Forget Unlearning: Towards True Data-Deletion in Machine Learning

Rishav Chourasia [1] **Neil Shah** [1]

Abstract

Unlearning algorithms aim to remove deleted data's influence from trained models at a cost lower than full retraining. However, prior guarantees of unlearning in literature are flawed and don't protect the privacy of deleted records. We show that when people delete their data as a function of published models, records in a database become interdependent. So, even retraining a fresh model after deletion of a record doesn't ensure its privacy. Secondly, unlearning algorithms that cache partial computations to speed up the processing can leak deleted information over a series of releases, violating the privacy of deleted records in the long run. To address these, we propose a sound deletion guarantee and show that ensuring the privacy of existing records is necessary for the privacy of deleted records. Under this notion, we propose an optimal, computationally efficient, and sound machine unlearning algorithm based on noisy gradient descent.

1. Introduction

Corporations today collect their customers' private information to train Machine Learning (ML) models that power a variety of services like recommendations, searches, targeted ads, etc. To prevent any unintended use of personal data, privacy policies, such as the General Data Protection Regulation and the California Consumer Privacy Act, require that these corporations provide people the "*Right to be Forgotten*" (RTBF)—if a person wants to revoke access to their data, an organization must comply by erasing all information about them without undue delay (usually a month). This includes ML models trained in standard ways as privacy attacks like membership inference (Shokri et al., 2017) and model inversion (Fredrikson et al., 2015) demonstrate that training data can be exfiltrated from them.

[1]Department of Computer Science, National University of Singapore, Singapore. Correspondence to: Rishav Chourasia <rishav1@comp.nus.edu.sg>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Retraining fresh ML models from scratch after deletions is computationally expensive and does not scale. As an alternative to retraining, there is a growing interest in designing cheap *Machine Unlearning* algorithms for erasing the influence of deleted data from already trained models. To quantify how well an unlearning algorithm deletes the requested information in the worst-case, Ginart et al. (2019) propose a *differential privacy* (DP) like (ε, δ)-indistinguishability certification between the unlearning algorithm's output and that of fresh retraining without the deleted records. Several unlearning certifications that follow the same intuition have since been proposed and used to certify numerous unlearning mechanisms (Izzo et al., 2021; Sekhari et al., 2021; Neel et al., 2021; Guo et al., 2019; Ullah et al., 2021).

However, *is indistinguishability from retraining a trustworthy guarantee of deletion privacy?* We argue that it is not. In the real world, a person's decision to remove their information is often influenced by what a deployed model reveals about them. Unfortunately, the same revealed information or attempts to censor it may also affect other people's decisions. This phenomenon is famously known as the *Streisand effect*, named after the American singer and actress Barbara Streisand's attempts to censor a picture of her cliff-top Malibu residence, originally taken to document the coastal erosion under the California Coastal Records Project. Prior to Streisand's lawsuit, the picture was downloaded only six times, two of which were by Streisand's attorneys. However, after the case gained public attention, the site received over 420,000 visitors in the following months (Adelman & Adelman, 2002). This kind of *adaptivity* in real-world interactions creates interdependencies among the records in a database—some patterns in stored records simply wouldn't exist unless specific information about a target record that influenced others was previously revealed. Even after removing the requested records from the training database and retraining a model from scratch, we demonstrate that an attacker can still identify the deleted records by solely examining the retrained model, all due to the predictability of adaptive, yet unseen, past interactions. As such, we argue that any deletion certification based on being indistinguishable from retraining, as done in all prior unlearning definitions including Gupta et al. (2021)'s *adaptive unlearning* certification, is fundamentally flawed.

Is indistinguishability from retraining a reliable measure of deletion privacy when requests are non-adaptive? Once again, we argue that it is not. A reliable guarantee of deletion privacy should ensure that deleted records cannot be recovered by an observer who has access to *multiple (potentially all)* model releases after processing the deletion request. However, approximate indistinguishability from retraining implies that deleted data cannot be accurately recovered from a *single* unlearned model only, which we assert is insufficient. We show that some unlearning algorithms can generate models indistinguishable from retraining beyond any arbitrarily small threshold while still exposing the deleted data over multiple releases. This vulnerability arises in algorithms that retain partial computations in the form of internal data-dependent states to expedite subsequent deletions. Such internal states can carry information about previously deleted records and influence several future releases, which may eventually expose the deleted records. Therefore, prior unlearning guarantees are myopic and unreliable in sequential deletion scenarios.

Lastly, we contend that *indistinguishability from retraining based notions of deletion privacy are also incomplete* because they disregard perfectly valid deletion mechanisms. For instance, a (useless) mechanism that outputs a fixed untrained model in response to any deletion request is a valid deletion algorithm. However, since its output is easy to tell apart from that of retraining done by any sensible learning algorithm, the mechanism cannot satisfy any deletion certification based on being indistinguishable from retraining.

This paper proposes a definition of *deletion privacy* that does not suffer from the aforementioned shortcomings. Our framework considers an unlearning mechanism reliable if **A)** its output is not influenced by any internal states dependent on previously deleted records; and if **B)** for any deletion request, one can demonstrate indistinguishability between its output and some random variable independent of the records deleted. Avoiding reliance on any internal states contaminated by previously deleted records makes the unlearning mechanism a *post-processing operation* for the deleted records—once records have been deleted in the ML model, they stay permanently deleted throughout all future steps. Additionally, our framework is reliable even for adaptive requests as we measure deletion privacy based on indistinguishability from a *random variable independent of the deleted records by construction*, rather than the output of retraining which may become dependent on deleted records under adaptivity.

Unlike prior works on machine unlearning, we provide a rigorous *proof of the soundness of our deletion privacy certification*. We show that if the certification holds, no attacker, regardless of the number of the post-unlearning releases observed or their understanding of how people might have responded in the past, can successfully disambiguate a deleted record. Furthermore, our deletion privacy certification can *certify valid deletion mechanisms that prior unlearning definitions cannot*. For example, it can certify the fixed-output mechanism mentioned earlier, thanks to the flexibility of design for a random variable construction in our definition.

We note that the concept of deletion privacy differs from the conventional notion of differential privacy in terms of the information they restrict. While standard differential privacy limits the information related to individual records *currently present in the database*, deletion privacy focuses on the information concerning records that have been *previously deleted from the database*. We explore the relationship between these two privacy certifications for an unlearning mechanism and present two complementary findings. First, we prove that when requests are adaptive, *an unlearning mechanism must preserve the privacy of the remaining records in order to ensure privacy for the deleted records*. Failure to do so can introduce undesired correlations among records, such as when a Streisand effect occurs, which can hinder deletion in the information-theoretic sense. On the other hand, we also establish that if *an unlearning algorithm satisfies deletion privacy guarantees for non-adaptive edit requests and is additionally differentially private, then it also satisfies deletion privacy for adaptive requests*. This reduction greatly simplifies the design of a reliable unlearning mechanism in the real-world setting, as designers can focus on creating unlearning mechanisms that are both differentially private and provide deletion privacy guarantees under the assumption that deletion requests are independent to models released in the past.

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 database, provided by a differential privacy guarantee, and a significantly smaller bound for the records previously deleted, ensured through a deletion privacy guarantee. Based on our findings, *we redefine the problem of data-deletion in ML* as designing a mechanism that **(1.)** satisfies a deletion privacy guarantee against non-adaptive deletion requests, **(2.)** is differentially private for remaining records, and **(3.)** has the same utility guarantee as retraining under identical differential privacy constraints. On top of these objectives, a data-deletion mechanism must also be computational cheaper than retraining for being useful.

We present a data-deletion solution that utilizes fine-tuning through *noisy gradient descent (Noisy-GD)*, a popular differentially-private learning algorithm (Bassily et al.,

2014; Abadi et al., 2016). Our solution achieves all the three objectives while providing substantial computation savings for both convex and non-convex losses. Notably, we demonstrate a powerful synergy between deletion privacy and differential privacy, where the same noise needed for privacy of present database records also ensures the erasure of information related to the deleted records. For convex and smooth losses, we certify that under a $(q, \varepsilon_{\rm dd})$-Rényi non-adaptive deletion-privacy and $(q, \varepsilon_{\rm dp})$-Rényi differential-privacy constraint, our Noisy-GD based data-deletion mechanism for d-dimensional models over n-sized databases with requests that modify up to r records at a time can maintain the pareto-optimal excess empirical risk of the order $O\big(\frac{qd}{\varepsilon_{\rm dp} n^2}\big)$ while being $\Omega(n \log(\min\{\frac{n}{r}, n\sqrt{\frac{\varepsilon_{\rm dd}}{qd}}\})$ cheaper than retraining in gradient complexity. For non-convex, bounded and smooth losses, we show a computational saving of $\Omega(dn \log \frac{n}{r})$ in gradient complexity under the same constraints with an excess risk of $\tilde{O}\big(\frac{qd}{\varepsilon_{\rm dp} n^2} + \frac{1}{n}\sqrt{\frac{q}{\varepsilon_{\rm dp}}}\big)$. Compared to our results, prior works offer a worse computation savings under the same utility constraints (Izzo et al., 2021; Guo et al., 2019; Sekhari et al., 2021; Ullah et al., 2021), violate RTBF under adaptivity (Bourtoule et al., 2021; Gupta et al., 2021), or require internal states that depend on previously deleted records for matching our utility bounds (Neel et al., 2021).

2. Preliminaries

2.1. Indistinguishability and Differential Privacy

We provide the basics of indistinguishability of random variables (with more details in Appendix B). Let Θ, Θ' be two random variables in space $\mathcal{O}$ with probability densities ν, ν' respectively.

Definition 2.1 ((ε, δ)-indistinguishability (Dwork et al., 2014))**.** *We say Θ and Θ' are* (ε, δ)-indistinguishable *(denoted by $\Theta \overset{\varepsilon,\delta}{\approx} \Theta'$) if, for all $O \subset \mathcal{O}$,*

$$\begin{aligned} \mathbb{P}\left[\Theta \in O\right] &\leq e^{\varepsilon}\mathbb{P}\left[\Theta' \in O\right] + \delta \quad \textit{and} \\ \mathbb{P}\left[\Theta' \in O\right] &\leq e^{\varepsilon}\mathbb{P}\left[\Theta \in O\right] + \delta. \end{aligned} \tag{1}$$

Definition 2.2 (Rényi divergence (Rényi et al., 1961))**.** *When ν is* absolutely continuous *w.r.t. ν' (denoted as $\nu \ll \nu'$),* Rényi divergence *of ν w.r.t. ν' is defined as*

$$\mathrm{R}_q\left(\nu \| \nu'\right) = \frac{1}{q-1} \log \mathrm{E}_q\left(\nu \| \nu'\right), \tag{2}$$

where order $q > 1$ and

$$\mathrm{E}_q\left(\nu \| \nu'\right) = \mathop{\mathbb{E}}_{\theta \sim \nu'}\left[\left(\frac{\nu(\theta)}{\nu'(\theta)}\right)^q\right]. \tag{3}$$

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

Definition 2.3 ((Rényi) Differential Privacy (Dwork et al., 2014; Mironov, 2017))**.** *A randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{O}$ is said to be* (ε, δ)-differentially private *if $\mathcal{M}(\mathcal{D}) \overset{\varepsilon,\delta}{\approx} \mathcal{M}(\mathcal{D}')$ for all* neighbouring *databases $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$. Similarly, $\mathcal{M}$ is* (q, ε)-Rényi differentially private *if* $\mathrm{R}_q\left(\mathcal{M}(\mathcal{D}) \| \mathcal{M}(\mathcal{D}')\right) \leq \varepsilon$.

2.2. (Adaptive) Machine Unlearning

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

Definition 2.4 (Edit request)**.** *A* replacement operation $\langle \mathrm{ind}, \mathbf{y} \rangle \in [n] \times \mathcal{X}$ *on a database $\mathcal{D} = (\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathcal{X}^n$ performs the following modification:*

$$\mathcal{D} \circ \langle \mathrm{ind}, \mathbf{y} \rangle = (\mathbf{x}_1, \cdots, \mathbf{x}_{\mathrm{ind}-1}, \mathbf{y}, \mathbf{x}_{\mathrm{ind}+1}, \cdots, \mathbf{x}_n). \tag{4}$$

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

$$\mathcal{D} \circ u = \mathcal{D} \circ \langle \mathrm{ind}_1, \mathbf{y}_1 \rangle \circ \cdots \circ \langle \mathrm{ind}_r, \mathbf{y}_r \rangle, \tag{5}$$

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

Similar to Ginart et al. (2019), we define a *deletion* or an *unlearning algorithm* as a (possibly stochastic) mapping $\bar{\mathrm{A}} : \mathcal{X}^n \times \mathcal{U}^r \times \mathcal{O} \to \mathcal{O}$. This algorithm takes a database $\mathcal{D} \in \mathcal{X}^n$, an edit request $u \in \mathcal{U}^r$ and the current model in $\mathcal{O}$, and produces an updated model in $\mathcal{O}$. Since edit requests needs to be processed monthly under RTBF guidelines, we adopt the online setting introduced by Neel et al. (2021) in which a stream of edit requests $(u_i)_{i \geq 1} \overset{\rm def}{=} (u_1, u_2, \cdots)$, with $u_i \in \mathcal{U}^r$, arrives sequentially. As per this formulation, the *data curator*, characterized by algorithms $(\mathrm{A}, \bar{\mathrm{A}}, f_{\rm pub})$, executes the learning algorithm A on the initial database $\mathcal{D}_0 \in \mathcal{X}^n$ during the setup stage before arrival of the first edit request to generate the initial model $\hat{\Theta}_0 \in \mathcal{O}$, i.e., $\hat{\Theta}_0 = \mathrm{A}(\mathcal{D}_0)$. Thereafter at any edit step $i \geq 1$, to reflect an incoming edit request $u_i \in \mathcal{U}^r$ that transforms $\mathcal{D}_{i-1} \circ u_i \to \mathcal{D}_i$, the curator executes the unlearning algorithm $\bar{\mathrm{A}}$ on current database $\mathcal{D}_{i-1}$, the edit request u_i, and the current model $\hat{\Theta}_{i-1}$ for generating the next model $\hat{\Theta}_i \in \mathcal{O}$, i.e., $\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) = \hat{\Theta}_i$. 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 $\phi_i = f_{\rm pub}(\hat{\Theta}_i)$ for all $i \geq 0$.

[1] Over a month, 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.

The publish function $f_{\text{pub}} : \mathcal{O} \to \Phi$ generates these publishable objects in some space Φ, which can represent any downstream usage such as making predictions.

Ginart et al. (2019) 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. (2021) model such an interactive environment through an *adaptive update requester*. We provide the following generalized definition of Gupta et al. (2021)'s update requester and describe its interaction with a data curator in Algorithm 1.

Definition 2.5 (Update requester (Gupta et al., 2021))**.** *The edit sequence $(u_i)_{i\geq 1}$ is generated by an update requester $\mathcal{Q}$ that inputs a subset of interaction history between herself and the curator $(\mathrm{A}, \bar{\mathrm{A}}, f_{\text{pub}})$, and outputs a new edit request for the current round. We quantify the strength of $\mathcal{Q}$ with two integers (p, r). Here p is the maximum number of prior published objects that the requester $\mathcal{Q}$ has access to for generating 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 $\mathcal{Q} : \Phi^{\leq p} \times \mathcal{U}^{r*} \to \mathcal{U}^r$. Given a sorted list of observable indices $\vec{s} = (s_1, \cdots, s_p) \in \mathbb{N}^p$ the i^{th} edit request u_i generated by $\mathcal{Q}$ on interaction with $(\mathrm{A}, \bar{\mathrm{A}})$ is defined as*

$$u_i = \mathcal{Q}(\underbrace{\phi_{s_1}, \phi_{s_2}, \cdots, \phi_{s_j}}_{\stackrel{\text{def}}{=} \phi_{\vec{s}<i}}; \underbrace{u_1, u_2, \cdots, u_{i-1}}_{\stackrel{\text{def}}{=} u_{<i}}), \quad (6)$$

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

Algorithm 1 Interacting curator $(\mathrm{A}, \bar{\mathrm{A}}, f_{\text{pub}})$ & requester $\mathcal{Q}$

Require: Database $\mathcal{D}_0 \in \mathcal{X}^n$, observable indices $\vec{s} \in \mathbb{N}^p$.
1: Initialize $\hat{\Theta}_0 \leftarrow \mathrm{A}(\mathcal{D}_0)$
2: Publish $\phi_0 \leftarrow f_{\text{pub}}(\hat{\Theta}_0)$
3: **for** $i = 1, 2, \cdots$ **do**
4: Get next request $u_i \leftarrow \mathcal{Q}(\phi_{\vec{s}<i}; u_{<i})$
5: Update model $\hat{\Theta}_i \leftarrow \bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1})$
6: Publish $\phi_i \leftarrow f_{\text{pub}}(\hat{\Theta}_i)$
7: Update database $\mathcal{D}_i \leftarrow \mathcal{D}_{i-1} \circ u_i$
8: **end for**

A 0-adaptive requester is considered as non-adaptive and by ∞-adaptivity we mean requesters that have access to the entire history of interaction transcript $(\phi_{<i}; u_{<i})$ at step i.

We present the definitions of *unlearning* and *adaptive unlearning*[2] proposed by Neel et al. (2021) and Gupta et al. (2021) respectively (along with other widely-used data deletion definitions provided in Appendix C.1).

[2]Definition 2.6 of adaptive unlearning is stronger than Gupta et al. (2021)'s since theirs require only one-sided indistinguishability with $(1-\gamma)$ probability over generated edit requests $u_{\leq i}$.

Definition 2.6 (Machine unlearning (Neel et al., 2021; Gupta et al., 2021))**.** *We say that $\bar{\mathrm{A}}$ is an (ε, δ)-*unlearning *algorithm for A under a publish function f_{pub}, if for all initial databases $\mathcal{D}_0 \in \mathcal{X}^n$ and all non-adaptive 1-requesters $\mathcal{Q}$, the following condition holds. For every edit step $i \geq 1$, and for all generated edit sequences $u_{\leq i} \stackrel{\text{def}}{=} (u_1, \cdots, u_i)$,*

$$f_{\text{pub}}(\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}))\big|_{u_{\leq i}} \stackrel{\varepsilon,\delta}{\approx} f_{\text{pub}}(\mathrm{A}(\mathcal{D}_i)). \quad (7)$$

*If (7) holds for all ∞-adaptive 1-requesters $\mathcal{Q}$, we say that $\bar{\mathrm{A}}$ is an (ε, δ)-*adaptive-unlearning *algorithm for A.*

3. Existing Data-Deletion and Unlearning Guarantees are Unsound and Incomplete

> *The controller shall have the obligation to erase personal data without undue delay where ... the data subject withdraws consent on which the processing is based ...*
>
> Article 17(1)(b), GDPR.

In this section we analyze the limitations of prior data deletion certifications intended to uphold the "Right to be Forgotten". We present a realistic threat model that data curators must address according to the standard interpretation of the RTBF guidelines. Subsequently, we highlight multiple reasons why both adaptive and non-adaptive machine unlearning, as described in Definition 2.6 (along with other established definitions detailed in Appendix C.1), are inadequate in addressing this threat model.

Building a 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 behaviour. However, it is crucial to ensure that an attacker cannot extract deleted information from un-

[3]Our threat model doesn't include attacks which involve comparing releases before and after deletion (such as Chen et al. (2021)'s). Succeeding in such attacks technically do not violate RTBF as they need information published before a request arrives.

learned models, regardless of their understanding of general human behaviour patterns. To account for the worst-case scenario, we model the attacker as having *complete knowledge about the adaptive requester* $\mathcal{Q}$ (as defined in Definition 2.5), which represents any form of dependence relationships between published outcomes and subsequent requests. However, the attacker does not observe the interaction transcript $(\phi_{<i}; u_{<i})$ of $\mathcal{Q}$.

Unsoundness due to adaptivity. We highlight the problem with certifying data deletion based on indistinguishability from retraining under adaptive requests with a simple example. For a data domain $\mathcal{X} = \{-2, -1, 1, 2\}$, consider the following learning and unlearning algorithms $(\mathrm{A}, \bar{\mathrm{A}})$. For any database $\mathcal{D} \subset \mathcal{X}$ and any subset $S \subset \mathcal{D}$ of records to be deleted,

$$\mathrm{A}(\mathcal{D}) = \sum_{\mathbf{x} \in \mathcal{D}} \mathbf{x}, \quad \text{and} \quad \bar{\mathrm{A}}(\mathcal{D}, S, \mathrm{A}(\mathcal{D})) = \sum_{\mathbf{x} \in \mathcal{D} \setminus S} \mathbf{x}. \tag{8}$$

Note that the unlearning algorithm $\bar{\mathrm{A}}$ perfectly imitates the learning algorithm A as $\bar{\mathrm{A}}(\mathcal{D}, S, \mathrm{A}(\mathcal{D})) = \mathrm{A}(\mathcal{D} \setminus S)$ for any $S \subset \mathcal{D}$. Now consider two neighbouring databases $\mathcal{D}_{-1} = \{-2, -1, 2\}$, $\mathcal{D}_1 = \{-2, 1, 2\}$ and the following dependence between the learned model $\mathrm{A}(\mathcal{D})$ and deletion request S:

$$S = \begin{cases} \{\mathbf{x} \in \mathcal{X} | \mathbf{x} < 0\} & \text{if } \mathrm{A}(\mathcal{D}) < 0, \\ \{\mathbf{x} \in \mathcal{X} | \mathbf{x} \geq 0\} & \text{otherwise.} \end{cases} \tag{9}$$

Knowing this dependence, an attacker can distinguish whether $\mathcal{D}$ is $\mathcal{D}_{-1}$ or $\mathcal{D}_1$ by looking only at $\bar{\mathrm{A}}(\mathcal{D}, \mathbf{x}, \mathrm{A}(\mathcal{D}))$. This is because if $\mathcal{D} = \mathcal{D}_{-1}$, then the output after deletion is positive, and if $\mathcal{D} = \mathcal{D}_1$ the output is negative. Note that even though $\bar{\mathrm{A}}$ perfectly imitates retraining via A and the attacker does not observe either the model $\mathrm{A}(\mathcal{D})$ or the request S, 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 database to have patterns specific to the identity of a target record being deleted, and **B)** an attacker knowing the relationship between unobserved 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 that people typically behave adaptively and malicious attackers exploit common behavioral patterns, we argue that several data deletion definitions in the literature, such as those proposed by Ginart et al. (2019), Guo et al. (2019), and Sekhari et al. (2021), which advocate for data deletion based on indistinguishability from retraining, do not provide reliable certifications for the "Right to be Forgotten" (further detailed in Appendix C.1). Furthermore, we present a theorem that shows how Definition 2.6 of adaptive unlearning guarantee by Gupta et al. (2021), specifically designed to ensure RTBF under adaptive deletion requests, also fails under adaptivity.

Theorem 3.1. *There exists an algorithm pair* $(\mathrm{A}, \bar{\mathrm{A}})$ *satisfying* $(0, 0)$*-adaptive-unlearning under publish function* $f_{\mathrm{pub}}(\theta) = \theta$ *such that by designing a* 1*-adaptive* 1*-requester* $\mathcal{Q}$*, 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* $\mathcal{Q}$*'s transcript* $(\phi_{<i}; u_{<i})$*.*

Unsoundness due to secret states. Both adaptive and non-adaptive unlearning guarantees in Definition 2.6 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 2.6, even when edit requests are non-adaptive. This vulnerability arises because Definition 2.6 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 database. 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.

Theorem 3.2. *For every* $\varepsilon > 0$*, there exists a pair* $(\mathrm{A}, \bar{\mathrm{A}})$ *of algorithms that satisfy* $(\varepsilon, 0)$*-unlearning under some publish function* f_{pub} *such that for all non-adaptive* 1*-requesters* $\mathcal{Q}$*, 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 the post-edit releases* $\phi_{\geq i}$*.*

Several unlearning definitions, such as those by Ginart et al. (2019), Guo et al. (2019) and Sekhari et al. (2021) (detailed in Appendix C.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. (2019), 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.

Incompleteness. Prior unlearning definitions measure the effectiveness of a deletion algorithm by assessing how closely its output resembles that of retraining. We argue that resembling the output of retraining is not necessary for erasing deleted information. Consequently, many valid deletion algorithms do not meet existing definitions. For example, let's consider a (useless) mechanism $\bar{\mathrm{A}}$ that out-

puts a fixed predetermined model $\theta \in \mathcal{O}$ regardless of its inputs. It is evident that $\bar{\mathrm{A}}$ is a valid deletion algorithm for any learning algorithm, as $\bar{\mathrm{A}}(\cdot,\cdot,\cdot)$ does not rely on the input database or the learned model (nor does $f_{\mathrm{pub}}(\bar{\mathrm{A}}(\cdot,\cdot,\cdot))$ for any f_{pub}). Yet $\bar{\mathrm{A}}$ fails to satisfy Definition 2.6 for any sensible learning algorithm A. Similarly, definitions by Ginart et al. (2019), Guo et al. (2019) and Sekhari et al. (2021) are also incomplete (refer to Appendix C.1).

4. 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 3. 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.

We noted previously in Theorem 3.2 that using partial computations 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 $f_{\mathrm{pub}}(\theta) = \theta$ in Algorithm 1). 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 database 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 4.1 ((q,ε)-deletion-privacy under p-adaptive r-requesters)**.** *Let $q > 1$, $\varepsilon \geq 0$, and $p, r \in \mathbb{N}$. We say that an algorithm pair $(\mathrm{A}, \bar{\mathrm{A}})$ satisfies (q,ε)-*deletion-privacy under p-adaptive r-requesters *if the following condition holds for all p-adaptive r-requester $\mathcal{Q}$. For every step $i \geq 1$, there exists a randomized mapping $\pi_i^{\mathcal{Q}} : \mathcal{X}^n \to \mathcal{O}$ such that for all initial databases $\mathcal{D}_0 \in \mathcal{X}^n$,*

$$\mathrm{R}_q\left(\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) \middle\| \pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle)\right) \leq \varepsilon, \quad (10)$$

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

Soundness. Definition 4.1 reliably safeguards the "Right to be Forgotten" as the random variable $\pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle)$ stays independent of the deleted record $\mathcal{D}_0[\mathrm{ind}]$ by design, even when the update requester $\mathcal{Q}$ is ∞-adaptive. When an attacker aims to identify a record at index 'ind' in $\mathcal{D}_0$ 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 bounded Rényi divergence in inequality (10) ensures that any observer of the unlearned model $\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1})$ cannot be overly certain that the observation was *not* $\pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle)$ instead[4]. Consequently, the unlearned model itself must possess minimal information regarding the deleted record $\mathcal{D}_0[\mathrm{ind}]$. This argument allows us to establish the following guarantee of soundness.

Theorem 4.1 (Definition 4.1 safeguards RTBF)**.** *If the algorithm pair $(\mathrm{A}, \bar{\mathrm{A}})$ satisfies (q,ε)-deletion-privacy guarantee under all p-adaptive r-requesters, then even with complete knowledge of a p-adaptive r-requester $\mathcal{Q}$ that interacts with the curator before a target record $\mathcal{D}_0[\mathrm{ind}]$ in the initial database $\mathcal{D}_0$ is deleted at step $i \geq 1$ by request u_i, any attacker MI : $\mathcal{O}^* \to \{0,1\}$ observing only the post-deletion models $\hat{\Theta}_{\geq i} = (\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots)$ has an advantage*

$$\mathit{Adv}(\mathit{MI}) \stackrel{\mathrm{def}}{=} \mathbb{P}\left[\mathit{MI}(\hat{\Theta}_{\geq i}){=}1 \middle| \mathbf{x}\right] - \mathbb{P}\left[\mathit{MI}(\hat{\Theta}_{\geq i}){=}1 \middle| \mathbf{x}'\right] \quad (11)$$

for disambiguating between two possible values $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ of the deleted record $\mathcal{D}_0[\mathrm{ind}]$ bounded as follows.

$$\mathit{Adv}(\mathit{MI}) \leq \min\left\{\sqrt{2\varepsilon}, \frac{q e^{\varepsilon(q-1)/q}}{q-1}[2(q-1)]^{\frac{1}{q}} - 1\right\} \quad (12)$$

As $q \to \infty$, r.h.s. of (12) approaches $\min\left\{\sqrt{2\varepsilon}, e^{\varepsilon} - 1\right\}$. Note that r.h.s. of (12) also goes to 0 as $\varepsilon \to 0$, implying Definition 4.1 is sound.

Remark 4.2 (Deletion-Privacy generalizes prior unlearning definitions under non-adaptivity)**.** *A non-adaptive requester $\mathcal{Q}$ is equivalent to fixing the request sequence $(u_i)_{i\geq 1}$ a-priori. Since $\mathcal{D}_i = (\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle) \circ u_1 \circ \cdots \circ u_i$ when $\langle \mathrm{ind}, \mathbf{y}\rangle \in u_i$, note that database $\mathcal{D}_i$ is a function of $\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle$ when $\mathcal{Q}$ is non-adaptive. Consequently, for non-adaptive $\mathcal{Q}$, we can set $\pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle) = \pi(\mathcal{D}_i)$ in* (10) *for any randomized map $\pi : \mathcal{X}^n \to \mathcal{O}$, including the learning algorithm* A.

Completeness. Our Definition 4.1 allows for the certification of fixed-output unlearning mechanisms described in Section 3, which prior unlearning definitions based on indistinguishability to retraining cannot accomplish. This

[4] Refer to Theorem B.1 in Appendix B to see that (10) implies a one-sided indistinguishability. This is enough for the purpose of ensuring RTBF as we only want to prevent events that makes an attacker confident that some deleted information was leaked; we don't care if attacker becomes confident that nothing was leaked.

flexibility stems from the freedom of choice for mechanism $\pi_i^{\mathcal{Q}}$ (see Remark 4.2), which can be set to consistently produce the same fixed output.

