FORGET UNLEARNING: TOWARDS TRUE DATA-DELETION IN MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unlearning has emerged as a technique to efficiently erase information of deleted records from learned models. We show, however, that the influence created by the original presence of a data point in the training set can still be detected after running certified unlearning algorithms (which can result in its reconstruction by an adversary). Thus, under realistic assumptions about the dynamics of model releases over time and in the presence of adaptive adversaries, we show that unlearning is not equivalent to data deletion and does not guarantee the "right to be forgotten." We then propose a more robust data-deletion guarantee and show that it is necessary to satisfy differential privacy to ensure true data deletion. Under our notion, we propose an accurate, computationally efficient, and secure data-deletion machine learning algorithm in the online setting based on noisy gradient descent algorithm.

1 INTRODUCTION

Many corporations today collect their customers' private information to train Machine Learning (ML) models that power a variety of services, encompassing recommendations, searches, targeted ads, and more. To prevent any unintended use of personal data, privacy policies, such as General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA), require that these corporations provide the "*right to be forgotten*" (RTBF) to their data subjects—if a user wishes to revoke access to their data, an organization must comply by erasing all information about the user without undue delay (which is typically a month). This includes ML models trained in standard ways as model inversion (Fredrikson et al., 2015) and membership inference attacks (Shokri et al., 2017; Carlini et al., 2019) demonstrate that individual training data can be exfiltrated from these models.

Periodic retraining of models after excluding deleted users can be costly. So, there is a growing interest in designing computationally cheap *Machine Unlearning* algorithms as an alternative to retraining for erasing the influence of deleted data from (and registering the influence of added data to) trained models. Since it is generally difficult to tell how a specific data point affects a model, Ginart et al. (2019) propose quantifying the worst-case information leakage from an unlearned model through an *unlearning guarantee* on the mechanism, defined as a *differential privacy* (DP) like (ε, δ)-indistinguishability between its output and that of retraining on the updated database. With some minor variations in this definition, several mechanisms have been proposed and certified as unlearning algorithms in literature (Ginart et al., 2019; Izzo et al., 2021; Sekhari et al., 2021; Neel et al., 2021; Guo et al., 2019; Ullah et al., 2021).

However, *is indistinguishability to retraining a sufficient guarantee of data deletion?* We argue that it is not. In the real world, a user's decision to remove his information is often affected by what a deployed model reveals about him. The same revealed information may also affect other users' decisions. Such *adaptive* requests make the records in a database interdependent, causing a retrained model to contain influences of a record even if the record is no longer in the training set. We demonstrate on an unlearning-via-retraining mechanism that if an adversary is allowed to design an *adaptive requester* that interactively generates database edit requests as a function of published models, she can duplicate a target record in the controller's database before its deletion. Therefore, we argue that under adaptive requests, measuring data-deletion via indistinguishability to retraining (as proposed by Gupta et al. (2021)) is fundamentally flawed because it does not capture the influence a record might have previously had on the rest of the database. Our example shows a clear violation

of RTBF since even on retraining on the database with the original record removed, a model can reveal substantial information about the deleted record due to the possibility of duplication by an interactive adversary.

Is an unlearning guarantee a sound and complete measure of data deletion when requests are non-adaptive? Again, we argue that it is neither. A sound data-deletion guarantee must ensure the non-recovery of deleted records from an *infinite* number of model releases after deletion. However, approximate indistinguishability to retraining implies an inability to accurately recover deleted data from a *singular* unlearned model only, which we argue is not sufficient. We show that certain algorithms can satisfy an unlearning guarantee yet blatantly reveal the deleted data eventually over multiple releases. The vulnerability arises in algorithms that maintain partial computations in internal data structures for speeding up subsequent deletions. These internal states can retain information even after record deletion and influence multiple future releases, making the myopic unlearning guarantee unreliable in an online setting. Several proposed unlearning algorithms in literature (Ginart et al., 2019; Neel et al., 2021) are stateful (rely on internal states) and, therefore, cannot be trusted. Secondly, unlearning is an incomplete notion of data deletion as it excludes valid data-deletion mechanisms that do not imitate retraining. For instance, a (useless) mechanism that outputs a fixed untrained model on any request is a valid deletion algorithm. However, since its output is easily distinguishable from retraining, it fails to satisfy any meaningful unlearning guarantees.

This paper proposes a truer definition of data deletion that does not suffer from the abovementioned shortcomings. According to our notion, a data-deletion mechanism is reliable if **A)** it is stateless (i.e., it maintains no internal data structures), and **B)** generates models that are indistinguishable from any distribution independent of the deleted records. Statelessness thwarts the danger of sustained information leakage through internal data structures after deletion. Moreover, by defining *data deletion* as indistinguishability with any deleted-record independent distribution as oppsed to the output of retraining, we ensure reliability from adaptive requesters who might have encoded/duplicated the deleted record among other records in the database earlier.

In general, we argue that data-deletion mechanisms must also be *differentially private with respect to the remaining records* to prevent unwanted record duplications by an adaptive requester. DP also protects against membership inference attacks that extract deleted records by looking at models before and after deletion (Chen et al., 2021). We emphasize that we are not advocating for doing data deletion through differentially-private mechanisms that cap the information content of all records equally, deleted or otherwise, which is known to be inefficient (Sekhari et al., 2021). Instead, data-deletion mechanisms should provide two differing information reattainment bounds; one for records currently in the database in the form of a differential privacy guarantee and the other for records previously deleted in the form of a data-deletion guarantee. We provide a general theorem that if a mechanism is differentially private with respect to the remaining records and satisfies a data-deletion guarantee under non-adaptive edit requests, then it also satisfies a data-deletion guarantee under adaptive requests.

Based on our findings, we redefine the problem of data-deletion as designing a mechanism that **(1.)** satisfies a data-deletion 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. We judge the usefulness of a data-deletion mechanism based on its computational savings over retraining.

We provide a data-deletion solution to this problem based on Noisy Gradient Descent (Noisy-GD), a popular differentially private learning algorithm (Bassily et al., 2014; Abadi et al., 2016; Chourasia et al., 2021). Our solution demonstrates a powerful synergy between data deletion and differential privacy as the same noise needed for the privacy of records in the database also rapidly erases information regarding records deleted from the database. We provide a data-deletion guarantee for Noisy-GD in terms of Rényi divergence (Rényi et al., 1961) bound (which implies (ε, δ)-indistinguishability (Mironov, 2017)). For convex and smooth losses, we certify that under a $(q, \varepsilon_{\mathrm{dd}})$-Rényi data-deletion and $(q, \varepsilon_{\mathrm{dp}})$-Rényi DP constraint, our Noisy-GD based deletion mechanism for d-dimensional models over n-sized databases under adaptive edit requests that modify no more than r records can maintain optimal excess empirical risk of the order $O\big(\frac{qd}{\varepsilon_{\mathrm{dp}} n^2}\big)$ while saving $\Omega(n \log(\min\{\frac{n}{r}, n\sqrt{\frac{\varepsilon_{\mathrm{dd}}}{qd}}\})$ computations in gradient complexity. Our utility guarantee matches the known lower bound on private empirical risk minimization under same privacy budget (Bassily et al.,

2014). We also provide data-deletion guarantee in the non-convex setting under the assumption that loss function is bounded and smooth, and show a computational saving of $\Omega(dn \log \frac{n}{r})$ in gradient complexity while maintaining an excess risk of $\tilde{O}\big(\frac{qd}{\varepsilon_{\mathrm{dp}} n^2} + \frac{1}{n}\sqrt{\frac{q}{\varepsilon_{\mathrm{dp}}}}\big)$. Compared to our results, prior works either have a worse excess risk guarantee under a computation budget (Izzo et al., 2021; Guo et al., 2019; Sekhari et al., 2021; Ullah et al., 2021), do not satisfy differential privacy (Bourtoule et al., 2021; Gupta et al., 2021), or require unsafe internal data structures for matching our utility bounds (Neel et al., 2021).

2 Model and Preliminaries

2.1 Indistinguishability Notions and Differential Privacy

We provide the basics of indistinguishability of random variables (with additionally details in Appendix D). Let $\Theta, \Theta' \in \mathcal{O}$ be two random variables 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 *and write $\Theta \overset{\varepsilon,\delta}{\approx} \Theta'$ if, for all $O \subset \mathcal{O}$,*

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

Definition 2.2 (Rényi divergence (Rényi et al., 1961))**.** Rényi divergence *of ν w.r.t. ν' of order $q > 1$ is defined as*

$$\mathrm{R}_q\left(\nu \| \nu'\right) = \frac{1}{q-1}\log \mathrm{E}_q\left(\nu \| \nu'\right), \quad \textit{where} \quad \mathrm{E}_q\left(\nu \| \nu'\right) = \mathop{\mathbb{E}}_{\theta \sim \nu'}\left[\left(\frac{\nu(\theta)}{\nu'(\theta)}\right)^q\right], \tag{2}$$

when ν is absolutely continuous *w.r.t. ν' (denoted as $\nu \ll \nu'$). If $\nu \not\ll \nu'$, we'll say $\mathrm{R}_q\left(\nu \| \nu'\right) = \infty$.*

The notion of differential privacy is by now a golden standard measure of privacy (Dwork et al., 2006; 2014; Mironov, 2017).

Definition 2.3 (Differential Privacy (Dwork et al., 2014; Mironov, 2017))**.** *A randomized algorithm $\mathcal{A}$ is (ε, δ)-*differentially private *((ε, δ)-DP) if, for all databases $\mathcal{D}, \mathcal{D}'$ that differ in a single record (aka.* neighbouring databases*) the distribution of random variables representing the outputs $\mathcal{A}(\mathcal{D})$ and $\mathcal{A}(\mathcal{D}')$ are (ε, δ)-indistinguishable. Similarly, algorithm $\mathcal{A}$ is (q, ε)-*Rényi differentially private *((q, ε)-RDP) if $\mathrm{R}_q\left(\mathcal{A}(\mathcal{D}) \| \mathcal{A}(\mathcal{D}')\right) \leq \varepsilon$ for all neighbouring databases $\mathcal{D}, \mathcal{D}'$.*

Remark 1. *Rényi divergence is assymetric (i.e. $\mathrm{R}_q\left(\nu \| \nu'\right) \neq \mathrm{R}_q\left(\nu' \| \nu\right)$) and implies indistinguishability only in one direction. Mironov (2017, Proposition 3) show that $\mathrm{R}_q\left(\nu \| \nu'\right) \leq \varepsilon_0$ implies $\mathbb{P}\left[\Theta \in O\right] \leq e^{\varepsilon}\mathbb{P}\left[\Theta' \in O\right] + \delta$ with $\varepsilon = \varepsilon_0 + \frac{\log 1/\delta}{q-1}$ and any $0 < \delta < 1$.*

2.2 Privacy Preserving Machine Learning

Modern machine learning typically uses deep neural networks, representing a parametric transformation of input features to output. These networks are typically trained via Stochastic Gradient Descent (SGD) to minimize a differentiable loss function $\boldsymbol{\ell}(\theta; \mathbf{x}) : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ over an available corpus $\mathcal{D}$ of n records from a data universe $\mathcal{X}$. Typically an $L2$ regularizer $\mathbf{r}(\theta) = \frac{\lambda \|\theta\|_2^2}{2}$ is also included to prevent overfitting the network to $\mathcal{D}$, i.e. the minimization objective is

$$\mathcal{L}_{\mathcal{D}}(\theta) = \frac{1}{|\mathcal{D}|}\sum_{\mathbf{x} \in \mathcal{D}} \boldsymbol{\ell}(\theta; \mathbf{x}) + \mathbf{r}(\theta). \tag{3}$$

The SGD algorithm learns an approximate minimizer by iteratively nudging model parameters Θ downhill along the gradient $-\nabla \mathcal{L}_{\mathcal{B}}(\Theta)$ on a mini-batch $\mathcal{B}$ sampled randomly from $\mathcal{D}$ without replacement. The training performance of a learned model Θ is measured in terms of its *excess empirical risk* $\mathrm{err}(\Theta; \mathcal{D})$, defined as

$$\mathrm{err}(\Theta; \mathcal{D}) \overset{\text{def}}{=} \mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta) - \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})\right], \tag{4}$$

where expectation is taken of coins of the algorithm and $\theta^*_{\mathcal{D}} \overset{\text{def}}{=} \mathop{\arg\min}_{\theta \in \mathbb{R}^d} \mathcal{L}_{\mathcal{D}}(\theta)$.

Algorithm 1 Noisy-GD: Noisy Gradient Descent

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

Unfortunately, neural networks trained by SGD pose a privacy risk to user data as they are known to memorize certain records within the model parameters (Feldman, 2019). Privacy attacks such as membership inference (Shokri et al., 2017; Carlini et al., 2019) and model inversion (Fredrikson et al., 2015) can extract this information about individual records from trained networks. Noisy-GD, described in Algorithm 1, is a popular privacy-preserving variant of SGD (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016). In Noisy-GD, the loss gradients are clipped per record to a magnitude L via $\mathrm{Clip}_L(\mathbf{v}) = \mathbf{v}/\max(1, \frac{\|\mathbf{v}\|_2}{L})$ and Gaussian noise is added. We recap the privacy bounds by Abadi et al. (2016) and Chourasia et al. (2021) in terms of RDP, adapted to Algorithm 1.

Theorem 1 (RDP guarantee for Noisy-GD Algorithm 1 (Abadi et al., 2016))**.** *If $\boldsymbol{\ell}(\theta; \mathbf{x})$ is L-lipschitz (or gradients are clipped to a magnitude of L), then Noisy-GD satisfies (q, ε)-RDP with $\varepsilon = \frac{qL^2}{\sigma^2 n^2} \cdot \eta K$.*

Theorem 2 (RDP guarantee for Noisy-GD Algorithm 1 (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)-RDP with $\varepsilon = \frac{4qL^2}{\lambda \sigma^2 n^2}\left(1 - e^{-\lambda \eta K/2}\right)$.*

Remark 2. *We assume throughout the paper that the loss $\boldsymbol{\ell}(\theta; \mathbf{x})$ is L-Lipschitz for all $\mathbf{x} \in \mathcal{X}$, which is common in DP literature (Bassily et al., 2014; Dimitrakakis et al., 2014; Feldman et al., 2020). However, our guarantees apply under clipping as well (discussed further in Appendix C). We restrict our analysis to full-batch setting, since tight privacy analysis for Noisy-SGD in Rényi divergence is still an open problem*[1]

2.3 Online Edit Requests and Machine Unlearning

A database $\mathcal{D} \in \mathcal{X}^n$ can be modified by *edit requests* that replaces r distinct records[2].

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

$$\mathcal{D} \circ \langle \mathrm{ind}, \mathbf{y} \rangle = (\mathbf{x}_1, \cdots, \mathbf{x}_{\mathrm{ind}-1}, \mathbf{y}, \mathbf{x}_{\mathrm{ind}+1}, \cdots, \mathbf{x}_n). \tag{5}$$

Let $r \leq n$ and $\mathcal{U}^r = [n]^r_{\neq} \times \mathcal{X}^r$. 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 batch of r replacement operations modifying distinct indices atomically, i.e.*

$$\mathcal{D} \circ u = \mathcal{D} \circ \langle \mathrm{ind}_1, \mathbf{y}_1 \rangle \circ \cdots \circ \langle \mathrm{ind}_r, \mathbf{y}_r \rangle, \quad \textit{where } \mathrm{ind}_i \neq \mathrm{ind}_j \textit{ for all } i \neq j. \tag{6}$$

We describe the standard setting of online data-deletion in ML adopted in Gupta et al. (2021) and Neel et al. (2021). Let $\mathcal{O}$ denote the space of internal model parameters plus any internal state variables or data-structures that may be leveraged for unlearning and Φ be the space of publishable outcomes such as sanitized models or predictions. We execute a *learning* or *retraining* algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{O}$ on the initial database $\mathcal{D}_0 \in \mathcal{X}^n$ to generate the initial (secret) model with accompanying state metadata $\hat{\Theta}_0 \in \mathcal{O}$. Thereafter at any edit step $i \geq 1$, to reflect an incoming edit request $u_i \in \mathcal{U}^r$ that transforms $\mathcal{D}_{i-1} \to \mathcal{D}_i$, we execute a *data-deletion* algorithm $\mathcal{R} : \mathcal{X}^n \times \mathcal{U}^r \times \mathcal{O} \to \mathcal{O}$ that maps the current database $\mathcal{D}_{i-1}$, the edit request u_i, and the current (secret) model with metadata $\hat{\Theta}_{i-1}$ to the next (secret) model with metadata $\hat{\Theta}_i \in \mathcal{O}$. At any step $i \geq 0$, to generate publishable outcome $\phi_i \in \Phi$ from a (secret) model with metadata $\hat{\Theta}_i$, we use a mapping $f_{\mathrm{pub}} : \mathcal{O} \to \Phi$.

