for all $q > 1$.*

Proof. The Gibbs distributions P, Q have a density

$$P(\theta) = \frac{1}{\Lambda} e^{-\mathcal{L}(\theta)/\sigma^2}, \quad \text{and} \quad Q(\theta) = \frac{1}{\Lambda'} e^{-\mathcal{L}'(\theta)/\sigma^2},$$

where Λ, Λ' are the respective normalization constants. If for all $\theta \in \mathbb{R}^d$, the potential difference $|\mathcal{L}(\theta) - \mathcal{L}'(\theta)| \leq c$, then

$$\begin{aligned}
\mathrm{R}_q\left(P \| Q\right) &= \frac{1}{q-1} \log \int \frac{P^q}{Q^{q-1}} \mathrm{d}\theta \\
&= \frac{1}{q-1} \log \int \left(\frac{\Lambda'}{\Lambda}\right)^{q-1} \exp\left(\frac{q-1}{\sigma^2}(\mathcal{L}'(\theta) - \mathcal{L}(\theta))\right) \times P(\theta) \mathrm{d}\theta \\
&\leq \frac{1}{q-1} \left\{ (q-1) \log \frac{\Lambda'}{\Lambda} + \log \exp\left(\frac{c(q-1)}{\sigma^2} \int P \mathrm{d}\theta\right) \right\} \\
&= \frac{1}{q-1} \left\{ (q-1) \log \frac{\int \exp\left(-\frac{\mathcal{L}(\theta)}{\sigma^2} + \frac{\mathcal{L}(\theta) - \mathcal{L}'(\theta)}{\sigma^2}\right) \mathrm{d}\theta}{\int \exp\left(-\frac{\mathcal{L}(\theta)}{\sigma^2}\right) \mathrm{d}\theta} + \frac{c(q-1)}{\sigma^2} \right\} \\
&\leq \frac{2c}{\sigma^2}.
\end{aligned}$$

□

Theorem 26 (Near optimality of Gibbs sampling). *If the loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is $\sigma^2 \log(B)/4$-bounded and β-smooth, the regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2} \|\theta\|_2^2$, then the excess empirical risk for a model $\bar{\Theta} \in \mathbb{R}^d$ sampled from the Gibbs distribution $\pi(D) \propto e^{-\mathcal{L}_D/\sigma^2}$ is*

$$\mathrm{err}(\bar{\Theta}; D) = \mathbb{E}\left[\mathcal{L}_D(\bar{\Theta}) - \mathcal{L}_D(\theta_D^*)\right] \leq \frac{d\sigma^2}{2} \left(\log \frac{\beta + \lambda}{\lambda} + \sqrt{B}\right). \tag{B.58}$$

Proof. We simplify expected loss as

$$\mathbb{E}\left[\mathcal{L}_D(\bar{\Theta})\right] = \int \mathcal{L}_D \pi(D) \mathrm{d}\theta = \sigma^2(\mathrm{H}(\pi(D)) - \log(\Lambda_D)), \tag{B.59}$$

where the differential entropy is given by

$$\mathrm{H}(\pi(D)) = -\int \pi(D) \log \pi(D) \mathrm{d}\theta = -\int \frac{e^{-\mathcal{L}_D/\sigma^2}}{\Lambda_D} \log \frac{e^{-\mathcal{L}_D/\sigma^2}}{\Lambda_D} \mathrm{d}\theta \tag{B.60}$$

is the differential entropy of $\pi(D)$, and $\Lambda_D = \int e^{-\mathcal{L}_D/\sigma^2}\mathrm{d}\theta$ is the normalization constant. Since the potential function $\mathcal{L}_D$ is $(\lambda+\beta)$-smooth, we have

$$\begin{aligned}
-\sigma^2\log(\Lambda_D) &= -\sigma^2\log\int e^{-\mathcal{L}_D/\sigma^2}\mathrm{d}\theta \\
&= \mathcal{L}_D(\theta_D^*) - \sigma^2\log\int e^{(\mathcal{L}_D(\theta_D^*)-\mathcal{L}_D(\theta))/\sigma^2}\mathrm{d}\theta \\
&\le \mathcal{L}_D(\theta_D^*) - \sigma^2\log\int e^{-(\beta+\lambda)\left\|\theta-\theta_D^*\right\|_2^2/2\sigma^2}\mathrm{d}\theta \\
&= \mathcal{L}_D(\theta_D^*) - \frac{d\sigma^2}{2}\log\left(\frac{2\pi\sigma^2}{\lambda+\beta}\right).
\end{aligned}$$

Since $\boldsymbol{\ell}(\theta;\mathbf{x})$ is $\sigma^2\log(B)/4$-bounded, note that for the Gaussian distribution $\rho \sim \mathcal{N}\left(0, \frac{\sigma^2}{\lambda}I_d\right)$, the density ratio lies in $\frac{\pi(D)(\theta)}{\rho(\theta)} \in \left[\frac{1}{\sqrt{B}}, \sqrt{B}\right]$ for all $\theta \in R^d$. We decompose entropy $\mathrm{H}(\pi(D))$ into cross-entropy and KL divergence to get

$$\begin{aligned}
\mathrm{H}(\pi(D)) &= -\int \pi(D)\log\rho\mathrm{d}\theta - \mathrm{KL}\left(\pi(D)\|\rho\right) \\
&\le -\int \pi(D)\log\left[\left(\frac{\lambda}{2\pi\sigma^2}\right)^{d/2} e^{-\frac{\lambda\|\theta\|_2^2}{2\sigma^2}}\right]\mathrm{d}\theta \qquad (\text{Since } \mathrm{KL}\left(\pi(D)\|\rho\right) \ge 0) \\
&= \frac{d}{2}\log\frac{2\pi\sigma^2}{\lambda} + \frac{\lambda}{2\sigma^2}\int \|\theta\|_2^2\,\pi(D)(\theta)\mathrm{d}\theta \\
&\le \frac{d}{2}\log\frac{2\pi\sigma^2}{\lambda} + \frac{\lambda\sqrt{B}}{2\sigma^2}\int \|\theta\|_2^2\,\rho(\theta)\mathrm{d}\theta \qquad (\text{Since } \tfrac{\pi(D)(\theta)}{\rho(\theta)} \in \left[\tfrac{1}{\sqrt{B}}, \sqrt{B}\right]) \\
&= \frac{d}{2}\log\frac{2\pi\sigma^2}{\lambda} + \frac{d\sqrt{B}}{2}.
\end{aligned}$$

On combining the bounds, we get

$$\mathrm{err}(\bar{\Theta}; D) = \mathbb{E}\left[\mathcal{L}_D(\bar{\Theta}) - \mathcal{L}_D(\theta_D^*)\right] \le \frac{d\sigma^2}{2}\left(\log\frac{\beta+\lambda}{\lambda} + \sqrt{B}\right). \tag{B.61}$$

□

Theorem 27 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda, \beta, L, \sigma^2, \eta > 0$, constants $q, B > 1$, and constants $d > \rho_{\mathrm{dp}} \ge \rho_{\mathrm{dd}} > 0$. Let the loss function $\boldsymbol{\ell}(\theta;\mathbf{x})$ be $\frac{\sigma^2\log(B)}{4}$-bounded, L-Lipschitz and β-smooth, the regularizer be $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, and the weight initialization distribution be $\rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda}I_d\right)$. Then,*

(1.) *both $\mathcal{A}_{\textit{Noisy-GD}}$ and $\bar{\mathcal{A}}_{\textit{Noisy-GD}}$ are (q, ρ_{dp})-Rényi DP for any $\eta \ge 0$ and any $K_{\mathcal{A}}, K_{\bar{\mathcal{A}}} \ge 0$ if*

$$\sigma^2 \ge \frac{qL^2}{\rho_{\mathrm{dp}}n^2}\cdot\eta\max\{K_{\mathcal{A}}, K_{\bar{\mathcal{A}}}\}, \tag{B.62}$$

(2.) *pair* $(\mathcal{A}_{\textit{Noisy-GD}}, \bar{\mathcal{A}}_{\textit{Noisy-GD}})$ *satisfies* (q, ρ_{dd}) *Rényi deletion privacy under all non-adaptive* r*-requesters for any* $\sigma^2 > 0$*, if learning rate is* $\eta \leq \frac{\lambda \rho_{\mathrm{dd}}}{64dqB(\beta+\lambda)^2}$ *and the number of iterations satisfy*

$$\begin{aligned} K_{\mathcal{A}} &\geq \frac{2B}{\lambda\eta} \log\left(\frac{q\log(B)}{\rho_{\mathrm{dd}}}\right), \text{ and} \\ K_{\bar{\mathcal{A}}} &\geq K_{\mathcal{A}} - \frac{2B}{\lambda\eta} \log\left(\frac{\log(B)}{2\left(\rho_{\mathrm{dd}} + \frac{r}{n}\log(B)\right)}\right), \end{aligned} \tag{B.63}$$

(3.) *and all models in the sequence* $(\hat{\Theta}_i)_{i\geq 0}$ *produced by interactions between* $(\mathcal{A}_{\textit{Noisy-GD}}, \bar{\mathcal{A}}_{\textit{Noisy-GD}})$ *and* $\mathbb{Q}$ *on any* $D_0 \in \mathcal{X}^n$*, where* $\mathbb{Q}$ *is an* r*-requester, have an excess empirical risk* $\mathrm{err}(\hat{\Theta}_i; D_i) = \tilde{O}\left(\frac{dq}{\rho_{\mathrm{dp}} n^2} + \frac{1}{n}\sqrt{\frac{q\rho_{\mathrm{dd}}}{\rho_{\mathrm{dp}}}}\right)$ *when inequalities in Equation B.63 and Equation B.62 are equalities.*

Proof. **(1.) Privacy.** By Theorem 39, Noisy-GD with K iterations on an L-Lipschitz loss function satisfies (q, ρ_{dp})-Rényi DP for any initial weight distribution ρ and learning rate $\eta \geq 0$ if $\sigma^2 = \frac{qL^2}{\rho_{\mathrm{dp}} n^2} \cdot \eta K$. Since, both $\mathcal{A}_{\text{Noisy-GD}}$ and $\bar{\mathcal{A}}_{\text{Noisy-GD}}$ run Noisy-GD for $K_{\mathcal{A}}$ and $K_{\bar{\mathcal{A}}}$ iterations respectively, setting the noise variance given in the Theorem statement ensures (q, ρ_{dp})-Rényi DP for both.