However, our Definition 4.1 may not be complete. 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 3.2, stateful unlearning algorithms of this nature can unveil new information to observers with each subsequent release.

4.1. Link to Differential Privacy

A differential privacy guarantee on A and $\bar{\mathrm{A}}$ sets a limit on the information present in an unlearned model regarding individual records that remain in the database. However, our concept of deletion privacy specifically restricts the information concerning only the deleted records. Nonetheless, these two properties are somewhat interconnected in the case of adaptive edit requests. We argue that when A and $\bar{\mathrm{A}}$ are both differentially private, they prevent an adaptive requester from establishing dependencies between records in the curator's database. By leveraging this property, we establish a reduction from adaptive to non-adaptive deletion privacy in the following theorem, assuming that A and $\bar{\mathrm{A}}$ also satisfy Rényi differential privacy.

Theorem 4.3 (From adaptive to non-adaptive deletion)**.** *If an algorithm pair $(\mathrm{A}, \bar{\mathrm{A}})$ satisfies $(q, \varepsilon_{\mathrm{dd}})$-deletion-privacy under all non-adaptive r-requesters and is also $(q, \varepsilon_{\mathrm{dp}})$-Rényi DP with respect to records not being deleted, then it also satisfies $(q, \varepsilon_{\mathrm{dd}} + p\varepsilon_{\mathrm{dp}})$-deletion-privacy under all p-adaptive r-requesters.*

Remark 4.4. *Gupta et al. (2021) also prove a reduction from adaptive to non-adaptive unlearning (Definition 2.6) under differential privacy. We remark that our reduction is fundamentally different from theirs as they require DP to hold with regard to a change of description of internal randomness as opposed to standard data item replacement in ours. We discuss the key differences in Appendix D.1.*

Theorem 4.3 simplifies the certification process for unlearning algorithms to ensure RTBF compliance. Under the assumption that deletion requests are independent of previous releases, showing that the algorithm is both differentially private and provides deletion privacy is sufficient.

Furthermore, we contend that in order to guarantee deletion privacy for erased records in the real-world setting, it is essential 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 4.5 (Privacy of remaining records is necessary for adaptive deletion privacy)**.** *Let Test $: \mathcal{O} \to \{0, 1\}$ be a membership inference test for* A *to distinguish between neighbouring databases $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$. Similarly, let $\overline{Test} : \mathcal{O} \to \{0, 1\}$ be a membership inference test for* $\bar{\mathrm{A}}$ *to distinguish between $\bar{\mathcal{D}}, \bar{\mathcal{D}}' \in \mathcal{X}^n$ that are neighbouring after applying edit $\bar{u} \in \mathcal{U}^1$. If Adv(Test) $> \delta$ and Adv($\overline{Test}$) $> \delta$, then the pair $(\mathrm{A}, \bar{\mathrm{A}})$ cannot satisfy (q, ε)-deletion-privacy under* 1*-adaptive* 1*-requester for any*

$$\varepsilon < \max\left\{\frac{\delta^4}{2}, \log(q-1) + \frac{q}{q-1}\log\left(\frac{1+\delta^2}{q2^{1/q}}\right)\right\}. \quad (13)$$

4.2. (Un)Learning Framework: ERM

Let space of model parameters be $\mathbb{R}^d$ 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 database $\mathcal{D} \in \mathcal{X}^n$ under $L2$ regularization, that is, the minimization objective is

$$\mathcal{L}_{\mathcal{D}}(\theta) = \frac{1}{n}\sum_{\mathbf{x}\in\mathcal{D}} \boldsymbol{\ell}(\theta; \mathbf{x}) + \mathbf{r}(\theta), \text{ with } \mathbf{r}(\theta) = \frac{\lambda \|\theta\|_2^2}{2}. \quad (14)$$

The *excess empirical risk* of a model Θ on $\mathcal{D}$ is defined as

$$\mathrm{err}(\Theta; \mathcal{D}) = \mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta) - \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})\right], \quad (15)$$

where $\theta^*_{\mathcal{D}} = \underset{\theta\in\mathbb{R}^d}{\arg\min}\, \mathcal{L}_{\mathcal{D}}(\theta)$, and expectation is over Θ.

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

- **(1.)** both A and $\bar{\mathrm{A}}$ satisfy $(q, \varepsilon_{\mathrm{dp}})$-Rényi DP with respect to individual records in the input database,
- **(2.)** pair $(\mathrm{A}, \bar{\mathrm{A}})$ satisfies $(q, \varepsilon_{\mathrm{dd}})$-deletion-privacy guarantee for all non-adaptive r-requesters $\mathcal{Q}$,
- **(3.)** and, all models $(\hat{\Theta}_i)_{i\ge 0}$ produced by $(\mathrm{A}, \bar{\mathrm{A}}, \mathcal{Q})$ on any $\mathcal{D}_0 \in \mathcal{X}^n$ have $\mathrm{err}(\hat{\Theta}_i; \mathcal{D}_i) \le \alpha$.

Objectives **(1.)** and **(2.)** together ensure that $(\mathrm{A}, \bar{\mathrm{A}})$ satisfies deletion-privacy for adaptive requests as well, and objective **(3.)** ensures (un)learned models are useful[5].

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

[5]Constraint α in **(3.)** should be close to the optimal excess risk attainable by ERM on $\mathcal{D}_i$ under $(q, \varepsilon_{\mathrm{dp}})$-Rényi DP.

5. Deletion Using Noisy Gradient Descent

This section proposes a simple and effective data-deletion solution based on Noisy-GD (Abadi et al., 2016), a popular privacy-preserving ERM mechanism described in Algorithm 2. Appendix G.3 provides its Rényi DP guarantees.

Algorithm 2 Noisy-GD: Noisy Gradient Descent

Require: Database $\mathcal{D} \in \mathcal{X}^n$, model $\Theta \in \mathbb{R}^d$, number of iterations $K \in \mathbb{N}$.
1: Initialize $\Theta_0 = \Theta$
2: **for** $k = 0, 1, \cdots, K-1$ **do**
3: $\quad \nabla \mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) = \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{D}} \nabla \boldsymbol{\ell}(\Theta_{\eta k}; \mathbf{x}) + \nabla \mathbf{r}(\Theta_{\eta k})$
4: $\quad \Theta_{\eta(k+1)} = \Theta_{\eta k} - \eta \nabla \mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) + \sqrt{2\eta} \mathcal{N}\left(0, \sigma^2 \mathbb{I}_d\right)$
5: **end for**
6: **return** $\Theta_{\eta K}$

Our proposed approach falls under the Descent-to-Delete framework proposed by Neel et al. (2021), wherein, after each deletion request u_i, we run Noisy-GD starting from the previous model $\hat{\Theta}_{i-1}$ and perform a small number of gradient descent steps over records in the modified database $\mathcal{D}_i = \mathcal{D}_{i-1} \circ u_i$; sufficient to erase information regarding deleted records in the subsequent model $\hat{\Theta}_i$. Our algorithms $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ is defined as follows.

Definition 5.1 (Noisy-GD based data-deletion solution)**.** *Let $K_{\mathrm{A}}, K_{\bar{\mathrm{A}}} \in \mathbb{N}$ and ρ be a Gaussian weight initialization distribution in $\mathbb{R}^d$. For any $\mathcal{D} \in \mathcal{X}^n$, our learning algorithm $\mathrm{A}_{\textit{Noisy-GD}} : \mathcal{X}^n \to \mathbb{R}^d$ is defined as*

$$\mathrm{A}_{\textit{Noisy-GD}}(\mathcal{D}) = \textit{Noisy-GD}(\mathcal{D}, \Theta, K_{\mathrm{A}}), \tag{16}$$

where $\Theta \sim \rho$. And, for any edit request $u \in \mathcal{U}^r$ on database $\mathcal{D} \in \mathcal{X}^n$ and any model $\Theta \in \mathbb{R}^d$, our unlearning algorithm $\bar{\mathrm{A}}_{\textit{Noisy-GD}} : \mathcal{X}^n \times \mathcal{U}^r \times \mathbb{R}^d \to \mathbb{R}^d$ is defined as

$$\bar{\mathrm{A}}_{\textit{Noisy-GD}}(\mathcal{D}, u, \Theta) = \textit{Noisy-GD}(\mathcal{D} \circ u_i, \Theta, K_{\bar{\mathrm{A}}}). \tag{17}$$

Our curator $(\mathrm{A}_{\textit{Noisy-GD}}, \bar{\mathrm{A}}_{\textit{Noisy-GD}})$ with any initial database $\mathcal{D}_0 \in \mathcal{X}^n$ interacts with any update requester $\mathcal{Q}$ as described in Algorithm 1 with publish function $f_{\mathrm{pub}}(\theta) = \theta$.

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

5.1. Deletion and Utility Under Convexity

We give the following guarantees on the algorithm pair $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ when loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is convex.

Theorem 5.1 (Utility, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda, \beta, L > 0$, constant $q > 1$, and constants $\varepsilon_{\mathrm{dp}} \geq \varepsilon_{\mathrm{dd}} > 0$. Define constant $\kappa = \frac{\lambda+\beta}{\lambda}$. 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 $\sigma^2 = \frac{4qL^2}{\lambda \varepsilon_{\mathrm{dp}} n^2}$, and the weight initialization distribution is $\rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda(1-\eta\lambda/2)\mathbb{I}_d}\right)$, then*

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

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

$$\textit{if} \quad K_{\bar{\mathrm{A}}} \geq 4\kappa \log \frac{\varepsilon_{\mathrm{dp}}}{\varepsilon_{\mathrm{dd}}}, \tag{18}$$

(3.) *and all models in sequence $(\hat{\Theta}_i)_{i \geq 0}$ produced by the interactions between $(\mathrm{A}_{\textit{Noisy-GD}}, \bar{\mathrm{A}}_{\textit{Noisy-GD}})$ and $\mathcal{Q}$ on any $\mathcal{D}_0 \in \mathcal{X}^n$, where $\mathcal{Q}$ is any r-requester, have an excess empirical risk $\mathrm{err}(\hat{\Theta}_i; \mathcal{D}_i) = O\left(\frac{qd}{\varepsilon_{\mathrm{dp}} n^2}\right)$ if*

$$\begin{aligned} K_{\mathrm{A}} &\geq 4\kappa \log\left(\frac{\varepsilon_{\mathrm{dp}} n^2}{4qd}\right), \quad \textit{and} \\ K_{\bar{\mathrm{A}}} &\geq 4\kappa \log \max\left\{5\kappa, \frac{8\varepsilon_{\mathrm{dp}} r^2}{qd}\right\}. \end{aligned} \tag{19}$$

Our utility upper bound in Theorem 5.1 matches the theoretical lower bound of $\Omega(\min\{1, \frac{d}{\varepsilon^2 n^2}\})$ by Bassily et al. (2014) on the optimal empirical risk attainable by any (ε, δ)-DP algorithms on Lipschitz, smooth, strongly-convex loss functions[6]. Consequently, our unlearning algorithm $\bar{\mathrm{A}}_{\text{Noisy-GD}}$ maintains the optimal privacy-utility tradeoffs of retraining with $\mathrm{A}_{\text{Noisy-GD}}$, while being a lot cheaper. Specifically, it offers a saving of $\Omega(n \log \min\{\frac{n}{r}, n\sqrt{\frac{\varepsilon_{\mathrm{dd}}}{qd}}\})$ in gradient complexity per-request (i.e., $n(K_{\mathrm{A}} - K_{\bar{\mathrm{A}}})$) while guaranteeing adaptive deletion privacy, differential privacy and optimal utility. This level of saving surpasses that of all existing unlearning algorithms known to us, and we provide a detailed comparison in Table 1.

It is worth noting that in order to satisfy $(q, \varepsilon_{\mathrm{dp}})$-Rényi differential privacy and $(q, \varepsilon_{\mathrm{dd}})$-deletion privacy for non-adaptive r-requesters, the necessary number of iterations $K_{\bar{\mathrm{A}}}$ remains constant regardless of the size, r, of the deletion batch, depending solely on the ratio $\frac{\varepsilon_{\mathrm{dd}}}{\varepsilon_{\mathrm{dp}}}$. 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 $r \leq \sqrt{\frac{qd}{\varepsilon_{\mathrm{dd}}}}$, performing an adequate number of unlearning iterations to satisfy the deletion privacy guarantee is also sufficient to ensure optimal utility of the unlearned model.

[6] Refer to Theorem B.1 to see that $(q, \varepsilon_{\mathrm{dp}})$-Rényi DP implies (ε, δ)-DP for $q = 1 + \frac{2}{\varepsilon}\log(1/\delta)$ and $\varepsilon_{\mathrm{dp}} = \varepsilon/2$. When $\varepsilon = \Theta(\log(1/\delta))$, one can evaluate that $\frac{q}{\varepsilon_{\mathrm{dp}}} = \Theta(\frac{\log(1/\delta)}{\varepsilon^2})$.

Unlearning Algorithm	Requires internal states that depend on deleted records?	Compute savings for ith edit
Noisy-m-A-SGD [Thm. 1, (Ullah et al., 2021)]	No	$\Omega\left(\sqrt{d}\left(1-\frac{\sqrt{d}}{n}\right)\right)$
Perturbed-GD [Thm. 9, (Neel et al., 2021)]	Yes	$\Omega\left(n\log\left(\frac{\varepsilon n}{\sqrt{d}}\right)\right)$
Perturbed-GD [Thm. 28, (Neel et al., 2021)]	No	$\Omega\left(n\log\left(\frac{\varepsilon n}{\log^2(id)\sqrt{d}}\right)\right)$
Noisy-GD [Thm. 5.1, Ours]	No	$\Omega\left(n\log\min\left\{n,\frac{\varepsilon n}{\sqrt{d}}\right\}\right)$

Table 1: Comparison 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 databases. 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 $(1+\frac{2}{\varepsilon}\log(1/\delta),\frac{\varepsilon}{2})$-deletion-privacy guarantee (which implies one-sided (ε,δ)-unlearning (cf. Theorem B.1 & Remark 4.2)) and have the same excess empirical risk $\alpha=O(1)$.

5.2. Deletion and Utility under Non-Convexity

For a non-convex loss function $\boldsymbol{\ell}(\theta;\mathbf{x})$, we provide the following set of guarantees on the pair $(\mathrm{A}_{\text{Noisy-GD}},\bar{\mathrm{A}}_{\text{Noisy-GD}})$.

Theorem 5.2 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda,\beta,L,\sigma^2,\eta>0$, constants $q,B>1$, and constants $d>\varepsilon_{\mathrm{dp}}\geq\varepsilon_{\mathrm{dd}}>0$. 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 $\rho=\mathcal{N}\left(0,\frac{\sigma^2}{\lambda}\mathbb{I}_d\right)$. Then,*

(1.) *both $\mathrm{A}_{\text{Noisy-GD}}$ and $\bar{\mathrm{A}}_{\text{Noisy-GD}}$ are $(q,\varepsilon_{\mathrm{dp}})$-Rényi DP for any $\eta\geq0$ and any $K_{\mathrm{A}},K_{\bar{\mathrm{A}}}\geq0$ if*

$$\sigma^2\geq\frac{qL^2}{\varepsilon_{\mathrm{dp}}n^2}\cdot\eta\max\{K_{\mathrm{A}},K_{\bar{\mathrm{A}}}\},\tag{20}$$

(2.) *pair $(\mathrm{A}_{\text{Noisy-GD}},\bar{\mathrm{A}}_{\text{Noisy-GD}})$ satisfies $(q,\varepsilon_{\mathrm{dd}})$-deletion-privacy under all non-adaptive r-requesters for any $\sigma^2>0$, if learning rate is $\eta\leq\frac{\lambda\varepsilon_{\mathrm{dd}}}{64dqB(\beta+\lambda)^2}$ and the number of iterations satisfy*

$$\begin{aligned}K_{\mathrm{A}}&\geq\frac{2B}{\lambda\eta}\log\left(\frac{q\log(B)}{\varepsilon_{\mathrm{dd}}}\right),\ \textit{and}\\ K_{\bar{\mathrm{A}}}&\geq K_{\mathrm{A}}-\frac{2B}{\lambda\eta}\log\left(\frac{\log(B)}{2\left(\varepsilon_{\mathrm{dd}}+\frac{r}{n}\log(B)\right)}\right),\end{aligned}\tag{21}$$

(3.) *and all models in the sequence $(\hat{\Theta}_i)_{i\geq0}$ produced by interactions between $(\mathrm{A}_{\text{Noisy-GD}},\bar{\mathrm{A}}_{\text{Noisy-GD}})$ and $\mathcal{Q}$ on any $\mathcal{D}_0\in\mathcal{X}^n$, where $\mathcal{Q}$ is an r-requester, have an excess empirical risk $\mathrm{err}(\hat{\Theta}_i;\mathcal{D}_i)=\tilde{O}\left(\frac{dq}{\varepsilon_{\mathrm{dp}}n^2}+\frac{1}{n}\sqrt{\frac{q\varepsilon_{\mathrm{dd}}}{\varepsilon_{\mathrm{dp}}}}\right)$ when inequalities in* (21) *and* (20) *are equalities.*

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 $\tilde{O}\big(\frac{1}{n}\sqrt{\frac{q\varepsilon_{\mathrm{dd}}}{\varepsilon_{\mathrm{dp}}}}\big)$. 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 $\bar{\mathrm{A}}_{\text{Noisy-GD}}$ in scenarios where the proportion of edited records in a single edit request satisfies $\frac{r}{n}\leq\frac{1}{2}-\frac{\varepsilon_{\mathrm{dd}}}{\log B}$. For instance, in the deletion regime where we desire $\varepsilon_{\mathrm{dd}}=\log(B)/4$, employing $\bar{\mathrm{A}}_{\text{Noisy-GD}}$ instead of retraining with $\mathrm{A}_{\text{Noisy-GD}}$ requires $\Omega(dn\log\frac{n}{r})$ fewer gradient steps.

Remark 5.3. *Both Theorems 5.1 and 5.2 also hold when gradients $\nabla\boldsymbol{\ell}(\theta;\mathbf{x})$ are clipped to L instead of assuming L-Lipschitzness. Appendix F.1 discusses how gradient clipping is compatible with other assumptions we make.*

6. Conclusions

We showed that prior unlearning certifications in literature are unreliable in real-world scenarios, and proposed a new deletion privacy guarantee that safeguards the "Right to be Forgotten". We also showed the perils of caching partial computations, 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.

Acknowledgements

The authors would like to thank Martin Strobel, Hannah Brown and anonymous reviewers for helpful discussions on earlier versions of this paper.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Adelman, K. and Adelman, G. California coastal records project, 2002.

Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473. IEEE, 2014.

Bobkov, S. G. On isoperimetric constants for log-concave probability distributions. In *Geometric aspects of functional analysis*, pp. 81–88. Springer, 2007.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

California Consumer Privacy Act. Title 1.81.5. california consumer privacy act of 2018 [1798.100 - 1798.199.100], 2018.

Chen, M., Zhang, Z., Wang, T., Backes, M., Humbert, M., and Zhang, Y. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 896–911, 2021.

Chewi, S., Erdogdu, M. A., Li, M. B., Shen, R., and Zhang, M. Analysis of langevin monte carlo from poincar\'e to log-sobolev. *arXiv preprint arXiv:2112.12662*, 2021.

Chourasia, R., Ye, J., and Shokri, R. Differential privacy dynamics of langevin diffusion and noisy gradient descent. *Advances in Neural Information Processing Systems*, 34, 2021.

Donsker, M. D. and Varadhan, S. S. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36 (2):183–212, 1983.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

General Data Protection Regulation. Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016, 2016.

Ginart, A., Guan, M., Valiant, G., and Zou, J. Y. Making ai forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Gross, L. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Gupta, V., Jung, C., Neel, S., Roth, A., Sharifi-Malvajerdi, S., and Waites, C. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34, 2021.

Holley, R. and Stroock, D. W. Logarithmic sobolev inequalities and stochastic ising models. 1986.

Izzo, Z., Smart, M. A., Chaudhuri, K., and Zou, J. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR, 2021.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Ledoux, M. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Neel, S., Roth, A., and Sharifi-Malvajerdi, S. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Nekvinda, A. and Zajíček, L. A simple proof of the rademacher theorem. *Časopis pro pěstování matematiky*, 113(4):337–341, 1988.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Otto, F. and Villani, C. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Rényi, A. 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.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. Remember what you want to forget: Algorithms for machine unlearning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18075–18086. Curran Associates, Inc., 2021.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Ullah, E., Mai, T., Rao, A., Rossi, R. A., and Arora, R. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pp. 4126–4142. PMLR, 2021.

Van Erven, T. and Harremos, P. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Vaserstein, L. N. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Wang, Y.-X., Fienberg, S., and Smola, A. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pp. 2493–2502. PMLR, 2015.

Appendix

Table of Contents

A. Table of Notations

Table 2: Symbol reference.

Symbol	Meaning
$\mathcal{O}$	Arbitrary model parameter space.
Φ	Space of publishable objects.
$d, \mathbb{R}^d$	Dimension of model parameters and d-dimensional Euclidean space.
n	Database size.
$\mathcal{X}, \mathcal{X}^n$	Data universe and Domain of all datasets of size n.
ν, ν', π, μ	Arbitrary distributions on $\mathcal{O}$ or on $\mathbb{R}^d$.
$\mathcal{Q}$	An edit requester.
r, p	Integers representing the power of an adaptive requester.
$\mathcal{U}, \mathcal{U}^r$	Space of singular and batched replacement edits in $[n] \times \mathcal{X}$.
u, u_i, U_i	Arbitrary edit request, i^{th} edit request in $\mathcal{U}^r$ and its random variable.
$\mathcal{D}, \mathcal{D}_i$	An example database and database after i^{th} update.
$\mathbf{x}, \mathbf{y}$	Singular data records from universe $\mathcal{X}$.
η	Step size or learning rate in Noisy-GD.
σ^2	Variance scaling used in weight initialization distribution or gradient noise.
$\boldsymbol{\ell}(\theta; \mathbf{x})$	Twice continuously differentiable loss function on models in $\mathbb{R}^d$.
$\mathbf{r}(\theta)$	$L2$ regularizer $\lambda \left\|\theta\right\|_2^2 / 2$.
$\mathcal{L}(\theta), \mathcal{L}_{\mathcal{D}}(\theta)$	Arbitrary optimization objective and an $\mathbf{r}(\theta)$ regularized objective on $\mathcal{D}$ over $\boldsymbol{\ell}(\theta; \mathbf{x})$.
$\mathrm{err}(\Theta; \mathcal{D})$	Excess empirical risk of random model Θ over objective $\mathcal{L}_{\mathcal{D}}$.
$\pi(\mathcal{D})$	An mapping from $\mathcal{X}^n$ to distributions on $\mathbb{R}^d$; sometimes distributions are Gibbs.
$\Lambda_{\mathcal{D}}$	Normalization constant of the Gibbs distribution $\pi(\mathcal{D})$.
$\pi_i^{\mathcal{Q}}$	An imaginary mechanism designed to prove deletion privacy of a data-deletion solution.
T_k	A map over $\mathbb{R}^d$.
ρ	Weight initialization distribution for Noisy-GD.
$\mathbf{v}, \mathbf{v}'$	Vector fields on $\mathbb{R}^d$.
$\theta^*_{\mathcal{D}}, \theta^*_{\mathcal{D}_i}$	Risk minimizer for $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_{\mathcal{D}_i}$.
q	Order of Rényi divergence.
$\varepsilon_{\mathrm{dp}}, \varepsilon_{\mathrm{dd}}$	Differential privacy budget and deletion privacy budget in q-Rényi divergence.
ε, δ	Parameters for DP-like indistinguishability.
$\mathrm{A}, \mathrm{A}_{\text{Noisy-GD}}$	Learning algorithm and Noisy-GD based learning algorithm respectively.
$\bar{\mathrm{A}}, \bar{\mathrm{A}}_{\text{Noisy-GD}}$	Data-deletion algorithm and Noisy-GD based unlearning algorithm respectively.
$K_{\mathrm{A}}, K_{\bar{\mathrm{A}}}$	Number of learning and data-deletion iterations in Noisy-GD.
k, t	Index of a Noisy-GD iteration and continuous time variable for tracing diffusions.
$\Theta_{\eta k}, \Theta'_{\eta k}$	Parameters at iteration k of Noisy-GD.
Θ_t, Θ'_t	Parameters at time t of tracing diffusion for Noisy-GD.
μ_t, μ'_t	Probability density for Θ_t, Θ'_t.
$\mathbb{I}_d$	d-dimensional identity matrix.
$\mathbf{Z}, \mathbf{Z}_k, \mathbf{Z}'_k$	Random variables taken from $\mathcal{N}(0, \mathbb{I}_d)$.
$\mathrm{d}\mathbf{Z}_t, \mathrm{d}\mathbf{Z}'_t$	Two independent Weiner process.
λ, β, B, L	$L2$ regularizer constant and smoothness, boundedness, and Lipschitzness constants.
$\mathrm{Clip}_L(\cdot)$	Operator that clips vectors in $\mathbb{R}^d$ to a magnitude of L.
$\mathrm{R}_q(\nu \Vert \nu'), \mathrm{E}_q(\nu \Vert \nu')$	Rényi divergence and q^{th} moment of likelihood ratio r.v. between ν and ν'.
$\mathrm{I}(\nu \Vert \nu'), \mathrm{I}_q(\nu \Vert \nu')$	Fisher and q-Rényi Information of distribution of ν w.r.t ν'.
$\mathrm{W}_2(\nu, \nu')$	Wasserstein distance between distribution ν and ν'.
$\mathrm{KL}(\nu \Vert \nu')$	Kullback-Leibler divergence of distribution ν w.r.t. ν'.
$P_t, \mathcal{G}, \mathcal{G}^*$	Markov semigroup, its infinitesimal generator, and its Fokker-Planck operator.
$\mathrm{Ent}_\pi(f^2)$	Entropy of function f^2 under any arbitrary distribution π.
$\mathrm{H}(\cdot)$	Differential entropy of a distribution.
$\mathrm{LS}(c)$	Log-sobolev inequality with constant c.

B. Divergence Measures and Their Properties

Let $\Theta, \Theta' \in \mathcal{O}$ be two random variables with probability measures ν, ν' respectively. We abuse the notations to denote respective probability densities with ν, ν' as well. We say that ν is absolutely continuous with respect to ν' (denoted by $\nu \ll \nu'$) if for all measurable sets $O \subset \mathcal{O}$, $\nu(O) = 0$ whenever $\nu'(O) = 0$.

Definition B.1 ((ε, δ)-indistinguishability (Dwork et al., 2014))**.** *We say ν and ν' are (ε, δ)-*indistinguishable *if for all $O \subset \mathcal{O}$,*

$$\mathop{\mathbb{P}}_{\Theta\sim\nu}[\Theta \in O] \leq e^{\varepsilon} \mathop{\mathbb{P}}_{\Theta'\sim\nu'}[\Theta' \in O] + \delta \quad \textit{and} \quad \mathop{\mathbb{P}}_{\Theta'\sim\nu'}[\Theta' \in O] \leq e^{\varepsilon} \mathop{\mathbb{P}}_{\Theta\sim\nu}[\Theta \in O] + \delta. \tag{22}$$

In this paper, we measure indistinguishability in terms of Rényi divergence.

Definition B.2 (Rényi divergence (Rényi et al., 1961))**.** Rényi divergence *of ν w.r.t. ν' of order $q > 1$ is defined as*

$$\mathrm{R}_q(\nu\|\nu') = \frac{1}{q-1}\log \mathrm{E}_q(\nu\|\nu'), \quad \textit{where} \quad \mathrm{E}_q(\nu\|\nu') = \mathop{\mathbb{E}}_{\theta\sim\nu'}\left[\left(\frac{\nu(\theta)}{\nu'(\theta)}\right)^q\right], \tag{23}$$

when ν is absolutely continuous *w.r.t. ν' (denoted as $\nu \ll \nu'$). If $\nu \not\ll \nu'$, we'll say $\mathrm{R}_q(\nu\|\nu') = \infty$. We abuse the notation $\mathrm{R}_q(\Theta\|\Theta')$ to denote divergence $\mathrm{R}_q(\nu\|\nu')$ between the measures of Θ, Θ'.*

A bound on Rényi divergence implies a one-directional (ε, δ)-indistinguishability as described below.

Theorem B.1 (Conversion theorem of Rényi divergence (Mironov, 2017, Proposition 3))**.** *Let $q > 1$ and $\varepsilon > 0$. If distributions ν, ν' satisfy $\mathrm{R}_q(\nu\|\nu') < \varepsilon_0$, then for any $O \subset \mathcal{O}$,*

$$\mathop{\mathbb{P}}_{\Theta\sim\nu}[\Theta \in O] \leq e^{\varepsilon} \mathop{\mathbb{P}}_{\Theta'\sim\nu'}[\Theta' \in O] + \delta, \tag{24}$$

for $\varepsilon = \varepsilon_0 + \frac{\log 1/\delta}{q-1}$ and any $0 < \delta < 1$.

We use the following properties of Rényi divergence in some of our proofs.