[1] Ye & Shokri (2022) show an improvement of $\approx \frac{b}{n}$ over Theorem 2 for Noisy-SGD with mini-batch size b, but typical mechanisms such as Gaussian get an amplification of $\approx \frac{b^2}{n^2}$ on subsampling (Mironov et al., 2019).

[2] Prior works define edits as addition or deletion of a record to the database(Neel et al., 2021; Ginart et al., 2019) which alters the database size. We only consider replacement edits for analytical simplicity.

An edit request u_i can either be independent of the history of prior published outcomes $\phi_{\leq i} = (\phi_1, \cdots, \phi_i)$, which we call a *non-adaptive* request, or be *adaptively* requested in response to the history $\phi_{\leq i}$ (e.g. a voter changing his inclination after seeing pre-election results).

Definition 2.5 (Update requester (Gupta et al., 2021))**.** *An* adaptive update requester *is a (possibly randomized) mapping $\mathcal{Q} : (\Phi \times \mathcal{U}^r)^* \to \mathcal{U}^r$ that takes as input the history of interaction between herself and the online algorithm, and outputs a new edit request for the current round. Given an update requester $\mathcal{Q}$, a start database $\mathcal{D}_0$, algorithms $(\mathcal{A}, \mathcal{R})$, and publishing function f_{pub}, the sequence of updates $(u_i)_{i\geq 1}$ can be written as*

$$u_1 = \mathcal{Q}(\phi_0), u_2 = \mathcal{Q}(\phi_0, u_1, \phi_1), \cdots, u_i = \mathcal{Q}(\phi_0, u_1, \cdots, u_{i-1}, \phi_{i-1}). \tag{7}$$

*We say $\mathcal{Q}$ is a p-*adaptive *r-*requester *if it depends on at most p of the published outcomes and issues edit requests in $\mathcal{U}^r$, i.e., there exists a mapping $\mathcal{Q}' : \Phi^{\leq p} \times \mathcal{U}^{r*} \to \mathcal{U}^r$ such that*

$$u_i = \mathcal{Q}(\phi_0, u_1, \cdots, u_{i-1}, \phi_{i-1}) = \mathcal{Q}'(u_1, u_2, \cdots, \phi_{s^1}, u_{s^1}, \cdots, \phi_{s^p}, u_{s^p}, \cdots, u_{i-1}), \tag{8}$$

for all $i \geq 1$, where $0 \leq s^1 < \cdots < s^p$. We refer to 0-adaptive requesters as non-adaptive.

Ginart et al. (2019) propose quantifying *data-deletion* ability of an algorithm $\mathcal{R}$ based on its (ε, δ)-indistinguishability w.r.t. the fresh-retraining algorithm $\mathcal{A}$ and called it an *unlearning guarantee*. Neel et al. (2021) extend this unlearning definition to the online setting of arbitrarily long edit sequences and Gupta et al. (2021) further extend it to handle adaptively chosen requests, which we state below.

Definition 2.6 ((ε, δ)-unlearning (Neel et al., 2021; Gupta et al., 2021))**.** *We say that $\mathcal{R}$ is an (ε, δ)-*unlearning *algorithm for $\mathcal{A}$ 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 $u_{\leq i} = (u_1, \cdots, u_i)$,*

$$f_{\text{pub}}(\mathcal{R}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1})) \big| u_{\leq i} \overset{\varepsilon,\delta}{\approx} f_{\text{pub}}(\mathcal{A}(\mathcal{D}_i)). \tag{9}$$

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

Remark 3. *The above definition of adaptive unlearning is stronger than the one in Gupta et al. (2021) since their definition requires satisfying only one-sided indistinguishability with at-least $(1 - \gamma)$ probability over generated edit requests $u_{\leq i}$.*

3 Unlearning Guarantees Are Unsound and Incomplete Measure of "Right to Be Forgotten"

Data deletion under the law of "right to be forgotten" (RTBF) is an obligation to *permanently erase* all information about an individual upon a verified request. In this section, we argue that unlearning guarantee in Definition 2.6 is *neither a sound nor a complete* measure of data-deletion from ML models that RTBF enforces.

Soundness. Unlearning guarantee is a bound on information leakage about a deleted record through a *single released outcome*. However, an adversary can observe multiple (potentially infinite) releases after deletion. We argue that satisfying an unlearning guarantee does not imply that an adversary is unaware of deleted records when she observes post-deletion releases. The vulnerability arises as an unlearning guarantee permits algorithms to maintain internal states while using a publishing function f_{pub} for releases. These internal states 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. *For every $\varepsilon > 0$, there exists a pair $(\mathcal{A}, \mathcal{R})$ of algorithms that satisfy $(\varepsilon, 0)$-unlearning under a publish function f_{pub} such that for all non-adaptive 1-requesters $\mathcal{Q}$, their exists an adversary 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 $\phi_{\geq i} = (\phi_i, \phi_{i+1}, \cdots)$.*

A more severe violation of RTBF may occur for even the best unlearning mechanisms that perfectly imitate retraining (i.e., satisfy an $(0, 0)$-adaptive unlearning guarantee) if the edit requests are adaptive. We demonstrate in the following theorem that an adversary can design an adaptive requester for a perfect unlearning mechanism that can extract a target record and send edit requests to duplicate it multiple times before its deletion. So, even after the original target record gets deleted, the database still has enough copies for the information to be extractable post-deletion, which violates RTBF.

Theorem 4. *There exists a pair $(\mathcal{A}, \mathcal{R})$ of algorithms that satisfy $(0, 0)$-adaptive-unlearning under a publish function f_{pub} such that by designing a 1-adaptive 1-requester $\mathcal{Q}$, an adversary, even with no access to $\mathcal{Q}$'s interaction transcript, can infer the identity of a record deleted at any arbitrary edit step $i > 3$ with probability at-least $1 - (1/2)^{i-3}$ from a single post-edit release ϕ_i.*

Such attacks are possible because unlearning guarantees protect the privacy of deleted records but not that of records currently present. Before a record's deletion, an unlearning algorithm may reveal all its information that can, in turn, be encoded back in the database through edit requests. Note that we do not assume that the adversary can witness the interaction between the requester and the data controller, as any information learned before the deletion does not violate RTBF.

We remark that lot of data-deletion algorithms in literature that claim to ensure RTBF via unlearning guarantees (Neel et al., 2021; Bourtoule et al., 2021; Gupta et al., 2021; Ginart et al., 2019; Ullah et al., 2021) have at least one of the above vulnerabilities.

Completeness. Another issue with unlearning guarantees is that valid data-deletion algorithms may fail to satisfy it. As per Definition 2.6, the publishable outcome generated by a data-deletion mechanism must imitate that of a retraining algorithm for satisfying an unlearning guarantee. However, imitating retraining is not necessary for data deletion. For instance, consider a (useless) mechanism $\mathcal{R}$ that outputs a fixed untrained model in $\mathcal{O}$ regardless of its inputs. This $\mathcal{R}$ would be an unacceptable unlearning algorithm for most re-training algorithms $\mathcal{A}$ under an identity publish function $f_{\text{pub}}(\hat{\Theta}) = \hat{\Theta}$. However, $f_{\text{pub}}(\mathcal{R}(\cdot))$ contains no information about the deleted records and should be acceptable under the data-deletion paradigm.

4 Redefining Data-Deletion in Machine Learning

In this section, we redefine *data deletion in Machine Learning* to address the problems with the notion of unlearning that we demonstrate in the preceding section. The first change we propose is to rule out the possibility of information leakage through internal data structures (as shown in Theorem 3) by requiring deletion mechanisms to be *stateless*. That is, the models produced by learning or the data-deletion algorithm are directly released without applying any publish function f_{pub}.

Secondly, the following definition of a *data-deletion guarantee* fixes the security blind spot of an adaptive unlearning guarantee. As demonstrated in Theorem 4, an adaptive requester can encode patterns specific to a target record in the database by making edit decisions in response to the observed outcomes. Thus, being indistinguishable from retraining on the edited database does not guarantee data deletion, as the target's information remains extractable even after the target record's deletion, potentially revealing its identity. In our definition, we account for an adaptive adversary's influence by measuring the indistinguishability of a data-deletion mechanism's output distribution from any distribution independent of the deleted record.

Definition 4.1 ((q, ε)-data-deletion under p-adaptive r-requesters)**.** *Let $q > 1$, $\varepsilon \geq 0$, and $p, r \in \mathbb{N}$. We say that an algorithm pair $(\mathcal{A}, \mathcal{R})$ satisfies an (q, ε)-data-deletion under p-adaptive r-requesters if, for all initial databases $\mathcal{D}_0 \in \mathcal{X}^n$ and any adaptive update requester that witnesses not more than p model releases and issues edit requests in $\mathcal{U}^r$, the following condition holds. For every step $i \geq 1$ and all edit edit request $u = \{\langle \text{ind}_u, \mathbf{y}_u \rangle\} \in \mathcal{U}^1$, their exists a distribution π_i^u that is independent of the record being deleted, i.e., $\pi_i^u \perp \mathcal{D}_{i-1}[\text{ind}_u]$ such that*

$$\mathrm{R}_q \left(\mathcal{R}(\mathcal{D}_{i-1}, u, \hat{\Theta}_{i-1}) \middle\| \pi_i^u \right) \leq \varepsilon. \tag{10}$$

By independence, we mean that the construction of the distribution π_i^u should be such that it is not affected by the original presence of any specific deleted record $\mathcal{D}_{i-1}[\text{ind}_u]$. We argue that the above definition is a reliable guarantee on data-deletion as for any edit request $u \in \mathcal{U}^1$, no adversary observing $\mathcal{R}(\mathcal{D}_{i-1}, u, \hat{\Theta}_{i-1})$ can be too confident that the observation was *not* from π_i^u, a distribution

that contains no information about the record that u deletes. The Rényi divergence bound in our definition implies precisely this one-sided indistinguishability, as pointed out in Remark 1.

Remark 4. *Under adaptive requests, the construction $\pi_i^u = \mathcal{A}(\mathcal{D}_{i-1} \circ u)$ does not qualify as being independent of the record that request u deletes in $\mathcal{D}_{i-1}$. Although the edit request u removes the record $\mathcal{D}_{i-1}[\mathrm{ind}_u]$ from $\mathcal{D}_{i-1}$, we highlight that it does not remove the record's influences from the rest of the database due to its original presence, as an adaptive requester could have successfully duplicated/encoded information of the target record. So, arguing data-deletion via indistinguishability to $\mathcal{A}(\mathcal{D}_{i-1} \circ u)$ is flawed under adaptivity.*

Remark 5. *Under non-adaptive requests, an adversary cannot introduce any target-dependent influence on the database, as it is equivalent to fixing the request sequence apriori. Hence, the original presence of $\mathcal{D}_{i-1}[\mathrm{ind}_u]$ does not influence other records in $\mathcal{D}_{i-1}$, so $\mathcal{D}_{i-1} \circ u$ is completely independent of it. Consequently, under non-adaptive requests, the distribution π_i^u in* (10) *can be the output $\pi(\mathcal{D}_{i-1} \circ u)$ of any randomized map $\pi : \mathcal{X}^n \to \mathcal{O}$, including the learning algorithm $\mathcal{A}$.*

Note that our definition of data deletion introduces a gradation on the power of an update requester $\mathcal{Q}$ based on the count of previous releases it is allowed to observe. We argue that this is necessary because the amount of dependence that an adaptive request can encode back in a database is proportional to the information it can extract about a target record from past observations. We formalize this idea in the following theorem.

Theorem 5 (Reduction from adaptive to non-adaptive data deletion)**.** *Let $0 \leq \varepsilon_{\mathrm{dd}} \leq \varepsilon_{\mathrm{dp}}$. If an algorithm pair $(\mathcal{A}, \mathcal{R})$ satisfies $(q, \varepsilon_{\mathrm{dd}})$-data-deletion under all non-adaptive r-requesters and $(q, \varepsilon_{\mathrm{dp}})$-Rényi differential privacy, then it also satisfies $(q, \varepsilon_{\mathrm{dd}} + p\varepsilon_{\mathrm{dp}})$-data-deletion under all p-adaptive r-requesters.*

This paper provides the following reformulation of the data deletion problem in Machine Learning.
Problem Definition. Let constants $q > 1$, $0 < \varepsilon_{\mathrm{dd}} \leq \varepsilon_{\mathrm{dp}}$, and $\alpha > 0$. The goal is to design a learning mechanism $\mathcal{A} : \mathcal{X}^n \to \mathcal{O}$ and a deletion mechanisms $\mathcal{R} : \mathcal{X}^n \times \mathcal{U}^r \times \mathcal{O} \to \mathcal{O}$ such that

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

We judge the benefit of data deletion based on the computation saving that $\mathcal{R}$ offers over $\mathcal{A}$ per request. That is, if we want $\mathrm{err}(\bar{\Theta}_i; \mathcal{D}_i) \leq \alpha$ for all $i \geq 0$, where $\bar{\Theta}_i$ is trained from scratch by $\mathcal{A}(\mathcal{D}_i)$, how large is $\mathrm{Cost}(\mathcal{A})$ compared to $\mathrm{Cost}(\mathcal{R})$. Note that α should ideally be the optimal excess risk attainable under the $(q, \varepsilon_{\mathrm{dp}})$-RDP constraint. In the third constraint, we also remark that one may use any reasonable measure of utility, such as population risk, instead of excess empirical risk.

5 Data Deletion Using Noisy Gradient Descent

This section proposes a simple and effective data-deletion solution based on Noisy-GD (Algorithm 1). We base our design on Noisy-GD because we only need to provide a data-deletion guarantee under non-adaptive requesters as its differential privacy property (in Theorem 1 and Theorem 2) ensures security against adaptive requesters (through our reduction Theorem 5). Our data-deletion solution is formally stated as follows.

Definition 5.1 (Data-deletion based on Noisy-GD)**.** *Let space of model parameters be $\mathcal{O} = \mathbb{R}^d$. Our learning algorithm $\mathcal{A}_{\textit{Noisy-GD}}(\mathcal{D}_0) \stackrel{\textit{def}}{=} \textit{Noisy-GD}(\mathcal{D}_0, \Theta_0, K_{\mathcal{A}})$ runs Noisy-GD algorithm for $K_{\mathcal{A}}$ iterations to generate the first learned model $\hat{\Theta}_0$ on the input database $\mathcal{D}_0 \in \mathcal{X}^n$, with Θ_0 sampled from a weight initialization distribution ρ. Our data-deletion algorithm $\mathcal{R}_{\textit{Noisy-GD}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) \stackrel{\textit{def}}{=} \textit{Noisy-GD}(\mathcal{D}_{i-1} \circ u_i, \hat{\Theta}_{i-1}, K_{\mathcal{R}})$ processes an edit request $u_i \in \mathcal{U}^r$ by running $K_{\mathcal{R}}$ iterations of Noisy-GD algorithm on the updated database $\mathcal{D}_i = \mathcal{D}_{i-1} \circ u_i$ to generate the updated model $\hat{\Theta}_i$ from the current model $\hat{\Theta}_{i-1} \in \mathbb{R}^d$.*

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

5.1 Data-Deletion and Accuracy Guarantees Under Convexity

We give the following guarantee for pair $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$ under convexity of loss $\boldsymbol{\ell}(\theta; \mathbf{x})$.

Theorem 6 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda, \beta, L > 0$, $q > 1$, and $0 < \varepsilon_{\text{dd}} \leq \varepsilon_{\text{dp}}$. 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, the regularizer be $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$. If the learning rate be $\eta = \frac{1}{2(\lambda+\beta)}$, the gradient noise variance is $\sigma^2 = \frac{4qL^2}{\lambda\varepsilon_{\text{dp}}n^2}$, 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 $\mathcal{A}_{\text{Noisy-GD}}$ and $\mathcal{R}_{\text{Noisy-GD}}$ are $(q, \varepsilon_{\text{dp}})$-RDP for any $K_{\mathcal{A}}, K_{\mathcal{R}} \geq 0$,*

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

$$\text{if} \quad K_{\mathcal{R}} \geq 4\kappa \log \frac{\varepsilon_{\text{dp}}}{\varepsilon_{\text{dd}}}, \tag{11}$$

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

$$\text{if} \quad K_{\mathcal{A}} \geq 4\kappa \log\left(\frac{\varepsilon_{\text{dp}}n^2}{4qd}\right), \quad \text{and} \quad K_{\mathcal{R}} \geq 4\kappa \log \max\left\{5\kappa, \frac{8\varepsilon_{\text{dp}}r^2}{qd}\right\}. \tag{12}$$

Proof of Theorem 6 can be found in Appendix G.3.