(2.) Deletion. For showing Rényi deletion-privacy under non-adaptive requests, recall that it is sufficient to show that there exists a map $\pi : \mathcal{X}^n \to \mathcal{Y}$ such that for all $i \geq 1$,

$$\mathrm{R}_q\left(\bar{\mathcal{A}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) \middle\| \pi(D_i)\right) \leq \rho_{\mathrm{dd}}, \tag{B.64}$$

for all edit sequences $(u_i)_{i\geq 1}$ from $\mathcal{U}^r$, where $(\hat{\Theta}_i)_{i\geq 0}$ is the sequence of models generated by the interaction of $(\mathcal{A}_{\text{Noisy-GD}}, \bar{\mathcal{A}}_{\text{Noisy-GD}}, \mathbb{Q})$ on any database $D_0 \in \mathcal{X}^n$. For all $i \geq 0$, let $\hat{P}_i$ denote the distribution of $\hat{\Theta}_i$. We prove Equation B.64 via induction.

Base step: Note that the initial weight distribution $\rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda} I_d\right)$ has a density proportional to $e^{-\mathbf{r}(\theta)/\sigma^2}$ and the distribution $\pi(D_0)$ has a density proportional to $e^{-\mathcal{L}_{D_0}(\theta)/\sigma^2}$. Since both of these are Gibbs distributions with their potential difference $|\mathcal{L}_{D_0}(\theta) - \mathbf{r}(\theta)| \leq \sigma^2 \log(B)/4$ for all $\theta \in \mathbb{R}^d$ due to boundedness assumption on $\boldsymbol{\ell}(\theta; \mathbf{x})$, we have from Lemma 23 that

$$\mathrm{R}_q\left(\rho \| \pi(D_0)\right) \leq \frac{2}{\sigma^2} \times \frac{\sigma^2 \log(B)}{4} = \frac{\log(B)}{2}. \tag{B.65}$$

Under the stated assumptions on loss $\boldsymbol{\ell}(\theta;\mathbf{x})$ and learning rate η, note that the convergence Theorem 25 holds. Since $\hat{\Theta}_0 = \mathcal{A}_{\text{Noisy-GD}}(D_0) = \text{Noisy-GD}(D_0, \Theta_0, K_{\mathcal{A}})$, where $\Theta_0 \sim \rho$, we have

$$\begin{aligned}
\mathrm{R}_q\left(\hat{P}_0 \middle\| \pi(D_0)\right) &\leq q\exp\left(-\frac{\lambda\eta K_{\mathcal{A}}}{2B}\right)\mathrm{R}_q\left(\rho\|\pi(D_0)\right) + \frac{32d\eta qB(\beta+\lambda)^2}{\lambda} \\
&\leq q\exp\left(-\frac{\lambda\eta K_{\mathcal{A}}}{2B}\right)\left(\frac{\log(B)}{2}\right) + \frac{\rho_{\text{dd}}}{2} \qquad (\text{Since } \eta \leq \tfrac{\lambda\rho_{\text{dd}}}{64dqB(\beta+\lambda)^2}) \\
&\leq \rho_{\text{dd}} \qquad (\text{Since } K_{\mathcal{A}} \geq \tfrac{2B}{\lambda\eta}\log\left(\tfrac{q\log(B)}{\rho_{\text{dd}}}\right))
\end{aligned}$$

Induction step: Suppose $\mathrm{R}_q\left(\hat{P}_{i-1}\middle\|\pi(D_{i-1})\right) \leq \rho_{\text{dd}}$. Again, from boundedness of $\boldsymbol{\ell}(\theta;\mathbf{x})$, we have $|\mathcal{L}_{D_{i-1}}(\theta) - \mathcal{L}_{D_i}(\theta)| \leq \frac{\rho\sigma^2\log B}{2n}$ for all $\theta \in \mathbb{R}^d$. Therefore, from Lemma 23 we have for all $q > 1$ that

$$\mathrm{R}_q\left(\pi(D_{i-1})\|\pi(D_i)\right) \leq \frac{\rho\log(B)}{n}. \tag{B.66}$$

So from the weak triangle inequality Theorem 8 of Rényi divergence,

$$\mathrm{R}_q\left(\hat{P}_{i-1}\middle\|\pi(D_i)\right) \leq \mathrm{R}_q\left(\hat{P}_{i-1}\middle\|\pi(D_{i-1})\right) + \mathrm{R}_\infty\left(\pi(D_{i-1})\|\pi(D_i)\right) \leq \rho_{\text{dd}} + \frac{\rho\log(B)}{n}. \tag{B.67}$$

Note that $K_{\bar{\mathcal{A}}} \geq K_{\mathcal{A}} - \frac{2B}{\lambda\eta}\log\left(\frac{\log(B)}{2\left(\rho_{\text{dd}}+\frac{\rho}{n}\log(B)\right)}\right) \geq \frac{2B}{\lambda\eta}\log\left(\frac{2q\left(\rho_{\text{dd}}+\frac{\rho}{n}\log(B)\right)}{\rho_{\text{dd}}}\right)$. Since $\hat{\Theta}_i = \bar{\mathcal{A}}_{\text{Noisy-GD}}(D_{i-1}, u_i, \hat{\Theta}_{i-1}) = \text{Noisy-GD}(D_i, \hat{\Theta}_{i-1}, K_{\bar{\mathcal{A}}})$, convergence Theorem 25 gives

$$\begin{aligned}
\mathrm{R}_q\left(\hat{P}_i\middle\|\pi(D_i)\right) &\leq q\exp\left(-\frac{\lambda\eta K_{\bar{\mathcal{A}}}}{2B}\right)\mathrm{R}_q\left(\hat{P}_{i-1}\middle\|\pi(D_i)\right) + \frac{32d\eta qB(\beta+\lambda)^2}{\lambda} \\
&\leq q\exp\left(-\frac{\lambda\eta K_{\bar{\mathcal{A}}}}{2B}\right)\left(\rho_{\text{dd}} + \frac{\rho\log(B)}{n}\right) + \frac{\rho_{\text{dd}}}{2} \\
&\qquad (\text{From Equation B.67 and constraint } \eta \leq \tfrac{\lambda\rho_{\text{dd}}}{64dqB(\beta+\lambda)^2}) \\
&\leq \rho_{\text{dd}}. \qquad (\text{Since } K_{\bar{\mathcal{A}}} \geq \tfrac{2B}{\lambda\eta}\log\left(\tfrac{2q\left(\rho_{\text{dd}}+\frac{\rho}{n}\log(B)\right)}{\rho_{\text{dd}}}\right))
\end{aligned}$$

Hence, by induction, $\mathrm{R}_q\left(\hat{P}_i\middle\|\pi(D_i)\right) \leq \rho_{\text{dd}}$ holds for all $i \geq 0$.

(3.) Accuracy. Let $\theta^*_{D_i} = \arg\min_{\theta\in\mathbb{R}^d} \mathcal{L}_{D_i}(\theta)$, and $\bar{\Theta}_i \sim \pi(D_i)$. We decompose the excess empirical risk of Noisy-GD as follows:

$$\text{err}(\hat{\Theta}_i; D_i) = \mathbb{E}\left[\mathcal{L}_{D_i}(\hat{\Theta}_i) - \mathcal{L}_{D_i}(\bar{\Theta}_i)\right] + \mathbb{E}\left[\mathcal{L}_{D_i}(\bar{\Theta}_i) - \mathcal{L}_{D_i}(\theta^*_{D_i})\right]. \tag{B.68}$$

The second term is the suboptimality of Gibbs distribution and by Theorem 26, it is bounded as

$$\mathbb{E}\left[\mathcal{L}_{D_i}(\bar{\Theta}_i) - \mathcal{L}_{D_i}(\theta^*_{D_i})\right] \leq \frac{d\sigma^2}{2}\left(\log\frac{\beta+\lambda}{\lambda} + \sqrt{B}\right). \tag{B.69}$$

From $(\lambda+\beta)$-smoothness of $\mathcal{L}_{D_i}$, for any coupling Π of $\hat{\Theta}_i$ and $\bar{\Theta}_i$, the first term satisfies

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{D_i}(\hat{\Theta}_i) - \mathcal{L}_{D_i}(\bar{\Theta}_i)\right] &\leq \mathop{\mathbb{E}}_{\Pi}\left[\left\langle \nabla\mathcal{L}_{D_i}(\bar{\Theta}_i), \hat{\Theta}_i - \bar{\Theta}_i\right\rangle + \frac{\lambda+\beta}{2}\left\|\hat{\Theta}_i - \Theta_i\right\|_2^2\right] \\
&= \mathop{\mathbb{E}}_{\Pi}\left[\left\langle \sum_{\mathbf{x}\in D_i}\nabla\boldsymbol{\ell}(\bar{\Theta}_i;\mathbf{x}) + \lambda\bar{\Theta}_i, \hat{\Theta}_i - \bar{\Theta}_i\right\rangle + \frac{\lambda+\beta}{2}\left\|\hat{\Theta}_i - \Theta_i\right\|_2^2\right]. \\
&\qquad \text{(From } L\text{-Lipschitzness of } \boldsymbol{\ell}(\theta;\mathbf{x}) \text{ and Jensen's inequality)} \\
&\leq L\sqrt{\mathop{\mathbb{E}}_{\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right]} + \lambda\mathop{\mathbb{E}}_{\Pi}\left[\left\langle\bar{\Theta}_i, \bar{\Theta}_i - \hat{\Theta}_i\right\rangle\right] + \frac{\lambda+\beta}{2}\mathop{\mathbb{E}}_{\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right] \\
&\qquad \text{(From Young's inequality Equation 2.25)} \\
&\leq L\sqrt{\mathop{\mathbb{E}}_{\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right]} + \frac{\lambda}{2}\mathop{\mathbb{E}}_{\bar{\Theta}_i\sim\pi(D_i)}\left[\left\|\bar{\Theta}_i\right\|_2^2\right] + \frac{2\lambda+\beta}{2}\mathop{\mathbb{E}}_{\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right].
\end{aligned}$$

Recall that the distribution $\pi(D)$ satisfies LSI(λ/B) inequality. On choosing the coupling Π to be the infimum, we get the following bound on Wasserstein's distance from Lemma 5.

$$\inf_{\Pi}\sqrt{\mathop{\mathbb{E}}_{\hat{\Theta}_i,\bar{\Theta}_i\sim\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right]} = \mathrm{W}_2\left(\hat{\Theta}_i, \bar{\Theta}_i\right) \leq \sqrt{\frac{2B\sigma^2}{\lambda}\mathrm{KL}\left(\hat{P}_i\middle\|\pi(D_i)\right)} \leq \sqrt{\frac{2\rho_{\mathrm{dd}}B\sigma^2}{\lambda}}. \tag{B.70}$$

The last inequality above follows from monotonicity of Rényi divergence in q and the fact that $\lim_{q\to1}\mathrm{R}_q\left(P\|Q\right) = \mathrm{KL}\left(P\|Q\right)$.