Theorem B.2 (Mononicity of Rényi divergence (Mironov, 2017, Proposition 9))**.** *For $1 \leq q_0 < q$, and arbitrary probability measures ν and ν' over $\mathcal{O}$, $\mathrm{R}_{q_0}(\nu\|\nu') \leq \mathrm{R}_q(\nu\|\nu')$.*

Theorem B.3 (Rényi composition (Mironov, 2017, Proposition 1))**.** *If $\mathrm{A}_1, \cdots, \mathrm{A}_k$ are randomized algorithms satisfying, respectively, (q, ε_1)-Rényi DP, $\cdots$, (q, ε_k)-Rényi DP then their composed mechanism defined as $(\mathrm{A}_1(\mathcal{D}), \cdots, \mathrm{A}_k(\mathcal{D}))$ is $(q, \varepsilon_1 + \cdots + \varepsilon_k)$-Rényi DP. Moreover, i^{th} algorithm can be chosen on the basis of the outputs of algorithms $\mathrm{A}_1, \cdots, \mathrm{A}_{i-1}$.*

Theorem B.4 (Weak triangle inequality of Rényi divergence (Mironov, 2017, Proposition 12))**.** *For any distribution ρ on $\mathcal{O}$, the Rényi divergence of ν w.r.t. ν' satisfies the following weak triangle inequality:*

$$\mathrm{R}_q(\nu\|\nu') \leq \mathrm{R}_q(\nu\|\rho) + \mathrm{R}_\infty(\rho\|\nu'). \tag{25}$$

Another popular notion of information divergence is the Kullback-Leibler divergence.

Definition B.3 (Kullback-Leibler divergence (Kullback & Leibler, 1951))**.** Kullback-Leibler *(KL) divergence* $\mathrm{KL}(\nu\|\nu')$ *of ν w.r.t. ν' is defined as*

$$\mathrm{KL}(\nu\|\nu') = \mathop{\mathbb{E}}_{\theta\sim\nu}\left[\log\frac{\nu(\theta)}{\nu'(\theta)}\right]. \tag{26}$$

Rényi divergence generalizes Kullback-Leibler divergence as $\lim_{q\to1}\mathrm{R}_q(\nu\|\nu') = \mathrm{KL}(\nu\|\nu')$ (Van Erven & Harremos, 2014).

Some other divergence notions that we rely on are the following.

Definition B.4 (Wasserstein distance (Vaserstein, 1969))**.** Wasserstein distance *between ν and ν' is*

$$\mathrm{W}_2(\nu, \nu') = \inf_{\Pi} \mathop{\mathbb{E}}_{\Theta,\Theta'\sim\Pi}\left[\|\Theta - \Theta'\|_2^2\right]^{\frac{1}{2}}, \tag{27}$$

where Π is any joint distribution on $\mathcal{O} \times \mathcal{O}$ with ν and ν' as its marginal distributions.

Definition B.5 (Relative Fisher information (Otto & Villani, 2000))**.** *If* $\nu \ll \nu'$ *and* $\frac{\nu}{\nu'}$ *is differentiable, then* relative Fisher information *of* ν *with respect to* ν' *is defined as*

$$\mathrm{I}\left(\nu \| \nu'\right) = \mathop{\mathbb{E}}_{\theta \sim \nu} \left[\left\| \nabla \log \frac{\nu(\theta)}{\nu'(\theta)} \right\|_2^2 \right]. \tag{28}$$

Definition B.6 (Relative Rényi information (Vempala & Wibisono, 2019))**.** *Let* $q > 1$*. If* $\nu \ll \nu'$ *and* $\frac{\nu}{\nu'}$ *is differentiable, then* relative Rényi information *of* ν *with respect to* ν' *is defined as*

$$\mathrm{I}_q\left(\nu \| \nu'\right) = \frac{4}{q^2} \mathop{\mathbb{E}}_{\theta \sim \nu'} \left[\left\| \nabla \left(\frac{\nu(\theta)}{\nu'(\theta)} \right)^{q/2} \right\|_2^2 \right] = \mathop{\mathbb{E}}_{\theta \sim \nu'} \left[\left(\frac{\nu(\theta)}{\nu'(\theta)} \right)^{q-2} \left\| \nabla \left(\frac{\nu(\theta)}{\nu'(\theta)} \right) \right\|_2^2 \right]. \tag{29}$$

C. Proofs for Section 3

Theorem 3.1. *There exists an algorithm pair* $(\mathrm{A}, \bar{\mathrm{A}})$ *satisfying* $(0,0)$*-adaptive-unlearning under publish function* $f_{\mathrm{pub}}(\theta) = \theta$ *such that by designing a* 1*-adaptive* 1*-requester* $\mathcal{Q}$*, 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* $\mathcal{Q}$*'s transcript* $(\phi_{<i}; u_{<i})$*.*

Proof. Let data universe $\mathcal{X}$, the internal model space $\mathcal{O}$, as well as publishable outcome space Φ be $\mathbb{R}$. Consider the task of releasing a sequence of medians using function $\mathrm{med} : \mathbb{R}^* \to \mathbb{R}$ in the online setting when the initial database $\mathcal{D}_0 \in \mathcal{X}^n$ is being modified by some adaptive requester $\mathcal{Q}$. Given a database $\mathcal{D} \in \mathcal{X}^n$, our learning algorithm is defined as $\mathrm{A}(\mathcal{D}) = \mathrm{med}(\mathcal{D})$. For an arbitrary edit request $u \in \mathcal{U}^r$, our unlearning algorithm is defined as $\bar{\mathrm{A}}(\mathcal{D}, u, \bullet) = \mathrm{med}(\mathcal{D} \circ u)$ for any $\bullet \in \mathcal{O}$. Let the publish function $f_{\mathrm{pub}} : \mathcal{O} \to \Phi$ be an identity function, i.e. $f_{\mathrm{pub}}(\theta) = \theta$.

For any initial database $\mathcal{D}_0 \in \mathcal{X}^n$ and an adaptive sequence $(u_i)_{i \geq 1}$ generated by any ∞-adaptive 1-requester $\mathcal{Q}$, note that

$$f_{\mathrm{pub}}(\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \bullet)) = f_{\mathrm{pub}}(\mathrm{A}(\mathcal{D}_i)), \quad \text{for all } i \geq 1 \text{ and any } \bullet \in \mathcal{O}. \tag{30}$$

Therefore, $\bar{\mathrm{A}}$ is a $(0,0)$-adaptive unlearning algorithm for A under f_{pub}.

Now suppose that n is odd and $\mathcal{D}_0$ consists of unique entries. W.L.O.G assume that the median record $\mathrm{med}(\mathcal{D}_0)$ is at index ind^m and its owner will be deleting it at step i by sending a non-adaptive edit request $u_i = \{\langle \mathrm{ind}^m, \mathbf{y} \rangle\}$ such that $\mathbf{y} \neq \mathrm{med}(\mathcal{D}_0)$. We design the following 1-adaptive 1-requester $\mathcal{Q}$ 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., $\mathrm{med}(\mathcal{D}_i) = \mathrm{med}(\mathcal{D}_0)$:

$$\mathcal{Q}(\phi_0, u_1, u_2, \cdots, u_{j-1}) = \{\langle \mathrm{ind}_j, \phi_0 \rangle\} \quad \forall\, 1 \leq j < i, \tag{31}$$

where ind_j is randomly sampled from $[n] \setminus \{\mathrm{ind}_1, \cdots, \mathrm{ind}_{j-1}\}$ without replacement. Note that by the end of interaction, $\mathcal{Q}$ replaces at-least $i-2$ unique records in $\mathcal{D}_0$ with $\phi_0 = \mathrm{med}(\mathcal{D}_0)$. If one of those original records was larger than $\mathrm{med}(\mathcal{D}_0)$ and another was smaller than $\mathrm{med}(\mathcal{D}_0)$, then it is guaranteed that $\mathrm{med}(\mathcal{D}_i) = \mathrm{med}(\mathcal{D}_0)$. Therefore, $\mathbb{P}\left[\mathrm{med}(\mathcal{D}_i) = \mathrm{med}(\mathcal{D}_0)\right]$ is at-least

$$\begin{aligned} \mathbb{P}\left[\exists \mathrm{ind}^l, \mathrm{ind}^u \in \{\mathrm{ind}_1, \cdots, \mathrm{ind}_{i-1}\} \text{ s.t. } \mathcal{D}_0[\mathrm{ind}^l] < \mathcal{D}_0[\mathrm{ind}^m] < \mathcal{D}_0[\mathrm{ind}^u] \right] &\geq 1 - 2 \times \binom{\lfloor n \rfloor / 2}{i-2} \Big/ \binom{n}{i-2} \\ &\geq 1 - \left(\frac{1}{2}\right)^{i-3}. \end{aligned}$$

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 $\bar{\mathrm{A}}$ for deletion.

□

Theorem 3.2. *For every* $\varepsilon > 0$*, there exists a pair* $(\mathrm{A}, \bar{\mathrm{A}})$ *of algorithms that satisfy* $(\varepsilon, 0)$*-unlearning under some publish function* f_{pub} *such that for all non-adaptive* 1*-requesters* $\mathcal{Q}$*, 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 the post-edit releases* $\phi_{\geq i}$*.*

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 $\mathcal{Q}$. Since $\mathcal{Q}$ is non-adaptive by assumption, it is equivalent to the entire edit sequence $\{u_i\}_{i\geq 1}$ being fixed before interaction. We design an algorithm pair $(\mathrm{A}, \bar{\mathrm{A}})$ for this task with secret model space being $\mathcal{O} = \mathbb{N}^3$ and published outcome space being $\Phi = \mathbb{R}$, with the publish function being $f_{\mathrm{pub}}(\langle a, b, c\rangle) = a + b/c + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right)$ (with the convention that $b/c = 0$ if $b = c = 0$). At any step $i \geq 0$, our internal model $\hat{\Theta}_i = \langle \mathrm{cnt}_i, \mathrm{del}_i, i\rangle$ encodes the current count of h on database $\mathcal{D}_i$, the count of h on records previously deleted by $u_{\leq i}$, and the current step index i. Our learning algorithm initializes the secret model as $\hat{\Theta}_0 = \mathrm{A}(\mathcal{D}_0) = \langle \sum_{\mathbf{x}\in\mathcal{D}_0} h(\mathbf{x}), 0, 0\rangle$, and, for an edit request $u_i = \{\langle \mathrm{ind}_i, \mathbf{y}_i\rangle\}$, our algorithm $\bar{\mathrm{A}}$ updates the secret model $\hat{\Theta}_{i-1} \to \hat{\Theta}_i$ following the rule

$$\hat{\Theta}_i = \bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) = \langle \mathrm{cnt}_i, \mathrm{del}_i, i\rangle \text{ where } \begin{cases} \mathrm{cnt}_i = \mathrm{cnt}_{i-1} + h(\mathbf{y}_i) - h(\mathcal{D}_{i-1}[\mathrm{ind}_i]), \\ \mathrm{del}_i = \mathrm{del}_{i-1} + h(\mathcal{D}_{i-1}[\mathrm{ind}_i]). \end{cases}$$

Note that $\forall i \geq 1$, $\Delta_i \stackrel{\text{def}}{=} \mathrm{del}_i/i \in [0, 1]$. Therefore, from properties of Laplace mechanism (Dwork et al., 2014), it is straightforward to see that for all $i \geq 1$,

$$\begin{aligned} f_{\mathrm{pub}}(\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}))\big| u_{\leq i} &= \sum_{\mathbf{x}\in\mathcal{D}_i} h(\mathbf{x}) + \Delta_i + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right) \\ &\stackrel{\varepsilon,0}{\approx} \sum_{\mathbf{x}\in\mathcal{D}_i} h(\mathbf{x}) + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right) = f_{\mathrm{pub}}(\mathrm{A}(\mathcal{D}_i)). \end{aligned}$$

Hence, $\bar{\mathrm{A}}$ is an $(\varepsilon, 0)$-unlearning algorithm for 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 $\mathcal{D}'_{i-1}$ that differs from $\mathcal{D}_{i-1}$ only at index ind_i such that $h(\mathcal{D}'_{i-1}[\mathrm{ind}_i]) \neq h(\mathcal{D}_{i-1}[\mathrm{ind}_i])$. Let random variable sequences $\phi_{\geq i}$ and $\phi'_{\geq i}$ denote the releases by $\bar{\mathrm{A}}$ in the scenarios that the $(i-1)^{\text{th}}$ database was $\mathcal{D}_{i-1}$ and $\mathcal{D}'_{i-1}$ 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. $\mathcal{D}_{j-1} \circ u_j = \mathcal{D}'_{j-1} \circ u_j$ 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,

$$\max_{O\subset\Phi^*} \log \frac{\mathbb{P}\left[\phi_{\geq i} \in O\right]}{\mathbb{P}\left[\phi'_{\geq i} \in O\right]} = \sum_{j=i}^{\infty} \max_{O_j\subset\mathbb{R}} \log \frac{\mathbb{P}\left[\phi_j \in O_j\right]}{\mathbb{P}\left[\phi'_j \in O_j\right]} = \sum_{j=i}^{\infty} \log e^{\varepsilon/j} = \infty.$$

□

C.1. Unsoundness and Incompleteness of Offline Unlearning Definitions

In this subsection, we show that our criticisms on trustworthiness of unlearning notions under adaptive requests in Section 3 also apply to the other popular deletion privacy notions like Ginart et al. (2019), Guo et al. (2019) and Sekhari et al. (2021).

Definition C.1 (Data deletion operation (Ginart et al., 2019))**.** *Algorithm* $\bar{\mathrm{A}}$ *is a* data deletion operation for a learning algorithm A *if* $\bar{\mathrm{A}}(\mathcal{D}, S, \mathrm{A}(\mathcal{D})) \stackrel{0,0}{\approx} \mathrm{A}(\mathcal{D} \setminus S)$ *for all* $\mathcal{D} \subset \mathcal{X}$ *and all subsets* $S \subset \mathcal{D}$ *that is* selected independently *of* $\mathrm{A}(\mathcal{D})$.

Definition C.2 ((ε, δ)-certified removal (Guo et al., 2019))**.** *A removal mechanism* $\bar{\mathrm{A}}$ *performs* (ε, δ)-certified removal *for learning algorithm* A *if for all databases* $\mathcal{D} \subset \mathcal{X}$ *and deletion subsets* $S \subset \mathcal{D}$*,*

$$\bar{\mathrm{A}}(\mathcal{D}, S, \mathrm{A}(\mathcal{D})) \stackrel{\varepsilon,\delta}{\approx} \mathrm{A}(\mathcal{D} \setminus S). \tag{32}$$

Definition C.3 ((ε, δ)-unlearning (Sekhari et al., 2021))**.** *For all* $\mathcal{D} \subset \mathcal{X}$ *of size* n *and deletion subsets* $S \subset \mathcal{D}$ *such that* $|S| \leq m$*, a learning algorithm* A *and an unlearning algorithm* $\bar{\mathrm{A}}$ *is* (ε, δ)-unlearning *if*

$$\bar{\mathrm{A}}(T(\mathcal{D}), S, \mathrm{A}(\mathcal{D})) \stackrel{\varepsilon,\delta}{\approx} \bar{\mathrm{A}}(T(\mathcal{D} \setminus S), \varnothing, \mathrm{A}(\mathcal{D} \setminus S)), \tag{33}$$

where $\varnothing$ *denotes the empty set and* $T(\mathcal{D})$ *denotes the data statistics available to* $\bar{\mathrm{A}}$ *about* $\mathcal{D}$*.*

Unsoundness. Definition C.1 explicitly assumes that there is no influence of the learned model $\mathrm{A}(\mathcal{D})$ 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 $\mathrm{A}(\mathcal{D})$ reveals about them. Therefore, Definition C.1's certification in these settings is trivially void. Unlike Definition C.1 however, Definitions C.2 and C.3 make no assumptions about dependence between the deletion request S and the learned model $\mathrm{A}(\mathcal{D})$. So, request S can depend on $\mathrm{A}(\mathcal{D})$ under these two certifications. We recall the example we provide in Section 3 to show that Definitions C.2 and C.3, are also unsound under adaptivity.

For the universe of records $\mathcal{X} = \{-2, -1, 1, 2\}$, consider the following learning and unlearning algorithms:

$$\mathrm{A}(\mathcal{D}) = \sum_{\mathbf{x}\in\mathcal{D}} \mathbf{x}, \quad \text{and} \quad \bar{\mathrm{A}}(\mathcal{D}, S, \mathrm{A}(\mathcal{D})) = \sum_{\mathbf{x}\in\mathcal{D}\setminus S} \mathbf{x}. \tag{34}$$

Note that for any $\mathcal{D} \subset \mathcal{X}$ and any $S \subset \mathcal{D}$, the above algorithm pair $(\mathrm{A}, \bar{\mathrm{A}})$ satisfies Definitions C.1, C.2 and C.3 for $\varepsilon = \delta = 0$ and $T(\mathcal{D}) = \mathcal{D}$. Suppose the adversary is aware that the following dependence holds between the learned model $\mathrm{A}(\mathcal{D})$ and deletion request S:

$$S = \begin{cases} \{\mathbf{x} < 0 : \forall \mathbf{x} \in \mathcal{X}\} & \text{if } \mathrm{A}(\mathcal{D}) < 0, \\ \{\mathbf{x} \geq 0 : \forall \mathbf{x} \in \mathcal{X}\} & \text{otherwise.} \end{cases} \tag{35}$$

Consider two neighbouring databases $\mathcal{D}_{-1} = \{-2, -1, 2\}$ and $\mathcal{D}_1 = \{-2, 1, 2\}$. Knowing the above dependence, an adversary can determine whether $\mathcal{D} = \mathcal{D}_{-1}$ or $\mathcal{D} = \mathcal{D}_1$ by looking only at $\bar{\mathrm{A}}(\mathcal{D}, S, \mathrm{A}(\mathcal{D}))$. This is because if $\mathcal{D} = \mathcal{D}_{-1}$, then the observation after unlearning is 2, and if $\mathcal{D} = \mathcal{D}_1$, the observation after unlearning is -2. So, even though $(\mathrm{A}, \bar{\mathrm{A}})$ satisfies the guarantees of Ginart et al. (2019), Guo et al. (2019) and Sekhari et al. (2021), it blatantly reveals the identity (-1 or 1) of a deleted record to an adversary observing only the post-deletion release.

Incompleteness. Definitions C.1, C.2 and C.3 are also incomplete. Consider an unlearning algorithm $\bar{\mathrm{A}}$ 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 $\bar{\mathrm{A}}$ is a valid deletion algorithm as its output does not depend on the input database $\mathcal{D}$ or the learned model $\mathrm{A}(\mathcal{D})$. However, note that $\bar{\mathrm{A}}$ does not satisfy the unlearning Definition C.3, for any learning algorithm A. And, for a learning algorithm $\mathrm{A}(\mathcal{D}) = \sum_{\mathbf{x}\in\mathcal{D}} \mathbf{x}$, one can verify that the pair $(\mathrm{A}, \bar{\mathrm{A}})$ does not satisfy Definitions C.1 and C.2 either.

D. Proofs for Section 4

Theorem 4.1 (Definition 4.1 safeguards RTBF)**.** *If the algorithm pair $(\mathrm{A}, \bar{\mathrm{A}})$ satisfies (q, ε)-deletion-privacy guarantee under all p-adaptive r-requesters, then even with complete knowledge of a p-adaptive r-requester $\mathcal{Q}$ that interacts with the curator before a target record $\mathcal{D}_0[\mathrm{ind}]$ in the initial database $\mathcal{D}_0$ is deleted at step $i \geq 1$ by request u_i, any attacker $MI : \mathcal{O}^* \to \{0, 1\}$ observing only the post-deletion models $\hat{\Theta}_{\geq i} = (\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots)$ has an advantage*

$$\mathit{Adv}(\mathit{MI}) \stackrel{\text{def}}{=} \mathbb{P}\left[\mathit{MI}(\hat{\Theta}_{\geq i}) = 1 \middle| \mathbf{x}\right] - \mathbb{P}\left[\mathit{MI}(\hat{\Theta}_{\geq i}) = 1 \middle| \mathbf{x}'\right] \tag{36}$$

for disambiguating between two possible values $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ of the deleted record $\mathcal{D}_0[\mathrm{ind}]$ bounded as follows.

$$\mathit{Adv}(\mathit{MI}) \leq \min\left\{\sqrt{2\varepsilon}, \frac{q e^{\varepsilon(q-1)/q}}{q-1}[2(q-1)]^{\frac{1}{q}} - 1\right\} \tag{37}$$

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 $\mathcal{D}_{i-1}$ with '$\mathbf{y}$'. In the worst case, this record $\mathcal{D}_{i-1}[\mathrm{ind}]$ might have been there from the start, i.e. $\mathcal{D}_0[\mathrm{ind}] = \mathcal{D}_0[\mathrm{ind}]$, and influenced all the decisions of the adaptive requester $\mathcal{Q}$ in the edit steps $1, \cdots, i-1$. To prove soundness, we need to show that if $(\mathrm{A}, \bar{\mathrm{A}})$ satisfies (q, ε)-deletion-privacy, then even in this worst-case scenario, no adaptive adversary can design a membership inference test $\mathrm{MI}(\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots) \in \{0, 1\}$ that can distinguish with high probability the null hypothesis $H_0 = \{\mathcal{D}_0[\mathrm{ind}] = \mathbf{x}\}$ from the alternate hypothesis $H_1 = \{\mathcal{D}_0[\mathrm{ind}] = \mathbf{x}'\}$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. That is, the advantage of any test MI, defined as

$$\mathrm{Adv}(\mathrm{MI}) \stackrel{\text{def}}{=} \mathbb{P}\left[\mathrm{MI}(\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots) = 1 | H_0\right] - \mathbb{P}\left[\mathrm{MI}(\hat{\Theta}_i, \hat{\Theta}_{i+1}, \cdots) = 1 | H_1\right], \tag{38}$$

must be small. Since after processing edit request u_i, the databases $\mathcal{D}_i, \mathcal{D}_{i+1}, \cdots$ no longer contain the deleted record $\mathcal{D}_{i-1}[\text{ind}]$, the data-processing inequality implies that future models $\hat{\Theta}_{i+1}, \hat{\Theta}_{i+2}, \cdots$ cannot have more information about $\mathcal{D}_{i-1}[\text{ind}]$ 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 $(\text{A}, \bar{\text{A}})$ satisfy (q,ε)-deletion-privacy for any p-adaptive r-requester $\mathcal{Q}$, we know from Definition 4.1 that there exists a mapping $\pi_i^{\mathcal{Q}}$ such that for all $\mathcal{D}_0 \in \mathcal{X}^n$, the model $\hat{\Theta}_i$ generated by the interaction between $(\text{A}, \bar{\text{A}}, \mathcal{Q})$ on $\mathcal{D}_0$ after ith edit satisfies the inequality $\text{R}_q\left(\hat{\Theta}_i \middle\| \pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \text{ind}, \mathbf{y}\rangle)\right) \leq \varepsilon$. As the database $\mathcal{D}_0 \circ \langle \text{ind}, \mathbf{y}\rangle$ is identical under both hypothesis H_0 and H_1, we have $\text{R}_q\left(\hat{\Theta}_i | H_b \middle\| \bar{\Theta}\right) \leq \varepsilon$ for $b \in \{0,1\}$, where $\bar{\Theta} = \pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \text{ind}, \mathbf{y}\rangle)$. From Rényi divergence to (ε,δ)-indistinguishability conversion described in Theorem B.1, we get

$$\mathbb{P}\left[\text{MI}^*(\hat{\Theta}_i) = 1 | H_0\right] \leq e^{\varepsilon'(\delta)} \mathbb{P}\left[\text{MI}^*(\bar{\Theta}) = 1\right] + \delta, \text{ and} \tag{39}$$

$$\mathbb{P}\left[\text{MI}^*(\hat{\Theta}_i) = 0 | H_1\right] \leq e^{\varepsilon'(\delta)} \mathbb{P}\left[\text{MI}^*(\bar{\Theta}) = 0\right] + \delta, \tag{40}$$

where $\varepsilon'(\delta) = \varepsilon + \frac{\log 1/\delta}{q-1}$ for any $0 < \delta < 1$. On adding the two inequalities, we get:

$$\begin{aligned}
\text{Adv}(\text{MI}) \leq \text{Adv}(\text{MI}^*) &= \mathbb{P}\left[\text{MI}^*(\hat{\Theta}_i) = 1 | H_0\right] - \mathbb{P}\left[\text{MI}^*(\hat{\Theta}_i) = 1 | H_1\right] \\
&\leq \min_{\delta} e^{\varepsilon'(\delta)} - 1 + 2\delta \\
&= \frac{q e^{\varepsilon(q-1)/q}}{q-1} [2(q-1)]^{1/q} - 1
\end{aligned}$$

Alternatively, from monotonicity of Rényi divergence w.r.t. order q and the fact that Rényi divergence converges to KL divergence as $q \to 1$, we have from $\text{R}_q\left(\hat{\Theta}_i | H_b \middle\| \bar{\Theta}\right) \leq \varepsilon$ for $b \in \{0,1\}$ that

$$\begin{aligned}
&\text{KL}\left(\hat{\Theta}_i | H_b \middle\| \bar{\Theta}\right) \leq \text{R}_q\left(\hat{\Theta}_i | H_b \middle\| \bar{\Theta}\right) \leq \varepsilon \\
\implies &\mathbf{TV}\left(\hat{\Theta}_i | H_b; \bar{\Theta}\right) \leq \sqrt{\frac{\varepsilon}{2}}, \qquad \text{(From Pinkser inequality)}
\end{aligned}$$

for $b \in \{0,1\}$. So, from triangle inequality on total variation distance, we have

$$\mathbf{TV}\left(\hat{\Theta}_i | H_0; \hat{\Theta}_i | H_1\right) \leq \mathbf{TV}\left(\hat{\Theta}_i | H_0; \bar{\Theta}\right) + \mathbf{TV}\left(\hat{\Theta}_i | H_0; \bar{\Theta}\right) \leq \sqrt{2\varepsilon}. \tag{41}$$

So, advantage of any membership inference attack MI must have an advantage satisfying

$$\text{Adv}(\text{MI}) = \mathbb{P}\left[\text{MI}(\hat{\Theta}_i) = 1 | H_0\right] - \mathbb{P}\left[\text{MI}(\hat{\Theta}_i) = 1 | H_1\right] \leq \sqrt{2\varepsilon}. \tag{42}$$

□

Theorem 4.3 (From adaptive to non-adaptive deletion)**.** *If an algorithm pair* $(\text{A}, \bar{\text{A}})$ *satisfies* $(q, \varepsilon_{\text{dd}})$*-deletion-privacy under all non-adaptive* r*-requesters and is also* $(q, \varepsilon_{\text{dp}})$*-Rényi DP with respect to records not being deleted, then it also satisfies* $(q, \varepsilon_{\text{dd}} + p\varepsilon_{\text{dp}})$*-deletion-privacy under all* p*-adaptive* r*-requesters.*

Proof. To prove this theorem, we need to show that for any p-adaptive r-requester $\mathcal{Q}$, there exists a construction for a map $\pi_i^{\mathcal{Q}} : \mathcal{X}^n \to \mathcal{O}$ such that for all $\mathcal{D}_0 \in \mathcal{X}^n$, the sequence of model $(\hat{\Theta}_i)_{i \geq 0}$ generated by the interaction between $(\mathcal{Q}, \text{A}, \bar{\text{A}})$ on $\mathcal{D}_0$ satisfies the following inequality for all $i \geq 1$:

$$\text{R}_q\left(\bar{\text{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) \middle\| \pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \text{ind}, \mathbf{y}\rangle)\right) \leq \varepsilon_{\text{dd}} + p\varepsilon_{\text{dp}}, \quad \text{for all } u_i \in \mathcal{U}^r \text{ and } \langle \text{ind}, \mathbf{y}\rangle \in u_i. \tag{43}$$

Fix a database $\mathcal{D}_0 \in \mathcal{X}^n$ and an edit request $u_i \in \mathcal{U}^r$. Let $\mathcal{D}_0' \in \mathcal{X}^n$ be a neighbouring database defined to be $\mathcal{D}_0' = \mathcal{D}_0 \circ \langle \text{ind}, \mathbf{y}\rangle$ for an arbitrary replacement operation $\langle \text{ind}, \mathbf{y}\rangle \in u_i$. Given any p-adaptive r-requester $\mathcal{Q}$, 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 $\mathcal{Q}$'s interaction with $(\mathrm{A}, \bar{\mathrm{A}})$ with initial database as $\mathcal{D}_0$. Similarly, let $(\hat{\Theta}'_i)_{i\geq 0}$ and $(U'_i)_{i\geq 1}$ be the corresponding sequences generated due to the interaction among $(\mathcal{Q}, \mathrm{A}, \bar{\mathrm{A}})$ on $\mathcal{D}'_0$.