Our utility upper bound for data-deletion matches the theoretical lower bound of $\Omega(\min\left\{1, \frac{d}{\varepsilon^2 n^2}\right\})$ in Bassily et al. (2014) for the best attainable utility of (ε, δ)-DP algorithms on Lipschitz, smooth, strongly-convex loss functions[3]. Thus, our data-deletion algorithm, $\mathcal{R}_{\text{Noisy-GD}}$, incurs no additional cost in utility but saves substantial computation costs. Our data-deletion algorithm offers a computation saving of $\Omega(n \log \min\{\frac{n}{r}, n\sqrt{\frac{\varepsilon_{\text{dd}}}{qd}}\})$ per request while guaranteeing privacy, deletion, and optimal utility and without unsafe internal data structures. This saving is better than all existing unlearning algorithms in literature that we know of, and we present a detailed comparison in Table 1.

Also, observe that for satisfying $(q, \varepsilon_{\text{dp}})$-RDP and $(q, \varepsilon_{\text{dd}})$-data-deletion for non-adaptive r-requesters, the number of iterations $K_{\mathcal{R}}$ needed is independent of the size, r, of the deletion batch, depending solely on the ratio $\frac{\varepsilon_{\text{dd}}}{\varepsilon_{\text{dp}}}$. However, the number of iterations required for ensuring optimal utility with differential privacy grows with r. As such, we highlight that when deletion batches are sufficiently large, i.e., $r \geq \sqrt{\frac{qd}{\varepsilon_{\text{dd}}}}$, ensuring optimal utility under $(q, \varepsilon_{\text{dp}})$-RDP results in $(q, \varepsilon_{\text{dd}})$-data-deletion for free, thus demonstrating synergy between privacy and data deletion.

5.2 Data-Deletion and Utility Guarantees Under Non-Convexity

We give the following guarantee for pair $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$ under non-convexity of loss $\boldsymbol{\ell}(\theta; \mathbf{x})$.

Theorem 7 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda, \beta, L, \sigma^2, \eta > 0$, $q, B > 1$, and $0 < \varepsilon_{\text{dd}} \leq \varepsilon_{\text{dp}} < d$. 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 $\mathcal{A}_{\text{Noisy-GD}}$ and $\mathcal{R}_{\text{Noisy-GD}}$ are $(q, \varepsilon_{\text{dp}})$-RDP for any $\eta \geq 0$ and any $K_{\mathcal{A}}, K_{\mathcal{R}} \geq 0$ if*

$$\sigma^2 \geq \frac{qL^2}{\varepsilon_{\text{dp}}n^2} \cdot \eta \max\{K_{\mathcal{A}}, K_{\mathcal{R}}\}, \tag{13}$$

[3]Recall from Remark 1 that $(q, \varepsilon_{\text{dp}})$-RDP implies (ε, δ)-DP for $q = 1 + \frac{2}{\varepsilon}\log(1/\delta)$ and $\varepsilon_{\text{dp}} = \varepsilon/2$. When $\varepsilon = \Theta(\log(1/\delta))$, one can evaluate that $\frac{q}{\varepsilon_{\text{dp}}} = \Theta(\frac{\log(1/\delta)}{\varepsilon^2})$.

Table 1: Comparison of the computation savings in gradient complexity per edit request and the required size of (unsafe) internal data structures with prior unlearning algorithms. For a fair comparison, we require that each of them satisfy an (ε, δ)-data-deletion guarantee (ignoring the statelessness requirement) and have the same excess empirical risk bound. Edit requests are non-adaptive and modify $r = 1$ record in n-sized databases. The models' dimension is d. We assume loss $\boldsymbol{\ell}(\theta; \mathbf{x})$ to be convex, 1-Lipschitz, and $O(1)$-smooth, and $L2$ regularization constant to be $O(1)$.

Unlearning Algorithm	**Internal (unsafe) data-structure size**	**Compute savings for ith edit**
Noisy-m-A-SGD [Thm. 1, (Ullah et al., 2021)]	$\Omega(d^{5/4})$	$\Omega\left(\sqrt{d}\left(1 - \frac{\sqrt{d}}{n}\right)\right)$
Perturbed-GD [Thm. 9, (Neel et al., 2021)]	$\Omega(d)$	$\Omega\left(n \log\left(\frac{\varepsilon n}{\sqrt{d}}\right)\right)$
Perturbed-GD [Thm. 28, (Neel et al., 2021)]	0	$\Omega\left(n \log\left(\frac{\varepsilon n}{\log^2(id)\sqrt{d}}\right)\right)$
Noisy-GD [Thm. 6, Ours]	0	$\Omega\left(n \log \min\left\{n, \frac{\varepsilon n}{\sqrt{d}}\right\}\right)$

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

$$K_{\mathcal{A}} \geq \frac{2B}{\lambda\eta} \log\left(\frac{q \log(B)}{\varepsilon_{\mathrm{dd}}}\right), \quad K_{\mathcal{R}} \geq K_{\mathcal{A}} - \frac{2B}{\lambda\eta} \log\left(\frac{\log(B)}{2\left(\varepsilon_{\mathrm{dd}} + \frac{r}{n}\log(B)\right)}\right), \tag{14}$$

(3.) *and all models in sequence $(\hat{\Theta}_i)_{i \geq 0}$ output by $(\mathcal{A}_{\textit{Noisy-GD}}, \mathcal{R}_{\textit{Noisy-GD}}, \mathcal{Q})$ on any $\mathcal{D}_0 \in \mathcal{X}^n$, where $\mathcal{Q}$ is an r-requester, satisfy $\mathrm{err}(\hat{\Theta}_i; \mathcal{D}_i) = \tilde{O}\left(\frac{dq}{\varepsilon_{\mathrm{dp}} n^2} + \frac{1}{n}\sqrt{\frac{q\varepsilon_{\mathrm{dd}}}{\varepsilon_{\mathrm{dp}}}}\right)$ when inequalities in* (14) *and* (13) *are equalities.*

Proof of Theorem 7 can be found in Appendix G.4.

For the privacy result, we plug in (Abadi et al., 2016)'s guarantee (in RDP as stated in Theorem 1). To prove data deletion and utility, we establish a Rényi indistinguishability of generated models $\hat{\Theta}_i$ with the Gibbs distribution $\pi(\mathcal{D}) \propto \exp(-\mathcal{L}_{\mathcal{D}}/\sigma^2)$, i.e., we show $\mathrm{R}_q\left(\hat{\Theta}_i \middle\| \pi(\mathcal{D}_i)\right) \leq \varepsilon_{\mathrm{dd}}$ for all $i \geq 0$. We highlight that the techniques we develop for analyzing data deletion differ tremendously from those used to argue differential privacy; satisfying RDP requires a bounded gradient sensitivity, while we need bounded loss sensitivity for our data-deletion guarantee.

Under non-convexity, all prior works on deletion have focused on empirical analysis for utility. As far as we know, we are the first to provide utility guarantees in this setting. Moreover, our non-convex utility bound exceeds the optimal privacy-preserving utility under convexity by only a factor of $\tilde{O}\left(\frac{1}{n}\sqrt{\frac{q\varepsilon_{\mathrm{dd}}}{\varepsilon_{\mathrm{dp}}}}\right)$, which becomes small for large databases or small deletion to privacy budget ratio.

Also, observe a strict computational benefit in using $\mathcal{R}_{\text{Noisy-GD}}$ whenever the fraction of edited records in a single update request satisfies $\frac{r}{n} \leq \frac{1}{2} - \frac{\varepsilon_{\mathrm{dd}}}{\log B}$. In the deletion regime where $\varepsilon_{\mathrm{dd}} = \log(B)/4$, relying on $\mathcal{R}_{\text{Noisy-GD}}$ rather than retraining with $\mathcal{A}_{\text{Noisy-GD}}$ is $\Omega(dn \log \frac{n}{r})$ cheaper.

6 Conclusions

We showed that unlearning guarantees are an unsuitable data deletion measure in the online setting of adaptive edit requests and proposed a proper notion of data deletion in line with the "right to be forgotten." We also showed that differential privacy is necessary for data deletion when requests are adaptive and proved a reduction from secure deletion under non-adaptive requests to adaptive requests under DP. Our theoretical results on Noisy-GD based data-deletion, for both convex and non-convex losses, show a substantial computation saving over retraining with a reliable deletion guarantee and at no additional cost in utility.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

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

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

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, 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.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 896–911, 2021.

Sinho Chewi, Murat A Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of langevin monte carlo from poincar\'e to log-sobolev. *arXiv preprint arXiv:2112.12662*, 2021.

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

Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein. Robust and private bayesian inference. In *International Conference on Algorithmic Learning Theory*, pp. 291–305. Springer, 2014.

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

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

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

Vitaly Feldman. Does learning require memorization. *A short tale about a long tail. CoRR, abs/1906.05271*, 2019.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 439–449, 2020.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 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.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34, 2021.

Richard Holley and Daniel W Stroock. Logarithmic sobolev inequalities and stochastic ising models. 1986.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR, 2021.

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

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

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

Ilya Mironov, Kunal Talwar, and Li Zhang. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

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

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

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

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

Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1. Berkeley, California, USA, 1961.

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

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

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pp. 4126–4142. PMLR, 2021.

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

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

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

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pp. 2493–2502. PMLR, 2015.

Jiayuan Ye and Reza Shokri. Differentially private learning needs hidden state (or much faster convergence). *arXiv preprint arXiv:2203.05363*, 2022.

Appendix

Table of Contents

A Table of Notations **14**

B Calculus Refresher **15**

C Loss Function Properties **15**

D Divergence measures and their properties **17**

E Proofs for Section 3 **18**

F Proofs for Section 4 **20**

G Additional Preliminaries and Proofs for Section 5 **21**

G.1 Langevin Diffusion and Markov Semigroups 21

G.2 Isoperimetric Inequalities and Their Properties 21

G.3 Proofs for Subsection 5.1 23

G.4 Proofs for Subsection 5.2 32

A TABLE OF NOTATIONS

Table 2: Symbol reference

Symbol	Meaning
$\mathcal{O}$	Arbitrary model parameter space.
Φ	Space of publishable outcomes.
d	Dimension of model parameters.
$\mathbb{R}^d$	Space of d-dimensional model parameters.
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	Number of records replaced per edit and number of releases observable by a 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 $\pi(\mathcal{D})$.
π_i^u	A distribution independent of record deleted by request u on database $\mathcal{D}_{i-1}$.
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 data-deletion budget in q-Rényi divergence.
ε, δ	Parameters for DP-like indistinguishability.
$\mathcal{A}, \mathcal{A}_{\text{Noisy-GD}}$	Learning algorithm and Noisy-GD based learning algorithm respectively.
$\mathcal{R}, \mathcal{R}_{\text{Noisy-GD}}$	Data-deletion algorithm and Noisy-GD based data-deletion algorithm respectively.
$K_{\mathcal{A}}, K_{\mathcal{R}}$	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.
$\Theta_{\eta k}, \Theta'_{\eta k}$	Random variables representing the iteration k parameter distribution in Noisy-GD.
Θ_t, Θ'_t	Random variables representing the time t distribution 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.
$\mathrm{R}_q(\nu \| \nu')$	Rényi divergence of distribution of ν w.r.t ν'.
$\mathrm{E}_q(\nu \| \nu')$	q^{th} moment of likelihood ratio r.v. between ν and ν'.
$\mathrm{I}(\nu \| \nu'), \mathrm{I}_q(\nu \| \nu')$	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'.
$\mathrm{KL}(\nu \| \nu')$	Kullback Leibler divergence of distribution ν w.r.t. ν'.
$P_t, \mathcal{G}, \mathcal{G}^*$	Markov semigroup, its infintesimal 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.
$\mathrm{prox}_{\mathcal{L}}$	Proxmial mapping for objective $\mathcal{L}$.

B 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

$$\nabla\mathcal{L}(\theta) = \left(\frac{\partial\mathcal{L}(\theta)}{\partial\theta_1}, \cdots, \frac{\partial\mathcal{L}(\theta)}{\partial\theta_2}\right). \tag{15}$$

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

$$\nabla^2\mathcal{L}(\theta) = \left(\frac{\partial^2\mathcal{L}(\theta)}{\partial\theta_i\theta_j}\right)_{1\le i,j\le d}. \tag{16}$$

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

$$\Delta\mathcal{L}(\theta) = \mathsf{Tr}\left(\nabla^2\mathcal{L}(\theta)\right). \tag{17}$$

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

$$\mathrm{div}\,(\mathbf{v})\,(\theta) = \sum_{i=1}^{d} \frac{\partial\mathbf{v}_i(\theta)}{\partial\theta_i}. \tag{18}$$

Some identities that we would rely on:

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

$$\mathrm{div}\,(\nabla\mathcal{L})\,(\theta) = \sum_{i=1}^{d} \frac{\partial^2\mathcal{L}(\theta)}{\partial\theta_i^2} = \Delta\mathcal{L}(\theta). \tag{19}$$

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 $\mathcal{O}$,

$$\int_{\mathcal{O}} \langle\mathbf{v}(\theta), \nabla f(\theta)\mathrm{d}\theta\rangle = -\int_{\mathcal{O}} f(\theta)(\mathrm{div}\,(\mathbf{v}))(\theta)\mathrm{d}\theta. \tag{20}$$

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

$$\int_{\mathcal{O}} f(\theta)\Delta g(\theta)\mathrm{d}\theta = -\int_{\mathcal{O}} \langle\nabla f(\theta), \nabla g(\theta)\rangle\,\mathrm{d}\theta = \int_{\mathcal{O}} g(\theta)\Delta f(\theta)\mathrm{d}\theta. \tag{21}$$

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.

$$\langle\mathbf{v}_1, \mathbf{v}_2\rangle\,(\theta) \le \frac{1}{2a}\left\|\mathbf{v}_1(\theta)\right\|_2^2 + \frac{1}{2b}\left\|\mathbf{v}_2(\theta)\right\|_2^2. \tag{22}$$

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

C 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 C.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 $\mathbf{x} \in \mathcal{X}$,*

$$\left|\boldsymbol{\ell}(\theta; \mathbf{x}) - \boldsymbol{\ell}(\theta'; \mathbf{x})\right| \le L\left\|\theta - \theta'\right\|_2. \tag{23}$$

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

Definition C.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 C.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 $\mathbf{x}\in\mathcal{X}$,*

$$\boldsymbol{\ell}(\theta';\mathbf{x})\geq\boldsymbol{\ell}(\theta;\mathbf{x})+\langle\nabla\boldsymbol{\ell}(\theta;\mathbf{x}),\theta'-\theta\rangle\,, \tag{24}$$

and is said to be λ-strongly convex if

$$\boldsymbol{\ell}(\theta';\mathbf{x})\geq\boldsymbol{\ell}(\theta;\mathbf{x})+\langle\nabla\boldsymbol{\ell}(\theta;\mathbf{x}),\theta'-\theta\rangle+\frac{\lambda}{2}\left\|\theta'-\theta\right\|_2^2. \tag{25}$$

Theorem 8 ((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 C.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 $\mathbf{x}\in\mathcal{X}$,*

$$\left\|\nabla\boldsymbol{\ell}(\theta;\mathbf{x})-\nabla\boldsymbol{\ell}(\theta';\mathbf{x})\right\|_2\leq\beta\left\|\theta-\theta'\right\|_2. \tag{26}$$

Theorem 9 ((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 $\mathbf{x}\in\mathcal{X}$,*

$$\nabla^2\boldsymbol{\ell}(\theta;\mathbf{x})\preccurlyeq\beta\mathbb{I}_d. \tag{27}$$

C.0.1 Effect of Gradient Clipping

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

$$\mathcal{L}_{\mathcal{D}}(\theta)=\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}}\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})+\mathbf{r}(\theta), \tag{28}$$

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 10 (Gradient clipping retains convexity)**.** *If $\boldsymbol{\ell}(\theta;\mathbf{x})$ is a twice continuously differentiable convex function for every $\mathbf{x}\in\mathbb{R}^d$, then surrogate loss $\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ resulting from gradient clipping is also convex for every $\mathbf{x}\in\mathbb{R}^d$.*

Proof. Note that the clip operation $\mathrm{Clip}_L(\mathbf{v})$ is a closed-form solution of the orthogonal projection onto a closed ball of radius L and centered around origin, i.e.

$$\mathrm{Clip}_L(\mathbf{v})=\operatorname*{arg\,min}_{\|\mathbf{v}'\|_2\leq L}\left\|\mathbf{v}-\mathbf{v}'\right\|_2. \tag{29}$$

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

$$\langle\mathbf{v}'-\mathrm{Clip}_L(\mathbf{v}),\mathbf{v}-\mathrm{Clip}_L(\mathbf{v})\rangle\leq 0\quad\text{if and only if }\|\mathbf{v}'\|_2\leq L. \tag{30}$$

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

$$\left\langle\nabla\bar{\boldsymbol{\ell}}(\theta+h\hat{\mathbf{v}};\mathbf{x})-\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}),\nabla\boldsymbol{\ell}(\theta;\mathbf{x})-\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})\right\rangle\leq 0, \tag{31}$$

$$\left\langle\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})-\nabla\bar{\boldsymbol{\ell}}(\theta+h\hat{\mathbf{v}};\mathbf{x}),\nabla\boldsymbol{\ell}(\theta+h\hat{\mathbf{v}};\mathbf{x})-\nabla\bar{\boldsymbol{\ell}}(\theta+h\hat{\mathbf{v}};\mathbf{x})\right\rangle\leq 0, \tag{32}$$

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

$$\left\langle\nabla_{\hat{\mathbf{v}}}\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x}),\nabla\boldsymbol{\ell}(\theta;\mathbf{x})-\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})\right\rangle=0, \tag{33}$$

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 11 (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 the 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 12 (Gradient clipping retains smoothness)**.** *If $\boldsymbol{\ell}(\theta;\mathbf{x})$ is a continuously differentiable and β-smooth function for every $\mathbf{x}\in\mathbb{R}^d$, then surrogate loss $\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})$ resulting from gradient clipping is also β-smooth for every $\mathbf{x}\in\mathbb{R}^d$.*

Proof. Note that the gradient clipping operation is equivalent to an orthogonal projection operation into a ball of radius L, i.e. $\mathrm{Clip}_L(\mathbf{v})=\arg\min_{\mathbf{v}'}\{\|\mathbf{v}'-\mathbf{v}\|_2:\mathbf{v}\in\mathbb{R}^d,\|\mathbf{v}'\|_2\le L\}$. Since orthogonal projection onto a closed convex set is a 1-Lipschitz operation, for any $\theta,\theta'\in\mathbb{R}^d$,

$$\left\|\nabla\bar{\boldsymbol{\ell}}(\theta;\mathbf{x})-\nabla\bar{\boldsymbol{\ell}}(\theta';\mathbf{x})\right\|_2\le\|\nabla\boldsymbol{\ell}(\theta;\mathbf{x})-\nabla\boldsymbol{\ell}(\theta';\mathbf{x})\|_2\le\beta\,\|\theta-\theta'\|_2\,. \tag{34}$$

□

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 13 (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.