Since $\pi(D_i)$ is the Gibbs distribution with density proportional to $e^{-\mathcal{L}_{D_i}/\sigma^2}$, we have that

$$\mathop{\mathbb{E}}_{\bar{\Theta}_i\sim\pi(D_i)}\left[\left\|\bar{\Theta}_i\right\|_2^2\right] = \frac{1}{\Lambda_{D_i}}\int\|\theta\|_2^2\, e^{-\mathcal{L}_{D_i}(\theta)/\sigma^2}\mathrm{d}\theta \quad \text{where } \Lambda_{D_i} = \int e^{-\mathcal{L}_{D_i}(\theta)/\sigma^2}\mathrm{d}\theta. \tag{B.71}$$

From $\frac{\sigma^2\log B}{4}$-boundedness of $\boldsymbol{\ell}(\theta;\mathbf{x})$, note that we have for every $\theta\in\mathbb{R}^d$ that

$$|\mathcal{L}_{D_i}(\theta) - \mathbf{r}(\theta)| \leq \frac{\sigma^2\log B}{4}. \tag{B.72}$$

Therefore,

$$\Lambda_{D_i} = \int e^{-\mathcal{L}_{D_i}(\theta)/\sigma^2}\mathrm{d}\theta \geq \frac{1}{\sqrt[4]{B}}\int e^{-\mathbf{r}(\theta)/\sigma^2}\mathrm{d}\theta, \tag{B.73}$$

and hence,

$$\begin{aligned}\mathop{\mathbb{E}}_{\bar{\Theta}_i\sim\pi(D_i)}\left[\left\|\bar{\Theta}_i\right\|_2^2\right] &\leq \frac{\sqrt[4]{B}}{\int e^{-\mathbf{r}(\theta)/\sigma^2}\mathrm{d}\theta}\times\int\|\theta\|_2^2\, e^{-\mathcal{L}_{D_i}(\theta)/\sigma^2}\mathrm{d}\theta\\ &\leq \sqrt{B}\times\frac{\int\|\theta\|_2^2\, e^{-\mathbf{r}(\theta)/\sigma^2}}{\int e^{-\mathbf{r}(\theta)/\sigma^2}}\\ &= \sqrt{B}\mathop{\mathbb{E}}_{W\sim\mathcal{N}\left(0,\frac{\sigma^2}{\lambda}I_d\right)}\left[\|W\|_2^2\right]\\ &= \frac{\sqrt{B}\sigma^2 d}{\lambda}.\end{aligned}$$

Therefore, on combining all the bounds we get

$$\mathrm{err}(\hat{\Theta};D)\leq L\sigma\sqrt{\frac{2\rho_{\mathrm{dd}}B}{\lambda}}+\frac{\rho_{\mathrm{dd}}B\sigma^2(2\lambda+\beta)}{\lambda}+\frac{d\sigma^2}{2}\left(\log\frac{\beta+\lambda}{\lambda}+2\sqrt{B}\right)=O\left(\sigma\sqrt{\rho_{\mathrm{dd}}}+d\sigma^2\right). \tag{B.74}$$

Note that if the constraints on $K_{\mathcal{A}}$ and $K_{\bar{\mathcal{A}}}$ in Equation B.63 and on σ^2 in Equation B.62 are equalities instead, we have

$$\sigma^2=\frac{2qBL^2}{\lambda\rho_{\mathrm{dp}}n^2}\log\left(\frac{q\log(B)}{\rho_{\mathrm{dd}}}\right)=\tilde{O}\left(\frac{q}{\rho_{\mathrm{dp}}n^2}\right), \tag{B.75}$$

where $\tilde{O}(\cdot)$ hides logarithmic factors. Therefore, the excess empirical risk has an order

$$\mathrm{err}(\hat{\Theta};D)=\tilde{O}\left(\frac{1}{n}\sqrt{\frac{q\rho_{\mathrm{dd}}}{\rho_{\mathrm{dp}}}}+\frac{dq}{\rho_{\mathrm{dp}}n^2}\right). \tag{B.76}$$

□

B.4 Effect of Gradient Clipping

First order optimization methods on a continuously differentiable loss function $\boldsymbol{\ell}(\theta;\mathbf{x})$ over a database $D\in\mathcal{X}^n$ with gradient clipping $\mathrm{Clip}_L(\mathbf{v})=\mathbf{v}/\max\left(1,\frac{\|\mathbf{v}\|_2}{L}\right)$ is equivalent to optimizing

$$\mathcal{L}_D(\theta)=\frac{1}{|D|}\sum_{\mathbf{x}\in D}\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})+\mathbf{r}(\theta), \tag{B.77}$$

where $\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ is a surrogate loss function that satisfies $\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})=\mathrm{Clip}_L(\nabla\boldsymbol{\ell}(\theta;\mathbf{x}))$. This surrogate loss function inherits convexity, boundedness, and smoothness properties of $\boldsymbol{\ell}(\theta;\mathbf{x})$, as shown below.

Lemma 26 (Gradient clipping retains convexity)**.** *If $\boldsymbol{\ell}(\theta;\mathbf{x})$ is a twice continuously differentiable convex function for every $\mathbf{x}\in\mathbb{R}^d$, then surrogate loss $\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ resulting from gradient clipping is also convex for every $\mathbf{x}\in\mathbb{R}^d$.*

Proof. Note that the clip operation $\mathrm{Clip}_L(\mathbf{v})$ is a closed-form solution of the orthogonal projection onto a closed ball of radius L and centered around origin, i.e.

$$\mathrm{Clip}_L(\mathbf{v}) = \operatorname*{arg\,min}_{\|\mathbf{v}'\|_2\le L} \|\mathbf{v}-\mathbf{v}'\|_2 . \tag{B.78}$$

By properties of orthogonal projections on closed convex sets, for every $\mathbf{v},\mathbf{v}'\in\mathbb{R}^d$,

$$\langle \mathbf{v}' - \mathrm{Clip}_L(\mathbf{v}), \mathbf{v} - \mathrm{Clip}_L(\mathbf{v})\rangle \le 0 \quad \text{if and only if } \|\mathbf{v}'\|_2 \le L. \tag{B.79}$$

Therefore, for any $\theta\in\mathbb{R}^d$, and $\mathbf{x}\in\mathcal{X}$, we have

$$\left\langle \nabla\bar{\boldsymbol{\ell}}(\theta+h\hat{\mathbf{v}};\mathbf{x}) - \nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}), \nabla\boldsymbol{\ell}(\theta;\mathbf{x}) - \nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})\right\rangle \le 0, \tag{B.80}$$

$$\left\langle \nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}) - \nabla\bar{\boldsymbol{\ell}}(\theta+h\hat{\mathbf{v}};\mathbf{x}), \nabla\boldsymbol{\ell}(\theta+h\hat{\mathbf{v}};\mathbf{x}) - \nabla\bar{\boldsymbol{\ell}}(\theta+h\hat{\mathbf{v}};\mathbf{x})\right\rangle \le 0, \tag{B.81}$$

for all unit vectors $\hat{\mathbf{v}}\in\mathbb{R}^d$ and magnitude $h>0$. For the directional derivative of vector field $\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ along $\hat{\mathbf{v}}$, defined as $\nabla_{\hat{\mathbf{v}}}\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}) = \lim_{h\to 0^+}\frac{\nabla\bar{\boldsymbol{\ell}}(\theta+h\hat{\mathbf{v}};\mathbf{x})-\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})}{h}$, the above two inequalities imply

$$\left\langle \nabla_{\hat{\mathbf{v}}}\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}), \nabla\boldsymbol{\ell}(\theta;\mathbf{x}) - \nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})\right\rangle = 0, \tag{B.82}$$

for all $\hat{\mathbf{v}}$. Therefore, when $\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}) \ne \nabla\boldsymbol{\ell}(\theta;\mathbf{x})$, we must have $\nabla^2\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}) = 0$. And, when $\nabla\boldsymbol{\ell}(\theta;\mathbf{x}) = \nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$, gradients aren't clipped, which implies the rate of change of $\boldsymbol{\ell}(\theta;\mathbf{x})$ along any direction $\hat{\mathbf{v}}$ is

$$\begin{aligned}
\nabla_{\hat{\mathbf{v}}}\cdot\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}) &= \lim_{h\to 0^+}\left\langle \frac{\nabla\bar{\boldsymbol{\ell}}(\theta+h\hat{\mathbf{v}};\mathbf{x}) - \nabla\boldsymbol{\ell}(\theta;\mathbf{x})}{h}, \hat{\mathbf{v}}\right\rangle \\
&= \begin{cases} \hat{\mathbf{v}}^\top\nabla^2\boldsymbol{\ell}(\theta;\mathbf{x})\hat{\mathbf{v}} & \text{if } \exists h>0 \text{ s.t. } \nabla\bar{\boldsymbol{\ell}}(\theta+h\hat{\mathbf{v}};\mathbf{x}) = \nabla\boldsymbol{\ell}(\theta+h\hat{\mathbf{v}};\mathbf{x}) \\ 0 & \text{otherwise} \end{cases} \\
&\ge 0.
\end{aligned}$$

□

Lemma 27 (Gradient clipping retains boundedness)**.** *If $\boldsymbol{\ell}(\theta;\mathbf{x})$ is a continuously differentiable and B-bounded function for every $\mathbf{x}\in\mathcal{X}$, then a surrogate loss $\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ resulting from gradient clipping is also B-bounded.*