Since $(\mathrm{A}, \bar{\mathrm{A}})$ is assumed to satisfy $(q, \varepsilon_{\mathrm{dd}})$-deletion-privacy guarantee under non-adaptive r-requesters, recall from Remark 4.2 that there exists a mapping $\pi : \mathcal{X}^n \to \mathcal{O}$ such that for any fixed edit sequence $u_{\leq i} \stackrel{\text{def}}{=} (u_1, u_2, \cdots, u_i)$,

$$\mathrm{R}_q\left(\hat{\Theta}_i|_{U_{\leq i}=u_{\leq i}} \middle\| \pi(\mathcal{D}_0 \circ u_{\leq i})\right) \leq \varepsilon_{\mathrm{dd}} \tag{44}$$

$$\implies \mathrm{R}_q\left(\bar{\mathrm{A}}(\mathcal{D}_0 \circ U_{<i}, u_i, \hat{\Theta}_i)|_{U_{<i}=u_{<i}} \middle\| \pi(\mathcal{D}_0 \circ U'_{<i} \circ u_i)|_{U'_{<i}=u_{<i}}\right) \leq \varepsilon_{\mathrm{dd}}. \tag{45}$$

Note that since the replacement operation $\langle \mathrm{ind}, \mathbf{y}\rangle$ is part of the edit request u_i, we have $\mathcal{D}_0 \circ U'_{<i} \circ u_i = \mathcal{D}'_0 \circ U'_{<i} \circ u_i$. Moreover, since the sequence $U'_{<i}$ of edit requests is generated by the interaction of $(\mathcal{Q}, \mathrm{A}, \bar{\mathrm{A}})$ on $\mathcal{D}'_0 = \mathcal{D}_0 \circ \langle \mathrm{ind}, u\rangle$ and the ith edit request u_i is fixed beforehand, we can define a valid construction of a map $\pi_i^{\mathcal{Q}} : \mathcal{X}^n \to \mathcal{O}$ as per Definition 4.1 as follows:

$$\pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle) = \pi(\mathcal{D}'_0 \circ U'_{<i} \circ u_i). \tag{46}$$

For brevity, let $\hat{\Theta}_u = \bar{\mathrm{A}}(\mathcal{D}_0 \circ U_{<i}, u_i, \hat{\Theta}_{i-1})$, and $\hat{\Theta}'_u = \pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle)$. For this construction, we prove the requisite bound in (43) as follows.

$$\begin{aligned}
\mathrm{R}_q\left(\hat{\Theta}_u \middle\| \hat{\Theta}'_u\right) &\leq \mathrm{R}_q\left((\hat{\Theta}_u, U_{<i}) \middle\| (\hat{\Theta}'_u, U'_{<i})\right) && \text{(Data processing inequality (Van Erven \& Harremos, 2014, Theorem 1))} \\
&= \frac{1}{q-1}\log \int_\theta \sum_{u_{<i}} \frac{J(\theta, u_{<i})^q}{J'(\theta, u_{<i})^{q-1}} \mathrm{d}\theta && (J \text{ \& } J' \text{ are joint PDFs of } (\hat{\Theta}_u, U_{<i}) \text{ \& } (\hat{\Theta}'_u, U'_{<i})) \\
&= \frac{1}{q-1}\log \sum_{u_{<i}} \frac{\mathbb{P}\left[U_{<i} = u_{<i}\right]^q}{\mathbb{P}\left[U'_{<i} = u_{<i}\right]^{q-1}} \left\{ \int_\theta \frac{p_{\hat{\Theta}_u | U_{<i}=u_{<i}}(\theta)^q}{p_{\hat{\Theta}'_u | U'_{<i}=u_{<i}}(\theta)^{q-1}} \mathrm{d}\theta \right\} \\
&\leq \frac{1}{q-1}\log \sum_{u_{<i}} \frac{\mathbb{P}\left[U_{<i} = u_{<i}\right]^q}{\mathbb{P}\left[U'_{<i} = u_{<i}\right]^{q-1}} \exp((q-1)\varepsilon_{\mathrm{dd}}) && \text{(From (45))} \\
&= \varepsilon_{\mathrm{dd}} + \mathrm{R}_q\left(U_{<i} \| U'_{<i}\right) \\
&\leq \varepsilon_{\mathrm{dd}} + \mathrm{R}_q\left(\left(\hat{\Theta}_{s^1}, \cdots, \hat{\Theta}_{s^p}\right) \middle\| \left(\hat{\Theta}'_{s^1}, \cdots, \hat{\Theta}'_{s^p}\right)\right) && \text{(If } \mathcal{Q} \text{ sees outputs at steps } s^1, \cdots, s^p) \\
&\leq \varepsilon_{\mathrm{dd}} + p\varepsilon_{\mathrm{dp}}. && \text{(Via Rényi composition)}
\end{aligned}$$

□

Theorem 4.5 (Privacy of remaining records is necessary for adaptive deletion privacy)**.** *Let Test* $: \mathcal{O} \to \{0, 1\}$ *be a membership inference test for* A *to distinguish between neighbouring databases* $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^n$*. Similarly, let* $\overline{\textit{Test}} : \mathcal{O} \to \{0, 1\}$ *be a membership inference test for* $\bar{\mathrm{A}}$ *to distinguish between* $\bar{\mathcal{D}}, \bar{\mathcal{D}}' \in \mathcal{X}^n$ *that are neighbouring after applying edit* $\bar{u} \in \mathcal{U}^1$*. If Adv(Test)* $> \delta$ *and Adv(*$\overline{\textit{Test}}$*)* $> \delta$*, then the pair* $(\mathrm{A}, \bar{\mathrm{A}})$ *cannot satisfy* (q, ε)*-deletion-privacy under* 1*-adaptive* 1*-requester for any*

$$\varepsilon < \max\left\{\frac{\delta^4}{2}, \log(q-1) + \frac{q}{q-1}\log\left(\frac{1+\delta^2}{q2^{1/q}}\right)\right\}. \tag{47}$$

Proof. By assumption, we know that there exists tests $\mathrm{Test}, \overline{\mathrm{Test}} : \mathcal{O} \to \{0, 1\}$ such that

$$\mathrm{Adv}(\mathrm{Test}) \stackrel{\text{def}}{=} \mathbb{P}\left[\mathrm{Test}(\mathrm{A}(\mathcal{D})) = 1\right] - \mathbb{P}\left[\mathrm{Test}(\mathrm{A}(\mathcal{D}')) = 1\right] > \delta, \tag{48}$$

and for all $\theta \in \mathcal{O}$,

$$\mathrm{Adv}(\overline{\mathrm{Test}}) \stackrel{\text{def}}{=} \mathbb{P}\left[\overline{\mathrm{Test}}(\bar{\mathrm{A}}(\bar{\mathcal{D}}, \bar{u}, \theta)) = 1\right] - \mathbb{P}\left[\overline{\mathrm{Test}}(\bar{\mathrm{A}}(\bar{\mathcal{D}}', \bar{u}, \theta)) = 1\right] > \delta. \tag{49}$$

Define $O' = \{\theta \in \mathcal{O} | \text{Test}(\theta) = 1\}$ and $\bar{O}' = \{\theta \in \mathcal{O} | \overline{\text{Test}}(\theta) = 1\}$. We have that the total variation distance between $\mathrm{A}(\mathcal{D})$ and $\mathrm{A}(\mathcal{D}')$ is lower bounded as

$$\begin{aligned}
\mathbf{TV}\left(\mathrm{A}(\mathcal{D}); \mathrm{A}(\mathcal{D}')\right) &= \sup_{O \subset \mathcal{O}} \left| \mathbb{P}\left[\mathrm{A}(\mathcal{D}) \in O\right] - \mathbb{P}\left[\mathrm{A}(\mathcal{D}') \in O\right] \right| && (50)\\
&> \mathbb{P}\left[\mathrm{A}(\mathcal{D}) \in O'\right] - \mathbb{P}\left[\mathrm{A}(\mathcal{D}') \in O'\right] && (51)\\
&= \mathbb{P}\left[\text{Test}(\mathrm{A}(\mathcal{D})) = 1\right] - \mathbb{P}\left[\text{Test}(\mathrm{A}(\mathcal{D}')) = 1\right] > \delta. && (52)
\end{aligned}$$

Similarly, we also have that for all $\theta \in \mathcal{O}$, the total variation distance between $\bar{\mathrm{A}}(\bar{\mathcal{D}}, \bar{u}, \theta)$ and $\bar{\mathrm{A}}(\bar{\mathcal{D}}', \bar{u}, \theta)$ is lower bounded as

$$\begin{aligned}
\mathbf{TV}\left(\bar{\mathrm{A}}(\bar{\mathcal{D}}, \bar{u}, \theta); \bar{\mathrm{A}}(\bar{\mathcal{D}}', \bar{u}, \theta)\right) &= \sup_{O \subset \mathcal{O}} \left| \mathbb{P}\left[\bar{\mathrm{A}}(\bar{\mathcal{D}}, \bar{u}, \theta) \in O\right] - \mathbb{P}\left[\bar{\mathrm{A}}(\bar{\mathcal{D}}', \bar{u}, \theta) \in O\right] \right| && (53)\\
&> \mathbb{P}\left[\bar{\mathrm{A}}(\bar{\mathcal{D}}, \bar{u}, \theta) \in \bar{O}'\right] - \mathbb{P}\left[\bar{\mathrm{A}}(\bar{\mathcal{D}}', \bar{u}, \theta) \in \bar{O}'\right] && (54)\\
&= \mathbb{P}\left[\overline{\text{Test}}(\bar{\mathrm{A}}(\bar{\mathcal{D}}, \bar{u}, \theta)) = 1\right] - \mathbb{P}\left[\overline{\text{Test}}(\bar{\mathrm{A}}(\bar{\mathcal{D}}', \bar{u}, \theta)) = 1\right] > \delta. && (55)
\end{aligned}$$

Assume W.L.O.G. that $\bar{u}$ replaces at index n and the edited databases $\bar{\mathcal{D}} \circ u, \bar{\mathcal{D}}' \circ u$ differs only at index 1. Also assume that $\mathcal{D}, \mathcal{D}'$ differs at index n.

Recall from Definition 4.1 that satisfying (q, ε)-deletion-privacy under 1-adaptive 1-requesters requires existence of a map $\pi_n^{\mathcal{Q}} : \mathcal{X}^n \to \mathcal{O}$ for each $\mathcal{Q}$ such that for all $\mathcal{D}_0 \in \mathcal{X}^n$,

$$\mathrm{R}_q\left(\bar{\mathrm{A}}(\mathcal{D}_{n-1}, u_n, \hat{\Theta}_{n-1}) \middle\| \pi_n^{\mathcal{Q}}(\mathcal{D}_0 \circ u_n)\right) \leq \varepsilon, \tag{56}$$

To prove the theorem statement, we show that for a starting database $\mathcal{D}_0 \in \{\mathcal{D}, \mathcal{D}'\}$ and an edit request $u_n = \bar{u}$ that deletes the differing record in choices of $\mathcal{D}_0$ at edit step n, there exists a 1-adaptive 1-requester $\mathcal{Q}$ that sends adaptive edit requests $u_1, \cdots, u_{n-1}$ in the first $n-1$ steps such that no map $\pi_n^{\mathcal{Q}}$ exists that satisfies (56) for both choices of $\mathcal{D}_0$ when ε follows inequality (47).

Consider the following construction of 1-adaptive 1-requester $\mathcal{Q}$ that only observes the first model $\hat{\Theta}_0 = \mathrm{A}(\mathcal{D}_0)$ and generates the edit requests $(u_1, \cdots, u_{n-1})$ as follows:

$$\mathcal{Q}(\hat{\Theta}_0; u_1, u_2, \cdots, u_{i-1}) = \begin{cases} \langle i, \bar{\mathcal{D}}[i] \rangle & \text{if Test}(\hat{\Theta}_0) = 1, \\ \langle i, \bar{\mathcal{D}}'[i] \rangle & \text{otherwise.} \end{cases} \tag{57}$$

This requester $\mathcal{Q}$ transforms any initial database $\mathcal{D}_0$ to $\mathcal{D}_{n-1} = \bar{\mathcal{D}}$ if the outcome $\text{Test}(\hat{\Theta}_0) = 1$, otherwise to $\mathcal{D}_{n-1} = \bar{\mathcal{D}}'$. Consider an adversary that does not observe the interaction transcript $(\hat{\Theta}_{<n}; u_{<n})$, but is interested in identifying whether $\mathcal{D}_0$ was $\mathcal{D}$ or $\mathcal{D}'$. The adversary gets to observe only the output $\hat{\Theta}_n = \bar{\mathrm{A}}(\mathcal{D}_{n-1}, u_n, \hat{\Theta}_{n-1})$ generated after processing the edit request $u_n = \bar{u}$. On this observation, the adversary runs the membership inference test $\text{MI}(\hat{\Theta}_n) = \overline{\text{Test}}(\hat{\Theta}_n)$. The membership inference advantage of MI is

$$\begin{aligned}
\text{Adv}(\text{MI}; \mathcal{D}, \mathcal{D}') &\stackrel{\text{def}}{=} \mathbb{P}\left[\text{MI}(\hat{\Theta}_n) = 1 | \mathcal{D}_0 = \mathcal{D}\right] - \mathbb{P}\left[\text{MI}(\hat{\Theta}_n) = 1 | \mathcal{D}_0 = \mathcal{D}'\right] \\
&= \sum_{b \in \{0,1\}} \mathbb{P}\left[\overline{\text{Test}}(\hat{\Theta}_n) = 1 | \text{Test}(\hat{\Theta}_0) = b\right] \times \mathbb{P}\left[\text{Test}(\hat{\Theta}_0) = b | \mathcal{D}_0 = \mathcal{D}\right] \\
&\quad - \sum_{b \in \{0,1\}} \mathbb{P}\left[\overline{\text{Test}}(\hat{\Theta}_n) = 1 | \text{Test}(\hat{\Theta}_0) = b\right] \times \mathbb{P}\left[\text{Test}(\hat{\Theta}_0) = b | \mathcal{D}_0 = \mathcal{D}'\right] \\
&= \left(\mathbb{P}\left[\overline{\text{Test}}(\hat{\Theta}_n) = 1 | \mathcal{D}_{n-1} = \bar{\mathcal{D}}\right] - \mathbb{P}\left[\overline{\text{Test}}(\hat{\Theta}_n) = 1 | \mathcal{D}_{n-1} = \bar{\mathcal{D}}'\right]\right) \text{Adv}(\text{Test}; \mathcal{D}, \mathcal{D}') \\
&= \text{Adv}(\overline{\text{Test}}; \bar{\mathcal{D}}, \bar{\mathcal{D}}', \bar{u}) \times \text{Adv}(\text{Test}; \mathcal{D}, \mathcal{D}') > \delta^2.
\end{aligned}$$

So, from the contrapositive of our soundness Theorem 4.1, we have that $(\mathrm{A}, \bar{\mathrm{A}})$ cannot be an (ε, q)-deletion-privacy algorithm for ε and q satisfying

$$\delta^2 > \min\left\{\sqrt{2\varepsilon}, \frac{qe^{\varepsilon(q-1)/q}}{q-1}[2(q-1)]^{1/q} - 1\right\} \tag{58}$$

$$\iff \varepsilon < \max\left\{\frac{\delta^4}{2}, \log(q-1) + \frac{q}{q-1}\log\left(\frac{1+\delta^2}{q2^{1/q}}\right)\right\}. \tag{59}$$

□

D.1. Our Reduction Theorem 4.3 versus Gupta et al. (2021)'s Reduction

Adaptive unlearning guarantee in (Gupta et al., 2021, Definition 2.3) is designed to ensure that no adaptive requester $\mathcal{Q}$ can force the output distribution of the unlearning algorithm $\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1})$ to diverge substantially from that of retraining algorithm $\mathrm{A}(\mathcal{D}_i)$ 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. (2021)'s SISA unlearning algorithm randomly partitions the initial database $\mathcal{D}_0$ during setup and uses the same partitioning for processing edit requests, deleting records from respective shards on request. Gupta et al. (2021) show that an adaptive update requester $\mathcal{Q}$ can interactively send deletion requests $u_1, \cdots, u_i$ to SISA so that after some time, the partitioning of remaining records in $\mathcal{D}_i = \mathcal{D}_0 \circ u_1 \cdots u_i$ follows a pattern that is unlikely to occur on repartitioning of $\mathcal{D}_i$ if we execute $\mathrm{A}(\mathcal{D}_i)$.

They provide a general reduction (Gupta et al., 2021, 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. (2021) 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. (2021) still does not guarantee deletion privacy as per the RTBF guidelines. In Theorem 3.1, we demonstrate that there exists an algorithm pair $(\mathrm{A}, \bar{\mathrm{A}})$ satisfying adaptive unlearning Definition 2.6 (a strictly stronger version of (Gupta et al., 2021, 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 4.3 from adaptive to non-adaptive requests relies on differential privacy with respect to the standard replacement of records instead.

E. Calculus Refresher

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

$$\nabla\mathcal{L}(\theta) = \left(\frac{\partial\mathcal{L}(\theta)}{\partial\theta_1}, \cdots, \frac{\partial\mathcal{L}(\theta)}{\partial\theta_2}\right). \tag{60}$$

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

$$\nabla^2\mathcal{L}(\theta) = \left(\frac{\partial^2\mathcal{L}(\theta)}{\partial\theta_i\theta_j}\right)_{1\le i,j\le d}. \tag{61}$$

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

$$\Delta\mathcal{L}(\theta) = \mathsf{Tr}\left(\nabla^2\mathcal{L}(\theta)\right). \tag{62}$$

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

$$\operatorname{div}(\mathbf{v})(\theta) = \sum_{i=1}^{d} \frac{\partial \mathbf{v}_i(\theta)}{\partial \theta_i}. \tag{63}$$

Some identities that we would rely on:

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

$$\operatorname{div}(\nabla \mathcal{L})(\theta) = \sum_{i=1}^{d} \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_i^2} = \Delta \mathcal{L}(\theta). \tag{64}$$

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

$$\int_{\mathcal{O}} \langle \mathbf{v}(\theta), \nabla f(\theta) \mathrm{d}\theta \rangle = -\int_{\mathcal{O}} f(\theta)(\operatorname{div}(\mathbf{v}))(\theta)\mathrm{d}\theta. \tag{65}$$

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

$$\int_{\mathcal{O}} f(\theta)\Delta g(\theta)\mathrm{d}\theta = -\int_{\mathcal{O}} \langle \nabla f(\theta), \nabla g(\theta) \rangle \, \mathrm{d}\theta = \int_{\mathcal{O}} g(\theta)\Delta f(\theta)\mathrm{d}\theta. \tag{66}$$

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

$$\langle \mathbf{v}_1, \mathbf{v}_2 \rangle (\theta) \leq \frac{1}{2a} \|\mathbf{v}_1(\theta)\|_2^2 + \frac{1}{2b} \|\mathbf{v}_2(\theta)\|_2^2. \tag{67}$$

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

F. Loss Function Properties

In this section, we provide the formal definition of various properties that we assume in the paper. Let $\boldsymbol{\ell}(\theta; \mathbf{x}) : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ be a loss function on $\mathbb{R}^d$ for any record $\mathbf{x} \in \mathcal{X}$.

Definition F.1 (Lipschitzness)**.** *A function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is said to be L* Lipschitz continuous *if for all $\theta, \theta' \in \mathbb{R}^d$ and any $\mathbf{x} \in \mathcal{X}$,*

$$|\boldsymbol{\ell}(\theta; \mathbf{x}) - \boldsymbol{\ell}(\theta'; \mathbf{x})| \leq L \|\theta - \theta'\|_2. \tag{68}$$

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

Definition F.2 (Boundedness)**.** *A function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is said to be B-*bounded *if for all $\mathbf{x} \in \mathcal{X}$, its output takes values in range $[-B, B]$.*

Definition F.3 (Convexity)**.** *A continuous differential function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is said to be* convex *if for all $\theta, \theta' \in \mathbb{R}^d$ and $\mathbf{x} \in \mathcal{X}$,*

$$\boldsymbol{\ell}(\theta'; \mathbf{x}) \geq \boldsymbol{\ell}(\theta; \mathbf{x}) + \langle \nabla \boldsymbol{\ell}(\theta; \mathbf{x}), \theta' - \theta \rangle, \tag{69}$$

and is said to be λ-strongly convex if

$$\boldsymbol{\ell}(\theta'; \mathbf{x}) \geq \boldsymbol{\ell}(\theta; \mathbf{x}) + \langle \nabla \boldsymbol{\ell}(\theta; \mathbf{x}), \theta' - \theta \rangle + \frac{\lambda}{2} \|\theta' - \theta\|_2^2. \tag{70}$$

Theorem F.1 ((Nesterov, 2003, Theorem 2.1.4))**.** *A twice continuously differentiable function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is* convex *if and only if for all $\theta \in \mathbb{R}^d$ and $\mathbf{x} \in \mathcal{X}$, its hessian matrix $\nabla^2 \boldsymbol{\ell}(\theta; \mathbf{x})$ is positive semidefinite, i.e., $\nabla^2 \boldsymbol{\ell}(\theta; \mathbf{x}) \succcurlyeq 0$ and is λ-strongly convex if its hessian matrix satisfies $\nabla^2 \boldsymbol{\ell}(\theta; \mathbf{x}) \succcurlyeq \lambda \mathbb{I}_d$.*

Definition F.4 (Smoothness)**.** *A continuously differentiable function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is said to be β-*smooth *if for all $\theta, \theta' \in \mathbb{R}^d$ and $\mathbf{x} \in \mathcal{X}$,*

$$\|\nabla \boldsymbol{\ell}(\theta; \mathbf{x}) - \nabla \boldsymbol{\ell}(\theta'; \mathbf{x})\|_2 \leq \beta \|\theta - \theta'\|_2. \tag{71}$$

Theorem F.2 ((Nesterov, 2003, Theorem 2.1.6))**.** *A twice continuously differentiable convex function $\boldsymbol{\ell}(\theta; \mathbf{x})$ is β-smooth if and only if for all $\theta \in \mathbb{R}^d$ and $\mathbf{x} \in \mathcal{X}$,*

$$\nabla^2 \boldsymbol{\ell}(\theta; \mathbf{x}) \preccurlyeq \beta \mathbb{I}_d. \tag{72}$$

F.1. Effect of Gradient Clipping

First order optimization methods on a continuously differentiable loss function $\boldsymbol{\ell}(\theta; \mathbf{x})$ over a database $\mathcal{D} \in \mathcal{X}^n$ with gradient clipping $\mathrm{Clip}_L(\mathbf{v}) = \mathbf{v} / \max\left(1, \frac{\|\mathbf{v}\|_2}{L}\right)$ is equivalent to optimizing

$$\mathcal{L}_{\mathcal{D}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \bar{\boldsymbol{\ell}}(\theta; \mathbf{x}) + \mathbf{r}(\theta), \tag{73}$$

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 F.3 (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.

$$\mathrm{Clip}_L(\mathbf{v}) = \underset{\|\mathbf{v}'\|_2 \le L}{\arg\min} \ \|\mathbf{v} - \mathbf{v}'\|_2 \,. \tag{74}$$

By properties of orthogonal projections on closed convex sets, for every $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^d$,

$$\langle \mathbf{v}' - \mathrm{Clip}_L(\mathbf{v}), \mathbf{v} - \mathrm{Clip}_L(\mathbf{v}) \rangle \le 0 \quad \text{if and only if } \|\mathbf{v}'\|_2 \le L. \tag{75}$$

Therefore, for any $\theta \in \mathbb{R}^d$, and $\mathbf{x} \in \mathcal{X}$, we have

$$\left\langle \nabla \bar{\boldsymbol{\ell}}(\theta + h\hat{\mathbf{v}}; \mathbf{x}) - \nabla \bar{\boldsymbol{\ell}}(\theta; \mathbf{x}), \nabla \boldsymbol{\ell}(\theta; \mathbf{x}) - \nabla \bar{\boldsymbol{\ell}}(\theta; \mathbf{x}) \right\rangle \le 0, \tag{76}$$

$$\left\langle \nabla \bar{\boldsymbol{\ell}}(\theta; \mathbf{x}) - \nabla \bar{\boldsymbol{\ell}}(\theta + h\hat{\mathbf{v}}; \mathbf{x}), \nabla \boldsymbol{\ell}(\theta + h\hat{\mathbf{v}}; \mathbf{x}) - \nabla \bar{\boldsymbol{\ell}}(\theta + h\hat{\mathbf{v}}; \mathbf{x}) \right\rangle \le 0, \tag{77}$$

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

$$\left\langle \nabla_{\hat{\mathbf{v}}} \nabla \bar{\boldsymbol{\ell}}(\theta; \mathbf{x}), \nabla \boldsymbol{\ell}(\theta; \mathbf{x}) - \nabla \bar{\boldsymbol{\ell}}(\theta; \mathbf{x}) \right\rangle = 0, \tag{78}$$

for all $\hat{\mathbf{v}}$. Therefore, when $\nabla \bar{\boldsymbol{\ell}}(\theta; \mathbf{x}) \neq \nabla \boldsymbol{\ell}(\theta; \mathbf{x})$, 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

$$\begin{aligned} \nabla_{\hat{\mathbf{v}}} \cdot \nabla \bar{\boldsymbol{\ell}}(\theta; \mathbf{x}) &= \lim_{h \to 0^+} \left\langle \frac{\nabla \bar{\boldsymbol{\ell}}(\theta + h\hat{\mathbf{v}}; \mathbf{x}) - \nabla \boldsymbol{\ell}(\theta; \mathbf{x})}{h}, \hat{\mathbf{v}} \right\rangle \\ &= \begin{cases} \hat{\mathbf{v}}^\top \nabla^2 \boldsymbol{\ell}(\theta; \mathbf{x}) \hat{\mathbf{v}} & \text{if } \exists h > 0 \text{ s.t. } \nabla \bar{\boldsymbol{\ell}}(\theta + h\hat{\mathbf{v}}; \mathbf{x}) = \nabla \boldsymbol{\ell}(\theta + h\hat{\mathbf{v}}; \mathbf{x}) \\ 0 & \text{otherwise} \end{cases} \geq 0. \end{aligned}$$

□

Lemma F.4 (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 F.5 (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. $\mathrm{Clip}_L(\mathbf{v}) = \arg\min_{\mathbf{v}'}\{\|\mathbf{v}' - \mathbf{v}\|_2 : \mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}'\|_2 \leq L\}$. Since orthogonal projection onto a closed convex set is a 1-Lipschitz operation, for any $\theta, \theta' \in \mathbb{R}^d$,

$$\left\|\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}) - \nabla\bar{\boldsymbol{\ell}}(\theta';\mathbf{x})\right\|_2 \leq \|\nabla\boldsymbol{\ell}(\theta;\mathbf{x}) - \nabla\boldsymbol{\ell}(\theta';\mathbf{x})\|_2 \leq \beta\|\theta - \theta'\|_2. \tag{79}$$

□

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 F.6 (Rademacher's Theorem (Nekvinda & Zajíček, 1988))**.** *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 Section 5 rely on the above four properties on losses and therefore apply with gradient clipping instead of the Lipschitzness assumption.

G. Additional Preliminaries and Proofs for Section 5

G.1. Langevin Diffusion and Markov Semigroups

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)

$$\mathrm{d}\Theta_t = -\nabla\mathcal{L}(\Theta_t)\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}\mathbf{Z}_t, \tag{80}$$

where $\mathrm{d}\mathbf{Z}_t = \mathbf{Z}_{t+\mathrm{d}t} - \mathbf{Z}_t \sim \sqrt{\mathrm{d}t}\mathcal{N}(0, \mathbb{I}_d)$ is the d-dimensional Weiner process.

We present some preliminaries on the diffusion theory used in our analysis. 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 (80), the associated Markov semigroup $\mathbf{P}$, is defined as a family of operators $(P_t)_{t\geq 0}$, such that an operator P_t sends any real-valued measurable function $f : \mathbb{R}^d \to \mathbb{R}$ to

$$P_t f(\theta_0) = \mathbb{E}\left[f(\Theta_t)|\Theta_0 = \theta_0\right] = \int f(\theta_t)p_t(\theta_0, \theta_t)\mathrm{d}\theta_t. \tag{81}$$

The infinitesimal generator $\mathcal{G} \stackrel{\text{def}}{=} \lim_{s\to 0}\frac{1}{s}\left[P_{t+s} - P_s\right]$ for this diffusion semigroup is

$$\mathcal{G}f = \sigma^2\Delta f - \langle\nabla\mathcal{L}, \nabla f\rangle. \tag{82}$$

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 (80) for $\mathrm{d}t$ time. That is,

$$\partial_t P_t f(\theta_0) = \int \partial_t p_t(\theta_0, \theta_t)f(\theta_t)\mathrm{d}\theta_t = \int p_t(\theta_0, \theta_t)\mathcal{G}f(\theta_t)\mathrm{d}\theta_t. \tag{83}$$

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

$$\int f\mathcal{G}^* g\mathrm{d}\theta = \int g\mathcal{G}f\mathrm{d}\theta, \tag{84}$$

for all real-valued measurable functions $f, g : \mathbb{R}^d \to \mathbb{R}$. Note from (83) that, this operator provides an alternative way to represent the rate of change of function f at time t:

$$\partial_t P_t f(\theta_0) = \int f(\theta_t)\mathcal{G}^* p_t(\theta_0, \theta_t)\mathrm{d}\theta_t. \tag{85}$$

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 (80), the Fokker-Planck operator is the following:

$$\partial_t p_t(\theta) = \mathcal{G}^* p_t(\theta) = \mathrm{div}\left(p_t(\theta)\nabla\mathcal{L}(\theta)\right) + \sigma^2\Delta p_t(\theta). \tag{86}$$

From this Fokker-Planck equation, one can verify that the stationary or invariant distribution π of Langevin diffusion, which is the solution of $\partial_t p_t = 0$, follows the Gibbs distribution

$$\pi(\theta) \propto e^{-\mathcal{L}(\theta)/\sigma^2}. \tag{87}$$

Since π is the stationary distribution, note that for any measurable function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathop{\mathbb{E}}_{\pi}[\mathcal{G}f] = \int f\mathcal{G}^*\pi \mathrm{d}\theta = 0. \tag{88}$$

G.2. 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 π (Bakry et al., 2014). One such property of interest is the *logarithmic Sobolev (*LS*) inequality* (Gross, 1975), which we define next.