D Divergence measures and their properties

Let $\Theta,\Theta'\in\mathcal{O}$ be two random variables with probability measures ν,ν' respectively. We abuse the notataions 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 D.1 ((ε,δ)-indistinguishability (Dwork et al., 2014))**.** *We say ν and ν' are (ε,δ)-*indistinguishable *if for all $O\subset\mathcal{O}$,*

$$\mathop{\mathbb{P}}_{\Theta\sim\nu}[\Theta\in O]\le e^{\varepsilon}\mathop{\mathbb{P}}_{\Theta'\sim\nu'}[\Theta'\in O]+\delta\quad\textit{and}\quad\mathop{\mathbb{P}}_{\Theta'\sim\nu'}[\Theta'\in O]\le e^{\varepsilon}\mathop{\mathbb{P}}_{\Theta\sim\nu}[\Theta\in O]+\delta. \tag{35}$$

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

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

$$\mathrm{R}_q(\nu\|\nu')=\frac{1}{q-1}\log\mathrm{E}_q(\nu\|\nu'),\quad\textit{where}\quad\mathrm{E}_q(\nu\|\nu')=\mathop{\mathbb{E}}_{\theta\sim\nu'}\left[\left(\frac{\nu(\theta)}{\nu'(\theta)}\right)^q\right], \tag{36}$$

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 14 (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 $O\subset\mathcal{O}$,*

$$\mathop{\mathbb{P}}_{\Theta\sim\nu}[\Theta\in O]\le e^{\varepsilon}\mathop{\mathbb{P}}_{\Theta'\sim\nu'}[\Theta'\in O]+\delta, \tag{37}$$

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 15 (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 16 (Rényi composition (Mironov, 2017, Proposition 1))**.** *If $\mathcal{A}_1, \cdots, \mathcal{A}_k$ are randomized algorithms satisfying, respectively, (q, ε_1)-RDP, $\cdots$, (q, ε_k)-RDP then their composed mechanism defined as $(\mathcal{A}_1(\mathcal{D}), \cdots, \mathcal{A}_k(\mathcal{D}))$ is $(q, \varepsilon_1 + \cdots + \varepsilon_k)$-RDP. Moreover, i^{th} algorithm can be chosen on the basis of the outputs of algorithms $\mathcal{A}_1, \cdots, \mathcal{A}_{i-1}$.*

Theorem 17 (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:*

$$\mathrm{R}_q(\nu \| \nu') \leq \mathrm{R}_q(\nu \| \rho) + \mathrm{R}_\infty(\rho \| \nu'). \tag{38}$$

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

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

$$\mathrm{KL}(\nu \| \nu') = \mathop{\mathbb{E}}_{\theta \sim \nu}\left[\log \frac{\nu(\theta)}{\nu'(\theta)}\right]. \tag{39}$$

Rényi divergence generalizes Kullback-Leibler divergence (Van Erven & Harremos, 2014) as $\lim_{q \to 1} \mathrm{R}_q(\nu \| \nu') = \mathrm{KL}(\nu \| \nu')$.

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

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

$$\mathrm{W}_2(\nu, \nu') = \inf_{\Pi} \mathop{\mathbb{E}}_{\Theta, \Theta' \sim \Pi}\left[\|\Theta - \Theta'\|_2^2\right]^{\frac{1}{2}}, \tag{40}$$

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

Definition D.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*

$$\mathrm{I}(\nu \| \nu') = \mathop{\mathbb{E}}_{\theta \sim \nu}\left[\left\|\nabla \log \frac{\nu(\theta)}{\nu'(\theta)}\right\|_2^2\right]. \tag{41}$$

Definition D.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*

$$\mathrm{I}_q(\nu \| \nu') = \frac{4}{q^2} \mathop{\mathbb{E}}_{\theta \sim \nu'}\left[\left\|\nabla \left(\frac{\nu(\theta)}{\nu'(\theta)}\right)^{q/2}\right\|_2^2\right] = \mathop{\mathbb{E}}_{\theta \sim \nu'}\left[\left(\frac{\nu(\theta)}{\nu'(\theta)}\right)^{q-2}\left\|\nabla \left(\frac{\nu(\theta)}{\nu'(\theta)}\right)\right\|_2^2\right]. \tag{42}$$

E Proofs for Section 3

Theorem 3. *For every $\varepsilon > 0$, there exists a pair $(\mathcal{A}, \mathcal{R})$ of algorithms that satisfy $(\varepsilon, 0)$-unlearning under a publish function f_{pub} such that for all non-adaptive 1-requesters $\mathcal{Q}$, their exists an adversary 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 $\phi_{\geq i} = (\phi_i, \phi_{i+1}, \cdots)$.*

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 $(\mathcal{A}, \mathcal{R})$ 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 $\hat{\Theta}_0 = \mathcal{A}(\mathcal{D}_0) = \langle \sum_{\mathbf{x} \in \mathcal{D}_0} h(\mathbf{x}), 0, 0 \rangle$,

and, for an edit request $u_i = \{\langle \mathrm{ind}_i, \mathbf{y}_i \rangle\}$, our algorithm $\mathcal{R}$ updates the secret model $\hat{\Theta}_{i-1} \to \hat{\Theta}_i$ following the rule

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

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 $i \geq 1$,

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

Hence, $\mathcal{R}$ is an $(\varepsilon, 0)$-unlearning algorithm for $\mathcal{A}$ under f_{pub}.

To show that an adversary can still infer the identity of record deleted by update $u_i = (\mathrm{ind}_i, \bullet)$, 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 $\mathcal{R}$ 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,

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

□

Theorem 4. *There exists a pair $(\mathcal{A}, \mathcal{R})$ of algorithms that satisfy $(0, 0)$-adaptive-unlearning under a publish function f_{pub} such that by designing a 1-adaptive 1-requester $\mathcal{Q}$, an adversary, even with no access to $\mathcal{Q}$'s interaction transcript, can infer the identity of a record deleted at any arbitrary edit step $i > 3$ with probability at-least $1 - (1/2)^{i-3}$ from a single post-edit release ϕ_i.*

Proof. Let data universe be $\mathcal{X}$, the internal state space $\mathcal{O}$, and 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 $\mathcal{A}(\mathcal{D}) = \mathrm{med}(\mathcal{D})$. For an arbitrary edit request $u \in \mathcal{U}^r$, our unlearning algorithm is defined as $\mathcal{R}(\mathcal{D}, u, \bullet) = \mathrm{med}(\mathcal{D} \circ u)$. Let the publish function $f_{\mathrm{pub}} : \mathcal{O} \to \Phi$ be an identity function, i.e. $f_{\mathrm{pub}}(\hat{\Theta}) = \hat{\Theta}$.

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

$$f_{\mathrm{pub}}(\mathcal{R}(\mathcal{D}_{i-1}, u_i, \bullet)) = f_{\mathrm{pub}}(\mathcal{A}(\mathcal{D}_i)), \quad \text{for all } i \geq 1 \text{ and any } \bullet \in \mathcal{O}. \tag{43}$$

Therefore, $\mathcal{R}$ is a $(0, 0)$-adaptive unlearning algorithm for $\mathcal{A}$ 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., $\mathrm{med}(\mathcal{D}_i) = \mathrm{med}(\mathcal{D}_0)$:

$$\mathcal{Q}(\phi_0, u_1, u_2, \cdots, u_{j-1}) = \{\langle \mathrm{ind}_j, \phi_0 \rangle\} \quad \forall\, 1 \leq j < i, \tag{44}$$

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

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

□

F Proofs for Section 4

Theorem 5 (Reduction from Adaptive to Non-adaptive Data Deletion)**.** *If an algorithm pair $(\mathcal{A}, \mathcal{R})$ satisfy $(q, \varepsilon_{\mathrm{dd}})$-data-deletion under all non-adaptive r-requesters and $(q, \varepsilon_{\mathrm{dp}})$-Rényi differential privacy, then it also satisfies $(q, \varepsilon_{\mathrm{dd}} + p\varepsilon_{\mathrm{dp}})$-data-deletion for all p-adaptive r-requesters.*

Proof. Let $\mathcal{D}_0 \in \mathcal{X}^n$ be any arbitrary database and for any arbitrary edit request $u = \langle \mathrm{ind}_u, \mathbf{y}_u \rangle \in \mathcal{U}^1$, let $\mathcal{D}_0' \in \mathcal{X}^n$ be a neighbouring database such that $\mathcal{D}_0 \circ u = \mathcal{D}_0' \circ u$. 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 $(\mathcal{A}, \mathcal{R})$ with inital 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 $(\mathcal{Q}, \mathcal{A}, \mathcal{R})$ on $\mathcal{D}_0'$.

Since $(\mathcal{A}, \mathcal{R})$ is assumed to satisfy $(q, \varepsilon_{\mathrm{dd}})$-data-deletion guarantee under non-adpative r-requesters, there exists a mapping $\pi : \mathcal{X}^n \to \mathcal{O}$ such that for all $i \geq 1$, any edit $u \in \mathcal{U}^1$ and any sequence $u_{<i} \in \mathcal{U}^r$,

$$\mathrm{R}_q\left(\mathcal{R}(\mathcal{D}_0 \circ u_{<i}, u, \hat{\Theta}_{i-1}) \middle\| \pi(\mathcal{D}_0 \circ u_{<i} \circ u)\right) \leq \varepsilon_{\mathrm{dd}} \tag{45}$$

$$\implies \mathrm{R}_q\left(\mathcal{R}(\mathcal{D}_0 \circ U_{<i}, u, \hat{\Theta}_{i-1})\big|_{U_{<i}=u_{<i}} \middle\| \pi(\mathcal{D}_0 \circ U_{<i}' \circ u)\big|_{U_{<i}'=u_{<i}}\right) \leq \varepsilon_{\mathrm{dd}}. \tag{46}$$

Note that since u modifies the differing record between $\mathcal{D}_0$ and $\mathcal{D}_0'$, we have $\mathcal{D}_0 \circ U_{<i}' \circ u = \mathcal{D}_0' \circ U_{<i}' \circ u$. Moreover, the distribution of $\hat{\Theta}_u' \stackrel{\text{def}}{=} \pi(\mathcal{D}_0 \circ U_{<i}' \circ u)$ is independent of $\mathcal{D}_0[\mathrm{ind}_u]$ as the random variable sub-sequence $U_{i<0}'$ was generated on a dataset $\mathcal{D}_0'$ that didn't contain it. Therefore, for $\hat{\Theta}_u \stackrel{\text{def}}{=} \mathcal{R}(\mathcal{D}_0 \circ U_{<i}, u, \hat{\Theta}_{i-1})$ we have

$$\mathrm{R}_q\left(\hat{\Theta}_u \middle\| \hat{\Theta}_u'\right) \leq \mathrm{R}_q\left((\hat{\Theta}_u, U_{<i}) \middle\| (\hat{\Theta}_u', U_{<i}')\right)$$

(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 & J' are joint PDFs of $(\hat{\Theta}_u, U_{<i})$ & $(\hat{\Theta}_u', U_{<i}')$)

$$= \frac{1}{q-1} \log \sum_{u_{<i}} \frac{\mathbb{P}\left[U_{<i} = u_{<i}\right]^q}{\mathbb{P}\left[U_{<i}' = u_{<i}\right]^{q-1}} \left\{\int_\theta \frac{\hat{\Theta}_u|_{U_{<i}=u_{<i}}(\theta)^q}{\hat{\Theta}_u'|_{U_{<i}'=u_{<i}}(\theta)^{q-1}} \mathrm{d}\theta\right\}$$

$$\leq \frac{1}{q-1} \log \sum_{u_{<i}} \frac{\mathbb{P}\left[U_{<i} = u_{<i}\right]^q}{\mathbb{P}\left[U_{<i}' = u_{<i}\right]^{q-1}} \exp((q-1)\varepsilon_{\mathrm{dd}}) \qquad \text{(From (46))}$$

$$= \varepsilon_{\mathrm{dd}} + \mathrm{R}_q\left(U_{<i} \| U_{<i}'\right)$$

$$\leq \varepsilon_{\mathrm{dd}} + \mathrm{R}_q\left(\left(\hat{\Theta}_{s^1}, \cdots, \hat{\Theta}_{s^p}\right) \middle\| \left(\hat{\Theta}_{s^1}', \cdots, \hat{\Theta}_{s^p}'\right)\right)$$

(If $\mathcal{Q}$ sees outputs at steps $s^1, \cdots, s^p$)

$$\leq \varepsilon_{\mathrm{dd}} + p\varepsilon_{\mathrm{dp}}. \qquad \text{(Via Rényi composition)}$$

□

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)

$$\mathrm{d}\Theta_t = -\nabla\mathcal{L}(\Theta_t)\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}\mathbf{Z}_t, \tag{47}$$

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 preliminary knowledge 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 (47), 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 bounded measurable function $f : \mathbb{R}^d \to \mathbb{R}$ to

$$P_t f(\theta_0) = \mathbb{E}\left[f(\Theta_t)|\Theta_0 = \theta_0\right] = \int f(\theta_t)p_t(\theta_0, \theta_t)\mathrm{d}\theta_t. \tag{48}$$

The infentisimal 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

$$\mathcal{G}f = \sigma^2\Delta f - \langle\nabla\mathcal{L}, \nabla f\rangle. \tag{49}$$

This generator $\mathcal{G}$, when applied on a function $f(\theta_t)$, gives the infentisimal change in the value of a function f when θ_t undergoes diffusion as per (47) for $\mathrm{d}t$ time. That is,

$$\partial_t P_t f(\theta_0) = \int \partial_t p_t(\theta_0, \theta_t) f(\theta_t)\mathrm{d}\theta_t = \int p_t(\theta_0, \theta_t)\mathcal{G}f(\theta_t)\mathrm{d}\theta_t. \tag{50}$$

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

$$\int f\mathcal{G}^* g\mathrm{d}\theta = \int g\mathcal{G}f\mathrm{d}\theta, \tag{51}$$

for all real-valued bounded measurable functions $f, g : \mathbb{R}^d \to \mathbb{R}$. Note from (50) that, this operator provides an alternative way to represent the rate of change of function f at time t:

$$\partial_t P_t f(\theta_0) = \int f(\theta_t)\mathcal{G}^* p_t(\theta_0, \theta_t)\mathrm{d}\theta_t. \tag{52}$$

To put it simply, Fokker-Planck operator gives the infentesimal change in the distribution of Θ_t with respect to time. For the Langevin diffusion SDE (47), the Fokker-Planck operator is the following:

$$\partial_t p_t(\theta) = \mathcal{G}^* p_t(\theta) = \mathrm{div}\left(p_t(\theta)\nabla\mathcal{L}(\theta)\right) + \sigma^2\Delta p_t(\theta). \tag{53}$$