Proof. Since $\boldsymbol{\ell}(\theta;\mathbf{x})$ is continuously differentiable, its B-boundedness implies path integral of $\nabla\boldsymbol{\ell}(\theta;\mathbf{x})$ along any curve between $\theta,\theta' \in \mathbb{R}^d$ is less than $2B$. Since $\mathrm{Clip}_L(\cdot)$ operation clips the gradient magnitude, the path integral of $\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ is also less than $2B$. That is, the maximum and minimum values that $\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ takes differ no more than $2B$. By adjusting the constant of path integral, we can always ensure $\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ takes values in range $[-B,B]$ without affecting first order optimization algorithms. □

Lemma 28 (Gradient clipping retains smoothness)**.** *If $\boldsymbol{\ell}(\theta;\mathbf{x})$ is a continuously differentiable and β-smooth function for every $\mathbf{x}\in\mathbb{R}^d$, then surrogate loss $\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ resulting from gradient clipping is also β-smooth for every $\mathbf{x}\in\mathbb{R}^d$.*

Proof. Note that the gradient clipping operation is equivalent to an orthogonal projection operation into a ball of radius L, i.e. $\mathrm{Clip}_L(\mathbf{v}) = \arg\min_{\mathbf{v}'}\{\|\mathbf{v}'-\mathbf{v}\|_2 : \mathbf{v}\in\mathbb{R}^d, \|\mathbf{v}'\|_2 \le L\}$. Since orthogonal projection onto a closed convex set is a 1-Lipschitz operation, for any $\theta,\theta'\in\mathbb{R}^d$,

$$\left\|\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}) - \nabla\bar{\boldsymbol{\ell}}(\theta';\mathbf{x})\right\|_2 \le \|\nabla\boldsymbol{\ell}(\theta;\mathbf{x}) - \nabla\boldsymbol{\ell}(\theta';\mathbf{x})\|_2 \le \beta\,\|\theta-\theta'\|_2\,. \tag{B.83}$$

□

Additionally, the surrogate loss $\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ is twice differentiable almost everywhere if $\boldsymbol{\ell}(\theta;\mathbf{x})$ is smooth, which follows from the following Rademacher's Theorem.

Theorem 40 (Rademacher's Theorem [73])**.** *If $f:\mathbb{R}^n\to\mathbb{R}^n$ is Lipschitz continuous, then f is differentiable almost everywhere in $\mathbb{R}^n$.*

All our results in Sections § 4.5.2 and § 4.5.3 rely on the above four properties on losses and therefore apply with gradient clipping instead of the Lipschitzness assumption.

Appendix C

Supplement for Chapter 5

C.1 Table of Properties of Laplace Transform

Table C.1: Properties of the Laplace Transform. Let $g(t)$ and $h(t)$ be two functions defined for $t \in \mathbb{R}$ and let $a, b \in \mathbb{R}$ be arbitrary constants.

Property	**Expression**	
Linearity :	$\mathcal{L}\{ag(t)+bh(t)\}(s) = a\mathcal{L}\{g(t)\}(s) + b\mathcal{L}\{h(t)\}(s)$ $\mathcal{B}\{ag(t)+bh(t)\}(s) = a\mathcal{B}\{g(t)\}(s) + b\mathcal{B}\{h(t)\}(s)$	(C.1)
Time-Shifting :	$\mathcal{L}\{g(t-a)\mathbb{1}\{t>a\}\}(s) = e^{-as}\mathcal{L}\{g(t)\}(s),$ for $a>0$ $\mathcal{B}\{g(t-a)\}(s) = e^{-as}\mathcal{B}\{g(t)\}(s),$ for $a \in \mathbb{R}$	(C.2)
Frequency-Shifting :	$\mathcal{L}\left\{e^{at}g(t)\right\}(s) = \mathcal{L}\{g(t)\}(s-a)$ $\mathcal{B}\left\{e^{at}g(t)\right\}(s) = \mathcal{B}\{g(t)\}(s-a)$	(C.3)
Time-Scaling :	$\mathcal{L}\{g(at)\}(s) = \frac{1}{a}\mathcal{L}\{g(t)\}\left(\frac{s}{a}\right)$ for $a>0$ $\mathcal{B}\{g(at)\}(s) = \frac{1}{\lvert a\rvert}\mathcal{B}\{g(t)\}\left(\frac{s}{a}\right)$ for $a \in \mathbb{R}$	(C.4)
Reversal :	$\mathcal{B}\{g(-t)\}(s) = \mathcal{B}\{g(t)\}(-s)$	(C.5)
Derivative :	$\mathcal{L}\{\dot{g}(t)\}(s) = s\mathcal{L}\{g(t)\}(s) - g(0^+)$ $\mathcal{B}\{\dot{g}(t)\}(s) = s\mathcal{B}\{g(t)\}(s)$	(C.6)
Integration :	$\mathcal{L}\left\{\int_0^t g(t)dt\right\}(s) = \frac{1}{s}\mathcal{L}\{g(t)\}(s),$ for $\mathfrak{Re}(s)>0$ $\mathcal{B}\left\{\int_{-\infty}^t g(t)dt\right\}(s) = \frac{1}{s}\mathcal{B}\{g(t)\}(s),$ for $\mathfrak{Re}(s)>0$	(C.7)
Convolution :	$\mathcal{L}\left\{\int_0^t g(\tau)h(t-\tau)d\tau\right\}(s) = \mathcal{L}\{g(t)\}(s)\cdot\mathcal{L}\{h(t)\}(s)$ $\mathcal{B}\left\{\int_{-\infty}^{\infty} g(\tau)h(t-\tau)d\tau\right\}(s) = \mathcal{B}\{g(t)\}(s)\cdot\mathcal{B}\{h(t)\}(s)$	(C.8)

C.2 Deferred Proofs for § 5.3

Theorem 30. *For a random variable X, let $F_X(t) \stackrel{def}{=} \Pr[X \leq t]$ denote its* cumulative distribution function *and $f_X(t)$ denote its* generalized probability density function. *Let P and Q be probability distributions and $Z \sim \mathrm{PLD}(P\|Q)$ and $Z' \sim \mathrm{PLD}(Q\|P)$ denote their privacy loss random variables. If $Z \sim \mathrm{PLD}(P\|Q)$ and $Z' \sim \mathrm{PLD}(Q\|P)$, then for all $\varepsilon \in \mathbb{R}$,*

$$\begin{aligned}
\delta_{P|Q}(\varepsilon) &= \mathcal{L}\left\{1 - F_Z(t+\varepsilon)\right\}(1) && \text{(C.9)}\\
&= e^{\varepsilon} \cdot \mathcal{L}\left\{F_{Z'}(-t-\varepsilon)\right\}(-1) && \text{(C.10)}\\
&= e^{\varepsilon} \cdot \mathcal{L}\left\{f_{Z'}(-t-\varepsilon)\right\}(-1) - \mathcal{L}\left\{f_Z(t+\varepsilon)\right\}(1) && \text{(C.11)}\\
&= \mathcal{L}\left\{f_Z(t+\varepsilon)\right\}(0) - e^{\varepsilon} \cdot \mathcal{L}\left\{f_{Z'}(-t-\varepsilon)\right\}(0). && \text{(C.12)}
\end{aligned}$$

And, for all $q \in \mathrm{ROC}_{\mathcal{B}}\{f_{Z'}\}$ (or equivalently, $1-q \in \mathrm{ROC}_{\mathcal{B}}\{f_Z\}$),

$$e^{(q-1)\cdot \mathrm{R}_q(P\|Q)} = \mathrm{E}_q(P\|Q) = \mathcal{B}\{f_Z(t)\}(1-q) = \mathcal{B}\{f_{Z'}(t)\}(q). \tag{C.13}$$

Proof. Denote the set where privacy loss $L_{P|Q}(\theta)$ exceeds ε as

$$S^*_{>\varepsilon} = \{\theta \in \Omega : P(\theta) > e^{\varepsilon} Q(\theta)\}. \tag{C.14}$$

Note that for all $S \subset \Omega$, and all $\varepsilon \in \mathbb{R}$, we have

$$P(S) - e^{\varepsilon} Q(S) \leq P(S^*_{>\varepsilon}) - e^{\varepsilon} Q(S^*_{>\varepsilon}) = \delta_{P|Q}(\varepsilon), \tag{C.15}$$

because $S^*_{>\varepsilon}$ includes any and all points where $P(\theta) > e^{\varepsilon} Q(\theta)$. We can express the probabilities $P(S^*_{>\varepsilon})$ and $Q(S^*_{>\varepsilon})$ using the Laplace transform as follows:

$$P(S^*_{>\varepsilon}) = \int_{S^*_{>\varepsilon}} P(\theta)d\theta \tag{C.16}$$
$$= \int_{S^*_{>\varepsilon}} \left(\frac{P(\theta)}{Q(\theta)}\right) Q(\theta)d\theta \tag{C.17}$$
$$= \int_{S^*_{>\varepsilon}} e^{-L_{Q|P}(\theta)} Q(\theta)d\theta \tag{C.18}$$
$$= \int_{\varepsilon^+}^{\infty} e^t \int_{\{\theta\in\Omega: L_{Q|P}(\theta)=-t\}} Q(\theta)d\theta \tag{C.19}$$
$$= \int_{\varepsilon^+}^{\infty} e^t \int_{-t^-}^{-t^+} F_{Z'}(u)du \tag{C.20}$$
$$= \int_{\varepsilon^+}^{\infty} e^t f_{Z'}(-t)dt \tag{C.21}$$
$$= e^{\varepsilon} \int_{0^+}^{\infty} e^{t'} f_{Z'}(-t'-\varepsilon)dt' \tag{C.22}$$
$$= e^{\varepsilon}\mathcal{L}\left\{f_{Z'}(-t-\varepsilon)\right\}(-1). \tag{C.23}$$

$$Q(S^*_{>\varepsilon}) = \int_{S^*_{>\varepsilon}} Q(\theta)d\theta \tag{C.24}$$
$$= \int_{S^*_{>\varepsilon}} \left(\frac{Q(\theta)}{P(\theta)}\right) P(\theta)d\theta \tag{C.25}$$
$$= \int_{S^*_{>\varepsilon}} e^{-L_{P|Q}(\theta)} P(\theta)d\theta \tag{C.26}$$
$$= \int_{\varepsilon^+}^{\infty} e^{-t} \int_{\{\theta\in\Omega: L_{P|Q}(\theta)=t\}} P(\theta)d\theta \tag{C.27}$$
$$= \int_{\varepsilon^+}^{\infty} e^{-t} \int_{t^-}^{t^+} F_Z(u)du \tag{C.28}$$
$$= \int_{\varepsilon^+}^{\infty} e^{-t} f_Z(t)dt \tag{C.29}$$
$$= e^{-\varepsilon} \int_{0^+}^{\infty} e^{-t'} f_Z(t'+\varepsilon)dt \tag{C.30}$$
$$= e^{-\varepsilon}\mathcal{L}\left\{f_Z(t+\varepsilon)\right\}(1). \tag{C.31}$$