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

$$\Gamma(f, g) = \frac{1}{2}\left[\mathcal{G}(fg) - f\mathcal{G}g - g\mathcal{G}f\right], \tag{89}$$

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

$$\Gamma(f, g) = \sigma^2 \langle \nabla f, \nabla g \rangle. \tag{90}$$

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

Definition G.1 (Logarithmic Sobolev Inequality (see Bakry et al. (2014, p. 24)))**.** *A distribution with probability density π is said to satisfy a* logarithmic Sobolev inequality *(*$\mathrm{LS}(c)$*) (with respect to Γ in* (90)*) if for all functions $f \in \mathbb{L}^2(\mu)$ with continuous derivatives ∇f,*

$$\mathrm{Ent}_\pi(f^2) \le \frac{1}{2c}\int \frac{\Gamma(f^2)}{f^2}\pi \mathrm{d}\theta = \frac{2\sigma^2}{c}\int \|\nabla f\|_2^2 \,\pi \mathrm{d}\theta, \tag{91}$$

where entropy Ent_π *is defined as*

$$\mathrm{Ent}_\pi(f^2) = \mathop{\mathbb{E}}_{\pi}\left[f^2 \log f^2\right] - \mathop{\mathbb{E}}_{\pi}\left[f^2\right] \log \mathop{\mathbb{E}}_{\pi}\left[f^2\right]. \tag{92}$$

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 LS inequality.

Lemma G.1 (LS inequality of Gaussian distributions (see Bakry et al. (2014, p. 258)))**.** *Let ρ be a Gaussian distribution on $\mathbb{R}^d$ with covariance σ^2/λ (i.e., the Gibbs distribution* (87) *with $\mathcal{L}(\cdot)$ being the $L2$ regularizer $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$). Then ρ satisfies* $\mathrm{LS}(\lambda)$ *tightly (with respect to Γ in* (90)*), i.e.*

$$\mathrm{Ent}_\rho(f^2) = \frac{2\sigma^2}{\lambda}\int \|\nabla f\|_2^2 \,\rho \mathrm{d}\theta. \tag{93}$$

Additionally, if μ is a distribution on $\mathbb{R}^d$ that satisfy $\mathrm{LS}(c)$*, then the convolution $\mu \circledast \rho$, defined as the distribution of $\Theta + \mathbf{Z}$ where $\Theta \sim \mu$ and $\mathbf{Z} \sim \pi$, satisfies* LS *inequality with constant* $\left(\frac{1}{c} + \frac{1}{\lambda}\right)^{-1}$.

Bobkov (2007) show that like Gaussians, all strongly log concave distributions (or more generally, log-concave distributions with finite second order moments) satisfy LS inequality (e.g. Gibbs distribution π with any strongly convex $\mathcal{L}$). LS inequality is also satisfied under non-log-concavity too. For example, LS inequality is stable under Lipschitz maps, although such maps can destroy log-concavity.

Lemma G.2 (LS inequality under Lipschitz maps (see Ledoux (2001)))**.** *If π is a distribution on $\mathbb{R}^d$ that satisfies* $\mathrm{LS}(c)$*, then for any L-Lipschitz map $T : \mathbb{R}^d \to \mathbb{R}^d$, the pushforward distribution $T_{\#\pi}$, representing the distribution of $T(\Theta)$ when $\Theta \sim \pi$, satisfies* $\mathrm{LS}(c/L^2)$.

LS inequality is also stable under bounded perturbations to the distribution, as shown in the following lemma by Holley & Stroock (1986).

Lemma G.3 (LS inequality under bounded perturbations (see Holley & Stroock (1986))). *If π is the probability density of a distribution that satisfies $\mathrm{LS}(c)$, then any probability distribution with density π' such that $\frac{1}{\sqrt{B}} \leq \frac{\pi(\theta)}{\pi'(\theta)} \leq \sqrt{B}$ everywhere in $\mathbb{R}^d$ for some constant $B > 1$ satisfies $\mathrm{LS}(c/B)$.*

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

Lemma G.4 (LS inequality in terms of KL divergence (Vempala & Wibisono, 2019)). *The distribution π satisfies $\mathrm{LS}(c)$ inequality (with respect to Γ in (90)) if and only if for all distributions μ on $\mathbb{R}^d$ such that $\frac{\mu}{\pi} \in \mathbb{L}^2(\pi)$ with continuous derivatives $\nabla \frac{\mu}{\pi}$,*

$$\mathrm{KL}\left(\mu \| \pi\right) \leq \frac{\sigma^2}{2c} \mathrm{I}\left(\mu \| \pi\right). \tag{94}$$

Proof. Set f^2 in (91) to $\frac{\mu}{\pi}$ to obtain (94). Alternatively, set $\mu = \frac{f^2 \pi}{\mathbb{E}_\pi[f^2]}$ in (94) to obtain (91). □

Lemma G.5 (Wasserstein distance bound under LS inequality (Otto & Villani, 2000, Theorem 1)). *If distribution π satisfies $\mathrm{LS}(c)$ inequality (with respect to Γ in (90)) then for all distributions μ on $\mathbb{R}^d$,*

$$\mathrm{W}_2\left(\mu, \pi\right)^2 \leq \frac{2\sigma^2}{c} \mathrm{KL}\left(\mu \| \pi\right). \tag{95}$$

Lemma G.6 (LS inequality in terms of Rényi Divergence (Vempala & Wibisono, 2019)). *The distribution π satisfies $\mathrm{LS}(c)$ inequality (with respect to Γ in (90)) if and only if for all distributions μ on $\mathbb{R}^d$ such that $\frac{\mu}{\pi} \in \mathbb{L}^2(\pi)$ with continuous derivatives $\nabla \frac{\mu}{\pi}$, and any $q > 1$,*

$$\mathrm{R}_q\left(\mu \| \pi\right) + q(q-1)\partial_q \mathrm{R}_q\left(\mu \| \pi\right) \leq \frac{q^2 \sigma^2}{2c} \frac{\mathrm{I}_q\left(\mu \| \pi\right)}{\mathrm{E}_q\left(\mu \| \pi\right)}. \tag{96}$$

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

$$\mathop{\mathbb{E}}_{\pi}\left[f^2\right] = \mathop{\mathbb{E}}_{\pi}\left[\left(\frac{\mu}{\pi}\right)^q\right] = E(q), \qquad \text{(From (23))}$$

and,

$$\mathop{\mathbb{E}}_{\pi}\left[f^2 \log f^2\right] = \mathop{\mathbb{E}}_{\pi}\left[\left(\frac{\mu}{\pi}\right)^q \log\left(\frac{\mu}{\pi}\right)^q\right] = q\partial_q \mathop{\mathbb{E}}_{\pi}\left[\int_q \left(\frac{\mu}{\pi}\right)^q \log\left(\frac{\mu}{\pi}\right) \mathrm{d}q\right] = q\partial_q \mathop{\mathbb{E}}_{\pi}\left[\left(\frac{\mu}{\pi}\right)^q\right] = q\partial_q E(q).$$

(From Lebniz rule and (23))

Moreover,

$$\mathop{\mathbb{E}}_{\pi}\left[\left\|\nabla f\right\|_2^2\right] = \mathop{\mathbb{E}}_{\pi}\left[\left\|\nabla\left(\frac{\mu}{\pi}\right)^{\frac{q}{2}}\right\|_2^2\right] = \frac{q^2}{4} I(q) \qquad \text{(From (29))}$$

On substituting (91) with the above equalities, we get:

$$\begin{aligned}
&\mathrm{Ent}_\pi(f^2) \le \frac{2\sigma^2}{c}\mathop{\mathbb{E}}_{\pi}\left[\|\nabla f\|_2^2\right] \\
\iff & q\partial_q E(q) - E(q)\log E(q) \le \frac{q^2\sigma^2}{2c} I(q) \\
\iff & q\partial_q \log E(q) - \log E(q) \le \frac{q^2\sigma^2}{2c}\frac{I(q)}{E(q)} \\
\iff & q\partial_q\left((q-1)R(q)\right) - (q-1)R(q) \le \frac{q^2\sigma^2}{2c}\frac{I(q)}{E(q)} \qquad \text{(From (23))} \\
\iff & R(q) + q(q-1)\partial_q R(q) \le \frac{q^2\sigma^2}{2c}\frac{I(q)}{E(q)}
\end{aligned}$$

□

G.3. (Rényi) Differential Privacy Guarantees on Noisy-GD

In this section, we recap the differential privacy bounds in literature for Noisy-GD Algorithm 2.

Theorem G.7 (Rényi DP guarantee for Noisy-GD Algorithm 2)**.** *If $\boldsymbol{\ell}(\theta;\mathbf{x})$ is L-Lipschitz, then Noisy-GD satisfies (q,ε)-Rényi DP with $\varepsilon = \frac{qL^2}{\sigma^2 n^2}\cdot\eta K$.*

Proof. The L_2 sensitivity of gradient $\nabla\mathcal{L}_{\mathcal{D}}(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{\mathbf{x}\in\mathcal{D}}\nabla\boldsymbol{\ell}(\theta;\mathbf{x}) + \nabla\mathbf{r}(\theta)$ computed in step 2 of Algorithm 2 for neighboring databases in $\mathcal{X}^n$ that differ in a single record is $\frac{2L}{n}$ since $\boldsymbol{\ell}(\theta;\mathbf{x})$ is L-Lipschitz.

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 $2\sigma^2/\eta$. So, for neighboring databases $\mathcal{D},\mathcal{D}' \in \mathcal{X}^n$, we have from the Rényi DP bound of Gaussian mechanisms proposed by Mironov (2017, Proposition 7) that

$$\mathrm{R}_q\left(\Theta_{\eta(k+1)}\mid_{\Theta_{\eta k}=\theta_k}\middle\|\,\Theta'_{\eta(k+1)}\mid_{\Theta'_{\eta k}=\theta_k}\right) \le \frac{\eta q L^2}{n^2\sigma^2}, \tag{97}$$

where $(\Theta_{\eta k})_{0\le k\le K}$ and $(\Theta'_{\eta k})_{0\le k\le K}$ are intermediate parameters in Algorithm 2 when run on databases $\mathcal{D}$ and $\mathcal{D}'$ respectively. Finally, from Rényi composition Mironov (2017, Proposition 1), we have

$$\begin{aligned}
\mathrm{R}_q\left(\Theta_{\eta K}\middle\|\Theta'_{\eta K}\right) &\le \mathrm{R}_q\left((\Theta_0,\Theta_\eta,\cdots,\Theta_{\eta K})\middle\|(\Theta'_0,\Theta'_\eta,\cdots,\Theta'_{\eta K})\right) \\
&\le \sum_{k=0}^{K-1}\mathrm{R}_q\left(\Theta_{\eta(k+1)}\mid_{\Theta_{\eta k}=\theta_k}\middle\|\,\Theta'_{\eta(k+1)}\mid_{\Theta'_{\eta k}=\theta_k}\right) \\
&\le \frac{qL^2}{n^2\sigma^2}\cdot\eta K.
\end{aligned}$$

□

Remark G.8. *Different papers discussing Noisy-GD variants adopt different notational conventions for the total noise added to the gradients. The noise variance in our Algorithm 2 is $2\eta\sigma^2$; but is $\frac{\eta^2\sigma^2L^2}{n^2}$ in the full-batch setting of DP-SGD by Abadi et al. (2016). To translate the bound in Theorem G.7, one can simply rescale σ across different conventions to have the same noise variance, i.e., $2\eta\sigma^2 = \frac{\eta^2\hat{\sigma}^2L^2}{n^2}$.*

Our Theorem G.7 is somewhat identical to Abadi et al. (2016)'s (ε,δ)-DP bound. To verify this, note from Rényi divergence to (ε,δ)-indistinguishability conversion in Theorem B.1 that $(1+\frac{2}{\varepsilon}\log\frac{1}{\delta},\frac{\varepsilon}{2})$-Rényi DP implies (ε,δ)-DP. So, setting the bound in Theorem G.7 to be smaller than $\frac{\varepsilon}{2}$ and substituting $q = 1+\frac{2}{\varepsilon}\log\frac{1}{\delta}$, we get

$$\left(\frac{\varepsilon+2\log\frac{1}{\delta}}{\varepsilon}\right)\frac{L^2}{n^2\sigma^2}\cdot\eta K \le \frac{\varepsilon}{2} \iff \frac{\sqrt{K(\varepsilon+2\log\frac{1}{\delta})}}{\varepsilon} \le \hat{\sigma}.$$

For $\varepsilon \leq 2\log\frac{1}{\delta}$, we get the same noise bound as in Abadi et al. (2016, Theorem 1) for their (full-batch) DP-SGD algorithm.

Next, we recap the tighter Rényi DP guarantee of Chourasia et al. (2021) under stronger assumptions on the loss function.

Theorem G.9 (Rényi DP guarantee for Noisy-GD Algorithm 2 (Chourasia et al., 2021))**.** *If $\boldsymbol{\ell}(\theta;\mathbf{x})$ is convex, L-Lipschitz, and β-smooth and $\mathbf{r}(\theta)$ is the $L2$ regularizer with constant λ, then Noisy-GD with learning rate $\eta < \frac{1}{\beta+\lambda}$ satisfies (q,ε)-Rényi DP with $\varepsilon = \frac{4qL^2}{\lambda\sigma^2 n^2}\left(1-e^{-\lambda\eta K/2}\right)$.*

G.4. Proofs for Subsection 5.1

In this appendix, we provide a proof of our Theorem 5.1 which applies to convex losses $\boldsymbol{\ell}(\theta;\mathbf{x})$ under $L2$ regularizer $\mathbf{r}(\theta)$. Let $\mathcal{D}_0 \in \mathcal{X}^n$ be any arbitrary database, and $\mathcal{Q}$ be any non-adaptive r-requester.

Our first goal in this section is to prove $(q,\varepsilon_{\rm dd})$-deletion-privacy guarantees on our proposed algorithm pair $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ (in Definition 5.1) under $\mathcal{Q}$. That is, if $(\hat{\Theta}_i)_{i\geq 0}$ is the sequence of models produced by the interaction between $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}}, \mathcal{Q})$ on $\mathcal{D}_0$, we need to show that their exists a mapping $\pi_i^{\mathcal{Q}}$ such that for all $i \geq 1$ and any $u_i \in \mathcal{U}^r$,

$$\mathrm{R}_q\left(\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1})\middle\|\pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle)\right) \leq \varepsilon_{\rm dd} \quad \text{for all } \langle \mathrm{ind}, \mathbf{y}\rangle \in u_i. \tag{98}$$

For an arbitrary replacement operation $\langle \mathrm{ind}, \mathbf{y}\rangle$ in u_i, we define a map $\pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle) = \hat{\Theta}'_i$, where the model sequence $(\hat{\Theta}'_i)_{i\geq 0}$ is produced by the interaction of between the same algorithms $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}}, \mathcal{Q})$ but on $\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle$. Since non-adaptive requester $\mathcal{Q}$ 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

$$\mathrm{R}_q\left(\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1})\middle\|\bar{\mathrm{A}}(\mathcal{D}'_{i-1}, u_i, \hat{\Theta}'_{i-1})\right) \leq \varepsilon_{\rm dd}, \tag{99}$$

for for all $u_{\leq i}$ and for all neighbouring databases $\mathcal{D}_0, \mathcal{D}'_0$ s.t. $\mathcal{D}'_0 = \mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle$ with $\langle \mathrm{ind}, \mathbf{y}\rangle \in u_i$.

Note from our Definition 5.1 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 $0 \leq k < K_{\mathrm{A}}$, the loss function is $\mathcal{L}_{\mathcal{D}_0}$,
2. for the iterations $K_{\mathrm{A}} + (j-1)K_{\bar{\mathrm{A}}} \leq k < K_{\mathrm{A}} + jK_{\bar{\mathrm{A}}}$ on any $1 \leq j \leq i-1$, the loss function is $\mathcal{L}_{\mathcal{D}_j}$, and
3. for the iterations $K_{\mathrm{A}} + (i-1)K_{\bar{\mathrm{A}}} \leq k < K_{\mathrm{A}} + iK_{\bar{\mathrm{A}}}$, the loss function is $\mathcal{L}_{\mathcal{D}_{i-1}\circ u_i}$.

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

$$\begin{cases}\Theta_{\eta(k+1)} = \Theta_{\eta k} - \eta\nabla\mathcal{L}_{\mathcal{D}(k)}(\Theta_{\eta k}) + \sqrt{2\eta\sigma^2}\mathbf{Z}_k \\ \Theta'_{\eta(k+1)} = \Theta'_{\eta k} - \eta\nabla\mathcal{L}_{\mathcal{D}'(k)}(\Theta'_{\eta k}) + \sqrt{2\eta\sigma^2}\mathbf{Z}'_k,\end{cases} \quad \text{where } \mathbf{Z}_k, \mathbf{Z}'_k \sim \mathcal{N}(0, \mathbb{I}_d), \tag{100}$$

and Θ_0 and Θ'_0 are sampled from same the weight initialization distribution ρ. To prove the bound in (99), we follow the approach proposed in Chourasia et al. (2021) of 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.

$$\begin{cases}\Theta_t = T_k(\Theta_{\eta k}) - \frac{(t-\eta k)}{2}\nabla\left(\mathcal{L}_{\mathcal{D}(k)}(\Theta_{\eta k}) - \mathcal{L}_{\mathcal{D}'(k)}(\Theta_{\eta k})\right) + \sqrt{2\sigma^2}(\mathbf{Z}_t - \mathbf{Z}_{\eta k}), \\ \Theta'_t = T_k(\Theta'_{\eta k}) + \frac{(t-\eta k)}{2}\nabla\left(\mathcal{L}_{\mathcal{D}(k)}(\Theta'_{\eta k}) - \mathcal{L}_{\mathcal{D}'(k)}(\Theta'_{\eta k})\right) + \sqrt{2\sigma^2}(\mathbf{Z}'_t - \mathbf{Z}'_{\eta k}),\end{cases} \tag{101}$$

where $\mathbf{Z}_t, \mathbf{Z}'_t$ are two independent Weiner processes, and T_k is a map on $\mathbb{R}^d$ defined as

$$T_k = \mathbb{I}_d - \frac{\eta}{2}\nabla\left(\mathcal{L}_{\mathcal{D}(k)} + \mathcal{L}_{\mathcal{D}'(k)}\right). \tag{102}$$

Note that equation (101) is identical to (100) when $t = \eta(k+1)$, and can be expressed by the following stochastic differential equations (SDEs):

$$\begin{cases} \mathrm{d}\Theta_t = -\mathbf{g}_k(\Theta_{\eta k})\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}\mathbf{Z}_t \\ \mathrm{d}\Theta'_t = +\mathbf{g}_k(\Theta'_{\eta k})\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}\mathbf{Z}'_t, \end{cases} \quad \text{where } \mathbf{g}_k(\Theta) = \frac{1}{2n}\nabla\left[\boldsymbol{\ell}(\Theta; \mathcal{D}(k)[\mathrm{ind}]) - \boldsymbol{\ell}(\Theta; \mathcal{D}'(k)[\mathrm{ind}])\right], \tag{103}$$

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

Lemma G.10 (Fokker-Planck equation for SDE (103))**.** *Fokker-Planck equation for SDE in* (103) *at time* $\eta k < t \leq \eta(k+1)$*, is*

$$\begin{cases} \partial_t \mu_t(\theta) &= \mathrm{div}\left(\mu_t(\theta)\mathbb{E}\left[\mathbf{g}_k(\Theta_{\eta k})|\Theta_t = \theta\right]\right) + \sigma^2\Delta\mu_t(\theta), \\ \partial_t \mu'_t(\theta) &= \mathrm{div}\left(\mu'_t(\theta)\mathbb{E}\left[-\mathbf{g}_k(\Theta'_{\eta k})\middle|\Theta'_t = \theta\right]\right) + \sigma^2\Delta\mu'_t(\theta), \end{cases} \tag{104}$$

where μ_t *and* μ'_t *are the densities of* Θ_t *and* Θ'_t *respectively.*

Proof. Conditioned on observing parameter $\Theta_{\eta k} = \theta_{\eta k}$, the process $(\Theta_t)_{\eta k < t \leq \eta(k+1)}$ is a Langevin diffusion along a constant Vector field (i.e. on conditioning, we get a Langevin SDE (80) with $\nabla\mathcal{L}(\theta) = \mathbf{g}_k(\theta_{\eta k})$ for all $\theta \in \mathbb{R}^d$). Therefore as per (86), the conditional probability density $\mu_{t|\eta k}(\cdot|\theta_{\eta k})$ of Θ_t given $\Theta_{\eta k}$ follows the following Fokker-Planck equation:

$$\partial_t \mu_{t|\eta k}(\cdot|\theta_{\eta k}) = \mathrm{div}\left(\mu_{t|\eta k}(\cdot|\theta_{\eta k})\mathbf{g}_k(\theta_{\eta k})\right) + \sigma^2\Delta\mu_{t|\eta k}(\cdot|\theta_{\eta k}) \tag{105}$$

Taking expectation over $\mu_{\eta k}$ which is the distribution of $\Theta_{\eta k}$,

$$\begin{aligned} \partial_t \mu_t(\cdot) &= \int \mu_{\eta k}(\theta_{\eta k})\left\{\mathrm{div}\left(\mu_{t|\eta k}(\cdot|\theta_{\eta k})\mathbf{g}_k(\theta_{\eta k})\right) + \sigma^2\Delta\mu_{t|\eta k}(\cdot|\theta_{\eta k})\right\}\mathrm{d}\theta_{\eta k} \\ &= \mathrm{div}\left(\int \mathbf{g}_k(\theta_{\eta k})\mu_{t,\eta k}(\cdot,\theta_{\eta k})\mathrm{d}\theta_{\eta k}\right) + \sigma^2\Delta\mu_t(\cdot) \\ &= \mathrm{div}\left(\mu_t(\cdot)\left\{\int \mathbf{g}_k(\theta_{\eta k})\mu_{\eta k|t}(\theta_{\eta k}|\cdot)\mathrm{d}\theta_{\eta k}\right\}\right) + \sigma^2\Delta\mu_t(\cdot) \\ &= \mathrm{div}\left(\mu_t(\cdot)\mathbb{E}\left[\mathbf{g}_k(\Theta_{\eta k})|\Theta_t = \cdot\right]\right) + \sigma^2\Delta\mu_t(\cdot). \end{aligned}$$

where $\mu_{\eta k,|t}$ is the conditional density of $\Theta_{\eta k}$ given Θ_t. Proof for second Fokker-Planck equation is similar. □

We provide an overview of how we bound equation (99) in Figure 1. 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_{A} and $K_{\bar{\mathrm{A}}}$, and in phase (II) we prove an exponential contraction in divergence $\mathrm{R}_q\left(\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}\middle\|\bar{\mathrm{A}}(\mathcal{D}'_{i-1}, u_i, \hat{\Theta}'_{i-1})\right)$ with number of iterations $K_{\bar{\mathrm{A}}}$.

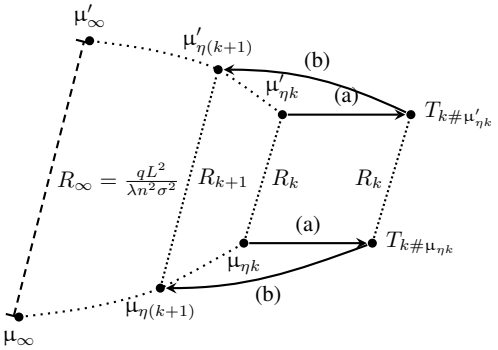
We first introduce a few lemmas that will be used in both phases. The first set of following lemmas show that the transformation $\Theta_{\eta k}, \Theta'_{\eta k} \to T_k(\Theta_{\eta k}), T_k(\Theta_{\eta k})$ preserves the Rényi divergence. To prove this property, we show that T_k is a differentiable bijective map in Lemma G.12 and apply the following Lemma from Vempala & Wibisono (2019).

Lemma G.11 (Vempala & Wibisono (2019, Lemma 15))**.** *If* $T : \mathbb{R}^d \to \mathbb{R}^d$ *is a differentiable bijective map, then for any random variables* $\Theta, \Theta' \in \mathbb{R}^d$*, and for all* $q > 0$*,*

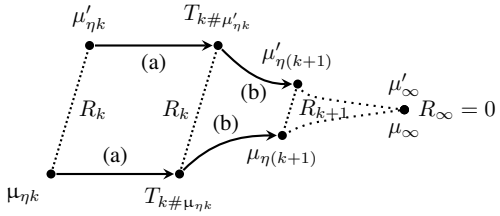
$$\mathrm{R}_q\left(T(\Theta)\|T(\Theta')\right) = \mathrm{R}_q\left(\Theta\|\Theta\right). \tag{106}$$

Lemma G.12. *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* T_k *defined in* (102) *is:*

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



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



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

Figure 1: Diagram illustrating the technical overview of Theorem G.15. Here $\mu_{\eta k}$ and $\mu_{\eta k'}$ 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 T_k (as defined in (102), and (b) a diffusion process. If divergence before descent step is $R_k = \mathrm{R}_q\left(\mu_{\eta k} \middle\| \mu'_{\eta k}\right)$, the stochastic mapping 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.

Proof. **Differentiable bijection.** To see that T_k is injective, assume $T_k(\theta) = T_k(\theta')$ for some $\theta, \theta' \in \mathbb{R}^d$. Then, by $(\beta+\lambda)$-smoothness of $\mathcal{L} \stackrel{\text{def}}{=} (\mathcal{L}_{\mathcal{D}(k)} + \mathcal{L}_{\mathcal{D}'(k)})/2$,

$$\begin{aligned}\|\theta - \theta'\|_2 &= \|T_k(\theta) + \eta\nabla\mathcal{L}(\theta) - T_k(\theta') - \eta\nabla\mathcal{L}(\theta')\|_2 \\ &= \eta\,\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2 \\ &\le \eta(\lambda+\beta)\,\|\theta - \theta'\|_2\,.\end{aligned}$$

Since $\eta < 1/(\lambda+\beta)$, we must have $\|\theta - \theta'\|_2 = 0$. For showing T_k is surjective, consider the proximal mapping

$$\mathrm{prox}_{\mathcal{L}}(\theta) = \arg\min_{\theta' \in \mathbb{R}^d} \frac{\|\theta' - \theta\|_2^2}{2} - \eta\mathcal{L}(\theta'). \tag{107}$$

Note that $\mathrm{prox}_{\mathcal{L}}(\cdot)$ is strongly convex for $\eta < \frac{1}{\lambda+\beta}$. Therefore, from KKT conditions, we have $\theta = \mathrm{prox}_{\mathcal{L}}(\theta) - \eta\nabla\mathcal{L}(\mathrm{prox}_{\mathcal{L}}(\theta)) = T_k(\mathrm{prox}_{\mathcal{L}}(\theta))$. Differentiability of T_k follows from the twice continuously differentiable assumption on $\boldsymbol{\ell}(\theta; \mathbf{x})$.

Lipschitzness. Let $\mathcal{L} \stackrel{\text{def}}{=} (\mathcal{L}_{\mathcal{D}(k)} + \mathcal{L}_{\mathcal{D}'(k)})/2$. For any $\theta, \theta' \in \mathbb{R}^d$,

$$\begin{aligned}\|T_k(\theta) - T_k(\theta')\|_2^2 &= \|\theta - \eta\nabla\mathcal{L}(\theta) - \theta' + \eta\nabla\mathcal{L}(\theta')\|_2^2 \\ &= \|\theta - \theta'\|_2^2 + \eta^2\,\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2^2 - 2\eta\,\langle\theta - \theta', \nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\rangle\,.\end{aligned}$$

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

$$\begin{aligned}&\langle\theta - \theta', \nabla g(\theta) - \nabla g(\theta')\rangle \ge \frac{1}{\beta}\,\|\nabla g(\theta) - \nabla g(\theta')\|_2^2 \\ \implies &\langle\theta - \theta', \nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\rangle - \lambda\,\|\theta - \theta'\|_2^2 \ge \frac{1}{\beta}\Big(\,\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2^2 + \lambda^2\,\|\theta - \theta'\|_2^2 \\ &\qquad\qquad - 2\lambda\,\langle\theta - \theta', \nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\rangle\,\Big) \\ \implies &\langle\theta - \theta', \nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\rangle \ge \frac{1}{2\lambda+\beta}\,\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2^2 + \frac{\lambda(\lambda+\beta)}{2\lambda+\beta}\,\|\theta - \theta'\|_2^2\,.\end{aligned}$$

Substituting this in the above inequality, and noting that $\eta \leq \frac{2}{2\lambda+\beta}$, we get

$$\begin{aligned}\|T_k(\theta) - T_k(\theta')\|_2^2 &\leq \left(1 - \frac{2\eta\lambda(\lambda+\beta)}{2\lambda+\beta}\right)\|\theta-\theta'\|_2^2 + \left(\eta^2 - \frac{2\eta}{\beta+2\lambda}\right)\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\|_2^2 \\ &\leq \left(1 - \frac{2\eta\lambda(\lambda+\beta)}{2\lambda+\beta}\right)\|\theta-\theta'\|_2^2 + \left(\eta^2\lambda^2 - \frac{2\eta\lambda^2}{\beta+2\lambda}\right)\|\theta-\theta'\|_2^2 \\ &= (1-\eta\lambda)^2\|\theta-\theta'\|_2^2.\end{aligned}$$

□

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

Lemma G.13 (Rate of change of Rényi divergence (Chourasia et al., 2021))**.** *Let $V_t, V_t' : \mathbb{R}^d \to \mathbb{R}^d$ be two time dependent vector field such that $\max_{\theta\in\mathbb{R}^d}\|V_t(\theta) - V_t'(\theta)\|_2 \leq L$ 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*

$$\begin{cases}\partial_t\mu_t(\theta) = \operatorname{div}(\mu_t(\theta)V_t(\theta)) + \sigma^2\Delta\mu_t(\theta) \quad \textit{and} \\ \partial_t\mu_t'(\theta) = \operatorname{div}(\mu_t'(\theta)V_t'(\theta)) + \sigma^2\Delta\mu_t'(\theta),\end{cases} \tag{108}$$

respectively, where μ_t and μ_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

$$\partial_t\mathrm{R}_q(\mu_t\|\mu_t') \leq \frac{qL^2}{2\sigma^2} - \frac{q\sigma^2}{2}\frac{\mathrm{I}_q(\mu_t\|\mu_t')}{\mathrm{E}_q(\mu_t\|\mu_t')}. \tag{109}$$

We will apply the above lemma to the Fokker-Planck equation (104) of our pair of tracing diffusion SDE (101) and solve the resulting differential inequality to prove the bound in (99). To assist our proof, we rely on the following lemma showing that our two tracing diffusion satisfy the LS inequality described in Definition G.1, which enables the use the inequality (96) in Lemma G.6.

