Descend or Rewind? Stochastic Gradient Descent Unlearning

Siqiao Mu
Department of Engineering Sciences and Applied Mathematics
Northwestern University
Evanston, IL, 60208
siqiaomu2026@u.northwestern.edu

Diego Klabjan
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL, 60208

Abstract

Machine unlearning algorithms aim to remove the impact of selected training data from a model without the computational expenses of retraining from scratch. Two first-order algorithms are "Descent-to-Delete" (D2D) and "Rewind-to-Delete" (R2D), full-batch gradient descent algorithms that are black-box, easy to implement and satisfy provable unlearning guarantees. In particular, the stochastic version of D2D is widely implemented as the "finetuning" unlearning baseline, despite lacking theoretical backing on nonconvex functions. In this work, we prove (ε, δ) certified unlearning guarantees for *stochastic* R2D and D2D for strongly convex, convex, and nonconvex loss functions, by analyzing unlearning through the lens of disturbed or biased gradient systems, which may be contracting, semi-contracting, or expansive respectively. Our argument relies on coupling the random behavior of the unlearning and retraining trajectories, resulting in a *tail probability bound* on the sensitivity that yields (ε, δ) unlearning. We determine that D2D can yield tighter guarantees for strongly convex functions, but R2D is more appropriate for convex and nonconvex functions. Finally, we compare the algorithms empirically, demonstrating the strengths and weaknesses of each approach.

1 Introduction

Machine unlearning algorithms aim to remove the influence of specific data from a trained model without retraining from scratch. First introduced in Cao and Yang [2015], machine unlearning has attracted significant attention in recent years, driven by heightened concerns over user privacy, data quality, and the energy expenditures of training massive deep learning models such as large language models (LLMs). Regulatory pressures also play a role: provisions in the European Union's General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and the Canadian Consumer Privacy Protection Act (CPPA), protect a user's "right to be forgotten" by requiring that individuals be able to request removal of their personal information, whether stored in databases or retained by models (Sekhari et al. [2021]). Since it is impractical to retrain from scratch every time a user requests removal of their data, we desire cost-saving methods for "forgetting" information from models after training.

To mathematically characterize the extent of unlearning, a rich line of work investigates (ε, δ)-certified unlearning algorithms, which are theoretically guaranteed to return model weights that are

Preprint.

probabilistically indistinguishable to the weights obtained from retraining from scratch, defined in Guo et al. [2020], Ginart et al. [2019]. This is typically achieved by bounding the distance between the weights after unlearning and after retraining, and injecting appropriately calibrated Gaussian noise, a technique from differential privacy (DP) known as the Gaussian mechanism (Dwork and Roth [2014]). Such provable guarantees can be a powerful alternative to empirical metrics of unlearning, which can be unreliable or misleading, as in Tu et al. [2025], Zhang et al. [2025].

However, many existing certified unlearning algorithms have limited practicality for modern deep learning settings. The vast majority are either second-order methods that require access to the Hessian (or Hessian vector products) as in Zhang et al. [2024], Qiao et al. [2025], Sekhari et al. [2021], Suriyakumar and Wilson [2022], Guo et al. [2020] or first-order methods that require computing the full gradient, as in Neel et al. [2021], Chien et al. [2024a], Mu and Klabjan [2026]. Both settings are computationally intractable for large-scale models, which are typically trained with *stochastic gradient descent* (SGD) algorithms. While two recent works are able to show certified unlearning with stochastic gradients, they either prove a weaker version of certified unlearning, as in Koloskova et al. [2025], or only apply to convex functions as in Chien et al. [2024b]. To reflect realistic practices, we desire SGD-based unlearning algorithms for *nonconvex* functions.