Combining the two, we get equation Equation C.11:

$$\delta_{P|Q}(\varepsilon) = e^{\varepsilon}\cdot\mathcal{L}\left\{f_{Z'}(-t-\varepsilon)\right\}(-1) - \mathcal{L}\left\{f_Z(t+\varepsilon)\right\}(1). \tag{C.32}$$

Alternatively, we can express the profile directly as:

$$P(S^*_{>\varepsilon}) - e^{\varepsilon}Q(S^*_{>\varepsilon}) = \int_{S^*_{>\varepsilon}} (1 - e^{\varepsilon}\frac{Q(\theta)}{P(\theta)})P(\theta)d\theta \tag{C.33}$$
$$= \int_{S^*_{>\varepsilon}} (1 - e^{\varepsilon - L_{P|Q}(\theta)})P(\theta)d\theta \tag{C.34}$$
$$= \int_{\varepsilon^+}^{\infty} (1 - e^{\varepsilon - t}) f_Z(t)dt \tag{C.35}$$
$$= \int_{\varepsilon^+}^{\infty} f_Z(t)dt - \int_{0^+}^{\infty} e^{-t'} f_Z(t'+\varepsilon)dt' \qquad (\text{Change } t = t'+\varepsilon)$$
$$= 1 - F_Z(\varepsilon^+) - \mathcal{L}\left\{f_Z(t+\varepsilon)\right\}(1) \tag{C.36}$$
$$\overset{\textit{Equation C.6}}{=} \mathcal{L}\left\{1 - F_Z(t+\varepsilon)\right\}(1). \tag{C.37}$$

Similarly, we can express it in terms of Z' as

$$P(S^*_{>\varepsilon}) - e^{\varepsilon}Q(S^*_{>\varepsilon}) = \int_{S^*_{>\varepsilon}} (\frac{P(\theta)}{Q(\theta)} - e^{\varepsilon})Q(\theta)d\theta \tag{C.38}$$

$$= \int_{S^*_{>\varepsilon}} (e^{L_{P|Q}(\theta)} - e^{\varepsilon})Q(\theta)d\theta \tag{C.39}$$

$$= \int_{\varepsilon^+}^{\infty} (e^t - e^{\varepsilon}) f_{Z'}(-t)dt \tag{C.40}$$

$$= e^{\varepsilon}\left(\int_{0^+}^{\infty} e^{t'} f_{Z'}(-t'-\varepsilon)dt' - \int_{\varepsilon^+}^{\infty} f_{Z'}(-t)dt\right)$$

$$\text{(Change } t = t' + \varepsilon\text{)}$$

$$= e^{\varepsilon}\left(\mathcal{L}\left\{f_{Z'}(-t-\varepsilon)\right\}(-1) - F_{Z'}(-\varepsilon^-)\right) \tag{C.41}$$

$$\overset{\textit{Equation C.6}}{=} e^{\varepsilon}\mathcal{L}\left\{F_{Z'}(-t-\varepsilon)\right\}(-1). \tag{C.42}$$

For showing Equation C.12, we apply the derivative property of Laplace transform to Equation C.9 and Equation C.10 to get

$$\delta_{P|Q}(\varepsilon) = \mathcal{L}\left\{1 - F_Z(t+\varepsilon)\right\}(1) \overset{\textit{Equation C.6}}{=} -\mathcal{L}\left\{f_Z(t+\varepsilon)\right\}(1) + 1 - F_Z(\varepsilon^+), \quad \text{and} \tag{C.43}$$

$$\delta_{P|Q}(\varepsilon) = e^{\varepsilon}\cdot\mathcal{L}\left\{F_{Z'}(-t-\varepsilon)\right\}(-1) \overset{\textit{Equation C.6}}{=} e^{\varepsilon}\cdot\left(\mathcal{L}\left\{f_{Z'}(-t-\varepsilon)\right\}(-1) - F_{Z'}(-\varepsilon^-)\right). \tag{C.44}$$

Adding the above two equations and subtracting Equation C.11 from it, we get

$$\delta_{P|Q}(\varepsilon) = 1 - F_Z(\varepsilon^+) - e^{\varepsilon}F_{Z'}(-\varepsilon^-) \tag{C.45}$$

$$= \Pr[Z > \varepsilon] - e^{\varepsilon}\Pr[Z' < -\varepsilon] \tag{C.46}$$

$$= \int_{0^+}^{\infty} e^{0\cdot t}\cdot f_Z(t+\varepsilon)dt - e^{\varepsilon}\cdot\int_{0^+}^{\infty} e^{0\cdot t} f_{Z'}(-t-\varepsilon)dt \tag{C.47}$$

$$= \mathcal{L}\left\{f_Z(t+\varepsilon)\right\}(0) - e^{\varepsilon}\cdot\mathcal{L}\left\{f_{Z'}(-t-\varepsilon)\right\}(0). \tag{C.48}$$

For the last part, recall from definition that Rényi divergence $\mathrm{R}_q(P\|Q) = \frac{1}{q-1}\log \mathrm{E}_q(P\|Q)$, for which we show the following two equivalences:

$$\mathrm{E}_q\left(P\|Q\right) = \int_\Omega \left(\frac{P(\theta)}{Q(\theta)}\right)^{q-1} P(\theta)d\theta \tag{C.49}$$

$$= \int_\Omega e^{(q-1)L_{P|Q}(\theta)} P(\theta)d\theta \tag{C.50}$$

$$= \int_{-\infty}^{\infty} e^{(q-1)t} f_Z(t)dt \tag{C.51}$$

$$= \mathcal{B}\left\{f_Z(t)\right\}(1-q). \tag{C.52}$$

$$\mathrm{E}_q\left(P\|Q\right) = \int_\Omega \left(\frac{P(\theta)}{Q(\theta)}\right)^{q} Q(\theta)d\theta \tag{C.53}$$

$$= \int_\Omega e^{-qL_{Q|P}(\theta)} Q(\theta)d\theta \tag{C.54}$$

$$= \int_{-\infty}^{\infty} e^{-qt} f_{Z'}(t)dt \tag{C.55}$$

$$= \mathcal{B}\left\{f_{Z'}(t)\right\}(q). \tag{C.56}$$

□

Theorem 32 (Privacy profile of randomized response)**.** *Fix $\varepsilon > 0$ and $\delta \in [0,1]$. Let $\mathcal{A}_{\mathrm{RR}} : \{0,1\} \to \{0,1\} \times \{\perp, \top\}$ be the randomized response mechanism, which has the following output probabilities.*

$$\mathcal{A}_{\mathrm{RR}}(0) = \begin{cases} (0,\perp) & \text{with probability } \delta, \\ (0,\top) & \text{with probability } \frac{(1-\delta)e^\varepsilon}{e^\varepsilon+1}, \\ (1,\top) & \text{with probability } \frac{(1-\delta)}{e^\varepsilon+1}, \\ (1,\perp) & \text{with probability } 0, \end{cases} \mathcal{A}_{\mathrm{RR}}(1) = \begin{cases} (0,\perp) & \text{with probability } 0, \\ (0,\top) & \text{with probability } \frac{(1-\delta)}{e^\varepsilon+1}, \\ (1,\top) & \text{with probability } \frac{(1-\delta)e^\varepsilon}{e^\varepsilon+1}, \\ (1,\perp) & \text{with probability } \delta. \end{cases} \tag{C.57}$$

For $P = \mathcal{A}_{\mathrm{RR}}(0)$ and $Q = \mathcal{A}_{\mathrm{RR}}(1)$, the privacy profiles are

$$\forall t \in \mathbb{R} \ : \ \delta_{P|Q}(t) = \delta_{Q|P}(t) = \begin{cases} \delta & \text{if } \varepsilon < t, \\ 1 - \frac{(e^t+1)}{e^\varepsilon+1}(1-\delta) & \text{if } -\varepsilon < t \le \varepsilon, \\ 1 - e^t(1-\delta) & \text{if } t \le -\varepsilon. \end{cases} \tag{C.58}$$

Proof. Let $S_1 = \{(0,\perp)\}$, $S_2 = S_1 \cup \{(1,\perp)\}$, and $S_3 = S_2 \cup \{(1,\top)\}$. From Equa-

tion 5.27,

$$\delta_{P|Q}(t) = \Pr_{Z \leftarrow \mathrm{PLD}(P\|Q)}[Z > t] - e^{\varepsilon} \cdot \Pr_{Z' \leftarrow \mathrm{PLD}(Q\|P)}[Z' < -t] \tag{C.59}$$

$$= \Pr_{P}[\log \frac{P(\Theta)}{Q(\Theta)} > t] - e^{\varepsilon} \cdot \Pr_{Q}[\log \frac{Q(\Theta)}{P(\Theta)} < -t] \tag{C.60}$$

$$= \Pr_{P}[P(\Theta) > e^t \cdot Q(\Theta)] - e^{\varepsilon} \cdot \Pr_{Q}[P(\Theta) < e^t \cdot Q(\Theta)] \tag{C.61}$$

$$= \begin{cases} P(S_1) - e^t \cdot Q(S_1) & \text{if } \varepsilon < t, \\ P(S_2) - e^t \cdot Q(S_2) & \text{if } -\varepsilon < t \leq \varepsilon, \\ P(S_3) - e^t \cdot Q(S_3) & \text{otherwise} \end{cases} \tag{C.62}$$

$$= \begin{cases} \delta & \text{if } \varepsilon < t, \\ \delta + \frac{1-\delta}{e^{\varepsilon}+1} \cdot (e^{\varepsilon} - e^t) & \text{if } -\varepsilon < t \leq \varepsilon, \\ 1 - e^t \cdot (1-\delta) & \text{otherwise} \end{cases} \tag{C.63}$$