Lemma G.14. *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 $(\Theta_t)_{0\leq t\leq\eta(K_\mathrm{A}+iK_{\bar{\mathrm{A}}})}$ and $(\Theta_t')_{0\leq t\leq\eta(K_\mathrm{A}+iK_{\bar{\mathrm{A}}})}$ defined in* (101) *with $\Theta_0, \Theta_0' \sim \rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda(1-\eta\lambda/2)}\mathbb{I}_d\right)$ satisfy* LS *inequality with constant $\lambda(1-\eta\lambda/2)$.*

Proof. For any iteration $0 \leq k < K_\mathrm{A} + iK_{\bar{\mathrm{A}}}$ in the extended run of Noisy-GD, and any $0 \leq s \leq \eta$, let's define two functions $\mathcal{L}_s, \mathcal{L}_s' : \mathbb{R}^d \to \mathbb{R}$ as follows:

$$\mathcal{L}_s = \frac{1+s/\eta}{2}\mathcal{L}_{\mathcal{D}(k)} + \frac{1-s/\eta}{2}\mathcal{L}_{\mathcal{D}'(k)}, \quad \text{and } \mathcal{L}_s' = \frac{1-s/\eta}{2}\mathcal{L}_{\mathcal{D}(k)} + \frac{1+s/\eta}{2}\mathcal{L}_{\mathcal{D}'(k)}. \tag{110}$$

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 $0 \leq s \leq \eta$ and any k. We define maps $T_s, T_s' : \mathbb{R}^d \to \mathbb{R}^d$ as

$$T_s(\theta) = \theta - \eta\nabla\mathcal{L}_s(\theta), \quad \text{and } T_s'(\theta) = \theta - \nabla\mathcal{L}_s'(\theta). \tag{111}$$

From a similar argument as in Lemma G.12, both T_s and T_s' are $(1-\eta\lambda)$-Lipschitz for learning rate $\eta \leq \frac{2}{2\lambda+\beta}$.

Note that the densities of Θ_t and Θ_t' of the tracing diffusion for $t = \eta k + s$ can be respectively expressed as

$$\mu_t = T_{s\#}(\mu_{\eta k}) \circledast \mathcal{N}\left(0, 2s\sigma^2\mathbb{I}_d\right), \quad \text{and } \mu_t' = T_{s\#}'(\mu_{\eta k}') \circledast \mathcal{N}\left(0, 2s\sigma^2\mathbb{I}_d\right), \tag{112}$$

where $\mu_{\eta k}$ and $\mu_{\eta k}'$ 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 G.1 they satisfy LS inequality with constant $\lambda(1-\eta\lambda/2)$.

Induction step: Suppose $\mu_{\eta k}$ and $\mu'_{\eta k}$ satisfy LS inequality with constant $\lambda(1-\eta\lambda/2)$. Since equation (112) shows that μ_t, μ'_t are both Gaussian convolution on a pushforward distribution of $\mu_{\eta k}, \mu'_{\eta k}$ respectively over a Lipschitz function, from Lemma G.1 and Lemma G.2, both μ_t, μ'_t satisfy LS inequality with constant

$$\left(\frac{(1-\eta\lambda)^2}{\lambda(1-\eta\lambda/2)} + 2s\right)^{-1} \geq \lambda(1-\eta\lambda/2) \times \underbrace{[(1-\eta\lambda)^2 + \lambda\eta(2-\eta\lambda)]^{-1}}_{=1}, \tag{113}$$

for all $\eta k \leq t \leq \eta(k+1)$. □

We are now ready to prove the deletion-privacy bound in (99).

Theorem G.15 (Deletion-Privacy guarantee on $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ under convexity)**.** *Let the weight initialization distribution be $\mathsf{p} = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda(1-\eta\lambda/2)}\right)$, 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 $(\mathrm{A}, \bar{\mathrm{A}})$ satisfies a $(q, \varepsilon_{\mathrm{dd}})$-deletion-privacy guarantee under all non-adaptive r-requesters for any noise variance $\sigma^2 > 0$ and $K_{\mathrm{A}} \geq 0$ if*

$$K_{\bar{\mathrm{A}}} \geq \frac{2}{\eta\lambda} \log\left(\frac{4qL^2}{\lambda\varepsilon_{\mathrm{dd}}\sigma^2 n^2}\right). \tag{114}$$

Proof. Following the preceding discussion, to prove this theorem, it suffices to show that the inequality (99) holds under the stated conditions. Consider the Fokker-Planck equation described in Lemma G.10 for the pair of tracing diffusions SDEs in (103): at any time t in duration $\eta k < t \leq \eta(k+1)$ for any iteration $0 \leq k < K_{\mathrm{A}} + iK_{\bar{\mathrm{A}}}$,

$$\begin{cases} \partial_t \mu_t(\theta) & = \operatorname{div}\left(\mu_t(\theta)\mathbb{E}\left[\mathbf{g}_k(\Theta_{\eta k})|\Theta_t = \theta\right]\right) + \sigma^2\Delta\mu_t(\theta), \\ \partial_t \mu'_t(\theta) & = \operatorname{div}\left(\mu'_t(\theta)\mathbb{E}\left[-\mathbf{g}_k(\Theta'_{\eta k})\middle|\Theta'_t = \theta\right]\right) + \sigma^2\Delta\mu'_t(\theta), \end{cases} \tag{115}$$

where μ_t and μ'_t are the distribution of Θ_t and Θ'_t. Since $\boldsymbol{\ell}(\theta; \mathbf{x})$ is L-Lipschitz and for any $K_{\mathrm{A}} + (i-1)K_{\bar{\mathrm{A}}} \leq k < K_{\mathrm{A}} + iK_{\bar{\mathrm{A}}}$ we have $\mathcal{D}(k)[\mathrm{ind}] = \mathcal{D}'(k)[\mathrm{ind}]$, note from the definition of $\mathbf{g}_k(\theta)$ in (103) that

$$\left\|\mathbb{E}\left[\mathbf{g}_k(\Theta_{\eta k})|\Theta_t = \theta\right] - \mathbb{E}\left[-\mathbf{g}_k(\Theta'_{\eta k})|\Theta'_t = \theta\right]\right\|_2 \leq \begin{cases} \frac{2L}{n} & \text{if } k < K_{\mathrm{A}} + (i-1)K_{\bar{\mathrm{A}}} \\ 0 & \text{otherwise} \end{cases}. \tag{116}$$

Therefore, applying Lemma G.13 to the above pair of Fokker-Planck equations gives that for any t in duration $\eta k < t \leq \eta(k+1)$,

$$\partial_t \mathrm{R}_q\left(\mu_t \| \mu'_t\right) \leq \frac{2qL^2}{\sigma^2 n^2}\mathbb{1}\left\{t \leq \eta(K_{\mathrm{A}} + (i-1)K_{\bar{\mathrm{A}}})\right\} - \frac{q\sigma^2}{2}\frac{\mathrm{I}_q\left(\mu_t \| \mu'_t\right)}{\mathrm{E}_q\left(\mu_t \| \mu'_t\right)}. \tag{117}$$

Equation (117) suggests a phase change in the dynamics at iteration $k = K_{\mathrm{A}} + (i-1)K_{\bar{\mathrm{A}}}$. In phase I, the divergence bound increases with time due to the effect of the differing record in database pairs $(\mathcal{D}_j, \mathcal{D}'_j)_{0 \leq j \leq i-1}$. In phase II however, the update request u_i makes $\mathcal{D}_{i-1} \circ u_i = \mathcal{D}'_{i-1} \circ u_i$, and so doing gradient descent rapidly shrinks the divergence bound. This phase change is illustrated in the Figure 1.

For brevity, we denote $R(q,t) = \mathrm{R}_q\left(\mu_t \| \mu'_t\right)$. Since $\eta < \frac{1}{\lambda+\beta} < \frac{2}{2\lambda+\beta}$, from Lemma G.14, the distribution μ'_t satisfies LS inequality with constant $\lambda(1-\lambda\eta/2)$. So, we can apply Lemma G.6 to simplify the above partial differential inequality as follows.

$$\partial_t R(q,t) + \lambda(1-\lambda\eta/2)\left(\frac{R(q,t)}{q} + (q-1)\partial_q R(q,t)\right) \leq \frac{2qL^2}{\sigma^2 n^2}\mathbb{1}\left\{t \leq \eta(K_{\mathrm{A}} + (i-1)K_{\bar{\mathrm{A}}})\right\}. \tag{118}$$

For brevity, let constant $c_1 = \lambda(1-\lambda\eta/2)$ and constant $c_2 = \frac{2L^2}{\sigma^2 n^2}$. We define $u(q,t) = \frac{R(q,t)}{q}$. Then,

$$\begin{aligned} &\partial_t R(q,t) + c_1\left(\frac{R(q,t)}{q} + (q-1)\partial_q R(q,t)\right) \leq c_2 q \times \mathbb{1}\left\{t \leq \eta(K_{\mathrm{A}} + (i-1)K_{\bar{\mathrm{A}}})\right\} \\ \implies &\partial_t u(q,t) + c_1 u(q,t) + c_1(q-1)\partial_q u(q,t) \leq c_2 \times \mathbb{1}\left\{t \leq \eta(K_{\mathrm{A}} + (i-1)K_{\bar{\mathrm{A}}})\right\}. \end{aligned}$$

For some constant $\bar{q} > 1$, let $q(s) = (\bar{q}-1)\exp\left[c_1\left\{s - \eta(K_{\mathrm{A}} + iK_{\bar{\mathrm{A}}})\right\}\right]+1$ 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 $\eta k < s \le \eta(k+1)$, the differential inequality followed along the path $u(s) = u(q(s), t(s))$ is

$$\frac{\mathrm{d}u(s)}{\mathrm{d}s} + c_1 u(s) \le c_2 \times \mathbb{1}\left\{t \le \eta(K_{\mathrm{A}} + (i-1)K_{\bar{\mathrm{A}}})\right\} \tag{119}$$

$$\implies \frac{\mathrm{d}}{\mathrm{d}s}\{e^{c_1 s}u(s)\} \le c_2 \times \mathbb{1}\left\{t \le \eta(K_{\mathrm{A}} + (i-1)K_{\bar{\mathrm{A}}})\right\}. \tag{120}$$

Since the map $T_k(\cdot)$ in (102) is a differentiable bijection for $\eta < \frac{1}{\lambda+\beta}$ as per Lemma G.12, note that Lemma G.11 implies that $\lim_{s\to\eta k^+} u(s) = u(\eta k)$. Therefore, we can directly integrate in the duration $0 \le t \le \eta(K_{\mathrm{A}} + iK_{\bar{\mathrm{A}}})$ to get

$$\begin{aligned}
&[e^{c_1 s}u(s)]_0^{\eta(K_{\mathrm{A}}+iK_{\bar{\mathrm{A}}})} \le \int_0^{\eta(K_{\mathrm{A}}+(i-1)K_{\bar{\mathrm{A}}})} c_2 e^{c_1 s}\mathrm{d}s \\
\implies & e^{c_1\eta(K_{\mathrm{A}}+iK_{\bar{\mathrm{A}}})}u(\eta(K_p + iK_u)) - u(0) \le \frac{c_2}{c_1}[e^{c_1\eta(K_{\mathrm{A}}+(i-1)K_{\bar{\mathrm{A}}})} - 1] \\
\implies & u(\eta(K_{\mathrm{A}} + iK_{\bar{\mathrm{A}}})) \le \frac{c_2}{c_1}e^{-c_1\eta K_{\bar{\mathrm{A}}}}. \qquad \text{(Since } u(0) = R(q(0),0)/q(0) = 0.\text{)}
\end{aligned}$$

Noting that $q(0) \ge 1$, on reverting the substitution, we get

$$\begin{aligned}
\mathrm{R}_{\bar{q}}\left(\mu_{\eta(K_{\mathrm{A}}+iK_{\bar{\mathrm{A}}})}\middle\|\mu'_{\eta(K_{\mathrm{A}}+iK_{\bar{\mathrm{A}}})}\right) &\le \frac{2\bar{q}L^2}{\lambda\sigma^2 n^2(1-\eta\lambda/2)}\exp\left(-\eta\lambda K_{\bar{\mathrm{A}}}(1-\eta\lambda/2)\right) \\
&\le \frac{4\bar{q}L^2}{\lambda\sigma^2 n^2}\exp\left(-\frac{\eta\lambda K_u}{2}\right) \qquad \text{(Since } \eta < \tfrac{1}{\lambda+\beta}\text{)}
\end{aligned}$$

Recall from our construction that $\mu_{\eta(K_{\mathrm{A}}+iK_{\bar{\mathrm{A}}})}$ and $\mu'_{\eta(K_{\mathrm{A}}+iK_{\bar{\mathrm{A}}})}$ are the distributions of outputs $\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1})$ and $\bar{\mathrm{A}}(\mathcal{D}'_{i-1}, u_i, \hat{\Theta}'_{i-1})$ respectively. Therefore, choosing $K_{\bar{\mathrm{A}}}$ as specified in the theorem statement concludes the proof. □

Our next goal in this section is to provide utility guarantees for the algorithm pair $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ in form of excess empirical risk bounds. For that, we introduce some additional auxiliary results first. The following Lemma G.16 shows that excess empirical risks does not increase too much on replacing r records in a database, and Lemma G.17 provides a convergence guarantee on the excess empirical risk of Noisy-GD algorithm under convexity.

Lemma G.16. *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 $\mathcal{D} \in \mathcal{X}^n$ after applying any edit request $u \in \mathcal{U}^r$ that modifies no more than r records is bounded as*

$$\mathrm{err}(\Theta; \mathcal{D}\circ u) \le \left(1 + \frac{\beta}{\lambda}\right)\left[2\,\mathrm{err}(\Theta;\mathcal{D}) + \frac{16r^2L^2}{\lambda n^2}\right]. \tag{121}$$

Proof. Let $\theta^*_{\mathcal{D}}$ and $\theta^*_{\mathcal{D}\circ u}$ be the minimizers of objectives $\mathcal{L}_{\mathcal{D}}(\cdot)$ and $\mathcal{L}_{\mathcal{D}\circ u}(\cdot)$ as defined in (14). From λ-strong convexity of the $\mathcal{L}_{\mathcal{D}}$,

$$\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}\circ u}) - \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}}) \ge \frac{\lambda}{2}\|\theta^*_{\mathcal{D}\circ u} - \theta^*_{\mathcal{D}}\|_2^2. \tag{122}$$

From optimality of $\theta^*_{\mathcal{D}\circ u}$ and L-Lipschitzness of $\boldsymbol{\ell}(\theta;\mathbf{x})$, we have

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}\circ u}) &= \mathcal{L}_{\mathcal{D}\circ u}(\theta^*_{\mathcal{D}\circ u}) + \frac{1}{n}\left(\sum_{\mathbf{x}\in\mathcal{D}}\boldsymbol{\ell}(\theta^*_{\mathcal{D}\circ u};\mathbf{x}) - \sum_{\mathbf{x}\in\mathcal{D}\circ u}\boldsymbol{\ell}(\theta^*_{\mathcal{D}\circ u};\mathbf{x})\right) \\
&\le \mathcal{L}_{\mathcal{D}\circ u}(\theta^*_{\mathcal{D}}) + \frac{1}{n}\left(\sum_{\mathbf{x}\in\mathcal{D}}\boldsymbol{\ell}(\theta^*_{\mathcal{D}\circ u};\mathbf{x}) - \sum_{\mathbf{x}\in\mathcal{D}\circ u}\boldsymbol{\ell}(\theta^*_{\mathcal{D}\circ u};\mathbf{x})\right) \\
&= \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}}) + \frac{1}{n}\sum_{\mathbf{x}\in\mathcal{D}}\left(\boldsymbol{\ell}(\theta^*_{\mathcal{D}\circ u};\mathbf{x}) - \boldsymbol{\ell}(\theta^*_{\mathcal{D}};\mathbf{x})\right) + \frac{1}{n}\sum_{\mathbf{x}\in\mathcal{D}\circ u}\left(\boldsymbol{\ell}(\theta^*_{\mathcal{D}};\mathbf{x}) - \boldsymbol{\ell}(\theta^*_{\mathcal{D}\circ u};\mathbf{x})\right) \\
&\le \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}}) + \frac{2rL}{n}\|\theta^*_{\mathcal{D}\circ u} - \theta^*_{\mathcal{D}}\|_2.
\end{aligned}$$

Combining the two inequalities give

$$\|\theta^*_{\mathcal{D}\circ u} - \theta^*_{\mathcal{D}}\|_2 \leq \frac{4rL}{\lambda n}. \tag{123}$$

Therefore, from $(\lambda+\beta)$-smoothness of $\mathcal{L}_{\mathcal{D}\circ u}$ and λ-strong convexity of $\mathcal{L}_{\mathcal{D}}$, we have

$$\begin{aligned}
\mathrm{err}(\Theta; \mathcal{D}\circ u) &= \mathbb{E}\left[\mathcal{L}_{\mathcal{D}\circ u}(\Theta) - \mathcal{L}_{\mathcal{D}\circ u}(\theta^*_{\mathcal{D}\circ u})\right] \\
&\leq \frac{\lambda+\beta}{2}\mathbb{E}\left[\|\Theta - \theta^*_{\mathcal{D}\circ u}\|_2^2\right] \\
&\leq (\lambda+\beta)\left[\mathbb{E}\left[\|\Theta - \theta^*_{\mathcal{D}}\|_2^2\right] + \|\theta^*_{\mathcal{D}} - \theta^*_{\mathcal{D}\circ u}\|_2^2\right] \\
&\leq \left(1+\frac{\beta}{\lambda}\right)\left[2\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta) - \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})\right] + \frac{16r^2L^2}{\lambda n^2}\right].
\end{aligned}$$

□

Lemma G.17 (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 $\Theta_{\eta K}$ = Noisy-GD$(\mathcal{D}, \Theta_0, K)$ for any $\mathcal{D} \in \mathcal{X}^n$ is bounded as*

$$\mathrm{err}(\Theta_{\eta K}; \mathcal{D}) \leq \mathrm{err}(\Theta_0; \mathcal{D})e^{-\lambda\eta K/2} + \left(1+\frac{\beta}{\lambda}\right)d\sigma^2. \tag{124}$$

Proof. Let $\Theta_{\eta k}$ denote the kth iteration parameter of Noisy-GD run. Recall that $k+1$th noisy gradient update step is

$$\Theta_{\eta(k+1)} = \Theta_{\eta k} - \eta\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) + \sqrt{2\eta\sigma^2}\mathbf{Z}_k. \tag{125}$$

From $(\beta+\lambda)$-smoothness of $\mathcal{L}_{\mathcal{D}}$, we have

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(\Theta_{\eta(k+1)}) &\leq \mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) + \left\langle \nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}), \Theta_{\eta(k+1)} - \Theta_{\eta k}\right\rangle + \frac{\beta+\lambda}{2}\left\|\Theta_{\eta(k+1)} - \Theta_{\eta k}\right\|_2^2 \\
&= \mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \eta\left\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\right\|_2^2 + \sqrt{2\eta\sigma^2}\left\langle\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}), \mathbf{Z}_k\right\rangle \\
&\quad + \frac{\eta^2(\beta+\lambda)}{2}\left\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\right\|_2^2 + \eta\sigma^2(\beta+\lambda)\left\|\mathbf{Z}_k\right\|_2^2 \\
&\quad - \eta\sqrt{2\eta\sigma^2}(\beta+\lambda)\left\langle\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}), \mathbf{Z}_k\right\rangle
\end{aligned}$$

On taking expectation over the joint distribution of $\Theta_{\eta k}, \Theta_{\eta(k+1)}, \mathbf{Z}_k$, the above simplifies to

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta_{\eta(k+1)})\right] \leq \mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\right] - \eta\left(1 - \frac{\eta(\lambda+\beta)}{2}\right)\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\right\|_2^2\right] + \eta d\sigma^2(\beta+\lambda). \tag{126}$$

Let $\theta^*_{\mathcal{D}} = \underset{\theta\in\mathbb{R}^d}{\arg\min}\ \mathcal{L}_{\mathcal{D}}(\theta)$. From λ-strong convexity of $\mathcal{L}_{\mathcal{D}}$, for any $\theta \in \mathbb{R}^d$, we have

$$\|\nabla\mathcal{L}_{\mathcal{D}}(\theta)\|_2^2 \geq 2\lambda(\mathcal{L}_{\mathcal{D}}(\theta) - \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})). \tag{127}$$

Let $\gamma = \lambda\eta(2 - \eta(\lambda+\beta))$. Plugging this in the above inequality, and subtracting $\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})$ on both sides, for $\eta < \frac{1}{\lambda+\beta}$, we get

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta_{\eta(k+1)}) - \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})\right] &\leq (1-\gamma)\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})\right] + \eta d\sigma^2(\beta+\lambda) \\
&\leq (1-\gamma)^{k+1}\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta_0) - \mathcal{L}_{\mathcal{D}}(\theta^*)\right] + \eta d\sigma^2(\beta+\lambda)(1 + \cdots + (1-\gamma)^{k+1}) \\
&\leq e^{-\gamma(k+1)/2}\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta_0) - \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})\right] + \frac{\eta d\sigma^2(\beta+\lambda)}{\gamma}.
\end{aligned}$$

For $\eta < \frac{1}{\lambda+\beta}$, note that $\gamma \geq \lambda\eta$, and so

$$\mathrm{err}(\Theta_{\eta K}; \mathcal{D}) \leq \mathrm{err}(\Theta_0; \mathcal{D})e^{-\lambda\eta K/2} + \left(1+\frac{\beta}{\lambda}\right)d\sigma^2. \tag{128}$$

□

Finally, we are ready to prove our main Theorem 5.1 showing that the algorithm pair $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ solves the data-deletion problem as described in Section 4. We basically combine the Rényi DP guarantee in Theorem G.9, non-adaptive deletion-privacy guarantee in Theorem G.15, and prove excess empirical risk bound using Lemma G.17 and Lemma G.16.

Theorem 5.1 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda, \beta, L > 0$, constant $q > 1$, and constants $\varepsilon_{\mathrm{dp}} \geq \varepsilon_{\mathrm{dd}} > 0$. Define constant $\kappa = \frac{\lambda+\beta}{\lambda}$. 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 $\sigma^2 = \frac{4qL^2}{\lambda\varepsilon_{\mathrm{dp}}n^2}$, and the weight initialization distribution is $\rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda(1-\eta\lambda/2)\mathbb{I}_d}\right)$, then*

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

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

$$\text{if} \quad K_{\bar{\mathrm{A}}} \geq 4\kappa \log \frac{\varepsilon_{\mathrm{dp}}}{\varepsilon_{\mathrm{dd}}}, \tag{129}$$

(3.) *and all models in sequence $(\hat{\Theta}_i)_{i\geq 0}$ produced by the interactions between $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ and $\mathcal{Q}$ on any $\mathcal{D}_0 \in \mathcal{X}^n$, where $\mathcal{Q}$ is any r-requester, have an excess empirical risk $\mathrm{err}(\hat{\Theta}_i; \mathcal{D}_i) = O\left(\frac{qd}{\varepsilon_{\mathrm{dp}}n^2}\right)$*

$$\text{if} \quad K_{\mathrm{A}} \geq 4\kappa \log\left(\frac{\varepsilon_{\mathrm{dp}}n^2}{4qd}\right), \quad \text{and} \quad K_{\bar{\mathrm{A}}} \geq 4\kappa \log\max\left\{5\kappa, \frac{8\varepsilon_{\mathrm{dp}}r^2}{qd}\right\}. \tag{130}$$

Proof. **(1.) Privacy.** By Theorem G.9, the Noisy-GD with K iterations will be $(q, \varepsilon_{\mathrm{dp}})$-Rényi DP for the stated choice of loss function, regularizer, and learning rate as long as $\sigma^2 \geq \frac{4qL^2}{\lambda\varepsilon_{\mathrm{dp}}n^2}\left(1 - e^{-\lambda\eta K/2}\right)$. Therefore, if we set $\sigma^2 = \frac{4qL^2}{\lambda\varepsilon_{\mathrm{dp}}n^2}$, Noisy-GD is $(q, \varepsilon_{\mathrm{dp}})$-Rényi DP for any K. For the same σ^2, both $\mathrm{A}_{\text{Noisy-GD}}$ and $\bar{\mathrm{A}}_{\text{Noisy-GD}}$ are also $(q, \varepsilon_{\mathrm{dp}})$-Rényi DP for any K_{A} and $K_{\bar{\mathrm{A}}}$ as they run Noisy-GD on respective databases for generating the output.

(2.) Deletion. By Theorem G.15, for the stated choice of loss function, regularizer, learning rate, and weight initialization distribution, the algorithm pair $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ satisfies $(q, \varepsilon_{\mathrm{dd}})$-deletion-privacy under all non-adaptive r-requesters $\mathcal{Q}$ if $K_{\bar{\mathrm{A}}} \geq \frac{2}{\eta\lambda}\log\left(\frac{4qL^2}{\lambda\varepsilon_{\mathrm{dd}}\sigma^2n^2}\right)$. By plugging in $\sigma^2 = \frac{4qL^2}{\lambda\varepsilon_{\mathrm{dp}}n^2}$ and $\eta = \frac{1}{2(\lambda+\beta)}$, this constraint simplifies to $K_{\bar{\mathrm{A}}} \geq 4\kappa\log\frac{\varepsilon_{\mathrm{dp}}}{\varepsilon_{\mathrm{dd}}}$.

(3.) Accuracy. We prove the induction hypothesis that under the conditions stated in the theorem, $\mathrm{err}(\hat{\Theta}_i; \mathcal{D}_i) \leq \frac{10\kappa qdL^2}{\lambda\varepsilon_{\mathrm{dp}}n^2}$ for all $i \geq 0$.