From this Fokker-Placnk 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

$$\pi(\theta) \propto e^{-\mathcal{L}(\theta)/\sigma^2}. \tag{54}$$

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

$$\mathop{\mathbb{E}}_{\pi}\left[\mathcal{G}f\right] = \int f\mathcal{G}^*\pi\mathrm{d}\theta = 0. \tag{55}$$

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 infintesimal generator $\mathcal{G}$ as

$$\Gamma(f, g) = \frac{1}{2}\left[\mathcal{G}(fg) - f\mathcal{G}g - g\mathcal{G}f\right], \tag{56}$$

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

$$\Gamma(f, g) = \sigma^2 \langle \nabla f, \nabla g \rangle . \tag{57}$$

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* (57)*) if for all functions $f \in \mathbb{L}^2(\mu)$ with continuous derivatives ∇f,*

$$\mathrm{Ent}_\pi(f^2) \leq \frac{1}{2c} \int \frac{\Gamma(f^2)}{f^2} \pi \mathrm{d}\theta = \frac{2\sigma^2}{c} \int \|\nabla f\|_2^2 \pi \mathrm{d}\theta, \tag{58}$$

where entropy Ent_π is defined as

$$\mathrm{Ent}_\pi(f^2) = \mathop{\mathbb{E}}_{\pi}\left[f^2 \log f^2\right] - \mathop{\mathbb{E}}_{\pi}\left[f^2\right] \log \mathop{\mathbb{E}}_{\pi}\left[f^2\right]. \tag{59}$$

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 18 (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* (54) *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* (57)*), i.e.*

$$\mathrm{Ent}_\rho(f^2) = \frac{2\sigma^2}{\lambda} \int \|\nabla f\|_2^2 \rho \mathrm{d}\theta. \tag{60}$$

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 19 (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 20 (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 proability 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 21 (LS inequality in terms of KL divergence (Vempala & Wibisono, 2019)). *The distribution π satisfies $\mathrm{LS}(c)$ inequality (with respect to Γ in* (57)*) 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 $\nabla \frac{\mu}{\pi}$,*

$$\mathrm{KL}\left(\mu \| \pi\right) \leq \frac{\sigma^2}{2c} \mathrm{I}\left(\mu \| \pi\right). \tag{61}$$

Proof. Set f^2 in (58) to $\frac{\mu}{\pi}$ to obtain (61). Alternatively, set $\mu = \frac{f^2 \pi}{\mathop{\mathbb{E}}_{\pi}[f^2]}$ in (61) to obtain (58). □

Lemma 22 (Wasserstein distance bound under LS inequality (Otto & Villani, 2000, Theorem 1)). *If distribution π satisfies $\mathrm{LS}(c)$ inequality (with respect to Γ in* (57)*) then for all distributions μ on $\mathbb{R}^d$,*

$$\mathrm{W}_2\left(\mu, \pi\right)^2 \leq \frac{2\sigma^2}{c} \mathrm{KL}\left(\mu \| \pi\right). \tag{62}$$

Lemma 23 (LS inequality in terms of Rényi Divergence (Vempala & Wibisono, 2019))**.** *The distribution π satisfies $\mathrm{LS}(c)$ inequality (with respect to Γ in (57)) 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 $q > 1$,*

$$\mathrm{R}_q\left(\mu \| \pi\right) + q(q-1)\partial_q \mathrm{R}_q\left(\mu \| \pi\right) \leq \frac{q^2\sigma^2}{2c} \frac{\mathrm{I}_q\left(\mu \| \pi\right)}{\mathrm{E}_q\left(\mu \| \pi\right)}. \tag{63}$$

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,

$$\mathop{\mathbb{E}}_{\pi}\left[f^2\right] = \mathop{\mathbb{E}}_{\pi}\left[\left(\frac{\mu}{\pi}\right)^q\right] = E(q), \tag{From (36)}$$

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 (36))

Moreover,

$$\mathop{\mathbb{E}}_{\pi}\left[\|\nabla f\|_2^2\right] = \mathop{\mathbb{E}}_{\pi}\left[\left\|\nabla\left(\frac{\mu}{\pi}\right)^{\frac{q}{2}}\right\|_2^2\right] = \frac{q^2}{4} I(q) \tag{From (42)}$$

On substituting (58) with the above equalities, we get:

$$\begin{aligned}
&\mathrm{Ent}_\pi(f^2) \leq \frac{2\sigma^2}{c} \mathop{\mathbb{E}}_{\pi}\left[\|\nabla f\|_2^2\right] \\
\iff & q\partial_q E(q) - E(q)\log E(q) \leq \frac{q^2\sigma^2}{2c} I(q) \\
\iff & q\partial_q \log E(q) - \log E(q) \leq \frac{q^2\sigma^2}{2c} \frac{I(q)}{E(q)} \\
\iff & q\partial_q\left((q-1)R(q)\right) - (q-1)R(q) \leq \frac{q^2\sigma^2}{2c} \frac{I(q)}{E(q)} && \text{(From (36))} \\
\iff & R(q) + q(q-1)\partial_q R(q) \leq \frac{q^2\sigma^2}{2c} \frac{I(q)}{E(q)}
\end{aligned}$$

□

G.3 Proofs for Subsection 5.1

In this appendix, we provide a proof of our Theorem 6 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 $(q, \varepsilon_{\mathrm{dd}})$-data-deletion guarantees on our proposed algorithm pair $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$ (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 $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}}, \mathcal{Q})$ on $\mathcal{D}_0$, we need to show that for all $i \geq 1$ and any $u \in \mathcal{U}^1$, their exists a distribution π_i^u that is independent of the record deleted by u such that

$$\mathrm{R}_q\left(\mathcal{R}(\mathcal{D}_{i-1}, u, \hat{\Theta}_{i-1}) \middle\| \pi_i^u\right) \leq \varepsilon_{\mathrm{dd}}. \tag{64}$$

Without loss in generality, suppose the record that $u = \langle \mathrm{ind}_u, \mathbf{y}_u \rangle$ deletes in database $\mathcal{D}_{i-1}$ existed in the start database $\mathcal{D}_0$ (i.e. $\mathrm{ind}_u \notin u_j$ for all $1 \leq j \leq i-1$). For a neighbouring database $\mathcal{D}_0' \in \mathcal{X}^n$, such that $\mathcal{D}_0 \circ u = \mathcal{D}_0' \circ u$, suppose $(\hat{\Theta}_i')_{i\geq 0}$ is the sequence of models produced by the interaction between $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}}, \mathcal{Q})$ on $\mathcal{D}_0'$. Note that, since $\mathcal{D}_0'$ never contained the record that u is erasing in $\mathcal{D}_{i-1}$, and since the non-adaptive requester $\mathcal{Q}$ is oblivious to the sequence of model releases (which is equivalent to saying that the edit sequence $(u_i)_{i\geq 1}$ is fixed before the interaction occurs), $\hat{\Theta}_{i-1}'$ is independent of the deleted record $\mathcal{D}_{i-1}[\mathrm{ind}_u]$. As such, we define π_i^u to be the distribution of

$\mathcal{R}(\mathcal{D}'_{i-1}, u, \hat{\Theta}'_{i-1})$. So, showing the data-deletion guarantee reduces to proving the following DP-like bound

$$\mathrm{R}_q\left(\mathcal{R}(\mathcal{D}_{i-1}, u, \hat{\Theta}_{i-1})\middle\|\mathcal{R}(\mathcal{D}'_{i-1}, u, \hat{\Theta}'_{i-1})\right) \leq \varepsilon_{\mathrm{dd}}, \tag{65}$$

for all neighbouring $\mathcal{D}_0, \mathcal{D}'_0$ s.t. $\mathcal{D}_0 \circ u = \mathcal{D}'_0 \circ u$, and for all $u_{<i}$ that does not edit index ind_u[4].

Note from our Definition 5.1 that the sequence of models $(\hat{\Theta}_0, \cdots, \hat{\Theta}_{i-1}, \mathcal{R}(\mathcal{D}_{i-1}, u, \hat{\Theta}_{i-1}))$ can be seen as being generated from a continuous run of Noisy-GD, where:

1. for iterations $0 \leq k < K_{\mathcal{A}}$, the loss function is $\mathcal{L}_{\mathcal{D}_0}$,
2. for the iterations $K_{\mathcal{A}} + (j-1)K_{\mathcal{R}} \leq k < K_{\mathcal{A}} + jK_{\mathcal{R}}$ on any $1 \leq j \leq i-1$, the loss function is $\mathcal{L}_{\mathcal{D}_j}$, and
3. for the iterations $K_{\mathcal{A}} + (i-1)K_{\mathcal{R}} \leq k < K_{\mathcal{A}} + iK_{\mathcal{R}}$, the loss function is $\mathcal{L}_{\mathcal{D}_{i-1} \circ u}$.

Let $(\Theta_{\eta k})_{0 \leq k \leq K_{\mathcal{A}} + iK_{\mathcal{R}}}$ 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 $\mathcal{D}'_0$. 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 $0 \leq k < K_{\mathcal{A}} + iK_{\mathcal{R}}$:

$$\begin{cases} \Theta_{\eta(k+1)} = \Theta_{\eta k} - \eta \nabla \mathcal{L}_{\mathcal{D}(k)}(\Theta_{\eta k}) + \sqrt{2\eta\sigma^2}\mathbf{Z}_k \\ \Theta'_{\eta(k+1)} = \Theta'_{\eta k} - \eta \nabla \mathcal{L}_{\mathcal{D}'(k)}(\Theta'_{\eta k}) + \sqrt{2\eta\sigma^2}\mathbf{Z}'_k, \end{cases} \quad \text{where } \mathbf{Z}_k, \mathbf{Z}'_k \sim \mathcal{N}(0, \mathbb{I}_d), \tag{66}$$

and Θ_0 and Θ'_0 are sampled from same the weight initialization distribution ρ. To prove the bound in (65), 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.

$$\begin{cases} \Theta_t = T_k(\Theta_{\eta k}) - \frac{(t-\eta k)}{2}\nabla\left(\mathcal{L}_{\mathcal{D}(k)}(\Theta_{\eta k}) - \mathcal{L}_{\mathcal{D}'(k)}(\Theta_{\eta k})\right) + \sqrt{2\sigma^2}(\mathbf{Z}_t - \mathbf{Z}_{\eta k}), \\ \Theta'_t = T_k(\Theta'_{\eta k}) + \frac{(t-\eta k)}{2}\nabla\left(\mathcal{L}_{\mathcal{D}(k)}(\Theta'_{\eta k}) - \mathcal{L}_{\mathcal{D}'(k)}(\Theta'_{\eta k})\right) + \sqrt{2\sigma^2}(\mathbf{Z}'_t - \mathbf{Z}'_{\eta k}), \end{cases} \tag{67}$$

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

$$T_k = \mathbb{I}_d - \frac{\eta}{2}\nabla\left(\mathcal{L}_{\mathcal{D}(k)} + \mathcal{L}_{\mathcal{D}'(k)}\right). \tag{68}$$

Note that equation (67) is identical to (66) when $t = \eta(k+1)$, and can be expressed by the following stochastic differential equations (SDEs):

$$\begin{cases} \mathrm{d}\Theta_t = -\mathbf{g}_k(\Theta_{\eta k})\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}\mathbf{Z}_t \\ \mathrm{d}\Theta'_t = +\mathbf{g}_k(\Theta'_{\eta k})\mathrm{d}t + \sqrt{2\sigma^2}\mathrm{d}\mathbf{Z}'_t, \end{cases} \quad \text{where } \mathbf{g}_k(\Theta) = \frac{1}{2n}\nabla\left[\boldsymbol{\ell}(\Theta; \mathcal{D}(k)[\mathrm{ind}_u]) - \boldsymbol{\ell}(\Theta; \mathcal{D}'(k)[\mathrm{ind}_u])\right], \tag{69}$$

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.

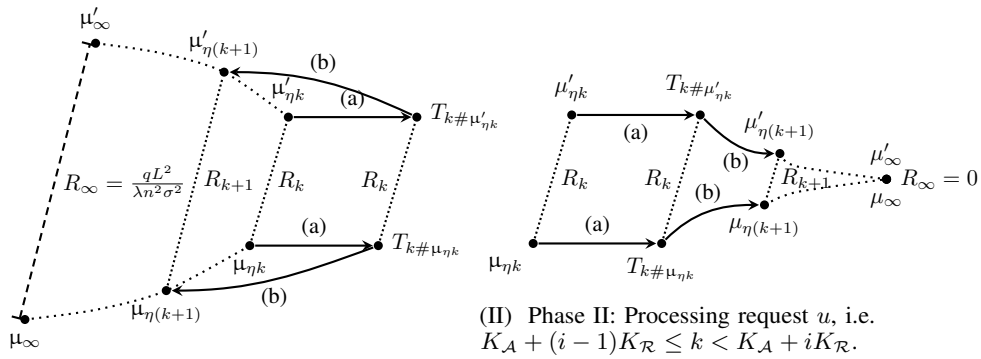
Lemma 24 (Fokker-Planck equation for SDE (69))**.** *Fokker-Planck equation for SDE in* (69) *at time* $\eta k < t \leq \eta(k+1)$*, is*

$$\begin{cases} \partial_t \mu_t(\theta) &= \mathrm{div}\left(\mu_t(\theta)\mathbb{E}\left[\mathbf{g}_k(\Theta_{\eta k})|\Theta_t = \theta\right]\right) + \sigma^2\Delta\mu_t(\theta), \\ \partial_t \mu'_t(\theta) &= \mathrm{div}\left(\mu'_t(\theta)\mathbb{E}\left[-\mathbf{g}_k(\Theta'_{\eta k})\middle|\Theta'_t = \theta\right]\right) + \sigma^2\Delta\mu'_t(\theta), \end{cases} \tag{70}$$

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 (47) with

[4]If a request in u_j for $j < i$ edits the record at index ind_u, we can construct a similar neighbouring database $\mathcal{D}'_j$ after the edit for our analysis.



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

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

Figure 1: Diagram illustrating the technical overview of Theorem 29. 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 (68), 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.

$\nabla\mathcal{L}(\theta) = \mathbf{g}_k(\theta_{\eta k})$ for all $\theta \in \mathbb{R}^d$). Therefore as per (53), 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:

$$\partial_t \mu_{t|\eta k}(\cdot|\theta_{\eta k}) = \mathrm{div}\left(\mu_{t|\eta k}(\cdot|\theta_{\eta k})\mathbf{g}_k(\theta_{\eta k})\right) + \sigma^2\Delta\mu_{t|\eta k}(\cdot|\theta_{\eta k}) \tag{71}$$

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 (65) 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 $K_{\mathcal{A}}$ and $K_{\mathcal{R}}$, and in phase (II) we prove an exponential contraction in the divergence $\mathrm{R}_q\left(\mathcal{R}(\mathcal{D}_{i-1}, u, \hat{\Theta}_{i-1}\middle\|\mathcal{R}(\mathcal{D}'_{i-1}, u, \hat{\Theta}'_{i-1})\right)$ with number of iterations $K_{\mathcal{R}}$.

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 26 and apply the following Lemma from Vempala & Wibisono (2019).

Lemma 25 (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 $q > 0$,*

$$\mathrm{R}_q\left(T(\Theta)\|T(\Theta')\right) = \mathrm{R}_q\left(\Theta\|\Theta\right). \tag{72}$$

Lemma 26. *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* (68) *is:*

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

Proof. **Differentiable bijection.** To see that 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 $\mathcal{L} \stackrel{\text{def}}{=} (\mathcal{L}_{\mathcal{D}(k)} + \mathcal{L}_{\mathcal{D}'(k)})/2$,

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

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

$$\operatorname{prox}_{\mathcal{L}}(\theta) = \arg\min_{\theta'\in\mathbb{R}^d} \frac{\|\theta'-\theta\|_2^2}{2} - \eta\mathcal{L}(\theta'). \tag{73}$$

Note that $\operatorname{prox}_{\mathcal{L}}(\cdot)$ is strongly convex for $\eta < \frac{1}{\lambda+\beta}$. Therefore, from KKT conditions, we have $\theta = \operatorname{prox}_{\mathcal{L}}(\theta) - \eta\nabla\mathcal{L}(\operatorname{prox}_{\mathcal{L}}(\theta)) = T_k(\operatorname{prox}_{\mathcal{L}}(\theta))$. 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

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

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

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

□

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 27 (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 $\max_{\theta\in\mathbb{R}^d} \|V_t(\theta) - V'_t(\theta)\|_2 \le L$ for all $\theta \in \mathbb{R}^d$ and $t \ge 0$. For a diffusion process $(\Theta_t)_{t\ge0}$ and $(\Theta'_t)_{t\ge0}$ defined by the Fokker-Planck equations*

$$\begin{cases} \partial_t \mu_t(\theta) = \operatorname{div}(\mu_t(\theta)V_t(\theta)) + \sigma^2\Delta\mu_t(\theta) & \text{and} \\ \partial_t \mu'_t(\theta) = \operatorname{div}(\mu'_t(\theta)V'_t(\theta)) + \sigma^2\Delta\mu'_t(\theta), \end{cases} \tag{74}$$

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 $t \ge 0$ is upper bounded as

$$\partial_t \mathrm{R}_q(\mu_t\|\mu'_t) \le \frac{qL^2}{2\sigma^2} - \frac{q\sigma^2}{2}\frac{\mathrm{I}_q(\mu_t\|\mu'_t)}{\mathrm{E}_q(\mu_t\|\mu'_t)}. \tag{75}$$

We will apply the above lemma to the Fokker-Planck equation (70) of our pair of tracing diffusion SDE (67) and solve the resulting differential inequality to prove the bound in (65). 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 (63) in Lemma 23.