With this goal in mind, we revisit two existing first-order certified unlearning algorithms: Descent-to-Delete (D2D), Neel et al. [2021], designed for (strongly) convex loss functions, and Rewind-to-Delete (R2D), Mu and Klabjan [2026], designed for nonconvex functions. Both algorithms perform learning via full-batch gradient descent (GD). Upon unlearning, both perform additional GD steps on the loss function of the retained dataset, but D2D "descends" from the final trained model, whereas R2D "rewinds" to an earlier saved checkpoint during training before performing the unlearning steps. Notably, although D2D is only theoretically supported for GD on strongly convex functions, its stochastic version is the basis for the "finetuning" unlearning baseline method, which is implemented in *virtually all* unlearning papers and is shown to underperform on deep neural networks. Mu and Klabjan [2026] argue that rewinding, instead of descending, is a more appropriate baseline method for nonconvex settings, but their analysis does not apply to stochastic gradients. Therefore, whether the stochastic versions of these methods can provably unlearn is an important open question.

In this work, we prove that SGD versions of D2D and R2D, denoted as SGD-D2D and SGD-R2D respectively, do achieve (ε, δ) certified unlearning, and we derive their privacy-utility-complexity tradeoffs under looser assumptions than the original works. We first examine SGD-R2D with projection, for which we can achieve clean guarantees with minimal assumptions on strongly convex, convex, and nonconvex functions. On an unbounded domain, we can also prove certified unlearning for SGD-R2D for strongly convex, convex, and nonconvex functions, assuming that the second moment of the noise is relatively bounded and the loss is finite at initialization. Finally, we prove (ε, δ) certified unlearning for SGD-D2D on strongly convex functions.

To analyze SGD-R2D, we leverage the properties of gradient systems to track the divergence between the training and retraining trajectories. By characterizing SGD on the original loss function as *biased* or *disturbed* SGD on the loss function of the retained data samples, we can achieve noise bounds that reflect the contractive, semi-contractive, and expansive properties of strongly convex, convex, and nonconvex functions respectively. For SGD-D2D, we leverage additional favorable properties of biased SGD on strongly convex functions that allow the bias to be "folded into" the classic convergence analysis, as long as the proportion of unlearned data is small enough. This approach is completely different from the original proof in Neel et al. [2021], which considers a highly constrained setting where the function is both strongly convex and Lipschitz continuous.

Our approach also relies on a key tail bound argument; by controlling the coupling between the two randomized trajectories, we achieve a *tail bound* on the sensitivity, with respect to this joint distribution. For projected SGD, we leverage Hoeffding's inequality to achieve a tight tail bound that scales advantageously with δ. For the unconstrained case, we can use Markov's inequality in combination with a first or second moment bound, leading to a weaker dependence on δ. This novel technique allows a more flexible analysis of SGD algorithms, for which deterministic sensitivity bounds, required by the classic Gaussian mechanism, can be challenging to obtain.

Our work reveals deeper insights into the difference between "rewinding" and "descending." On strongly convex functions, SGD-D2D yields tighter probabilistic bounds than unconstrained SGD-R2D. However, SGD-D2D is not guaranteed to be more efficient than retraining from scratch if the initial point is sufficiently close to the global minimum. In contrast, SGD-R2D is *always* more

Table 1: Comparison of first-order algorithms for certified unlearning.

Algorithm	Method	Noise	Loss Function
D2D (Neel et al. [2021])	GD w/ regularization	At the end	Convex
R2D (Mu and Klabjan [2026])	GD	At the end	Nonconvex
Langevin Unlearning (Chien et al. [2024a])	Projected GD	At every step	Nonconvex
Langevin SGD (Chien et al. [2024b])	Projected SGD	At every step	Convex
PSGD-R2D (Theorem 4.3)	Projected SGD	At the end	Nonconvex
SGD-R2D (Theorem 4.7)	SGD	At the end	Nonconvex
SGD-D2D (Theorem 4.8)	SGD	At the end	Strongly convex

efficient than retraining from scratch. In fact, for strongly convex functions and constant noise, the number of unlearning iterations K is *better than sublinear* in T, the number of training iterations, as K converges to a constant for arbitrarily large T.