The same holds for $\delta_{Q|P}$ due to symmetry of Equation C.57. □

Theorem 33 (Rényi DP of $(\varepsilon, 0)$-Randomized Response)**.** *For any $\varepsilon > 0$ and $\delta = 0$, the output distributions of randomized response mechanism in Theorem 32 exhibit the following Rényi divergence for all orders $q \in \mathbb{C}$ such that $\mathfrak{Re}(q) \notin \{0, 1\}$:*

$$\mathrm{R}_q\left(P\|Q\right) = \frac{1}{q-1} \log\left(\frac{e^{\varepsilon}}{1+e^{\varepsilon}} e^{-q\varepsilon} + \frac{1}{1+e^{\varepsilon}} e^{q\varepsilon}\right). \tag{C.64}$$

Proof. From Theorem 32, when $\delta = 0$, the privacy profile of randomized response algorithm's output-distributions P and Q is

$$\delta_{P|Q}(t) = \begin{cases} 0 & \text{if } \varepsilon < t, \\ \frac{e^{\varepsilon}-e^t}{1+e^{\varepsilon}} & \text{if } -\varepsilon < t \leq \varepsilon, \\ 1 - e^t & \text{otherwise.} \end{cases} \tag{C.65}$$

From the equivalence Equation 5.24 of Theorem 31,

$$\frac{e^{(q-1)R_q(P\|Q)}}{q(q-1)} = \mathcal{B}\left\{\delta_{P|Q}(t)\right\}(1-q) \tag{C.66}$$

$$= \int_{-\infty}^{\infty} e^{(q-1)t}\delta_{P|Q}(t)\mathrm{d}t \tag{C.67}$$

$$= \int_{-\infty}^{-\varepsilon} e^{(q-1)t}\cdot(1-e^t)\mathrm{d}t + \int_{-\varepsilon}^{\varepsilon} e^{(q-1)t}\cdot\frac{e^\varepsilon - e^t}{1+e^\varepsilon}\mathrm{d}t \tag{C.68}$$

$$= \left[\frac{e^{(q-1)t}}{q-1} - \frac{e^{qt}}{q}\right]_{-\infty}^{-\varepsilon} + \frac{1}{1+e^\varepsilon}\left[\frac{e^\varepsilon\cdot e^{(q-1)t}}{q-1} - \frac{e^{qt}}{q}\right]_{-\varepsilon}^{\varepsilon} \tag{C.69}$$

$$= \left(\frac{e^{-(q-1)\varepsilon}}{q-1} - \frac{e^{-q\varepsilon}}{q}\right) + \frac{1}{1+e^\varepsilon}\left[\left(\frac{e^{q\varepsilon}}{q-1} - \frac{e^{q\varepsilon}}{q}\right) - \left(\frac{e^\varepsilon\cdot e^{-(q-1)\varepsilon}}{q-1} - \frac{e^{-q\varepsilon}}{q}\right)\right] \tag{C.70}$$

$$= \frac{e^{-(q-1)\varepsilon}}{q-1}\cdot\left(1-\frac{e^\varepsilon}{1+e^\varepsilon}\right) - \frac{e^{-q\varepsilon}}{q}\cdot\left(1-\frac{1}{1+e^\varepsilon}\right) + \frac{e^{q\varepsilon}}{1+e^\varepsilon}\cdot\left(\frac{1}{q-1}-\frac{1}{q}\right) \tag{C.71}$$

$$= \frac{e^{-(q-1)\varepsilon}}{1+e^\varepsilon}\cdot\left(\frac{1}{q-1}-\frac{1}{q}\right) + \frac{e^{q\varepsilon}}{1+e^\varepsilon}\cdot\left(\frac{1}{q-1}-\frac{1}{q}\right) \tag{C.72}$$

$$= \frac{e^\varepsilon\cdot e^{-q\varepsilon} + e^{q\varepsilon}}{1+e^\varepsilon}\cdot\frac{1}{q(q-1)}. \tag{C.73}$$

Therefore, for any $q \in \mathbb{R}\setminus\{0,1\}$ we can cancel $q(q-1)$ from the denominator in both sides, which proves the theorem statement for real orders. From dominated convergence theorem the theorem statement holds for complex orders as well on corresponding real orders (cf. § 2.3.2). □

C.3 Deferred Proofs for § 5.4

Lemma 29 ([99, Corollary 24]). *Let P and Q be probability distributions over Ω. Fix $\varepsilon \geq 0$ and $\delta \in [0,1]$. Suppose P, Q satisfy (ϵ,δ)-differential privacy. Then there exists distributions A, B, P', Q' over Ω such that*

$$P = (1-\delta)\frac{e^\varepsilon}{e^\varepsilon+1}A + (1-\delta)\frac{1}{e^\varepsilon+1}B + \delta P', \tag{C.74}$$

$$Q = (1-\delta)\frac{e^\varepsilon}{e^\varepsilon+1}B + (1-\delta)\frac{1}{e^\varepsilon+1}A + \delta Q'. \tag{C.75}$$

Theorem 36 (Dominating Privacy Profile under (ε,δ)-DP). *Fix $\varepsilon \geq 0$ and $\delta \in [0,1]$. Suppose distributions P and Q over Ω satisfy (ε,δ)-differential privacy. Then,*

$$\forall t \in \mathbb{R} \;:\; \delta_{P|Q}(t) \leq \delta_{\mathrm{RR}}(t) \quad \textit{and} \quad \delta_{Q|P}(t) \leq \delta_{\mathrm{RR}}(t), \tag{C.76}$$

where $\delta_{\mathrm{RR}}(t)$ is the privacy profile of the randomized response mechanism $\mathcal{A}_{\mathrm{RR}}^{\varepsilon,\delta}$.

Proof. Since the output distributions P and Q are (ε, δ)-differentially private, Lemma 29 from Steinke [99] tells us that we can simulate these two distributions as post-processing of the randomized response mechanism $\mathcal{A}_{\mathrm{RR}}^{\varepsilon,\delta}$. To see this, imagine that $P = \mathcal{A}(D)$ and $Q = \mathcal{A}(D')$ are the output distributions of some mechanism $\mathcal{A}$. Define another mechanism $\mathcal{G} : \{0,1\} \times \{\bot, \top\} \to \Omega$ with output distribution:

$$\mathcal{G}(u) = \begin{cases} P' & \text{if } u = (0, \bot), \\ A & \text{if } u = (0, \top), \\ B & \text{if } u = (1, \top), \\ Q' & \text{if } u = (1, \bot). \end{cases} \tag{C.77}$$

From Lemma 29, see that $\mathcal{G}(\mathcal{A}_{\mathrm{RR}}^{\varepsilon,\delta}(0)) = \mathcal{A}(D)$ and $\mathcal{G}(\mathcal{A}_{\mathrm{RR}}^{\varepsilon,\delta}(1)) = \mathcal{A}(D')$. Therefore, by expressing P and Q in terms of distributions P', A, B, Q', we can conclude that for all $t \in \mathbb{R}$,

$$\delta_{P|Q}(t) = \sup_{S \subset \Omega} P(S) - e^t \cdot Q(S) \tag{C.78}$$

$$= \sup_{S \subset \Omega} \left(\delta \cdot P'(S) + \frac{(1-\delta)(e^\varepsilon - e^t)}{e^\varepsilon + 1} \cdot A(S) + \frac{(1-\delta)(1-e^{\varepsilon+t})}{e^\varepsilon + 1} \cdot B(S) - \delta e^t \cdot Q'(S) \right) \tag{C.79}$$

$$\leq \delta + \begin{cases} 0 & \text{if } \varepsilon < t, \\ \frac{(1-\delta)(e^\varepsilon - e^t)}{e^\varepsilon + 1} & \text{if } -\varepsilon < t \leq \varepsilon, \\ (1-\delta)(1-e^t) & \text{if } t \leq -\epsilon, \end{cases} \tag{C.80}$$

$$= \begin{cases} \delta & \text{if } \varepsilon < t, \\ 1 - \frac{(e^t+1)(1-\delta)}{e^\varepsilon + 1} & \text{if } -\varepsilon < t \leq \varepsilon, \\ 1 - e^t(1-\delta) & \text{if } t \leq -\varepsilon. \end{cases} \tag{C.81}$$

Note that the expression on the right is the privacy profile of the randomized response mechanism $\mathcal{A}_{\mathrm{RR}}^{\varepsilon,\delta}$. An identical bound follows for $\delta_{Q|P}(t)$ as well, with the switched roles: $P' \leftrightarrow Q'$ and $A \leftrightarrow B$. □

Theorem 37 (Tight Composition for (ε, δ)-DP)**.** *For any $\varepsilon_i \geq 0$, $\delta_i \in [0,1]$ for $i \in \{1, \cdots, k\}$, the k-fold composition of $(\varepsilon_i, \delta_i)$-differentially private mechanisms*

satisfies $(\varepsilon, \delta^{\otimes k}(\varepsilon))$-DP for all ε, defined recursively as

$$\forall t \in \mathbb{R} \ : \ \delta^{\otimes l}(t) = \delta_l + \frac{(1-\delta_l)}{e^{\varepsilon_l}+1}\left[e^{\varepsilon_l} \cdot \delta^{\otimes l-1}(t-\varepsilon_l) + \delta^{\otimes l-1}(t+\varepsilon_l)\right], \tag{C.82}$$

with $\delta^{\otimes 0}(t) = [1-e^t]_+$.

Proof. Let $P_{1:k}, Q_{1:k}$ be the joint output distributions of the k-fold composed mechanism on neighboring inputs. To prove the statement, we need to show that

$$\forall t \in \mathbb{R} \ : \ \delta_{P_{1:k}|Q_{1:k}}(t) \leq \delta^{\otimes}(t). \tag{C.83}$$

Let's define $P_i^{x_{<i}}, Q_i^{x_{<i}}$ be the output distributions of the i^{th} mechanism, conditioned on the preceding $i-1$ mechanisms' output being $x_{<i}$. From Theorem 36 and from Lemma 24, we know that under adaptive $(\varepsilon_i, \delta_i)$-DP, the privacy profiles for conditional distributions are dominated as follows.

$$\forall i \in \{1, \cdots, k\} \ : \ \sup_{x_{<i} \in \Omega_1 \times \cdots \times \Omega_{i-1}} \delta_{P_i^{x_{<i}}|Q_i^{x_{<i}}}(t) \leq \delta_{\mathrm{RR}}^{\varepsilon_i,\delta_i}(t), \tag{C.84}$$

where $\delta_{\mathrm{RR}}^{\varepsilon_i,\delta_i}(t)$ is the privacy profile of $\mathcal{A}_{\mathbb{RR}}^{\varepsilon_i,\delta_i}$.