Lemma 28. *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 $\eta \le \frac{2}{2\lambda+\beta}$, then the tracing diffusion $(\Theta_t)_{0\le t\le\eta(K_{\mathcal{A}}+iK_{\mathcal{R}})}$ and $(\Theta'_t)_{0\le t\le\eta(K_{\mathcal{A}}+iK_{\mathcal{R}})}$ defined in* (67) *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 $0 \le k < K_{\mathcal{A}} + iK_{\mathcal{R}}$ in the extended run of Noisy-GD, and any $0 \le s \le \eta$, let's define two functions $\mathcal{L}_s, \mathcal{L}'_s : \mathbb{R}^d \to \mathbb{R}$ as follows:

$$\mathcal{L}_s = \frac{1+s/\eta}{2}\mathcal{L}_{\mathcal{D}(k)} + \frac{1-s/\eta}{2}\mathcal{L}_{\mathcal{D}'(k)}, \quad \text{and } \mathcal{L}'_s = \frac{1-s/\eta}{2}\mathcal{L}_{\mathcal{D}(k)} + \frac{1+s/\eta}{2}\mathcal{L}_{\mathcal{D}'(k)}. \tag{76}$$

Since $\mathbf{r}(\cdot)$ is the $L2(\lambda)$ regularizer and $\boldsymbol{\ell}(\theta;\mathbf{x})$ is convex and β-smoothness, both $\mathcal{L}_s$ and $\mathcal{L}'_s$ are λ-strongly convex and $(\lambda+\beta)$-smooth for all $0 \le s \le \eta$ and any k. We define maps $T_s, T'_s : \mathbb{R}^d \to \mathbb{R}^d$ as

$$T_s(\theta) = \theta - \eta\nabla\mathcal{L}_s(\theta), \quad \text{and } T'_s(\theta) = \theta - \nabla\mathcal{L}'_s(\theta). \tag{77}$$

From a similar argument as in Lemma 26, both T_s and T'_s are $(1-\eta\lambda)$-Lipschitz for learning rate $\eta \le \frac{2}{2\lambda+\beta}$.

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

$$\mu_t = T_{s\#}(\mu_{\eta k}) \circledast \mathcal{N}\left(0, 2s\sigma^2\mathbb{I}_d\right), \quad \text{and } \mu'_t = T'_{s\#}(\mu'_{\eta k}) \circledast \mathcal{N}\left(0, 2s\sigma^2\mathbb{I}_d\right), \tag{78}$$

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 18 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 (78) 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 18 and Lemma 19, both μ_t, μ'_t satisfy LS inequality with constant

$$\left(\frac{(1-\eta\lambda)^2}{\lambda(1-\eta\lambda/2)} + 2s\right)^{-1} \ge \lambda(1-\eta\lambda/2) \times \underbrace{[(1-\eta\lambda)^2 + \lambda\eta(2-\eta\lambda)]^{-1}}_{=1}, \tag{79}$$

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

We are now ready to prove the data-deletion bound in (65).

Theorem 29 (Data-Deletion guarantee on $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$ under convexity)**.** *Let the weight initialization distribution be $\rho = \mathcal{N}\left(0, \frac{\sigma^2}{\lambda(1-\eta\lambda/2)}\right)$, 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* $(\mathcal{A}, \mathcal{R})$ *satisfies a* $(q, \varepsilon_{\mathrm{dd}})$*-data-deletion guarantee under all non-adaptive* r*-requesters for any noise variance* $\sigma^2 > 0$ *and* $K_{\mathcal{A}} \geq 0$ *if*

$$K_{\mathcal{R}} \geq \frac{2}{\eta\lambda} \log\left(\frac{4qL^2}{\lambda\varepsilon_{\mathrm{dd}}\sigma^2 n^2}\right). \tag{80}$$

Proof. Following the preceding discussion, to prove this theorem, it suffices to show that the inequality (65) holds under the stated conditions. Consider the Fokker-Planck equation described in Lemma 24 for the pair of tracing diffusions SDEs in (69): at any time t in duration $\eta k < t \leq \eta(k+1)$ for any iteration $0 \leq k < K_{\mathcal{A}} + iK_{\mathcal{R}}$,

$$\begin{cases} \partial_t \mu_t(\theta) &= \operatorname{div}\left(\mu_t(\theta)\mathbb{E}\left[\mathbf{g}_k(\Theta_{\eta k})|\Theta_t = \theta\right]\right) + \sigma^2\Delta\mu_t(\theta), \\ \partial_t \mu'_t(\theta) &= \operatorname{div}\left(\mu'_t(\theta)\mathbb{E}\left[-\mathbf{g}_k(\Theta'_{\eta k})\middle|\Theta'_t = \theta\right]\right) + \sigma^2\Delta\mu'_t(\theta), \end{cases} \tag{81}$$

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 $K_{\mathcal{A}} + (i-1)K_{\mathcal{R}} \leq k < K_{\mathcal{A}} + iK_{\mathcal{R}}$ we have $\mathcal{D}(k)[\mathrm{ind}_u] = \mathcal{D}'(k)[\mathrm{ind}_u]$, note from the definition of $\mathbf{g}_k(\theta)$ in (69) that

$$\left\|\mathbb{E}\left[\mathbf{g}_k(\Theta_{\eta k})|\Theta_t = \theta\right] - \mathbb{E}\left[-\mathbf{g}_k(\Theta'_{\eta k})\middle|\Theta'_t = \theta\right]\right\|_2 \leq \begin{cases} \frac{2L}{n} & \text{if } k < K_{\mathcal{A}} + (i-1)K_{\mathcal{R}} \\ 0 & \text{otherwise} \end{cases}. \tag{82}$$

Therefore, applying Lemma 27 to the above pair of Fokker-Planck equations gives that for any t in duration $\eta k < t \leq \eta(k+1)$,

$$\partial_t \mathrm{R}_q\left(\mu_t \| \mu'_t\right) \leq \frac{2qL^2}{\sigma^2 n^2}\mathbb{1}\left\{t \leq \eta(K_{\mathcal{A}} + (i-1)K_{\mathcal{R}})\right\} - \frac{q\sigma^2}{2}\frac{\mathrm{I}_q\left(\mu_t \| \mu'_t\right)}{\mathrm{E}_q\left(\mu_t \| \mu'_t\right)}. \tag{83}$$

Equation (83) suggests a phase change in the dynamics at iteration $k = K_{\mathcal{A}} + (i-1)K_{\mathcal{R}}$. 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 makes $\mathcal{D}_{i-1} \circ u = \mathcal{D}'_{i-1} \circ u$, 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 28, the distribution μ'_t satisfies LS inequality with constant $\lambda(1 - \lambda\eta/2)$. So, we can apply Lemma 23 to simplify the above partial differential inequality as follows.

$$\partial_t R(q,t) + \lambda(1 - \lambda\eta/2)\left(\frac{R(q,t)}{q} + (q-1)\partial_q R(q,t)\right) \leq \frac{2qL^2}{\sigma^2 n^2}\mathbb{1}\left\{t \leq \eta(K_{\mathcal{A}} + (i-1)K_{\mathcal{R}})\right\}. \tag{84}$$

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,

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

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

$$\frac{\mathrm{d}u(s)}{\mathrm{d}s} + c_1 u(s) \leq c_2 \times \mathbb{1}\left\{t \leq \eta(K_{\mathcal{A}} + (i-1)K_{\mathcal{R}})\right\} \tag{85}$$

$$\implies \frac{\mathrm{d}}{\mathrm{d}s}\{e^{c_1 s}u(s)\} \leq c_2 \times \mathbb{1}\left\{t \leq \eta(K_{\mathcal{A}} + (i-1)K_{\mathcal{R}})\right\}. \tag{86}$$

Since the map $T_k(\cdot)$ in (68) is a differentiable bijection for $\eta < \frac{1}{\lambda+\beta}$ as per Lemma 26, note that Lemma 25 implies that $\lim_{s \to \eta k^+} u(s) = u(\eta k)$. Therefore, we can directly integrate in the duration

$0 \leq t \leq \eta(K_{\mathcal{A}} + iK_{\mathcal{R}})$ to get

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

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

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

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

Our next goal in this section is to provide utility guarantees for the algorithm pair $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$ in form of excess empirical risk bounds. For that, we introduce some additional auxiliary results first. The following Lemma 30 shows that excess empirical risks does not increase too much on replacing r records in a database, and Lemma 31 provides a convergence guarantee on the excess empirical risk of Noisy-GD algorithm under convexity.

Lemma 30. *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 modifes no more than r records is bounded as*

$$
\operatorname{err}(\Theta; \mathcal{D} \circ u) \leq \left(1 + \frac{\beta}{\lambda}\right)\left[2\operatorname{err}(\Theta; \mathcal{D}) + \frac{16r^2L^2}{\lambda n^2}\right]. \tag{87}
$$

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 (3). From λ-strong convexity of the $\mathcal{L}_{\mathcal{D}}$,

$$
\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}\circ u}) - \mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}}) \geq \frac{\lambda}{2}\|\theta^*_{\mathcal{D}\circ u} - \theta^*_{\mathcal{D}}\|_2^2. \tag{88}
$$

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

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

Combining the two inequalities give

$$
\|\theta^*_{\mathcal{D}\circ u} - \theta^*_{\mathcal{D}}\|_2 \leq \frac{4rL}{\lambda n}. \tag{89}
$$

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

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

□

Lemma 31 (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*

$$\mathrm{err}(\Theta_{\eta K};\mathcal{D}) \le \mathrm{err}(\Theta_0;\mathcal{D})e^{-\lambda\eta K/2} + \left(1+\frac{\beta}{\lambda}\right)d\sigma^2. \tag{90}$$

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

$$\Theta_{\eta(k+1)} = \Theta_{\eta k} - \eta\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) + \sqrt{2\eta\sigma^2}\mathbf{Z}_k. \tag{91}$$

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

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

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

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta_{\eta(k+1)})\right] \le \mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\right] - \eta\left(1-\frac{\eta(\lambda+\beta)}{2}\right)\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\|_2^2\right] + \eta d\sigma^2(\beta+\lambda). \tag{92}$$

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

$$\|\nabla\mathcal{L}_{\mathcal{D}}(\theta)\|_2^2 \ge 2\lambda(\mathcal{L}_{\mathcal{D}}(\theta)-\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})). \tag{93}$$

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

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

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

$$\mathrm{err}(\Theta_{\eta K};\mathcal{D}) \le \mathrm{err}(\Theta_0;\mathcal{D})e^{-\lambda\eta K/2} + \left(1+\frac{\beta}{\lambda}\right)d\sigma^2. \tag{94}$$

□

Finally, we are ready to prove our main Theorem 6 showing that algorithm pair $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$ solves the data-deletion problem as described in Section 4. We basically combine the RDP guarantee in Theorem 2, non-adaptive data-deletion guarantee in Theorem 29, and prove excess empirical risk bound using Lemma 31 and Lemma 30.

Theorem 6 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda, \beta, L > 0$, $q > 1$, and $0 < \varepsilon_{\text{dd}} \leq \varepsilon_{\text{dp}}$. 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, the regularizer be $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$. If the learning rate be $\eta = \frac{1}{2(\lambda+\beta)}$, the gradient noise variance is $\sigma^2 = \frac{4qL^2}{\lambda\varepsilon_{\text{dp}}n^2}$, 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 $\mathcal{A}_{\text{Noisy-GD}}$ and $\mathcal{R}_{\text{Noisy-GD}}$ are $(q, \varepsilon_{\text{dp}})$-RDP for any $K_{\mathcal{A}}, K_{\mathcal{R}} \geq 0$,*

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

$$\text{if} \quad K_{\mathcal{R}} \geq 4\kappa \log \frac{\varepsilon_{\text{dp}}}{\varepsilon_{\text{dd}}}, \tag{95}$$

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

$$\text{if} \quad K_{\mathcal{A}} \geq 4\kappa \log\left(\frac{\varepsilon_{\text{dp}}n^2}{4qd}\right), \quad \text{and} \quad K_{\mathcal{R}} \geq 4\kappa \log \max\left\{5\kappa, \frac{8\varepsilon_{\text{dp}}r^2}{qd}\right\}. \tag{96}$$

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

(2.) Deletion. By Theorem 29, for the stated choice of loss function, regularizer, learning rate, and weight initialization distribution, the algorithm pair $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$ satisfies $(q, \varepsilon_{\text{dd}})$-data-deletion under all non-adaptive r-requesters $\mathcal{Q}$ if $K_{\mathcal{R}} \geq \frac{2}{\eta\lambda}\log\left(\frac{4qL^2}{\lambda\varepsilon_{\text{dd}}\sigma^2n^2}\right)$. By plugging in $\sigma^2 = \frac{4qL^2}{\lambda\varepsilon_{\text{dp}}n^2}$ and $\eta = \frac{1}{2(\lambda+\beta)}$, this constraint simplifies to $K_{\mathcal{R}} \geq 4\kappa \log \frac{\varepsilon_{\text{dp}}}{\varepsilon_{\text{dd}}}$.

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

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

$$\nabla\mathcal{L}_{\mathcal{D}_0}(\theta^*_{\mathcal{D}_0}) = \frac{1}{n}\sum_{\mathbf{x}\in\mathcal{D}_0}\nabla\boldsymbol{\ell}(\theta^*_{\mathcal{D}_0}; \mathbf{x}) - \lambda\theta^*_{\mathcal{D}_0} = 0 \implies \left\|\theta^*_{\mathcal{D}_0}\right\|_2 \leq \frac{L}{\lambda}. \tag{97}$$

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

$$\begin{aligned}
\text{err}(\Theta_0; \mathcal{D}_0) &= \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_0}(\Theta_0) - \mathcal{L}_{\mathcal{D}_0}(\theta^*_{\mathcal{D}_0})\right] \\
&\leq \frac{(\lambda+\beta)}{2}\mathbb{E}\left[\left\|\Theta_0 - \theta^*_{\mathcal{D}_0}\right\|_2^2\right] && \text{(From } (\lambda+\beta)\text{-smoothness of } \mathcal{L}_{\mathcal{D}_0}) \\
&= \frac{(\lambda+\beta)}{2}\left[\left\|\theta^*_{\mathcal{D}_0}\right\|_2^2 + \mathbb{E}\left[\|\Theta_0\|_2^2\right] - 2\mathbb{E}\left[\left\langle\theta^*_{\mathcal{D}_0}, \Theta_0\right\rangle\right]\right] \\
&\leq \left(1 + \frac{\beta}{\lambda}\right)\left[\frac{L^2}{2\lambda} + \frac{\sigma^2 d}{2-\lambda\eta}\right] && \text{(From (97) and } \mathbb{E}\left[\|\mathbf{Z}\|_2^2\right] = d \text{ if } \mathbf{Z} \sim \mathcal{N}(0, \mathbb{I}_d).) \\
&\leq \kappa\left[\frac{L^2}{2\lambda} + d\sigma^2\right].
\end{aligned}$$

Since $\hat{\Theta}_0 = \mathcal{A}_{\text{Noisy-GD}}(\mathcal{D}_0) = \text{Noisy-GD}(\mathcal{D}_0, \Theta_0, K_{\mathcal{A}})$, by Lemma 31, running $K_{\mathcal{A}} \geq 2\kappa \log\left(\frac{\varepsilon_{\text{dp}} n^2}{4qd}\right)$ iterations gives

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

Induction step: Assume that $\text{err}(\hat{\Theta}_{i-1}; \mathcal{D}_{i-1}) \leq \frac{10\kappa q d L^2}{\lambda \varepsilon_{\text{dp}} n^2}$. Since $\hat{\Theta}_i = \mathcal{R}_{\text{Noisy-GD}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) = \text{Noisy-GD}(\mathcal{D}_i, \hat{\Theta}_{i-1}, K_{\mathcal{R}})$, by Lemma 31 and Lemma 30, running $K_{\mathcal{R}} \geq 2\kappa \log \max\left\{5\kappa, \frac{8r^2}{qd}\right\}$ iterations gives

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

□

G.4 Proofs for Subsection 5.2

In this Appendix, we provide a proof of our data-deletion and utility guarantee in Theorem 7 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 $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}}, \mathcal{Q})$. Our first goal will be to prove $(q, \varepsilon_{\text{dd}})$-data deletion guarantee on $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$ and we will later use it for arguing utility as well. Recall from Definition 4.1 that to prove $(q, \varepsilon_{\text{dd}})$-data-deletion, we need to construct a distribution π_i^u for every edit request $u \in \mathcal{U}^1$ and step $i \geq 1$ such that π_i^u is independent of the original influences of the record deleted by u in $\mathcal{D}_{i-1}$, and show that on running the deletion algorithm to apply u onto $\hat{\Theta}_{i-1}$ we get a model distribution that is indistinguishable to π_i^u, that is $\text{R}_q\left(\mathcal{R}_{\text{Noisy-GD}}(\mathcal{D}_{i-1}, u, \hat{\Theta}_{i-1}) \middle\| \pi_i^u\right) \leq \varepsilon_{\text{dd}}$.