We establish privacy-utility-complexity tradeoffs for SGD-R2D and SGD-D2D, which are easy to implement and are "black-box," in that they only require noise injection at the end and do not require special algorithmic procedures during training. In addition, we perform experiments demonstrating the strengths of each approach, and illustrating that R2D can indeed circumvent the downsides of D2D in nonconvex settings, such as stalling in a stationary point. Our contributions are as follows,

- We prove (ε, δ) certified unlearning for SGD-R2D with and without projection on strongly convex, convex, and nonconvex loss functions.
- We prove (ε, δ) certified unlearning for SGD-D2D on strongly convex functions, using a novel proof approach that circumvents the limiting assumptions of the original D2D work.
- We conduct experiments comparing SGD-R2D and SGD-D2D on real-world datasets, confirming that rewinding is more appropriate for nonconvex settings.

Code is open-sourced at the anonymous GitHub repository `https://anonymous.4open.science/r/r2d2-3753/`.

2 Related Work

Certified unlearning. The term machine unlearning was first coined in Cao and Yang [2015] to address deterministic data deletion. The works Ginart et al. [2019], Guo et al. [2020] introduce a probabilistic notion of unlearning that uses differential privacy to theoretically certify the level of deletion. While there exists many second-order certified unlearning algorithms, as in Zhang et al. [2024], Qiao et al. [2025], Sekhari et al. [2021], Suriyakumar and Wilson [2022], Basaran et al. [2025], that require computation of the Hessian or Hessian vector-products, this work focuses on first-order methods (Table 1), which only require access to the gradient and are far more tractable for large-scale problems in terms of computation and storage. Existing first-order methods, detailed in Table 1, include full gradient methods as well as two SGD methods which we now discuss in detail. For additional discussion of related works, see Appendix C.

First, Chien et al. [2024b] analyzes noisy projected stochastic gradient descent and obtains theoretical guarantees on strongly convex and convex functions, by leveraging the uniqueness of the limiting distribution of the training process. However, their algorithm does not apply to nonconvex functions. Second, Koloskova et al. [2025] proposes performing SGD with clipping, regularization, and noise on the retain set during unlearning, which advantageously does not require the function to be Lipschitz smooth or convex. However, their algorithm only satisfies a *weaker* "post-processing" definition of unlearning, which ensures that the model obtained from unlearning one subset is indistinguishable from the model obtained from unlearning another, as in Allouah et al. [2025], Sekhari et al. [2021], Basaran et al. [2025]. In fact, since the analysis in Koloskova et al. [2025] relies on privacy amplification by iteration (PABI), it likely cannot be extended to obtain the stronger result in this paper, which achieves indistinguishability between the unlearned and retrained models. Both Chien et al. [2024b] and Koloskova et al. [2025] require noise injection at every iteration. In contrast, the D2D and R2D frameworks are "black-box" in that they only require noise once after training and once after unlearning. Therefore, they can be applied to models trained without any special procedures.

3 Algorithm

Let $\mathcal{D} = \{z_1, ..., z_n\}$ be a training dataset of n data points drawn from the data distribution $\mathcal{Z}$. Let $A : \mathcal{Z}^n \to \mathbb{R}^d$ be a randomized learning algorithm that trains on $\mathcal{D}$ and outputs a model with weight parameters $\theta \in \mathbb{R}^d$. Typically, the goal of a learning algorithm is to minimize $\mathcal{L}_\mathcal{D}(\theta)$, the empirical loss on $\mathcal{D}$, defined as $\mathcal{L}_\mathcal{D}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(z_i;\theta)$, where $\ell(z_i;\theta)$ represents the loss on the sample z_i. A standard approach for minimizing $\mathcal{L}_\mathcal{D}(\theta)$ is gradient descent, where the iterates are updated at time t as $\theta_t = \theta_{t-1} - \eta\nabla\mathcal{L}_\mathcal{D}(\theta_{t-1})$. However, computing the full gradient can be computationally intractable for large models and datasets, motivating SGD algorithms where at each iteration t we uniformly sample $\mathcal{B}_t$, a mini-batch of size b *with replacement* from $\mathcal{D}$ and construct a random gradient estimator $g_{\mathcal{B}_t}(\theta) = \frac{1}{b}\sum_{i\in\mathcal{B}_t}\nabla\ell(z_i;\theta)$, where $\mathbb{E}[g_{\mathcal{B}_t}(\theta)] = \nabla\mathcal{L}_\mathcal{D}(\theta)$. Vanilla SGD algorithms update θ with this gradient estimator:

$$\theta_t = \theta_{t-1} - \eta g_{\mathcal{B}_t}(\theta_{t-1}). \tag{1}$$

Finally, we project the iterates onto a nonempty and convex set $\mathcal{C} \subseteq \mathbb{R}^d$. When $\mathcal{C} = \mathbb{R}^d$, this is vanilla SGD, and when $\mathcal{C}$ is closed, this is projected SGD (PSGD). We update θ as follows,

$$\theta_t = \Pi_\mathcal{C}(\theta_{t-1} - \eta g_{\mathcal{B}_t}(\theta_{t-1})), \tag{2}$$

where $\Pi_\mathcal{C}(x)$ denotes the (unique) projection of θ onto $\mathcal{C}$, $\Pi_\mathcal{C}(x) = \arg\min_{x'\in\mathcal{C}}\|x - x'\|$.

Let $Z \subset \mathcal{D}$ denote a subset of size m we would like to unlearn, which we call the unlearned set, and let $\mathcal{D}' = \mathcal{D}\backslash Z$ denote the retained set. Certified unlearning produces an unlearning algorithm U that removes the influence of Z from the output of the learning algorithm $A(\mathcal{D})$, such that the model parameters obtained from $U(A(\mathcal{D}), \mathcal{D}, Z)$ are probabilistically *indistinguishable* from the output of $A(\mathcal{D}')$. In particular, such algorithmic outputs are (ε, δ)-indistinguishable, a concept originating in differential privacy.

Throughout, for a fixed dataset $\mathcal{D}$ and unlearning set Z, all probabilities and expectations are taken with respect to the internal randomness of the learning and unlearning algorithms (mini-batch index sampling and additive Gaussian noise). We assume that these random choices are generated from a fixed distribution that does not depend on the data values in $\mathcal{D}$.

Definition 3.1. (Dwork and Roth [2014], Neel et al. [2021]) Let X and Y be outputs of randomized algorithms. We say X and Y are (ε, δ)-indistinguishable if for all $S \subseteq \Omega$, we have

$$\mathbb{P}[X \in S] \leq e^\varepsilon \mathbb{P}[Y \in S] + \delta,$$
$$\mathbb{P}[Y \in S] \leq e^\varepsilon \mathbb{P}[X \in S] + \delta.$$

Definition 3.2. Let A be a randomized learning algorithm and U a randomized unlearning algorithm. Then U is an (ε, δ) certified unlearning algorithm for A if for all $Z \subset \mathcal{D}$, $U(A(\mathcal{D}), \mathcal{D}, Z)$ and $A(\mathcal{D}\backslash Z)$ are (ε, δ)-indistinguishable with respect to the algorithmic randomization.

Now we describe the R2D and D2D unlearning frameworks. For both, the learning algorithm A involves T training iterations on $\mathcal{L}_\mathcal{D}$, and the unlearning algorithm U involves K iterations on $\mathcal{L}_{\mathcal{D}'}$. The difference lies in the initialization of U: for R2D, U is initialized at the $T - K$th iterate of the learning trajectory, whereas for D2D, U is initialized at the Tth or last iterate of the learning trajectory. Finally, Gaussian noise is added at the end to achieve (ε, δ)-indistinguishability. These frameworks are easily combined with various gradient methods, including SGD and PSGD.

Algorithms 3 and 4 represent the learning and unlearning algorithms for SGD-D2D, and Algorithms 1 and 2 represent the learning and unlearning algorithms for SGD-D2D. When $\mathcal{C}$ is a closed set, we denote the algorithm as PSGD-R2D.