Base case: The minimizer $\theta^*_{\mathcal{D}_0}$ of $\mathcal{L}_{\mathcal{D}_0}$ satisfies

$$\nabla\mathcal{L}_{\mathcal{D}_0}(\theta^*_{\mathcal{D}_0}) = \frac{1}{n}\sum_{\mathbf{x}\in\mathcal{D}_0}\nabla\boldsymbol{\ell}(\theta^*_{\mathcal{D}_0}; \mathbf{x}) - \lambda\theta^*_{\mathcal{D}_0} = 0 \implies \left\|\theta^*_{\mathcal{D}_0}\right\|_2 \leq \frac{L}{\lambda}. \tag{131}$$

As a result, the excess empirical risk of initialization weights $\Theta_0 \sim \rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda(1-\eta\lambda/2)\mathbb{I}_d}\right)$ on $\mathcal{L}_{\mathcal{D}_0}$ is bounded as

$$\begin{aligned}
\mathrm{err}(\Theta_0; \mathcal{D}_0) &= \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_0}(\Theta_0) - \mathcal{L}_{\mathcal{D}_0}(\theta^*_{\mathcal{D}_0})\right] \\
&\leq \frac{(\lambda+\beta)}{2}\mathbb{E}\left[\left\|\Theta_0 - \theta^*_{\mathcal{D}_0}\right\|_2^2\right] && \text{(From } (\lambda+\beta)\text{-smoothness of } \mathcal{L}_{\mathcal{D}_0}) \\
&= \frac{(\lambda+\beta)}{2}\left[\left\|\theta^*_{\mathcal{D}_0}\right\|_2^2 + \mathbb{E}\left[\|\Theta_0\|_2^2\right] - 2\mathbb{E}\left[\left\langle\theta^*_{\mathcal{D}_0}, \Theta_0\right\rangle\right]\right] \\
&\leq \left(1 + \frac{\beta}{\lambda}\right)\left[\frac{L^2}{2\lambda} + \frac{\sigma^2 d}{2-\lambda\eta}\right] && \text{(From (131) and } \mathbb{E}\left[\|\mathbf{Z}\|_2^2\right] = d \text{ if } \mathbf{Z} \sim \mathcal{N}(0, \mathbb{I}_d).) \\
&\leq \kappa\left[\frac{L^2}{2\lambda} + d\sigma^2\right].
\end{aligned}$$

Since $\hat{\Theta}_0 = \mathrm{A}_{\text{Noisy-GD}}(\mathcal{D}_0) = \text{Noisy-GD}(\mathcal{D}_0, \Theta_0, K_{\mathrm{A}})$, by Lemma G.17, running $K_{\mathrm{A}} \geq 2\kappa \log\left(\frac{\varepsilon_{\mathrm{dp}} n^2}{4qd}\right)$ iterations gives

$$\begin{aligned}
\mathrm{err}(\hat{\Theta}_0; \mathcal{D}_0) &\leq \mathrm{err}(\Theta_0; \mathcal{D}_0) e^{-\lambda\eta K_{\mathrm{A}}/2} + \kappa d\sigma^2 \\
&\leq \kappa \left[\frac{L^2}{2\lambda} + d\sigma^2\right] e^{-\lambda\eta K_{\mathrm{A}}/2} + \kappa d\sigma^2 \\
&\leq \frac{\kappa L^2}{2\lambda} e^{-\lambda\eta K_{\mathrm{A}}/2} + \frac{8\kappa q d L^2}{\lambda \varepsilon_{\mathrm{dp}} n^2} && \text{(On substituting } \sigma^2 = \tfrac{4qL^2}{\lambda\varepsilon_{\mathrm{dp}} n^2}\text{)} \\
&\leq \frac{10\kappa q d L^2}{\lambda \varepsilon_{\mathrm{dp}} n^2} && \text{(Since } K_{\mathrm{A}} \geq 4\kappa \log\left(\tfrac{\varepsilon_{\mathrm{dp}} n^2}{4qd}\right)\text{)}
\end{aligned}$$

Induction step: Assume that $\mathrm{err}(\hat{\Theta}_{i-1}; \mathcal{D}_{i-1}) \leq \frac{10\kappa q d L^2}{\lambda \varepsilon_{\mathrm{dp}} n^2}$. Since $\hat{\Theta}_i = \bar{\mathrm{A}}_{\text{Noisy-GD}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) = \text{Noisy-GD}(\mathcal{D}_i, \hat{\Theta}_{i-1}, K_{\bar{\mathrm{A}}})$, by Lemma G.17 and Lemma G.16, running $K_{\bar{\mathrm{A}}} \geq 2\kappa \log \max\left\{5\kappa, \frac{8r^2}{qd}\right\}$ iterations gives

$$\begin{aligned}
\mathrm{err}(\hat{\Theta}_i; \mathcal{D}_i) &\leq \kappa \left[2\mathrm{err}(\hat{\Theta}_{i-1}; \mathcal{D}_{i-1}) + \frac{16 r^2 L^2}{\lambda n^2}\right] e^{-\lambda\eta K_{\bar{\mathrm{A}}}/2} + \kappa d\sigma^2 \\
&\leq \kappa \left[\frac{20\kappa q d L^2}{\lambda \varepsilon_{\mathrm{dp}} n^2} + \frac{16 r^2 L^2}{\lambda n^2}\right] e^{-\lambda\eta K_{\bar{\mathrm{A}}}/2} + \frac{4\kappa q d L^2}{\lambda \varepsilon_{\mathrm{dp}} n^2} && \text{(Substituting } \sigma^2\text{)} \\
&\leq \frac{16\kappa r^2 L^2}{\lambda n^2} e^{-\lambda\eta K_{\bar{\mathrm{A}}}/2} + \frac{8\kappa q d L^2}{\lambda \varepsilon_{\mathrm{dp}} n^2} && \text{(Since } K_{\bar{\mathrm{A}}} \geq 4\kappa \log(5\kappa)\text{)} \\
&\leq \frac{10\kappa q d L^2}{\lambda \varepsilon_{\mathrm{dp}} n^2} && \text{(Since } K_{\bar{\mathrm{A}}} \geq 4\kappa \log \tfrac{8\varepsilon_{\mathrm{dp}} r^2}{qd}\text{)}
\end{aligned}$$

□

G.5. Proofs for Subsection 5.2

In this Appendix, we provide a proof of our deletion-privacy and utility guarantee in Theorem 5.2 which applies to non-convex but bounded losses $\boldsymbol{\ell}(\theta; \mathbf{x})$ under $L2$ regularizer $\mathbf{r}(\theta)$. Suppose $\mathcal{D}_0 \in \mathcal{X}^n$ is an arbitrary database, $\mathcal{Q}$ is any non-adaptive r-requester, and $(\hat{\Theta}_i)_{i\geq 0}$ is the model sequence generated by the interaction of $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}}, \mathcal{Q})$. Our first goal will be to prove $(q, \varepsilon_{\mathrm{dd}})$-deletion-privacy guarantee on $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ and we will later use it for arguing utility as well. Recall from Definition 4.1 that to prove $(q, \varepsilon_{\mathrm{dd}})$-deletion-privacy, we need to construct a map $\pi_i^{\mathcal{Q}} : \mathcal{X}^n \to \mathcal{O}$ such that for all $i \geq 1$ and any $u_i \in \mathcal{U}^r$,

$$\mathrm{R}_q\left(\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) \middle\| \pi_i^{\mathcal{Q}}(\mathcal{D}_0 \circ \langle \mathrm{ind}, \mathbf{y}\rangle)\right) \leq \varepsilon_{\mathrm{dd}} \quad \text{for all } \langle \mathrm{ind}, \mathbf{y}\rangle \in u_i. \tag{132}$$

Our construction of $\pi_i^{\mathcal{Q}}$ for this proof is completely different from the one described in Appendix G.4. As discussed in Remark 4.2, since $\mathcal{Q}$ is non-adaptive, it suffices to show that there exists a map $\pi : \mathcal{X}^n \to \mathcal{O}$ such that for all $i \geq 1$,

$$\mathrm{R}_q\left(\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) \middle\| \pi(\mathcal{D}_i)\right) \leq \varepsilon_{\mathrm{dd}}, \tag{133}$$

for all $\mathcal{D}_0 \in \mathcal{X}^n$ 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:

$$\pi(\mathcal{D})(\theta) \propto e^{-\mathcal{L}_{\mathcal{D}}(\theta)/\sigma^2}. \tag{134}$$

The high-level intuition for this construction is that Noisy-GD can be interpreted as Unadjusted Langevin Algorithm (ULA) (Roberts & Tweedie, 1996), which is a discretization of the Langevin diffusion (described in eqn. (80)) that eventually converges to this Gibbs distribution (see Appendix G.1 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 & Wibisono (2019) followed by Chewi et al. (2021) resolved this problem with an elegant argument, relying solely on isoperimetric assumptions over (134) 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. (2021)'s convergence argument adapted to our Noisy-GD formulation, with a slightly 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 when run for K iterations on a database $\mathcal{D}$ with an arbitrary start distribution $\Theta_0 \sim \mu_0$. 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

$$\Theta_t = \Theta_{\eta k} - (t-\eta k)\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) + \sqrt{2\sigma^2}(\mathbf{Z}_t - \mathbf{Z}_{\eta k}), \tag{135}$$

where $\mathbf{Z}_t$ is a Weiner process. 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 (101)). We denote the distribution of random variable Θ_t with μ_t. The tracing diffusion during the duration $\eta k \leq t \leq \eta(k+1)$ is characterized by the following Fokker-Planck equation.

Lemma G.18 (Proposition 14 (Chewi et al., 2021))**.** *For tracing diffusion Θ_t defined in* (135)*, the equivalent Fokker-Planck equation in the interval $\eta k \leq t \leq \eta(k+1)$ is*

$$\partial_t \mu_t(\theta) = \operatorname{div}\left(\left\{\mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)|\Theta_t = \theta\right] + \sigma^2\nabla\log\frac{\mu_t(\theta)}{\pi(\mathcal{D})(\theta)}\right\}\mu_t(\theta)\right), \tag{136}$$

where $\pi(\mathcal{D})$ is the Gibbs distribution defined in (134)*.*

Proof. Conditioned on observing parameter $\Theta_{\eta k} = \theta_{\eta k}$, the process $(\Theta_t)_{\eta k\leq t\leq\eta(k+1)}$ is a Langevin diffusion along a constant Vector field $\nabla\mathcal{L}_{\mathcal{D}}(\theta_{\eta k})$. Therefore, the conditional probability density $\mu_{t|\eta k}(\cdot|\theta_{\eta k})$ of Θ_t given $\theta_{\eta k}$ follows the following Fokker-Planck equation.

$$\partial_t \mu_{t|\eta k}(\cdot|\theta_{\eta k}) = \sigma^2\Delta\mu_{t|\eta k}(\cdot|\theta_{\eta k}) + \operatorname{div}\left(\mu_{t|\eta k}(\cdot|\theta_{\eta k})\nabla\mathcal{L}_{\mathcal{D}}(\theta_{\eta k})\right) \tag{137}$$

Taking expectation over $\Theta_{\eta k}$, we have

$$\begin{aligned}
\partial_t\mu_t(\cdot) &= \int \mu_{\eta k}(\theta_{\eta k})\left\{\sigma^2\Delta\mu_{t|\eta k}(\cdot|\theta_{\eta k}) + \operatorname{div}\left(\mu_{t|\eta k}(\cdot|\theta_{\eta k})\nabla\mathcal{L}_{\mathcal{D}}(\theta_{\eta k})\right)\right\}\mathrm{d}\theta_{\eta k} \\
&= \sigma^2\Delta\mu_t(\cdot) + \operatorname{div}\left(\mu_t(\cdot)\nabla\mathcal{L}_{\mathcal{D}}(\cdot)\right) + \operatorname{div}\left(\mu_t(\cdot)\int\left[\nabla\mathcal{L}_{\mathcal{D}}(\theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\cdot)\right]\mu_{\eta k|t}(\theta_{\eta k}|\cdot)\mathrm{d}\theta_{\eta k}\right) \\
&= \sigma^2\operatorname{div}\left(\mu_t(\cdot)\nabla\log\frac{\mu_t(\cdot)}{\pi(\mathcal{D})(\cdot)}\right) + \operatorname{div}\left(\mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\cdot)|\Theta_t = \cdot\right]\mu_t(\cdot)\right),
\end{aligned}$$

where $\mu_{\eta k|t}$ is the conditional density of $\Theta_{\eta k}$ given Θ_t. For the last equality, we have used the fact that $\nabla\mathcal{L}_{\mathcal{D}} = -\sigma^2\nabla\log\pi(\mathcal{D})$ from (134). □

The following lemma provides a partial differential inequality that bounds the rate of change in Rényi divergence $\mathrm{R}_q\left(\mu_t\|\pi(\mathcal{D})\right)$ using Fokker-Planck equation (136) of Noisy GD's tracing diffusion.

Lemma G.19 ((Chewi et al., 2021, Proposition 15))**.** *Let $\rho_t := \mu_t/\pi(\mathcal{D})$ where $\pi(\mathcal{D})$ is the Gibbs distribution defined in* (134) *and $\psi_t := \rho_t^{q-1}/\mathrm{E}_q\left(\rho_t\|\pi(\mathcal{D})\right)$. The rate of change in $\mathrm{R}_q\left(\mu_t\|\pi(\mathcal{D})\right)$ along racing diffusion in time $\eta k \leq t \leq \eta(k+1)$ is bounded as*

$$\partial_t\mathrm{R}_q\left(\mu_t\|\pi(\mathcal{D})\right) \leq -\frac{3q\sigma^2}{4}\frac{\mathrm{I}_q\left(\mu_t\|\pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)} + \frac{q}{\sigma^2}\mathbb{E}\left[\psi_t(\Theta_t)\left\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)\right\|_2^2\right]. \tag{138}$$

Proof. For brevity, let $\Delta_t(\cdot) = \mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)|\Theta_t = \cdot\right]$ in context of this proof. From Lebinz integral rule, we have

$$\begin{aligned}
\partial_t \mathrm{R}_q\left(\mu_t\|\pi(\mathcal{D})\right) &= \frac{q}{(q-1)\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)}\int\left(\frac{\mu_t}{\pi(\mathcal{D})}\right)^{q-1}\partial_t\mu_t \mathrm{d}\theta \\
&= \frac{q}{(q-1)\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)}\int \rho_t^{q-1}\mathrm{div}\left(\left\{\Delta_t + \sigma^2\nabla\log\rho_t\right\}\mu_t\right)\mathrm{d}\theta && \text{(From (136))} \\
&= -\frac{q}{(q-1)\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)}\int\left\langle\nabla\left(\rho_t^{q-1}\right), \Delta_t + \sigma^2\nabla\log\rho_t\right\rangle\mu_t\mathrm{d}\theta \\
&= -\frac{q}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)}\int\rho_t^{q-2}\left\langle\nabla\rho_t, \Delta_t + \sigma^2\frac{\nabla\rho_t}{\rho_t}\right\rangle\mu_t\mathrm{d}\theta \\
&= -\frac{q}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)}\left\{\sigma^2\mathrm{I}_q\left(\mu_t\|\pi(\mathcal{D})\right) + \frac{2}{q}\underbrace{\mathop{\mathbb{E}}_{\mu_t}\left[\rho_t^{q/2-1}\left\langle\nabla\left(\rho_t^{q/2}\right), \Delta_t\right\rangle\right]}_{\stackrel{\text{def}}{=}F_1}\right\} && \text{(From (29))}
\end{aligned}$$

Note that the expectation in $\Delta_t(\cdot)$ is over the conditional distribution $\mu_{\eta k|t}$ while the expectation in F_1 is over μ_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.

$$\begin{aligned}
-F_1 &= \mathop{\mathbb{E}}_{\Theta_t\sim\mu_t}\left[\rho_t^{q/2-1}(\Theta_t)\left\langle\nabla\left(\rho_t^{q/2}\right)(\Theta_t), \mathop{\mathbb{E}}_{\Theta_{\eta k}\sim\mu_{\eta k|t}}\left[\nabla\mathcal{L}_{\mathcal{D}}(\Theta_t) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\right]\right\rangle\right] \\
&= \mathop{\mathbb{E}}_{\mu_{\eta k,t}}\left[\rho_t^{q/2-1}(\Theta_t)\left\langle\nabla\left(\rho_t^{q/2}\right)(\Theta_t), \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\right\rangle\right] \\
&\le \frac{\sigma^2}{2q}\mathbb{E}\left[\rho_t^{-1}(\Theta_t)\left\|\nabla\left(\rho_t^{q/2}\right)(\Theta_t)\right\|_2^2\right] + \frac{q}{2\sigma^2}\mathbb{E}\left[\rho_t^{q-1}(\Theta_t)\left\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_t) - \nabla\mathcal{L}_{B_k}(\Theta_{\eta k})\right\|_2^2\right] \\
&= \frac{q\sigma^2}{8}\mathrm{I}_q\left(\rho_t\|\mu\right) + \frac{q}{2\sigma^2}\mathbb{E}\left[\rho_t^{q-1}(\Theta_t)\left\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_t) - \nabla\mathcal{L}_{B_k}(\Theta_{\eta k})\right\|_2^2\right] && \text{(From (29))}
\end{aligned}$$

Substituting it in the preceding inequality proves the proposition. □

We need to solve the PDI (138) 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. (2021).

Lemma G.20 (Change of measure inequality (Chewi et al., 2021))**.** *If $\ell(\theta;\mathbf{x})$ is β-smooth, and regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, then for any probability density μ on $\mathbb{R}^d$,*

$$\mathop{\mathbb{E}}_{\mu}\left[\|\nabla\mathcal{L}_{\mathcal{D}}\|_2^2\right] \le 4\sigma^4\mathop{\mathbb{E}}_{\pi(\mathcal{D})}\left[\left\|\nabla\sqrt{\frac{\mu}{\pi(\mathcal{D})}}\right\|_2^2\right] + 2d\sigma^2(\beta+\lambda), \tag{139}$$

where $\pi(\mathcal{D})$ is the Gibbs distribution defined in (134)*.*

Proof. Consider the Langevin diffusion (80) described in Appendix G.1 over the potential $\mathcal{L}_{\mathcal{D}}$. The Gibbs distribution $\pi(\mathcal{D})$ is its stationary distribution, and the diffusion's infinitesimal generator $\mathcal{G}$ applied on the $\mathcal{L}_{\mathcal{D}}$ gives

$$\mathcal{G}\mathcal{L}_{\mathcal{D}} = \sigma^2\Delta\mathcal{L}_{\mathcal{D}} - \|\nabla\mathcal{L}_{\mathcal{D}}\|_2^2. \tag{140}$$

Therefore,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mu}\left[\|\nabla\mathcal{L}_{\mathcal{D}}\|_2^2\right] &= \sigma^2\mathop{\mathbb{E}}_{\mu}\left[\Delta\mathcal{L}_{\mathcal{D}}\right] - \mathop{\mathbb{E}}_{\mu}\left[\mathcal{G}\mathcal{L}_{\mathcal{D}}\right] && \text{(From (140))}\\
&\le d\sigma^2(\beta+\lambda) - \int \mathcal{G}\mathcal{L}_{\mathcal{D}}\left(\frac{\mu}{\pi(\mathcal{D})}-1\right)\pi(\mathcal{D})\mathrm{d}\theta && \text{(From } \beta\text{-smoothness and (88))}\\
&= d\beta\sigma^2(\beta+\lambda) + \int\left[\|\nabla\mathcal{L}_{\mathcal{D}}\|_2^2 - \sigma^2\Delta\mathcal{L}_{\mathcal{D}}\right]\left(\frac{\mu}{\pi(\mathcal{D})}-1\right)\pi(\mathcal{D})\mathrm{d}\theta\\
&= d\beta\sigma^2(\beta+\lambda) + \int\|\nabla\mathcal{L}_{\mathcal{D}}\|_2^2\left(\mu-\pi(\mathcal{D})\right)\mathrm{d}\theta\\
&\quad + \sigma^2\int\left\langle\nabla\mathcal{L}_{\mathcal{D}}, \nabla\left[\left(\frac{\mu}{\pi(\mathcal{D})}-1\right)\pi(\mathcal{D})\right]\right\rangle\mathrm{d}\theta && \text{(From (66))}\\
&= d\beta\sigma^2(\beta+\lambda) + \int\|\nabla\mathcal{L}_{\mathcal{D}}\|_2^2\left(\mu-\pi(\mathcal{D})\right)\mathrm{d}\theta + \sigma^2\int\left\langle\nabla\mathcal{L}_{\mathcal{D}}, -\frac{\nabla\mathcal{L}_{\mathcal{D}}}{\sigma^2}\right\rangle\left(\mu-\pi(\mathcal{D})\right)\mathrm{d}\theta\\
&\quad + \sigma^2\int\left\langle\nabla\mathcal{L}_{\mathcal{D}}, \nabla\frac{\mu}{\pi(\mathcal{D})}\right\rangle\pi(\mathcal{D})\mathrm{d}\theta && \text{(Since } \nabla\pi(\mathcal{D}) = -\tfrac{\nabla\mathcal{L}_{\mathcal{D}}}{\sigma^2}\pi(\mathcal{D})\text{)}\\
&= d\beta\sigma^2(\beta+\lambda) + 0 + 2\sigma^2\int\left\langle\sqrt{\frac{\mu}{\pi(\mathcal{D})}}\nabla\mathcal{L}_{\mathcal{D}}, \nabla\sqrt{\frac{\mu}{\pi(\mathcal{D})}}\right\rangle\pi(\mathcal{D})\mathrm{d}\theta\\
&\le d\beta\sigma^2(\beta+\lambda) + \frac{1}{2}\mathop{\mathbb{E}}_{\mu}\left[\|\nabla\mathcal{L}_{\mathcal{D}}\|_2^2\right] + 2\sigma^4\mathop{\mathbb{E}}_{\pi(\mathcal{D})}\left[\left\|\nabla\sqrt{\frac{\mu}{\pi(\mathcal{D})}}\right\|_2^2\right] && \text{(From (67) with } a=2\sigma^2\text{)}
\end{aligned}$$

□

Another change in measure inequality needed for the proof is the Donsker-Varadhan variational principle.

Lemma G.21 (Donsker-Varadhan Variational principle (Donsker & Varadhan, 1983))**.** *If ν and ν' are two distributions on $\mathbb{R}^d$ such that $\nu \ll \nu'$, then for all functions $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathop{\mathbb{E}}_{\Theta\sim\nu}\left[f(\Theta)\right] \le \mathrm{KL}\left(\nu\|\nu'\right) + \log\mathop{\mathbb{E}}_{\Theta'\sim\nu'}\left[\exp(f(\Theta'))\right]. \tag{141}$$

We are now ready to prove the rate of convergence guarantee for Noisy-GD following Chewi et al. (2021)'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 (Chewi et al., 2021, Theorem 4)).

Theorem G.22 (Convergence of Noisy-GD in Rényi divergence)**.** *Let constants $\beta, \lambda, \sigma^2 > 0$ and $q, B > 1$. Suppose the loss function $\ell(\theta;\mathbf{x})$ 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 $\eta \le \frac{\lambda}{64Bq^2(\beta+\lambda)^2}$, then for any database $\mathcal{D}\in\mathcal{X}^n$ and any weight initialization distribution μ_0 for Θ_0, the Rényi divergence of distribution $\mu_{\eta K}$ of output model $\Theta_{\eta K}$ = Noisy-GD$(\mathcal{D}, \Theta_0, K)$ with respect to the Gibbs distribution $\pi(\mathcal{D})$ defined in* (134) *shrinks as follows:*

$$\mathrm{R}_q\left(\mu_{\eta K}\|\pi(\mathcal{D})\right) \le q\exp\left(-\frac{\lambda\eta K}{2B}\right)\mathrm{R}_q\left(\mu_0\|\pi(\mathcal{D})\right) + \frac{32d\eta qB(\beta+\lambda)^2}{\lambda}. \tag{142}$$

Proof. From $(\beta+\lambda)$-smoothness of loss $\mathcal{L}_{\mathcal{D}}$ we have that for any $\eta k \le t \le \eta(k+1)$,

$$\begin{aligned}
\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)\|_2^2 &\le (\beta+\lambda)^2\|\Theta_{\eta k} - \Theta_t\|_2^2\\
&= (\beta+\lambda)^2\left\|(t-\eta k)\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \sqrt{2(t-\eta k)\sigma^2}\mathbf{Z}_k\right\|_2^2 && \text{(From (135))}\\
&\le 2\eta^2(\beta+\lambda)^2\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\|_2^2 + 4\eta\sigma^2(\beta+\lambda)^2\|\mathbf{Z}_k\|_2^2\\
&\le 4\eta^2(\beta+\lambda)^2\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)\|_2^2\\
&\quad + 4\eta^2(\beta+\lambda)^2\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)\|_2^2 + 4\eta\sigma^2(\beta+\lambda)^2\|\mathbf{Z}_k\|_2^2
\end{aligned}$$

Let $\rho_t := \frac{\mu_t}{\pi(\mathcal{D})}$ and $\psi_t := \rho_t^{q-1}/\mathrm{E}_q\left(\rho_t \| \pi(\mathcal{D})\right)$. 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 (138):

$$\mathbb{E}\left[\psi_t(\Theta_t)\left\|\nabla\mathcal{L}_{\mathcal{B}_k}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)\right\|_2^2\right] \leq 8\eta^2(\beta+\lambda)^2 \underbrace{\mathbb{E}\left[\psi_t(\Theta_t)\left\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)\right\|_2^2\right]}_{\stackrel{\text{def}}{=} F_1} + 32\eta\sigma^2(\beta+\lambda)^2 \underbrace{\mathbb{E}\left[\psi_t(\Theta_t)\left\|\mathbf{Z}_k\right\|_2^2/4\right]}_{\stackrel{\text{def}}{=} F_2}.$$

Hence, for solving the PDI (138), we have to bound the three expectations F_1 and F_2.

1. **Bounding F_1.** Note that $\mathop{\mathbb{E}}_{\Theta_t\sim\mu_t}[\psi_t(\Theta_t)] = \int \psi_t(\theta)\mu_t(\theta)\mathrm{d}\theta = \frac{1}{\mathrm{E}_q(\rho_t\|\pi(\mathcal{D}))}\int \frac{\mu_t^q}{\pi(\mathcal{D})^{q-1}}\mathrm{d}\theta = 1$. So, $\psi_t\mu_t(\theta) := \psi_t(\theta)\mu_t(\theta)$ is a probability density function on $\mathbb{R}^d$. On applying the measure change Lemma G.20 on it, we get

$$\begin{aligned}
F_1 = \mathop{\mathbb{E}}_{\psi_t\mu_t}\left[\left\|\nabla\mathcal{L}_{\mathcal{D}}\right\|_2^2\right] &\leq 4\sigma^4 \mathop{\mathbb{E}}_{\pi(\mathcal{D})}\left[\left\|\nabla\sqrt{\frac{\psi_t\mu_t}{\pi(\mathcal{D})}}\right\|_2^2\right] + 2d\sigma^2(\beta+\lambda) && \text{(From (139))}\\
&= 4\sigma^4 \mathop{\mathbb{E}}_{\pi(\mathcal{D})}\left[\frac{\left\|\nabla\sqrt{\rho_t^q}\right\|_2^2}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)}\right] + 2d\sigma^2(\beta+\lambda)\\
&= \sigma^4 q^2 \frac{\mathrm{I}_q\left(\mu_t\|\pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)} + 2d\sigma^2(\beta+\lambda). && \text{(From (29))}
\end{aligned}$$

2. **Bounding F_2.** Since $\psi_t\mu_t$ is a valid density on $\mathbb{R}^d$, the joint density $\psi_t\mu_{t,z}(\theta,z) := \psi_t(\theta)\mu_{t,z}(\theta,z)$ where $\mu_{t,z}$ is the joint density of Θ_t and $\mathbf{Z}_k$ is also a valid density. Note that the F_2 is an expectation on $\|\mathbf{Z}_k\|_2^2$ taken over the joint density $\psi_t\mu_{t,z}$. We can perform a measure change operation using Donsker-Varadhan principle to get

$$F_2 = \mathop{\mathbb{E}}_{\psi_t\mu_{t,z}}\left[\left\|\mathbf{Z}_k\right\|_2^2/4\right] \leq \mathrm{KL}\left(\psi_t\mu_{t,z}\|\mu_{t,z}\right) + \log\mathop{\mathbb{E}}_{\mu_z}\left[\exp(\left\|\mathbf{Z}_k\right\|_2^2/4)\right],$$

where we simplified the second term using the fact that the marginal μ_z of $\mu_{t,z}$ is a standard normal Gaussian. The random variable $\|\mathbf{Z}_k\|_2^2$ 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

$$\log\mathop{\mathbb{E}}_{\mu_z}\left[\exp(\left\|\mathbf{Z}_k\right\|_2^2/4)\right] = \log\mathrm{M}_{\chi_d^2}\left(\frac{1}{4}\right) = \frac{d\log 2}{2}. \tag{143}$$

The KL divergence term can be simplified as follows.

$$\begin{aligned}
\mathrm{KL}\left(\psi_t\mu_{t,z}\|\mu_{t,z}\right) &= \int\int \psi_t\mu_{t,z}(\theta_t,z)\log\psi_t(\theta_t)\mathrm{d}\theta_t\mathrm{d}z\\
&= \int \psi_t\mu_t\log\frac{\rho_t^{q-1}}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)}\mathrm{d}\theta_t && \text{(On marginalization of } z)\\
&= \frac{q-1}{q}\int \mu_t\psi_t\log\left\{\frac{\rho_t^q}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)} - \log\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)^{1/(q-1)}\right\}\mathrm{d}\theta_t\\
&= \frac{q-1}{q}\left\{\mathrm{KL}\left(\mu_t\psi_t\|\pi(\mathcal{D})\right) - \mathrm{R}_q\left(\mu_t\|\pi(\mathcal{D})\right)\right\}\\
&\leq \mathrm{KL}\left(\mu_t\psi_t\|\pi(\mathcal{D})\right) && \text{(Since } \mathrm{R}_q\left(\mu_t\|\pi(\mathcal{D})\right) > 0)
\end{aligned}$$

Note that under the assumptions of the Theorem, $\pi(\mathcal{D})$ satisfies log-Sobolev inequality (91) with constant λ/B (i.e. satisfies $\mathrm{LS}(\lambda/B)$). To see this, recall from Lemma G.1 that the Gaussian distribution $\rho(\theta) = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda}\mathbb{I}_d\right)$ satisfies $\mathrm{LS}(\lambda)$ inequality. Since loss $\ell(\theta; \mathbf{x})$ 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 G.3. Using this inequality, from Lemma G.4 we have

$$\begin{aligned}
\mathrm{KL}\left(\mu_t\psi_t \| \pi(\mathcal{D})\right) &\leq \frac{\sigma^2 B}{2\lambda} \int \mu_t\psi_t \left\| \nabla \log\left(\frac{\mu_t\psi_t}{\pi(\mathcal{D})}\right)\right\|_2^2 \mathrm{d}\theta_t \\
&= \frac{\sigma^2 B}{2\lambda} \int \frac{\rho_t^q}{\mathrm{E}_q\left(\mu_t \| \pi(\mathcal{D})\right)} \left\|\nabla \log(\rho_t^q)\right\|_2^2 \pi(\mathcal{D}) \mathrm{d}\theta_t \\
&= \frac{2\sigma^2 B}{\lambda} \frac{1}{\mathrm{E}_q\left(\mu_t \| \pi(\mathcal{D})\right)} \int \left\|\nabla(\rho_t^{q/2})\right\|_2^2 \pi(\mathcal{D}) \mathrm{d}\theta_t \\
&= \frac{q^2\sigma^2 B}{2\lambda} \frac{\mathrm{I}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t \| \pi(\mathcal{D})\right)}
\end{aligned}$$

On combining all the two bounds on F_1 and F_2 and rearranging, we have

$$\begin{aligned}
\mathbb{E}\left[\psi_t(\Theta_t) \left\|\nabla \mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla \mathcal{L}_{\mathcal{D}}(\Theta_t)\right\|_2^2\right] &\leq 8\eta q^2 \sigma^4 (\beta+\lambda)^2 \frac{\mathrm{I}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t \| \pi(\mathcal{D})\right)} \left(\eta + \frac{2B}{\lambda}\right) \\
&\quad + 16\eta d\sigma^2 (\beta+\lambda)^2 \left(\eta(\beta+\lambda) + \log 2\right)
\end{aligned}$$

Let step size be $\eta \leq \min\left\{\frac{2B}{\lambda}, \frac{\lambda}{64Bq^2(\beta+\lambda)^2}\right\}$. Then, the first term above is bounded as

$$8\eta q^2 \sigma^4 (\beta+\lambda)^2 \frac{\mathrm{I}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t \| \pi(\mathcal{D})\right)} \left(\eta + \frac{2B}{\lambda}\right) \leq \frac{\sigma^4}{2} \frac{\mathrm{I}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t \| \pi(\mathcal{D})\right)}. \tag{144}$$

Let $\eta \leq \frac{1}{4(\beta+\lambda)}$. Then, in the third term, $(\eta(\beta+\lambda) + \log 2) \leq 1$. Plugging the bound on discretization error back in the PDI (138), we get

$$\partial_t \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right) \leq -\frac{q\sigma^2}{4} \frac{\mathrm{I}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t \| \pi(\mathcal{D})\right)} + 16\eta d q (\beta+\lambda)^2. \tag{145}$$

Since $\pi(\mathcal{D})$ satisfies $\mathrm{LS}(\lambda/B)$ inequality, from Lemma G.6 this PDI reduces to

$$\partial_t \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right) + \frac{\lambda}{2B}\left(\frac{\mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{q} + (q-1)\partial_q \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right)\right) \leq 16 d\eta q (\beta+\lambda)^2. \tag{146}$$

Let $c_1 = \frac{\lambda}{2B}$ and $c_2 = 16 d\eta(\beta+\lambda)^2$. Additionally, let $u(q,t) = \frac{\mathrm{R}_q(\mu_t \| \pi(\mathcal{D}))}{q}$. Then,

$$\begin{aligned}
&\partial_t \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right) + c_1\left(\frac{\mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{q} + (q-1)\partial_q \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right)\right) \leq c_2 q \\
\implies &\frac{\partial_t \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{q} + c_1 \frac{\mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{q} + c_1(q-1)\left(\frac{\partial_q \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{q} - \frac{\mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{q^2}\right) \leq c_2 \\
\implies &\partial_t u(q,t) + c_1 u(q,t) + c_1(q-1)\partial_q u(q,t) \leq c_2.
\end{aligned}$$

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

$$\begin{aligned}\mathrm{R}_{q(\eta K)}\left(\mu_{\eta K}\|\pi(\mathcal{D})\right) &\leq \frac{q(\eta K)}{q(0)} e^{-c_1\eta K}\mathrm{R}_{q(0)}\left(\mu_0\|\pi(\mathcal{D})\right) + \frac{c_2}{c_1}q(\eta K) \\ &\leq \frac{q(\eta K)}{q(0)}\exp\left(-\frac{\lambda\eta K}{2B}\right)\mathrm{R}_{q(0)}\left(\mu_0\|\pi(\mathcal{D})\right) + \frac{32d\eta B(\beta+\lambda)^2}{\lambda}q(\eta K).\end{aligned}$$

Since $q(0) > 1$ and $\bar{q} = q(\eta K) > q(0)$, from monotonicity of Rényi divergence in q, we get

$$\mathrm{R}_{\bar{q}}\left(\mu_{\eta K}\|\pi(\mathcal{D})\right) \leq \bar{q}\exp\left(-\frac{\lambda\eta K}{2B}\right)\mathrm{R}_{\bar{q}}\left(\mu_0\|\pi(\mathcal{D})\right) + \frac{32d\eta\bar{q}B(\beta+\lambda)^2}{\lambda}. \tag{147}$$

Finally, noting that for constants $B, q > 1$ and $\beta, \lambda > 0$,

$$\eta \leq \min\{\frac{1}{2\sqrt{2}(\beta+\lambda)}, \frac{1}{4(\beta+\lambda)}, \frac{2B}{\lambda}, \frac{\lambda}{64Bq^2(\beta+\lambda)^2}\} = \frac{\lambda}{64Bq^2(\beta+\lambda)^2}, \tag{148}$$

completes the proof. □

We will use Theorem G.22 for proving the deletion-privacy and utility guarantee on the pair $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$. We need the following 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 (Wang et al., 2015; Dwork et al., 2014)).