Our construction of π_i^u for this proof is completely different from the one described in Appendix G.3. We make use of Remark 5 that under non-adaptivity, π_i^u can be the output $\pi(\mathcal{D}_{i-1} \circ u)$ of any randomized map $\pi : \mathcal{X}^n \to \mathbb{R}^d$. Our mapping of choice for the purpose is the following Gibbs distribution:

$$\pi(\mathcal{D})(\theta) \propto e^{-\mathcal{L}_{\mathcal{D}}(\theta)/\sigma^2}. \tag{98}$$

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. (47)) 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 (98) that hold for non-convex losses. Our data-deletion 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 convegence argument adapted to our Noisy-GD formulation, with a slighlty 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

$$\Theta_t = \Theta_{\eta k} - (t-\eta k)\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) + \sqrt{2\sigma^2}(\mathbf{Z}_t - \mathbf{Z}_{\eta k}), \tag{99}$$

where $\mathbf{Z}_t$ is a Weiner process. Note that if $\Theta_{\eta k}$ models the parameter distribution after the kth 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 (67)). 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 32 (Proposition 14 (Chewi et al., 2021))**.** *For tracing diffusion Θ_t defined in* (99)*, the equivalent Fokker-Planck equation in the interval $\eta k \leq t \leq \eta(k+1)$ is*

$$\partial_t \mu_t(\theta) = \operatorname{div}\left(\left\{\mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)|\Theta_t = \theta\right] + \sigma^2\nabla\log\frac{\mu_t(\theta)}{\pi(\mathcal{D})(\theta)}\right\}\mu_t(\theta)\right), \tag{100}$$

where $\pi(\mathcal{D})$ is the Gibbs distribution defined in (98)*.*

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.

$$\partial_t \mu_{t|\eta k}(\cdot|\theta_{\eta k}) = \sigma^2\Delta\mu_{t|\eta k}(\cdot|\theta_{\eta k}) + \operatorname{div}\left(\mu_{t|\eta k}(\cdot|\theta_{\eta k})\nabla\mathcal{L}_{\mathcal{D}}(\theta_{\eta k})\right) \tag{101}$$

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 (98). □

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 (100) of Noisy GD's tracing diffusion.

Lemma 33 (Proposition 15 (Chewi et al., 2021))**.** *Let $\rho_t := \mu_t/\pi(\mathcal{D})$ where $\pi(\mathcal{D})$ is the Gibbs distribution defined in* (98) *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*

$$\partial_t\mathrm{R}_q\left(\mu_t\|\pi(\mathcal{D})\right) \leq -\frac{3q\sigma^2}{4}\frac{\mathrm{I}_q\left(\mu_t\|\pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)} + \frac{q}{\sigma^2}\mathbb{E}\left[\psi_t(\Theta_t)\left\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)\right\|_2^2\right]. \tag{102}$$

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

$$\begin{aligned}
\partial_t \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right) &= \frac{q}{(q-1)\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)} \int \left(\frac{\mu_t}{\pi(\mathcal{D})}\right)^{q-1} \partial_t \mu_t \mathrm{d}\theta \\
&= \frac{q}{(q-1)\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)} \int \rho_t^{q-1} \mathrm{div}\left(\left\{\Delta_t + \sigma^2 \nabla \log \rho_t\right\} \mu_t\right) \mathrm{d}\theta \quad \text{(From (100))} \\
&= -\frac{q}{(q-1)\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)} \int \left\langle \nabla\left(\rho_t^{q-1}\right), \Delta_t + \sigma^2 \nabla \log \rho_t \right\rangle \mu_t \mathrm{d}\theta \\
&= -\frac{q}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)} \int \rho_t^{q-2} \left\langle \nabla \rho_t, \Delta_t + \sigma^2 \frac{\nabla \rho_t}{\rho_t} \right\rangle \mu_t \mathrm{d}\theta \\
&= -\frac{q}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)} \left\{ \sigma^2 \mathrm{I}_q\left(\mu_t\|\pi(\mathcal{D})\right) + \frac{2}{q} \underbrace{\mathop{\mathbb{E}}_{\mu_t}\left[\rho_t^{q/2-1}\left\langle \nabla\left(\rho_t^{q/2}\right), \Delta_t \right\rangle\right]}_{\stackrel{\text{def}}{=} F_1} \right\} \\
&\qquad \text{(From (42))}
\end{aligned}$$

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.

$$\begin{aligned}
-F_1 &= \mathop{\mathbb{E}}_{\Theta_t \sim \mu_t}\left[\rho_t^{q/2-1}(\Theta_t)\left\langle \nabla\left(\rho_t^{q/2}\right)(\Theta_t), \mathop{\mathbb{E}}_{\Theta_{\eta k}\sim\mu_{\eta k|t}}\left[\nabla\mathcal{L}_{\mathcal{D}}(\Theta_t) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\right]\right\rangle\right] \\
&= \mathop{\mathbb{E}}_{\mu_{\eta k,t}}\left[\rho_t^{q/2-1}(\Theta_t)\left\langle \nabla\left(\rho_t^{q/2}\right)(\Theta_t), \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\right\rangle\right] \\
&\leq \frac{\sigma^2}{2q}\mathbb{E}\left[\rho_t^{-1}(\Theta_t)\left\|\nabla\left(\rho_t^{q/2}\right)(\Theta_t)\right\|_2^2\right] + \frac{q}{2\sigma^2}\mathbb{E}\left[\rho_t^{q-1}(\Theta_t)\left\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_t) - \nabla\mathcal{L}_{B_k}(\Theta_{\eta k})\right\|_2^2\right] \\
&= \frac{q\sigma^2}{8}\mathrm{I}_q\left(\rho_t\|\mu\right) + \frac{q}{2\sigma^2}\mathbb{E}\left[\rho_t^{q-1}(\Theta_t)\left\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_t) - \nabla\mathcal{L}_{B_k}(\Theta_{\eta k})\right\|_2^2\right] \qquad \text{(From (42))}
\end{aligned}$$

Substituting it in the preceding inequality proves the proposition. □

We need to solve the PDI (102) 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 34 (Change of measure inequality (Chewi et al., 2021))**.** *If $\boldsymbol{\ell}(\theta;\mathbf{x})$ is β-smooth, and regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$, then for any probability density μ on $\mathbb{R}^d$,*

$$\mathop{\mathbb{E}}_{\mu}\left[\|\nabla\mathcal{L}_{\mathcal{D}}\|_2^2\right] \leq 4\sigma^4 \mathop{\mathbb{E}}_{\pi(\mathcal{D})}\left[\left\|\nabla\sqrt{\frac{\mu}{\pi(\mathcal{D})}}\right\|_2^2\right] + 2d\sigma^2(\beta+\lambda), \tag{103}$$

where $\pi(\mathcal{D})$ is the Gibbs distribution defined in (98)*.*

Proof. Consider the Langevin diffusion (47) 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 infintesimal generator $\mathcal{G}$ applied on the $\mathcal{L}_{\mathcal{D}}$ gives

$$\mathcal{G}\mathcal{L}_{\mathcal{D}} = \sigma^2\Delta\mathcal{L}_{\mathcal{D}} - \|\nabla\mathcal{L}_{\mathcal{D}}\|_2^2. \tag{104}$$

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 (104))}\\
&\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 (55))}\\
&= 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 (21))}\\
&= 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}))\\
&= 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]\\
&\qquad\qquad && \text{(From (22) with } a = 2\sigma^2)
\end{aligned}$$

□

Another change in measure inequality needed for the proof is the Donkser-Vardhan variational principle.

Lemma 35 (Donsker-Vardhan 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 $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathop{\mathbb{E}}_{\theta\sim\nu}\left[f(\theta)\right] \le \mathrm{KL}\left(\nu\|\nu'\right) + \log\mathop{\mathbb{E}}_{\theta'\sim\nu'}\left[\exp(f(\theta'))\right]. \tag{105}$$

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 36 (Convergence of Noisy-GD in Rényi divergence)**.** *Let constants $\beta, \lambda, \sigma^2 \ge 0$ and $q, B > 1$. Suppose the loss function $\boldsymbol{\ell}(\theta;\mathbf{x})$ is $(\sigma^2\log(B)/4)$-bounded and β-smooth, and regularizer is $\mathbf{r}(\theta) = \frac{\lambda}{2}\|\theta\|_2^2$. If step size is $\eta \le \frac{\lambda}{64Bq^2(\beta+\lambda)^2}$, 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* (98) *shrinks as follows:*

$$\mathrm{R}_q\left(\mu_{\eta K}\|\pi(\mathcal{D})\right) \le q\exp\left(-\frac{\lambda\eta K}{2B}\right)\mathrm{R}_q\left(\mu_0\|\pi(\mathcal{D})\right) + \frac{32d\eta qB(\beta+\lambda)^2}{\lambda}. \tag{106}$$

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

$$\begin{aligned}
\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)\|_2^2 &\le (\beta+\lambda)^2\|\Theta_{\eta k} - \Theta_t\|_2^2\\
&= (\beta+\lambda)^2\left\|(t-\eta k)\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \sqrt{2(t-\eta k)\sigma^2}\mathbf{Z}_k\right\|_2^2\\
&\qquad\qquad \text{(From (99))}\\
&\le 2\eta^2(\beta+\lambda)^2\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k})\|_2^2 + 4\eta\sigma^2(\beta+\lambda)^2\|\mathbf{Z}_k\|_2^2\\
&\le 4\eta^2(\beta+\lambda)^2\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_{\eta k}) - \nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)\|_2^2\\
&\quad + 4\eta^2(\beta+\lambda)^2\|\nabla\mathcal{L}_{\mathcal{D}}(\Theta_t)\|_2^2 + 4\eta\sigma^2(\beta+\lambda)^2\|\mathbf{Z}_k\|_2^2
\end{aligned}$$

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 (102):

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

Hence, for solving the PDI (102), we have to bound the three expectations F_1 and F_2.

1. **Bounding F_1.** Note that $\mathop{\mathbb{E}}_{\Theta_t\sim\mu_t}\left[\psi_t(\Theta_t)\right] = \int \psi_t(\theta)\mu_t(\theta)\mathrm{d}\theta = \frac{1}{\mathrm{E}_q(\rho_t\|\pi(\mathcal{D}))}\int \frac{\mu_t^q}{\pi(\mathcal{D})^{q-1}}\mathrm{d}\theta = 1$. So, $\psi_t\mu_t(\theta) := \psi_t(\theta)\mu_t(\theta)$ is a probablitly density function on $\mathbb{R}^d$. On applying the measure change Lemma 34 on it, we get

$$\begin{aligned}
F_1 = \mathop{\mathbb{E}}_{\psi_t\mu_t}\left[\left\|\nabla\mathcal{L}_{\mathcal{D}}\right\|_2^2\right] &\leq 4\sigma^4 \mathop{\mathbb{E}}_{\pi(\mathcal{D})}\left[\left\|\nabla\sqrt{\frac{\psi_t\mu_t}{\pi(\mathcal{D})}}\right\|_2^2\right] + 2d\sigma^2(\beta+\lambda) && \text{(From (103))}\\
&= 4\sigma^4 \mathop{\mathbb{E}}_{\pi(\mathcal{D})}\left[\frac{\left\|\nabla\sqrt{\rho_t^q}\right\|_2^2}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)}\right] + 2d\sigma^2(\beta+\lambda)\\
&= \sigma^4 q^2 \frac{\mathrm{I}_q\left(\mu_t\|\pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t\|\pi(\mathcal{D})\right)} + 2d\sigma^2(\beta+\lambda). && \text{(From (42))}
\end{aligned}$$

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-Vardhan principle to get

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

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 $\mathrm{M}_{\chi_d^2}(t) = \mathop{\mathbb{E}}_{X\sim\chi_d^2}\left[\exp(tX)\right] = (1-2t)^{-d/2}$ for $t < \frac{1}{2}$, we can simplify the second term in F_2 as

$$\log \mathop{\mathbb{E}}_{\mu_z}\left[\exp(\|\mathbf{Z}_k\|_2^2/4)\right] = \log \mathrm{M}_{\chi_d^2}\left(\frac{1}{4}\right) = \frac{d\log 2}{2}. \tag{107}$$

The KL divergence term can be simplified as follows.

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

Note that under the assumptions of the Theorem, $\pi(\mathcal{D})$ satisfies log-Sobolev inequality (58) with constant λ/B (i.e. satisfies $\mathrm{LS}(\lambda/B)$). To see this, recall from Lemma 18 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 $\boldsymbol{\ell}(\theta; \mathbf{x})$ is $(\sigma^2 \log(B)/4)$-bounded, the density ratio $\frac{\pi(D)(\theta)}{\rho(\theta)} \in \left[\frac{1}{\sqrt{B}}, \sqrt{B}\right]$. The claim therefore follows from Lemma 20. Using this inequality, from Lemma 21 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

$$8\eta q^2\sigma^4(\beta+\lambda)^2 \frac{\mathrm{I}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t \| \pi(\mathcal{D})\right)} \left(\eta + \frac{2B}{\lambda}\right) \leq \frac{\sigma^4}{2} \frac{\mathrm{I}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t \| \pi(\mathcal{D})\right)}. \tag{108}$$

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 (102), we get

$$\partial_t \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right) \leq -\frac{q\sigma^2}{4} \frac{\mathrm{I}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{\mathrm{E}_q\left(\mu_t \| \pi(\mathcal{D})\right)} + 16\eta d q(\beta+\lambda)^2. \tag{109}$$

Since $\pi(\mathcal{D})$ satisfies $\mathrm{LS}(\lambda/B)$ inequality, from Lemma 23 this PDI reduces to

$$\partial_t \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right) + \frac{\lambda}{2B}\left(\frac{\mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right)}{q} + (q-1)\partial_q \mathrm{R}_q\left(\mu_t \| \pi(\mathcal{D})\right)\right) \leq 16 d\eta q(\beta+\lambda)^2. \tag{110}$$

Let $c_1 = \frac{\lambda}{2B}$ and $c_2 = 16d\eta(\beta+\lambda)^2$. 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

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

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

$$\mathrm{R}_{\bar{q}}\left(\mu_{\eta K}\|\pi(\mathcal{D})\right) \leq \bar{q} \exp\left(-\frac{\lambda \eta K}{2B}\right) \mathrm{R}_{\bar{q}}\left(\mu_0\|\pi(\mathcal{D})\right) + \frac{32 d\eta \bar{q} B(\beta+\lambda)^2}{\lambda}. \tag{111}$$

Finally, noting that for constants $B, q > 1$ and $\beta, \lambda \geq 0$, step size $\eta \leq \min\{\frac{1}{2\sqrt{2}(\beta+\lambda)}, \frac{1}{4(\beta+\lambda)}, \frac{2B}{\lambda}, \frac{\lambda}{64Bq^2(\beta+\lambda)^2}\} = \frac{\lambda}{64Bq^2(\beta+\lambda)^2}$ completes the proof. □

We will use Theorem 36 for proving the data-deletion and utility gaurantee on $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$. 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 37 (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 7, we show that $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$ solves the data-deletion problem described in Section 4 even for non-convex losses. Our proof uses the convergence Theorem 36 and indistinguishability for bounded perturbation Lemma 37 to show that the data-deletion algorithm $\mathcal{R}_{\text{Noisy-GD}}$ can consistently produce models indistinguishable to the corresponding Gibbs distribution (98) in the online setting at a fraction of computation cost of $\mathcal{A}_{\text{Noisy-GD}}$. As discussed in Remark 5, such an indistinguishability is sufficient for ensuring data-deletion for non-adaptive requests. As for adaptive requests, the well-known RDP guarantee of Abadi et al. (2016) combined with our reduction Theorem 5 offers a data-deletion guarantee for $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}})$ under adaptivity.