Our analysis relies on comparing the training iterates with the retraining iterates, denoted as $\theta'_t = \Pi_\mathcal{C}(\theta'_{t-1} - \eta g_{\mathcal{B}'_t}(\theta'_{t-1}))$, where $\theta'_0 = \theta_0$ and $\mathcal{B}'_t$ is sampled from $\mathcal{D}'$. Specifically, we have that $\{\theta_t\}_{t=0}^T$ and $\{\theta'_t\}_{t=0}^T$ represent the learning iterates on $\mathcal{L}_\mathcal{D}$ and $\mathcal{L}_{\mathcal{D}'}$ respectively, both starting from $\theta'_0 = \theta_0$. In addition, $\{\theta''_t\}_{t=0}^K$ represents the unlearning iterates on $\mathcal{L}_{\mathcal{D}'}$, where for D2D we have $\theta''_0 = \theta_T$ and for R2D we have $\theta''_0 = \theta_{T-K}$. The goal is to bound the distance between θ'_T and θ''_K.

Algorithm 1 A: SGD-R2D

Require: dataset $\mathcal{D}$, initial point θ_0, domain $\mathcal{C}$
  for $t = 1, 2, ..., T$ **do**
    Uniformly sample with replacement $\mathcal{B}_t \sim \mathcal{D}$
    $\theta_t = \Pi_{\mathcal{C}}(\theta_{t-1} - \eta g_{\mathcal{B}_t}(\theta_{t-1}))$
  end for
  Save checkpoint θ_{T-K}
  Use $\tilde{\theta} = \theta_T + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2)$, for inference
  Upon unlearning, execute Algorithm 2

Algorithm 2 U: SGD-R2D

Require: dataset $\mathcal{D}'$, model checkpoint θ_{T-K}, domain $\mathcal{C}$
  $\theta''_0 = \theta_{T-K}$
  for $t = 1, ..., K$ **do**
    Uniformly sample with replacement $\mathcal{B}'_t \sim \mathcal{D}'$
    $\theta''_t = \Pi_{\mathcal{C}}(\theta''_{t-1} - \eta g_{\mathcal{B}'_t}(\theta''_{t-1}))$
  end for
  Use $\tilde{\theta} = \theta''_K + \xi'$, where $\xi' \sim \mathcal{N}(0, \sigma^2)$, for inference

Algorithm 3 A: SGD-D2D

Require: dataset $\mathcal{D}$, initial point θ_0
  for $t = 1, 2, ..., T$ **do**
    Uniformly sample with replacement $\mathcal{B}_t \sim \mathcal{D}$
    $\theta_t = \theta_{t-1} - \eta g_{\mathcal{B}_t}(\theta_{t-1})$
  end for
  Use $\tilde{\theta} = \theta_T + \xi$, where $\xi \sim \mathcal{N}(0, \sigma^2)$, for inference
  Upon unlearning, execute Algorithm 4.

Algorithm 4 U: SGD-D2D

Require: dataset $\mathcal{D}'$, model checkpoint θ_T
  $\theta''_0 = \theta_T$
  for $t = 1, ..., K$ **do**
    Uniformly sample with replacement $\mathcal{B}'_t \sim \mathcal{D}'$
    $\theta''_t = \theta''_{t-1} - \eta g_{\mathcal{B}'_t}(\theta''_{t-1})$
  end for
  Use $\tilde{\theta} = \theta''_K + \xi'$, where $\xi' \sim \mathcal{N}(0, \sigma^2)$, for inference

4 Analyses

For all proofs, we assume that the loss function ℓ is differentiable and Lipschitz smooth (Assumption 4.1), standard requirements for analyses of SGD algorithms.

Assumption 4.1. For all $z \in \mathcal{Z}$, the function $\ell(z; \theta)$ is differentiable in θ and Lipschitz smooth in θ with constant $L > 0$ such that for any $\theta_1, \theta_2 \in \mathbb{R}^d$, $\|\nabla \ell(z; \theta_1) - \nabla \ell(z; \theta_2)\| \leq L\|\theta_1 - \theta_2\|$.