Lemma G.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 $\nu \propto e^{-\mathcal{L}/\sigma^2}$ and $\nu' \propto e^{-\mathcal{L}'/\sigma^2}$ be the respective Gibbs distributions. If $|\mathcal{L}(\theta) - \mathcal{L}'(\theta)| \leq c$ for all $\theta \in \mathbb{R}^d$, then $\mathrm{R}_q\left(\nu\|\nu'\right) \leq \frac{2c}{\sigma^2}$ for all $q > 1$.*

Proof. The Gibbs distributions ν, ν' have a density

$$\nu(\theta) = \frac{1}{\Lambda}e^{-\mathcal{L}(\theta)/\sigma^2}, \quad \text{and} \quad \nu'(\theta) = \frac{1}{\Lambda'}e^{-\mathcal{L}'(\theta)/\sigma^2},$$

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

$$\begin{aligned}\mathrm{R}_q\left(\nu\|\nu'\right) &= \frac{1}{q-1}\log\int\frac{\nu^q}{\nu'^{q-1}}\mathrm{d}\theta \\ &= \frac{1}{q-1}\log\int\left(\frac{\Lambda'}{\Lambda}\right)^{q-1}\exp\left(\frac{q-1}{\sigma^2}(\mathcal{L}'(\theta)-\mathcal{L}(\theta))\right)\times\nu(\theta)\mathrm{d}\theta \\ &\leq \frac{1}{q-1}\left\{(q-1)\log\frac{\Lambda'}{\Lambda} + \log\exp\left(\frac{c(q-1)}{\sigma^2}\int\nu\mathrm{d}\theta\right)\right\} \\ &= \frac{1}{q-1}\left\{(q-1)\log\frac{\int\exp\left(-\frac{\mathcal{L}(\theta)}{\sigma^2}+\frac{\mathcal{L}(\theta)-\mathcal{L}'(\theta)}{\sigma^2}\right)\mathrm{d}\theta}{\int\exp\left(-\frac{\mathcal{L}(\theta)}{\sigma^2}\right)\mathrm{d}\theta} + \frac{c(q-1)}{\sigma^2}\right\} \\ &\leq \frac{2c}{\sigma^2}.\end{aligned}$$

□

In Theorem 5.2, we show that $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ solves the data-deletion problem described in Section 4 even for non-convex losses. Our proof uses the convergence Theorem G.22 and indistinguishability for bounded perturbation Lemma G.23 to show that the unlearning algorithm $\bar{\mathrm{A}}_{\text{Noisy-GD}}$ can consistently produce models indistinguishable to the corresponding Gibbs distribution (134) in the online setting at a fraction of computation cost of $\mathrm{A}_{\text{Noisy-GD}}$. As discussed in Remark 4.2, such an indistinguishability is sufficient for ensuring deletion-privacy for non-adaptive requests. As for adaptive requests, the well-known Rényi DP guarantee of Abadi et al. (2016) combined with our reduction Theorem 4.3 offers a deletion-privacy guarantee for $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}})$ under adaptivity.

Our proof of accuracy for the unlearned models leverages the fact that Gibbs distribution (134) is an almost excess risk minimizer as shown in the following Theorem G.24. Since our deletion-privacy guarantee is based on near-indistinguishability to (134), this property also ensures near-optimal excess risk of unlearned models.

Theorem G.24 (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 $\bar{\Theta}$ sampled from the Gibbs distribution $\pi(\mathcal{D})\propto e^{-\mathcal{L}_{\mathcal{D}}/\sigma^2}$ is*

$$\operatorname{err}(\bar{\Theta};\mathcal{D})=\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\bar{\Theta})-\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})\right]\leq\frac{d\sigma^2}{2}\left(\log\frac{\beta+\lambda}{\lambda}+\sqrt{B}\right). \tag{149}$$

Proof. We simplify expected loss as

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\bar{\Theta})\right]=\int\mathcal{L}_{\mathcal{D}}\pi(\mathcal{D})\mathrm{d}\theta=\sigma^2(\mathrm{H}(\pi(\mathcal{D}))-\log(\Lambda_{\mathcal{D}})), \tag{150}$$

where

$$\mathrm{H}(\pi(\mathcal{D}))=-\int\pi(\mathcal{D})\log\pi(\mathcal{D})\mathrm{d}\theta=-\int\frac{e^{-\mathcal{L}_{\mathcal{D}}/\sigma^2}}{\Lambda_{\mathcal{D}}}\log\frac{e^{-\mathcal{L}_{\mathcal{D}}/\sigma^2}}{\Lambda_{\mathcal{D}}}\mathrm{d}\theta \tag{151}$$

is the differential entropy of $\pi(\mathcal{D})$, and $\Lambda_{\mathcal{D}}=\int e^{-\mathcal{L}_{\mathcal{D}}/\sigma^2}\mathrm{d}\theta$ is the normalization constant. Since the potential function $\mathcal{L}_{\mathcal{D}}$ is $(\lambda+\beta)$-smooth, we have

$$\begin{aligned}
-\sigma^2\log(\Lambda_{\mathcal{D}})&=-\sigma^2\log\int e^{-\mathcal{L}_{\mathcal{D}}/\sigma^2}\mathrm{d}\theta\\
&=\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})-\sigma^2\log\int e^{(\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})-\mathcal{L}_{\mathcal{D}}(\theta))/\sigma^2}\mathrm{d}\theta\\
&\leq\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})-\sigma^2\log\int e^{-(\beta+\lambda)\|\theta-\theta^*_{\mathcal{D}}\|_2^2/2\sigma^2}\mathrm{d}\theta\\
&=\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})-\frac{d\sigma^2}{2}\log\left(\frac{2\pi\sigma^2}{\lambda+\beta}\right).
\end{aligned}$$

Since $\boldsymbol{\ell}(\theta;\mathbf{x})$ is $\sigma^2\log(B)/4$-bounded, note that for the Gaussian distribution $\rho\sim\mathcal{N}\left(0,\frac{\sigma^2}{\lambda}\mathbb{I}_d\right)$, the density ratio lies in $\frac{\pi(\mathcal{D})(\theta)}{\rho(\theta)}\in\left[\frac{1}{\sqrt{B}},\sqrt{B}\right]$ for all $\theta\in\mathbb{R}^d$. We decompose entropy $\mathrm{H}(\pi(\mathcal{D}))$ into cross-entropy and KL divergence to get

$$\begin{aligned}
\mathrm{H}(\pi(\mathcal{D}))&=-\int\pi(\mathcal{D})\log\rho\mathrm{d}\theta-\mathrm{KL}\left(\pi(\mathcal{D})\|\rho\right)\\
&\leq-\int\pi(\mathcal{D})\log\left[\left(\frac{\lambda}{2\pi\sigma^2}\right)^{d/2}e^{-\frac{\lambda\|\theta\|_2^2}{2\sigma^2}}\right]\mathrm{d}\theta && \text{(Since } \mathrm{KL}\left(\pi(\mathcal{D})\|\rho\right)\geq 0\text{)}\\
&=\frac{d}{2}\log\frac{2\pi\sigma^2}{\lambda}+\frac{\lambda}{2\sigma^2}\int\|\theta\|_2^2\,\pi(\mathcal{D})(\theta)\mathrm{d}\theta\\
&\leq\frac{d}{2}\log\frac{2\pi\sigma^2}{\lambda}+\frac{\lambda\sqrt{B}}{2\sigma^2}\int\|\theta\|_2^2\,\rho(\theta)\mathrm{d}\theta && \text{(Since } \tfrac{\pi(\mathcal{D})(\theta)}{\rho(\theta)}\in\left[\tfrac{1}{\sqrt{B}},\sqrt{B}\right]\text{)}\\
&=\frac{d}{2}\log\frac{2\pi\sigma^2}{\lambda}+\frac{d\sqrt{B}}{2}.
\end{aligned}$$

On combining the bounds, we get

$$\operatorname{err}(\bar{\Theta};\mathcal{D})=\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\bar{\Theta})-\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})\right]\leq\frac{d\sigma^2}{2}\left(\log\frac{\beta+\lambda}{\lambda}+\sqrt{B}\right). \tag{152}$$

□

Theorem 5.2 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda,\beta,L,\sigma^2,\eta>0$, constants $q,B>1$, and constants $d>\varepsilon_{\mathrm{dp}}\geq\varepsilon_{\mathrm{dd}}>0$. 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 $\rho=\mathcal{N}\left(0,\frac{\sigma^2}{\lambda}\mathbb{I}_d\right)$. Then,*

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

$$\sigma^2 \geq \frac{qL^2}{\varepsilon_{\mathrm{dp}} n^2} \cdot \eta \max\{K_{\mathrm{A}}, K_{\bar{\mathrm{A}}}\}, \tag{153}$$

(2.) *pair* $(\mathrm{A}_{\textit{Noisy-GD}}, \bar{\mathrm{A}}_{\textit{Noisy-GD}})$ *satisfy* $(q, \varepsilon_{\mathrm{dd}})$*-deletion-privacy under all non-adaptive* r*-requesters for any* $\sigma^2 > 0$*, if learning rate is* $\eta \leq \frac{\lambda \varepsilon_{\mathrm{dd}}}{64 dqB(\beta+\lambda)^2}$ *and number of iterations satisfy*

$$K_{\mathrm{A}} \geq \frac{2B}{\lambda\eta} \log\left(\frac{q \log(B)}{\varepsilon_{\mathrm{dd}}}\right), \quad K_{\bar{\mathrm{A}}} \geq K_{\mathrm{A}} - \frac{2B}{\lambda\eta} \log\left(\frac{\log(B)}{2\left(\varepsilon_{\mathrm{dd}} + \frac{r}{n}\log(B)\right)}\right), \tag{154}$$

(3.) *and all models in the sequence* $(\hat{\Theta}_i)_{i\geq 0}$ *produced by interactions between* $(\mathrm{A}_{\textit{Noisy-GD}}, \bar{\mathrm{A}}_{\textit{Noisy-GD}})$ *and* $\mathcal{Q}$ *on any* $\mathcal{D}_0 \in \mathcal{X}^n$*, where* $\mathcal{Q}$ *is an* r*-requester, have an excess empirical risk* $\mathrm{err}(\hat{\Theta}_i; \mathcal{D}_i) = \tilde{O}\left(\frac{dq}{\varepsilon_{\mathrm{dp}} n^2} + \frac{1}{n}\sqrt{\frac{q\varepsilon_{\mathrm{dd}}}{\varepsilon_{\mathrm{dp}}}}\right)$ *when inequalities in* (154) *and* (153) *are equalities.*

Proof. **(1.) Privacy.** By Theorem G.7, Noisy-GD with K iterations on an L-Lipschitz loss function satisfies $(q, \varepsilon_{\mathrm{dp}})$-Rényi DP for any initial weight distribution ρ and learning rate $\eta \geq 0$ if $\sigma^2 = \frac{qL^2}{\varepsilon_{\mathrm{dp}} n^2} \cdot \eta K$. Since, both $\mathrm{A}_{\text{Noisy-GD}}$ and $\bar{\mathrm{A}}_{\text{Noisy-GD}}$ run Noisy-GD for K_{A} and $K_{\bar{\mathrm{A}}}$ iterations respectively, setting the noise variance given in the Theorem statement ensures $(q, \varepsilon_{\mathrm{dp}})$-Rényi DP for both.

(2.) Deletion. For showing deletion-privacy under non-adaptive requests, recall that it is sufficient to show that there exists a map $\pi : \mathcal{X}^n \to \mathcal{O}$ such that for all $i \geq 1$,

$$\mathrm{R}_q\left(\bar{\mathrm{A}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) \middle\| \pi(\mathcal{D}_i)\right) \leq \varepsilon_{\mathrm{dd}}, \tag{155}$$

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 $(\mathrm{A}_{\text{Noisy-GD}}, \bar{\mathrm{A}}_{\text{Noisy-GD}}, \mathcal{Q})$ on any database $\mathcal{D}_0 \in \mathcal{X}^n$. For all $i \geq 0$, let $\hat{\mu}_i$ denote the distribution of $\hat{\Theta}_i$. We prove (155) via induction.

Base step: Note that the initial weight distribution $\rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda}\mathbb{I}_d\right)$ has a density proportional to $e^{-\mathbf{r}(\theta)/\sigma^2}$ and the distribution $\pi(\mathcal{D}_0)$ has a density proportional to $e^{-\mathcal{L}_{\mathcal{D}_0}(\theta)/\sigma^2}$. Since both of these are Gibbs distributions with their potential difference $|\mathcal{L}_{\mathcal{D}_0}(\theta) - \mathbf{r}(\theta)| \leq \sigma^2 \log(B)/4$ for all $\theta \in \mathbb{R}^d$ due to boundedness assumption on $\boldsymbol{\ell}(\theta; \mathbf{x})$, we have from Lemma G.23 that

$$\mathrm{R}_q\left(\rho \| \pi(\mathcal{D}_0)\right) \leq \frac{2}{\sigma^2} \times \frac{\sigma^2 \log(B)}{4} = \frac{\log(B)}{2}. \tag{156}$$

Under the stated assumptions on loss $\boldsymbol{\ell}(\theta; \mathbf{x})$ and learning rate η, note that the convergence Theorem G.22 holds. Since $\hat{\Theta}_0 = \mathrm{A}_{\text{Noisy-GD}}(\mathcal{D}_0) = \text{Noisy-GD}(\mathcal{D}_0, \Theta_0, K_{\mathrm{A}})$, where $\Theta_0 \sim \rho$, we have

$$\begin{aligned}
\mathrm{R}_q\left(\hat{\mu}_0 \| \pi(\mathcal{D}_0)\right) &\leq q \exp\left(-\frac{\lambda\eta K_{\mathrm{A}}}{2B}\right) \mathrm{R}_q\left(\rho \| \pi(\mathcal{D}_0)\right) + \frac{32 d\eta q B(\beta+\lambda)^2}{\lambda} \\
&\leq q \exp\left(-\frac{\lambda\eta K_{\mathrm{A}}}{2B}\right)\left(\frac{\log(B)}{2}\right) + \frac{\varepsilon_{\mathrm{dd}}}{2} && \left(\text{Since } \eta \leq \tfrac{\lambda\varepsilon_{\mathrm{dd}}}{64 dqB(\beta+\lambda)^2}\right) \\
&\leq \varepsilon_{\mathrm{dd}} && \left(\text{Since } K_{\mathrm{A}} \geq \tfrac{2B}{\lambda\eta}\log\left(\tfrac{q\log(B)}{\varepsilon_{\mathrm{dd}}}\right)\right)
\end{aligned}$$

Induction step: Suppose $\mathrm{R}_q\left(\hat{\mu}_{i-1} \| \pi(\mathcal{D}_{i-1})\right) \leq \varepsilon_{\mathrm{dd}}$. Again, from boundedness of $\boldsymbol{\ell}(\theta; \mathbf{x})$, we have $|\mathcal{L}_{\mathcal{D}_{i-1}}(\theta) - \mathcal{L}_{\mathcal{D}_i}(\theta)| \leq \frac{r\sigma^2 \log B}{2n}$ for all $\theta \in \mathbb{R}^d$. Therefore, from Lemma G.23 we have for all $q > 1$ that

$$\mathrm{R}_q\left(\pi(\mathcal{D}_{i-1}) \| \pi(\mathcal{D}_i)\right) \leq \frac{r\log(B)}{n}. \tag{157}$$

So from the weak triangle inequality Theorem B.4 of Rényi divergence,

$$\mathrm{R}_q\left(\hat{\mu}_{i-1} \| \pi(\mathcal{D}_i)\right) \leq \mathrm{R}_q\left(\hat{\mu}_{i-1} \| \pi(\mathcal{D}_{i-1})\right) + \mathrm{R}_\infty\left(\pi(\mathcal{D}_{i-1}) \| \pi(\mathcal{D}_i)\right) \leq \varepsilon_{\mathrm{dd}} + \frac{r\log(B)}{n}. \tag{158}$$

Note that $K_{\bar{\mathrm{A}}} \geq K_{\mathrm{A}} - \frac{2B}{\lambda\eta}\log\left(\frac{\log(B)}{2\left(\varepsilon_{\mathrm{dd}}+\frac{r}{n}\log(B)\right)}\right) \geq \frac{2B}{\lambda\eta}\log\left(\frac{2q\left(\varepsilon_{\mathrm{dd}}+\frac{r}{n}\log(B)\right)}{\varepsilon_{\mathrm{dd}}}\right)$. Since $\hat{\Theta}_i = \bar{\mathrm{A}}_{\text{Noisy-GD}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) = \text{Noisy-GD}(\mathcal{D}_i, \hat{\Theta}_{i-1}, K_{\bar{\mathrm{A}}})$, convergence Theorem G.22 gives

$$\begin{aligned}
\mathrm{R}_q\left(\hat{\mu}_i \| \pi(\mathcal{D}_i)\right) &\leq q\exp\left(-\frac{\lambda\eta K_{\bar{\mathrm{A}}}}{2B}\right)\mathrm{R}_q\left(\hat{\mu}_{i-1}\|\pi(\mathcal{D}_i)\right) + \frac{32d\eta q B(\beta+\lambda)^2}{\lambda} \\
&\leq q\exp\left(-\frac{\lambda\eta K_{\bar{\mathrm{A}}}}{2B}\right)\left(\varepsilon_{\mathrm{dd}} + \frac{r\log(B)}{n}\right) + \frac{\varepsilon_{\mathrm{dd}}}{2} \qquad \text{(From (158) and constraint } \eta \leq \tfrac{\lambda\varepsilon_{\mathrm{dd}}}{64dqB(\beta+\lambda)^2}\text{)} \\
&\leq \varepsilon_{\mathrm{dd}}. \qquad \text{(Since } K_{\bar{\mathrm{A}}} \geq \tfrac{2B}{\lambda\eta}\log\left(\tfrac{2q\left(\varepsilon_{\mathrm{dd}}+\frac{r}{n}\log(B)\right)}{\varepsilon_{\mathrm{dd}}}\right)\text{)}
\end{aligned}$$

Hence, by induction, $\mathrm{R}_q\left(\hat{\mu}_i\|\pi(\mathcal{D}_i)\right) \leq \varepsilon_{\mathrm{dd}}$ holds for all $i \geq 0$.

(3.) Accuracy. Let $\theta^*_{\mathcal{D}_i} = \arg\min_{\theta\in\mathbb{R}^d} \mathcal{L}_{\mathcal{D}_i}(\theta)$, and $\bar{\Theta}_i \sim \pi(\mathcal{D}_i)$. We decompose the excess empirical risk of Noisy-GD as follows:

$$\mathrm{err}(\hat{\Theta}_i; \mathcal{D}_i) = \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_i}(\hat{\Theta}_i) - \mathcal{L}_{\mathcal{D}_i}(\bar{\Theta}_i)\right] + \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_i}(\bar{\Theta}_i) - \mathcal{L}_{\mathcal{D}_i}(\theta^*_{\mathcal{D}_i})\right]. \tag{159}$$

The second term is the suboptimality of Gibbs distribution and by Theorem G.24, it is bounded as

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_i}(\bar{\Theta}_i) - \mathcal{L}_{\mathcal{D}_i}(\theta^*_{\mathcal{D}_i})\right] \leq \frac{d\sigma^2}{2}\left(\log\frac{\beta+\lambda}{\lambda} + \sqrt{B}\right). \tag{160}$$

From $(\lambda+\beta)$-smoothness of $\mathcal{L}_{\mathcal{D}_i}$, for any coupling Π of $\hat{\Theta}_i$ and $\bar{\Theta}_i$, the first term satisfies

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_i}(\hat{\Theta}_i) - \mathcal{L}_{\mathcal{D}_i}(\bar{\Theta}_i)\right] &\leq \mathop{\mathbb{E}}_{\Pi}\left[\left\langle \nabla\mathcal{L}_{\mathcal{D}_i}(\bar{\Theta}_i), \hat{\Theta}_i - \bar{\Theta}_i\right\rangle + \frac{\lambda+\beta}{2}\left\|\hat{\Theta}_i - \Theta_i\right\|_2^2\right] \\
&= \mathop{\mathbb{E}}_{\Pi}\left[\left\langle \sum_{\mathbf{x}\in\mathcal{D}_i}\nabla\boldsymbol{\ell}(\bar{\Theta}_i;\mathbf{x}) + \lambda\bar{\Theta}_i, \hat{\Theta}_i - \bar{\Theta}_i\right\rangle + \frac{\lambda+\beta}{2}\left\|\hat{\Theta}_i - \Theta_i\right\|_2^2\right]. \\
&\qquad \text{(From } L\text{-Lipschitzness of } \boldsymbol{\ell}(\theta;\mathbf{x}) \text{ and Jensen's inequality)} \\
&\leq L\sqrt{\mathop{\mathbb{E}}_{\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right]} + \lambda\mathop{\mathbb{E}}_{\Pi}\left[\left\langle\bar{\Theta}_i, \bar{\Theta}_i - \hat{\Theta}_i\right\rangle\right] + \frac{\lambda+\beta}{2}\mathop{\mathbb{E}}_{\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right] \\
&\qquad \text{(From Young's inequality (67))} \\
&\leq L\sqrt{\mathop{\mathbb{E}}_{\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right]} + \frac{\lambda}{2}\mathop{\mathbb{E}}_{\bar{\Theta}_i\sim\pi(\mathcal{D}_i)}\left[\left\|\bar{\Theta}_i\right\|_2^2\right] + \frac{2\lambda+\beta}{2}\mathop{\mathbb{E}}_{\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right].
\end{aligned}$$

Recall that the distribution $\pi(\mathcal{D})$ satisfies $\mathrm{LS}(\lambda/B)$ inequality. On choosing the coupling Π to be the infimum, we get the following bound on Wasserstein's distance from Lemma G.5.

$$\inf_{\Pi}\sqrt{\mathop{\mathbb{E}}_{\hat{\Theta}_i,\bar{\Theta}_i\sim\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right]} = \mathrm{W}_2\left(\hat{\Theta}_i, \bar{\Theta}_i\right) \leq \sqrt{\frac{2B\sigma^2}{\lambda}\mathrm{KL}\left(\mu_i\|\pi(\mathcal{D}_i)\right)} \leq \sqrt{\frac{2\varepsilon_{\mathrm{dd}}B\sigma^2}{\lambda}}. \tag{161}$$

The last inequality above follows from monotonicity of Rényi divergence in q and the fact that $\lim_{q\to1}\mathrm{R}_q\left(\nu\|\nu'\right) = \mathrm{KL}\left(\nu\|\nu'\right)$.

Since $\pi(\mathcal{D}_i)$ is the Gibbs distribution with density proportional to $e^{-\mathcal{L}_{\mathcal{D}_i}/\sigma^2}$, we have that

$$\mathop{\mathbb{E}}_{\bar{\Theta}_i\sim\pi(\mathcal{D}_i)}\left[\left\|\bar{\Theta}_i\right\|_2^2\right] = \frac{1}{\Lambda_{\mathcal{D}_i}}\int\|\theta\|_2^2\, e^{-\mathcal{L}_{\mathcal{D}_i}(\theta)/\sigma^2}\mathrm{d}\theta \quad \text{where } \Lambda_{\mathcal{D}_i} = \int e^{-\mathcal{L}_{\mathcal{D}_i}(\theta)/\sigma^2}\mathrm{d}\theta. \tag{162}$$

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

$$\left|\mathcal{L}_{\mathcal{D}_i}(\theta) - \mathbf{r}(\theta)\right| \leq \frac{\sigma^2\log B}{4}. \tag{163}$$

Therefore,

$$\Lambda_{\mathcal{D}_i} = \int e^{-\mathcal{L}_{\mathcal{D}_i}(\theta)/\sigma^2} \mathrm{d}\theta \geq \frac{1}{\sqrt[4]{B}} \int e^{-\mathbf{r}(\theta)/\sigma^2} \mathrm{d}\theta, \tag{164}$$

and hence,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\bar{\Theta}_i \sim \pi(\mathcal{D}_i)} \left[\left\| \bar{\Theta}_i \right\|_2^2 \right] &\leq \frac{\sqrt[4]{B}}{\int e^{-\mathbf{r}(\theta)/\sigma^2} \mathrm{d}\theta} \times \int \|\theta\|_2^2 \, e^{-\mathcal{L}_{\mathcal{D}_i}(\theta)/\sigma^2} \mathrm{d}\theta \\
&\leq \sqrt{B} \times \frac{\int \|\theta\|_2^2 \, e^{-\mathbf{r}(\theta)/\sigma^2}}{\int e^{-\mathbf{r}(\theta)/\sigma^2}} \\
&= \sqrt{B} \mathop{\mathbb{E}}_{\mathbf{Z} \sim \mathcal{N}\left(0, \frac{\sigma^2}{\lambda} \mathbb{I}_d\right)} \left[\|\mathbf{Z}\|_2^2 \right] \\
&= \frac{\sqrt{B} \sigma^2 d}{\lambda}.
\end{aligned}$$

Therefore, on combining all the bounds we get

$$\mathrm{err}(\hat{\Theta}; \mathcal{D}) \leq L\sigma \sqrt{\frac{2\varepsilon_{\mathrm{dd}} B}{\lambda}} + \frac{\varepsilon_{\mathrm{dd}} B \sigma^2 (2\lambda + \beta)}{\lambda} + \frac{d\sigma^2}{2} \left(\log \frac{\beta + \lambda}{\lambda} + 2\sqrt{B} \right) = O\left(\sigma \sqrt{\varepsilon_{\mathrm{dd}}} + d\sigma^2 \right). \tag{165}$$

Note that if the constraints on K_{A} and $K_{\bar{\mathrm{A}}}$ in (154) and on σ^2 in (153) are equalities instead, we have

$$\sigma^2 = \frac{2qBL^2}{\lambda \varepsilon_{\mathrm{dp}} n^2} \log \left(\frac{q \log(B)}{\varepsilon_{\mathrm{dd}}} \right) = \tilde{O}\left(\frac{q}{\varepsilon_{\mathrm{dp}} n^2} \right), \tag{166}$$

where $\tilde{O}(\cdot)$ hides logarithmic factors. Therefore, the excess empirical risk has an order

$$\mathrm{err}(\hat{\Theta}; \mathcal{D}) = \tilde{O}\left(\frac{1}{n} \sqrt{\frac{q\varepsilon_{\mathrm{dd}}}{\varepsilon_{\mathrm{dp}}}} + \frac{dq}{\varepsilon_{\mathrm{dp}} n^2} \right). \tag{167}$$

□