Using this, we prove the theorem statement inductively.

Base step. Let's denote the Heaveside step function as $H(t) := \mathbb{I}\{t > 0\}$. Then, we can write $\delta^{\otimes 0}(t) = (1 - H(t)) \cdot (1 - e^t)$. Using this, we can express

$$\delta^{\otimes 1}(t) = \delta_1 + \frac{1-\delta_1}{e^{\varepsilon_1}+1}\left[e^{\varepsilon_1} \cdot \delta^{\otimes 0}(t-\varepsilon_1) + \delta^{\otimes 0}(t+\varepsilon_1)\right] \tag{C.85}$$

$$= \delta_1 + \frac{1-\delta_1}{e^{\varepsilon_1}+1}\left[e_1^{\varepsilon} \cdot (1 - H(t-\varepsilon_1)) \cdot (1 - e^{t-\varepsilon_1}) + (1 - H(t+\varepsilon_1)) \cdot (1 - e^{t+\varepsilon_1})\right] \tag{C.86}$$

$$= 1 + e^x(1-\delta_1) + H(t-\varepsilon_1) \cdot \frac{(1-\delta_1)(e^t - e^{\varepsilon_1})}{e^{\varepsilon_1}+1} + H(t+\varepsilon_1) \cdot \frac{(1-\delta_1)(e^{t+\varepsilon_1}-1)}{e^{\varepsilon_1}+1} \tag{C.87}$$

$$= \begin{cases} \delta_1 & \text{if } \varepsilon_1 < t, \\ 1 - \frac{(e^t+1)(1-\delta_1)}{e_1^{\varepsilon}+1} & \text{if } -\varepsilon_1 < t \leq \varepsilon_1, \\ 1 - e^t(1-\delta_1) & \text{if } t \leq -\varepsilon_1. \end{cases} \tag{C.88}$$

$$= \delta_{\mathrm{RR}}^{\varepsilon_1,\delta_1}(x) \tag{C.89}$$

From Equation C.84, we therefore get that $\delta_{P_1|Q_1}(t) \leq \delta^{\otimes 1}(t)$ for all $t \in \mathbb{R}$.

Induction step. Suppose for any $l \in \{2, \cdots, k\}$ the composition of first $l-1$ mechanisms have a privacy profile dominated by $\delta^{\otimes l-1}$. More precisely,

$$\forall t \in \mathbb{R} \ : \ \delta_{P_{1:l-1}|Q_{1:l-1}}(t) \leq \delta^{\otimes l-1}(t). \tag{C.90}$$

We need to show that

$$\forall t \in \mathbb{R} \ : \ \delta_{P_{1:l}|Q_{1:l}}(t) \leq \delta^{\otimes l}(t). \tag{C.91}$$

Recall that from the adaptive $(\varepsilon_l, \delta_l)$-DP assumption on the l^{th} mechanism, Equation C.84 says that

$$\sup_{x_{<l} \in \Omega_1 \times \cdots \times \Omega_{l-1}} \delta_{P_l^{x_{<l}}|Q_l^{x_{<l}}}(t) \leq \delta_{\mathrm{RR}}^{\varepsilon_l, \delta_l}(t). \tag{C.92}$$

Therefore, from Theorem 35, Lemma 24 and the induction assumption, we have that

$$\begin{aligned}
\delta_{P_{1:l}|Q_{1:l}}(t) &\leq \left(\delta_{P_{1:l-1}|Q_{1:l-1}} \circledast \left(\ddot{\delta}_{\mathrm{RR}}^{\varepsilon_l, \delta_l} - \dot{\delta}_{\mathrm{RR}}^{\varepsilon_l, \delta_l}\right)\right)(t) && \text{(C.93)} \\
&\leq \left(\delta^{\otimes l-1} \circledast \left(\ddot{\delta}_{\mathrm{RR}}^{\varepsilon_l, \delta_l} - \dot{\delta}_{\mathrm{RR}}^{\varepsilon_l, \delta_l}\right)\right)(t) && \text{(C.94)} \\
&= \int_{-\infty}^{\infty} \delta^{\otimes l-1}(t-\tau) \times \left(\ddot{\delta}_{\mathrm{RR}}^{\varepsilon_l, \delta_l}(\tau) - \dot{\delta}_{\mathrm{RR}}^{\varepsilon_l, \delta_l}(\tau)\right) \mathrm{d}\tau. && \text{(C.95)}
\end{aligned}$$

To differentiate properly, in a manner that handles discontinuity, we state the function $\delta_{\mathrm{RR}}^{\varepsilon,\delta}(t)$ in terms of Heaviside functions as in Equation C.87:

$$\delta_{\mathrm{RR}}^{\varepsilon,\delta}(t) = \underbrace{1 - e^t(1-\delta)}_{I_1(t)} + H(t-\varepsilon) \cdot \underbrace{\frac{(e^t - e^\varepsilon)(1-\delta)}{e^\varepsilon + 1}}_{I_2(t)} + H(t+\varepsilon) \cdot \underbrace{\frac{(e^{t+\varepsilon} - 1)(1-\delta)}{e^\varepsilon + 1}}_{I_3(t)}. \tag{C.96}$$

Then, by chain rule, its first and second derivatives are:

$$\dot{\delta}_{\mathrm{RR}}^{\varepsilon,\delta}(t) = \dot{I}_1(t) + \left(H(t-\varepsilon) \cdot \dot{I}_2(t) + \underbrace{\Delta(t-\varepsilon) \cdot I_2(t)}_{J_2(t)}\right) + \left(H(t+\varepsilon) \cdot \dot{I}_3(t) + \underbrace{\Delta(t+\varepsilon) \cdot I_3(t)}_{J_3(t)}\right) \tag{C.97}$$

$$\begin{aligned}
\ddot{\delta}_{\mathrm{RR}}^{\varepsilon,\delta}(t) = \ddot{I}_1(t) &+ \left(H(t-\varepsilon) \cdot \ddot{I}_2(t) + \Delta(t-\varepsilon) \cdot \dot{I}_2(t) + \dot{J}_2(t)\right) \\
&+ \left(H(t+\varepsilon) \cdot \ddot{I}_3(t) + \Delta(t+\varepsilon) \cdot \dot{I}_3(t) + \dot{J}_3(t)\right)
\end{aligned} \tag{C.98}$$

Note that $\dot{I}_1(t) = \ddot{I}_1(t) = -e^t(1-\delta)$, $\dot{I}_2(t) = \ddot{I}_2(t) = e^t \cdot \frac{(1-\delta)}{e^\varepsilon+1}$ and $\dot{I}_3(t) = \ddot{I}_3(t) = e^t \cdot \frac{e^\varepsilon(1-\delta)}{e^\varepsilon+1}$. Therefore, on subtracting the two, a lot of terms cancel out, and we get:

$$\ddot{\delta}^{\varepsilon,\delta}_{\mathrm{RR}}(t) - \dot{\delta}^{\varepsilon,\delta}_{\mathrm{RR}}(t) = \Delta(t-\varepsilon)\cdot(\dot{I}_2(t) - I_2(t)) + \dot{J}_2(t) + \Delta(t+\varepsilon)\cdot(\dot{I}_3(t) - I_3(t)) + \dot{J}_3(t) \tag{C.99}$$

$$= \Delta(t-\varepsilon)\cdot\frac{(1-\delta)e^\varepsilon}{e^\varepsilon+1} + \dot{J}_2(t) + \Delta(t+\varepsilon)\cdot\frac{(1-\delta)}{e^\varepsilon+1} + \dot{J}_3(t). \tag{C.100}$$

Note that $J_2(t) = 0$ everywhere except at $t = \varepsilon$. Similarly, $J_3(t) = 0$ everywhere except $t = -\varepsilon$.

On substituting and convolving, we get

$$\delta_{P_{1:l}|Q_{1:l}}(t) \le \int_{-\infty}^{\infty} \delta^{\otimes l-1}(t-\tau) \times \left(\ddot{\delta}^{\varepsilon_l,\delta_l}_{\mathrm{RR}}(\tau) - \dot{\delta}^{\varepsilon_l,\delta_l}_{\mathrm{RR}}(\tau)\right) \mathrm{d}\tau \tag{C.101}$$

$$= \int_{-\infty}^{\infty} \delta^{\otimes l-1}(t-\tau) \times \left(\Delta(\tau-\varepsilon_l)\cdot\frac{(1-\delta_l)e^{\varepsilon_l}}{e^{\varepsilon_l}+1} + \Delta(\tau+\varepsilon_l)\cdot\frac{(1-\delta_l)}{e^{\varepsilon_l}+1}\right)\mathrm{d}\tau$$

$$+ \int_{-\infty}^{\infty} \delta^{\otimes l-1}(t-\tau) \times \left(\dot{J}_2(\tau) + \dot{J}_3(\tau)\right)\mathrm{d}\tau \tag{C.102}$$

$$= \delta^{\otimes l-1}(t-\varepsilon_l)\cdot\frac{(1-\delta_l)e^{\varepsilon_l}}{e^{\varepsilon_l}+1} + \delta^{\otimes l-1}(t+\varepsilon_l)\cdot\frac{1-\delta_l}{e^{\varepsilon_l}+1} \tag{C.103}$$

$$+ \underbrace{\int_{-\infty}^{\infty} \delta^{\otimes l-1}(t-\tau)\times\dot{J}_2(\tau)\mathrm{d}\tau}_{K_2(t)} + \underbrace{\int_{-\infty}^{\infty} \delta^{\otimes l-1}(t-\tau)\times\dot{J}_3(\tau)\mathrm{d}\tau}_{K_3(t)} \tag{C.104}$$