Our proof of accuracy for the data-deleted models leverages the fact that Gibbs distribution (98) is an almost excess risk minimizer as shown in the following Theorem 38. Since our data-deletion guarantee is based on near-indistinguishability to (98), this property also ensures near-optimal excess risk of data-deleted models.

Theorem 38 (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*

$$\mathrm{err}(\bar{\Theta};\mathcal{D})=\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\bar{\Theta})-\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})\right]\leq\frac{d\sigma^2}{2}\left(\log\frac{\beta+\lambda}{\lambda}+\sqrt{B}\right). \tag{112}$$

Proof. We simplify expected loss as

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\bar{\Theta})\right]=\int\mathcal{L}_{\mathcal{D}}\pi(\mathcal{D})\mathrm{d}\theta=\sigma^2(\mathrm{H}(\pi(\mathcal{D}))-\log(\Lambda_{\mathcal{D}})), \tag{113}$$

where

$$\mathrm{H}(\pi(\mathcal{D}))=-\int\pi(\mathcal{D})\log\pi(\mathcal{D})\mathrm{d}\theta=-\int\frac{e^{-\mathcal{L}_{\mathcal{D}}/\sigma^2}}{\Lambda_{\mathcal{D}}}\log\frac{e^{-\mathcal{L}_{\mathcal{D}}/\sigma^2}}{\Lambda_{\mathcal{D}}}\mathrm{d}\theta \tag{114}$$

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

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

On combining the bounds, we get

$$\mathrm{err}(\bar{\Theta};\mathcal{D})=\mathbb{E}\left[\mathcal{L}_{\mathcal{D}}(\bar{\Theta})-\mathcal{L}_{\mathcal{D}}(\theta^*_{\mathcal{D}})\right]\leq\frac{d\sigma^2}{2}\left(\log\frac{\beta+\lambda}{\lambda}+\sqrt{B}\right). \tag{115}$$

□

Theorem 7 (Accuracy, privacy, deletion, and computation tradeoffs)**.** *Let constants $\lambda,\beta,L,\sigma^2,\eta>0$, $q,B>1$, and $0<\varepsilon_{\mathrm{dd}}\leq\varepsilon_{\mathrm{dp}}<d$. 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 $\mathcal{A}_{\text{Noisy-GD}}$ and $\mathcal{R}_{\text{Noisy-GD}}$ are $(q, \varepsilon_{\text{dp}})$-RDP for any $\eta \geq 0$ and any $K_{\mathcal{A}}, K_{\mathcal{R}} \geq 0$ if*

$$\sigma^2 \geq \frac{qL^2}{\varepsilon_{\text{dp}} n^2} \cdot \eta \max\{K_{\mathcal{A}}, K_{\mathcal{R}}\}, \tag{116}$$

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

$$K_{\mathcal{A}} \geq \frac{2B}{\lambda\eta} \log\left(\frac{q\log(B)}{\varepsilon_{\text{dd}}}\right), \quad K_{\mathcal{R}} \geq K_{\mathcal{A}} - \frac{2B}{\lambda\eta} \log\left(\frac{\log(B)}{2\left(\varepsilon_{\text{dd}} + \frac{r}{n}\log(B)\right)}\right), \tag{117}$$

(3.) *and all models in sequence $(\hat{\Theta}_i)_{i\geq 0}$ output by $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}}, \mathcal{Q})$ on any $\mathcal{D}_0 \in \mathcal{X}^n$, where $\mathcal{Q}$ is an r-requester, satisfy $\text{err}(\hat{\Theta}_i; \mathcal{D}_i) = \tilde{O}\left(\frac{dq}{\varepsilon_{\text{dp}} n^2} + \frac{1}{n}\sqrt{\frac{q\varepsilon_{\text{dd}}}{\varepsilon_{\text{dp}}}}\right)$ when inequalities in* (117) *and* (116) *are equalities.*

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

(2.) Deletion. For showing data-deletion under non-adaptive requests, recall that it is sufficient to show

$$\forall i \geq 1 \text{ and } \forall u \in \mathcal{U}^1, \quad \mathrm{R}_q\left(\mathcal{R}_{\text{Noisy-GD}}(\mathcal{D}_{i-1}, u, \hat{\Theta}_{i-1}) \middle\| \pi(\mathcal{D}_{i-1} \circ u)\right) \leq \varepsilon_{\text{dd}}, \tag{118}$$

where $(\hat{\Theta}_i)_{i\geq 0}$ is the sequence of models generated by the interaction of $(\mathcal{A}_{\text{Noisy-GD}}, \mathcal{R}_{\text{Noisy-GD}}, \mathcal{Q})$ on any database $\mathcal{D}_0 \in \mathcal{X}^n$. For all $i \geq 0$, let $\hat{\mu}_i$ denote the distribution of $\hat{\Theta}_i$. To show (118) holds, we first prove via induction that $\mathrm{R}_q\left(\hat{\mu}_i \| \pi(\mathcal{D}_i)\right) \leq \varepsilon_{\text{dd}}$ for all $i \geq 0$.

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 37 that

$$\mathrm{R}_q\left(\rho \| \pi(\mathcal{D}_0)\right) \leq \frac{2}{\sigma^2} \times \frac{\sigma^2 \log(B)}{4} = \frac{\log(B)}{2}. \tag{119}$$

Under the stated assumptions on loss $\boldsymbol{\ell}(\theta; \mathbf{x})$ and learning rate η, note that the convergence Theorem 36 holds. Since $\hat{\Theta}_0 = \mathcal{A}_{\text{Noisy-GD}}(\mathcal{D}_0) = \text{Noisy-GD}(\mathcal{D}_0, \Theta_0, K_{\mathcal{A}})$, where $\Theta_0 \sim \rho$, we have

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

Induction step: Suppose $\mathrm{R}_q\left(\hat{\mu}_{i-1} \| \pi(\mathcal{D}_{i-1})\right) \leq \varepsilon_{\text{dd}}$. 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 37 we have for all $q > 1$ that

$$\mathrm{R}_q\left(\pi(\mathcal{D}_{i-1}) \| \pi(\mathcal{D}_i)\right) \leq \frac{r\log(B)}{n}. \tag{120}$$

So from the weak triangle inequality Theorem 17 of Rényi divergence,

$$\mathrm{R}_q\left(\hat{\mu}_{i-1} \| \pi(\mathcal{D}_i)\right) \leq \mathrm{R}_q\left(\hat{\mu}_{i-1} \| \pi(\mathcal{D}_{i-1})\right) + \mathrm{R}_\infty\left(\pi(\mathcal{D}_{i-1}) \| \pi(\mathcal{D}_i)\right) \leq \varepsilon_{\text{dd}} + \frac{r\log(B)}{n}. \tag{121}$$

Note that $K_{\mathcal{R}} \geq K_{\mathcal{A}} - \frac{2B}{\lambda\eta}\log\left(\frac{\log(B)}{2\left(\varepsilon_{\mathrm{dd}}+\frac{r}{n}\log(B)\right)}\right) \geq \frac{2B}{\lambda\eta}\log\left(\frac{2q\left(\varepsilon_{\mathrm{dd}}+\frac{r}{n}\log(B)\right)}{\varepsilon_{\mathrm{dd}}}\right)$. Since $\hat{\Theta}_i = \mathcal{R}_{\text{Noisy-GD}}(\mathcal{D}_{i-1}, u_i, \hat{\Theta}_{i-1}) = \text{Noisy-GD}(\mathcal{D}_i, \hat{\Theta}_{i-1}, K_{\mathcal{R}})$, from convergence Theorem 36 we have

$$\begin{aligned}
\mathrm{R}_q\left(\hat{\mu}_i \| \pi(\mathcal{D}_i)\right) &\leq q\exp\left(-\frac{\lambda\eta K_{\mathcal{R}}}{2B}\right)\mathrm{R}_q\left(\hat{\mu}_{i-1}\|\pi(\mathcal{D}_i)\right) + \frac{32d\eta qB(\beta+\lambda)^2}{\lambda} \\
&\leq q\exp\left(-\frac{\lambda\eta K_{\mathcal{R}}}{2B}\right)\left(\varepsilon_{\mathrm{dd}} + \frac{r\log(B)}{n}\right) + \frac{\varepsilon_{\mathrm{dd}}}{2} \\
&\qquad\qquad \text{(From (121) and constraint } \eta \leq \tfrac{\lambda\varepsilon_{\mathrm{dd}}}{64dqB(\beta+\lambda)^2}) \\
&\leq \varepsilon_{\mathrm{dd}}. \qquad\qquad \text{(Since } K_{\mathcal{R}} \geq \tfrac{2B}{\lambda\eta}\log\left(\tfrac{2q\left(\varepsilon_{\mathrm{dd}}+\frac{r}{n}\log(B)\right)}{\varepsilon_{\mathrm{dd}}}\right))
\end{aligned}$$

Hence, by induction, $\mathrm{R}_q\left(\hat{\mu}_i\|\pi(\mathcal{D}_i)\right) \leq \varepsilon_{\mathrm{dd}}$ holds for all $i \geq 0$. Using it, we prove (118). Let u be any edit request in $\mathcal{U}^1$. Note from boundedness of $\boldsymbol{\ell}(\theta;\mathbf{x})$ that for all $i \geq 1$, $|\mathcal{L}_{\mathcal{D}_{i-1}\circ u}(\theta) - \mathcal{L}_{\mathcal{D}_{i-1}}(\theta)| \leq \frac{\sigma^2\log B}{n}$ for all $\theta \in \mathbb{R}^d$. Therefore, from Lemma 37 we have for all $q > 1$ that

$$\mathrm{R}_q\left(\pi(\mathcal{D}_{i-1})\|\pi(\mathcal{D}_{i-1}\circ u)\right) \leq \frac{\log(B)}{n}. \tag{122}$$

So from the weak triangle inequality Theorem 17 of Rényi divergence,

$$\mathrm{R}_q\left(\hat{\mu}_{i-1}\|\pi(\mathcal{D}_{i-1}\circ u)\right) \leq \mathrm{R}_q\left(\hat{\mu}_{i-1}\|\pi(\mathcal{D}_{i-1})\right) + \mathrm{R}_\infty\left(\pi(\mathcal{D}_{i-1})\|\pi(\mathcal{D}_{i-1}\circ u)\right) \leq \varepsilon_{\mathrm{dd}} + \frac{\log(B)}{n}. \tag{123}$$

Since $\mathcal{R}_{\text{Noisy-GD}}(\mathcal{D}_{i-1}, u, \hat{\Theta}_{i-1}) = \text{Noisy-GD}(\mathcal{D}_{i-1}\circ u, \hat{\Theta}_{i-1}, K_{\mathcal{R}})$, from convergence Theorem 36 we have

$$\begin{aligned}
\mathrm{R}_q\left(\mathcal{R}_{\text{Noisy-GD}}(\mathcal{D}_{i-1}, u, \hat{\Theta}_{i-1})\middle\|\pi(\mathcal{D}_{i-1}\circ u)\right) &\leq q\exp\left(-\frac{\lambda\eta K_{\mathcal{R}}}{2B}\right)\mathrm{R}_q\left(\hat{\mu}_{i-1}\|\pi(\mathcal{D}_{i-1}\circ u)\right) \\
&\quad + \frac{32d\eta qB(\beta+\lambda)^2}{\lambda} \\
&\leq q\exp\left(-\frac{\lambda\eta K'}{2B}\right)\left(\varepsilon_{\mathrm{dd}} + \frac{\log(B)}{n}\right) + \frac{\varepsilon_{\mathrm{dd}}}{2} \\
&\qquad \text{(From (123) and constraint } \eta \leq \tfrac{\lambda\varepsilon_{\mathrm{dd}}}{64dqB(\beta+\lambda)^2}) \\
&\leq \varepsilon_{\mathrm{dd}} \quad \text{(Since } K_{\mathcal{A}} \geq \tfrac{2B}{\lambda\eta}\log\left(\tfrac{2q\left(\varepsilon_{\mathrm{dd}}+\frac{r}{n}\log(B)\right)}{\varepsilon_{\mathrm{dd}}}\right))
\end{aligned}$$

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

$$\mathrm{err}(\hat{\Theta}_i;\mathcal{D}_i) = \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_i}(\hat{\Theta}_i) - \mathcal{L}_{\mathcal{D}_i}(\bar{\Theta}_i)\right] + \mathbb{E}\left[\mathcal{L}_{\mathcal{D}_i}(\bar{\Theta}_i) - \mathcal{L}_{\mathcal{D}_i}(\theta^*_{\mathcal{D}_i})\right]. \tag{124}$$

The second term is the suboptimality of Gibbs distribution and by Theorem 38, it is bounded as

$$\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_i}(\bar{\Theta}_i) - \mathcal{L}_{\mathcal{D}_i}(\theta^*_{\mathcal{D}_i})\right] \leq \frac{d\sigma^2}{2}\left(\log\frac{\beta+\lambda}{\lambda} + \sqrt{B}\right). \tag{125}$$

Due to L-Lipschitzness and $(\lambda+\beta)$-smoothness of $\mathcal{L}_{\mathcal{D}_i}$, for any coupling Π of $\hat{\Theta}_i$ and $\bar{\Theta}_i$, the first term is bounded as

$$\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{\mathcal{D}_i}(\hat{\Theta}_i) - \mathcal{L}_{\mathcal{D}_i}(\bar{\Theta}_i)\right] &\leq \mathop{\mathbb{E}}_{\Pi}\left[\left\langle\nabla\mathcal{L}_{\mathcal{D}_i}(\bar{\Theta}_i), \hat{\Theta}_i - \bar{\Theta}_i\right\rangle + \frac{\lambda+\beta}{2}\left\|\hat{\Theta}_i - \Theta_i\right\|_2^2\right] \\
&\leq L\sqrt{\mathop{\mathbb{E}}_{\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right]} + \frac{\lambda+\beta}{2}\mathop{\mathbb{E}}_{\Pi}\left[\left\|\hat{\Theta}_i - \bar{\Theta}_i\right\|_2^2\right]. \\
&\qquad\qquad \text{(From Jensen's inequality)}
\end{aligned}$$

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 22.

$$\inf_{\Pi} \sqrt{\mathop{\mathbb{E}}_{\hat{\Theta}_i, \bar{\Theta}_i \sim \Pi} \left[\left\| \hat{\Theta}_i - \bar{\Theta}_i \right\|_2^2 \right]} = \mathrm{W}_2 \left(\hat{\Theta}_i, \bar{\Theta}_i \right) \leq \sqrt{\frac{2B\sigma^2}{\lambda} \mathrm{KL} \left(\mu_i \| \pi(\mathcal{D}_i) \right)} \leq \sqrt{\frac{2\varepsilon_{\mathrm{dd}} B\sigma^2}{\lambda}}. \tag{126}$$

The last inequality above follows from monotonicity of Rényi divergence in q and the fact that $\lim_{q \to 1} \mathrm{R}_q \left(\nu \| \nu' \right) = \mathrm{KL} \left(\nu \| \nu' \right)$. Therefore, on combining all the bounds we get

$$\mathrm{err}(\hat{\Theta}; \mathcal{D}) \leq L\sigma \sqrt{\frac{2\varepsilon_{\mathrm{dd}} B}{\lambda}} + \frac{\varepsilon_{\mathrm{dd}} B\sigma^2(\lambda + \beta)}{\lambda} + \frac{d\sigma^2}{2} \left(\log \frac{\beta + \lambda}{\lambda} + \sqrt{B} \right) = O \left(\sigma \sqrt{\varepsilon_{\mathrm{dd}}} + d\sigma^2 \right). \tag{127}$$

Note that if the constraints on $K_{\mathcal{A}}$ and $K_{\mathcal{R}}$ in (117) and on σ^2 in (116) are equalities instead, we have

$$\sigma^2 = \frac{2qBL^2}{\lambda \varepsilon_{\mathrm{dp}} n^2} \log \left(\frac{q \log(B)}{\varepsilon_{\mathrm{dd}}} \right) = \tilde{O} \left(\frac{q}{\varepsilon_{\mathrm{dp}} n^2} \right), \tag{128}$$

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

$$\mathrm{err}(\hat{\Theta}; \mathcal{D}) = \tilde{O} \left(\frac{1}{n} \sqrt{\frac{q\varepsilon_{\mathrm{dd}}}{\varepsilon_{\mathrm{dp}}}} + \frac{dq}{\varepsilon_{\mathrm{dp}} n^2} \right). \tag{129}$$

□