For the last integral, apply integration by parts to get

$$K_2(t) = \int_{-\infty}^{\infty} \delta^{\otimes l-1}(t-\tau)\dot{I}_2(\tau)\cdot\Delta(\tau-\varepsilon_l)\mathrm{d}\tau + \int_{-\infty}^{\infty} \delta^{\otimes l-1}(t-\tau)I_2(\tau)\cdot\dot{\Delta}(\tau-\varepsilon_l)\mathrm{d}\tau \tag{C.105}$$

$$= \delta^{\otimes l-1}(t-\varepsilon_l)\dot{I}_2(\varepsilon_l) - \left(\delta^{\otimes l-1}(t-\varepsilon_l)\dot{I}_2(\varepsilon_l) + \dot{\delta}^{\otimes l-1}(t-\varepsilon_l)I_2(\varepsilon_l)\right) \tag{C.106}$$

$$= 0 - 0 \tag{C.107}$$

since $I_2(\varepsilon_l) = 0$ and for any function f on $\mathbb{R}$ it holds that, $\int_{\mathbb{R}} f\cdot\dot{\Delta}\mathrm{d}\tau = -\int_{\mathbb{R}}\dot{f}\cdot\Delta\mathrm{d}\tau$. Similarly,

$$K_3(t) = \int_{-\infty}^{\infty} \delta^{\otimes l-1}(t-\tau)\dot{I}_3(\tau)\cdot\Delta(\tau+\varepsilon_l)\mathrm{d}\tau + \int_{-\infty}^{\infty} \delta^{\otimes l-1}(t-\tau)I_3(\tau)\cdot\dot{\Delta}(\tau+\varepsilon_l)\mathrm{d}\tau \tag{C.108}$$

$$= \delta^{\otimes l-1}(t+\varepsilon_l)\dot{I}_3(-\varepsilon_l) - \delta^{\otimes l-1}(t+\varepsilon_l)\dot{I}_3(-\varepsilon_l) - \dot{\delta}^{\otimes l-1}(t-\varepsilon_l)I_3(-\varepsilon_l) \tag{C.109}$$

$$= 0 - 0 \tag{C.110}$$

since $I_3(-\varepsilon_l) = 0$. Therefore, we have

$$\delta_{P_{1:l}|Q_{1:l}}(t) \leq \frac{(1-\delta_l)}{e^{\varepsilon_l}+1}\left[e^{\varepsilon_l}\cdot\delta^{\otimes l-1}(t-\varepsilon_l) + \delta^{\otimes l-1}(t+\varepsilon_l)\right] \leq \delta^{\otimes l}(t). \tag{C.111}$$

Hence, the induction statement holds. □

Corollary 5. *For any $\varepsilon \geq 0$, $\delta \in [0,1]$, the k-fold composition of (ε,δ)-DP mechanisms satisfies $(\varepsilon, \delta^{\otimes k}(\varepsilon))$-DP for all ε, where*

$$\forall t \in \mathbb{R} \;:\; \delta^{\otimes k}(t) = 1-(1-\delta)^k\left(1 - \mathop{\mathbb{E}}_{Y\leftarrow\text{Binomial}\left(k,\frac{e^\varepsilon}{1+e^\varepsilon}\right)}\left[1-e^{t-\varepsilon\cdot(2Y-k)}\right]_+\right). \tag{C.112}$$

Proof. We just have to show that recurrence relationship in Theorem 37, that is $\forall l \in \{1,\cdots,k\}$

$$\forall t \in \mathbb{R} \;:\; \delta^{\otimes l}(t) = \delta_l + \frac{(1-\delta_l)}{e^{\varepsilon_l}+1}\left[e^{\varepsilon_l}\cdot\delta^{\otimes l-1}(t-\varepsilon_l) + \delta^{\otimes l-1}(t+\varepsilon_l)\right], \quad \text{where} \quad \delta^{\otimes 0}(t) = [1-e^t]_+, \tag{C.113}$$

simplifies to the theorem statement when $\varepsilon_i = \varepsilon_j = \varepsilon$ and $\delta_i = \delta_j = \delta$ for all $i,j \in \{1,\cdots,k\}$. Let's define $p = \frac{e^\varepsilon}{e^\varepsilon+1}$. The recurrence can then be stated as

$$\delta^{\otimes l}(t) = \delta + (1-\delta)\mathop{\mathbb{E}}_{Y_l\leftarrow\text{Bernoulli}(p)}\left[\delta^{\otimes l-1}(t-\varepsilon(2Y_l-1))\right] \tag{C.114}$$

$$= \delta + (1-\delta)\mathop{\mathbb{E}}_{Y_l\leftarrow\text{Bernoulli}(p)}\left[\delta + (1-\delta)\mathop{\mathbb{E}}_{Y_{l-1}\leftarrow\text{Bernoulli}(p)}\left[\delta^{\otimes l-2}(t-\varepsilon(2Y_l-1)-\varepsilon(2Y_{l-1}-1))\right]\right] \tag{C.115}$$

$$= \delta\sum_{i=1}^{2}(1-\delta)^{i-1} + (1-\delta)^2\mathop{\mathbb{E}}_{\substack{Y_l\leftarrow\text{Bernoulli}(p)\\ Y_{l-1}\leftarrow\text{Bernoulli(p)}}}\left[\delta^{\otimes l-2}(t-\sum_{i=l-1}^{l}\varepsilon(2\cdot Y_i-1))\right] \tag{C.116}$$

$$= \delta\sum_{i=1}^{l}(1-\delta)^{i-1} + (1-\delta)^l\mathop{\mathbb{E}}_{\substack{Y_l\leftarrow\text{Bernoulli}(p)\\ Y_{l-1}\leftarrow\text{Bernoulli(p)}}}\left[\delta^{\otimes 0}(t-\sum_{i=1}^{l}\varepsilon(2\cdot Y_i-1))\right] \tag{C.117}$$

$$= 1-(1-\delta)^l + (1-\delta)^l\mathop{\mathbb{E}}_{Y\leftarrow\text{Binomial}(k,p)}\left[\delta^{\otimes 0}(t-\varepsilon(2\cdot Y-l))\right] \tag{C.118}$$

$$= 1-(1-\delta)^l\left(1-\mathop{\mathbb{E}}_{Y\leftarrow\text{Binomial}(k,p)}\left[1-e^{t-\varepsilon\cdot(2Y-k)}\right]_+\right). \tag{C.119}$$

□

C.3.1 Deferred Proofs for § 5.5

Theorem 38 (Poisson subsampling)**.** *Let* $0 \le \lambda \le 1$*. For any two distributions* P *and* Q *on* Ω*,*

$$\delta_{\lambda P+(1-\lambda)|Q}(\varepsilon) = \begin{cases} \lambda\delta_{P|Q}(\log(1+(e^{\varepsilon}-1)/\lambda)) & \text{if } \varepsilon > \log(1-\lambda) \\ 1-e^{\varepsilon} & \text{otherwise} \end{cases}. \tag{C.120}$$

Proof. Recall from Definition 26 that $\mathrm{PLD}(Q\|P)$ and $\mathrm{PLD}(Q\|\lambda P+(1-\lambda)Q)$ are the distributions of $L_{Q|P}(\Theta)$ and $L_{Q|P\lambda+(1-\lambda)Q}(\Theta)$ respectively with $\Theta \sim Q$. Since for any $\theta \in \Omega$,

$$L_{Q|\lambda P+(1-\lambda)Q}(\theta) = -\log\frac{\lambda P(\theta)+(1-\lambda)Q(\theta)}{Q(\theta)} = -\log(1-\lambda+\lambda\cdot e^{-L_{Q|P}(\theta)}), \tag{C.121}$$

the random variables $Z'_\lambda \sim \mathrm{PLD}(Q\|\lambda P+(1-\lambda)Q)$ and $Z' \sim \mathrm{PLD}(Q\|P)$ are related as

$$Z'_\lambda = -\log(1+\lambda(e^{-Z'}-1)). \tag{C.122}$$

Using Theorem 30 that we can express the privacy profile as:

$$\begin{aligned}
\delta_{\lambda P+(1-\lambda)Q|Q}(\varepsilon) &= e^{\varepsilon}\mathcal{L}\left\{F_{Z'_\lambda}(-t-\varepsilon)\right\}(-1) && \text{(C.123)}\\
&\overset{\text{Equation C.6}}{=} e^{\varepsilon}\left[\mathcal{L}\left\{f_{Z'_\lambda}(-t-\varepsilon)\right\}(-1) - F_{Z'_\lambda}(\varepsilon^-)\right] && \text{(C.124)}\\
&= e^{\varepsilon}\left[\int_{0^+}^{\infty} e^t f_{Z'_\lambda}(-t-\varepsilon)dt - \int_{-\infty}^{0^-} f_{Z'_\lambda}(t-\varepsilon)dt\right] && \text{(C.125)}\\
&= e^{\varepsilon}\int_{-\infty}^{0^-}(e^{-t}-1)f_{Z'_\lambda}(t-\varepsilon)dt && \text{(C.126)}\\
&= \int_{-\infty}^{-\varepsilon^-}\left(e^{-t'}-e^{\varepsilon}\right)f_{Z'_\lambda}(t')dt' && \text{(C.127)}\\
&= \mathbb{E}\left[e^{-Z'_\lambda}-e^{\varepsilon}\right]_+, && \text{(C.128)}
\end{aligned}$$

where $[x]_+ \stackrel{\text{def}}{=} \max\{0, x\}$. On substituting Z'_λ, we get:

$$\begin{aligned}
\delta_{\lambda P+(1-\lambda)Q|Q}(\varepsilon) &= \mathbb{E}\left[e^{-Z'_\lambda} - e^{\varepsilon}\right]_+ && \text{(C.129)}\\
&= \mathbb{E}\left[1-\lambda+\lambda e^{-Z'} - e^{\varepsilon}\right]_+ && \text{(C.130)}\\
&= \lambda\mathbb{E}\left[e^{-Z'} - \frac{e^{\varepsilon}+\lambda-1}{\lambda}\right]_+ && \text{(C.131)}\\
&= \begin{cases} \lambda\mathbb{E}\left[e^{-Z'} - e^{\log(1+(e^{\varepsilon}-1)/\lambda)}\right]_+ & \text{if } \varepsilon > \log(1-\lambda) \\ \lambda\mathbb{E}\left[e^{-Z'}\right] + 1 - \lambda - e^{\varepsilon} & \text{otherwise} \end{cases} && \text{(C.132)}\\
&= \begin{cases} \lambda\delta_{P|Q}(\log(1+(e^{\varepsilon}-1)/\lambda)) & \text{if } \varepsilon > \log(1-\lambda) \\ 1 - e^{\varepsilon} & \text{otherwise} \end{cases}. \\
& && (\because \mathbb{E}\left[e^{-Z'}\right] = 1)
\end{aligned}$$

□