As highlighted in the introduction, our analysis relies on a key coupling argument leading to a tail bound that holds with respect to this coupling. We couple the randomization (the "coin flips") of the learning and unlearning algorithms to minimize the distances between θ_t and θ'_t, as well as θ'_t and θ''_t, by choosing the batches sampled at each time step so that they coincide as much as possible. This is a standard technique in differential privacy, as in Abadi et al. [2016]. However, prior works evaluate the deviation on the full datasets $\mathcal{D}$ and $\mathcal{D}'$ and combine privacy amplification by subsampling with noise at every step to achieve DP. In contrast, to maintain the output perturbation structure (which is amenable to black-box unlearning), we evaluate the sensitivity over the coupling over all steps, obtaining a tail bound on the sensitivity $\|\theta'_T - \theta''_K\|$ that holds over the joint distribution of the runs, with probability $1 - \delta$. This produces an $(\varepsilon, 2\delta)$ indistinguishability guarantee.

Lemma 4.2. *Let x and y be random variables over some domain Ω, and let ξ, ξ' be draws from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Then for $0 < \varepsilon \leq 1$ and $\delta > 0$, we have the following.*

1. *In general, if the tail bound exists such that $\mathbb{P}[\| x - y \| \geq \Sigma] \leq \delta$, then $X = x + \xi$, $Y = y + \xi'$, are $(\epsilon, 2\delta)$-indistinguishable if $\sigma = \frac{\Sigma\sqrt{2\log(1.25/\delta)}}{\varepsilon}$.*
2. *If the second moment of the distance between x and y is bounded as $\mathbb{E}[\|x - y\|^2] \leq \Sigma^2$, then $X = x + \xi$, $Y = y + \xi'$ are $(\varepsilon, 2\delta)$-indistinguishable if $\sigma = \frac{\Sigma}{\varepsilon}\sqrt{\frac{2\log(1.25/\delta)}{\delta}}$.*
3. *If the first moment of the distance between x and y is bounded as $\mathbb{E}[\|x - y\|] \leq \Sigma$, then $X = x + \xi$, $Y = y + \xi'$ are $(\varepsilon, 2\delta)$-indistinguishable if $\sigma = \frac{\Sigma\sqrt{2\log(1.25/\delta)}}{\varepsilon\delta}$.*

Proof. See Appendix A.3.

Lemma 4.2 preserves differential privacy because the algorithmic randomness (for SGD, the mini-batch indices sampled uniformly from $\{1, ..., n\}$) does not depend on the actual data values. Like classic DP methods, we require that the bound Σ also does not encode information about the specific dataset, thereby preventing data leakage. We emphasize that our approach differs from $(\varepsilon, \delta, \gamma)$-*random differential privacy* (Hall et al. [2013], Rubinstein and Aldà [2017]), which considers whether (ε, δ)-privacy holds with probability $1 - \gamma$ *with respect to the sampling of the dataset* from a well-behaved data distribution.

Broadly speaking, our analysis of the R2D framework relies on characterizing unlearning as *biased* or *disturbed* SGD on a gradient system. The SGD iterates on $\mathcal{L}_{\mathcal{D}}$ are biased SGD iterates on $\mathcal{L}_{\mathcal{D}'}$, where the *unlearning bias* is the difference between the gradients on $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_{\mathcal{D}'}$. Upon unlearning, we perform *unbiased* SGD on $\mathcal{L}_{\mathcal{D}'}$. Our results on strongly convex, convex, and nonconvex functions align with nonlinear contraction theory established in Kozachkov et al. [2023], Sontag [2022] that predicts that θ_t and θ'_t will stay close for perturbed contracting systems (strongly convex functions), diverge linearly on perturbed semi-contracting systems (convex functions), and diverge exponentially otherwise (nonconvex functions). Like in Mu and Klabjan [2026], rewinding reverses the accumulation of these disturbances, drawing the unlearning and retraining trajectories close. Moreover, rewinding erases the impact of noise during unlearning through tight coupling.

In contrast, the analysis of D2D relies on additional tricks available for the analysis of biased gradient descent on strongly convex functions. By relying on the existence and attractivity of the global minimum of strongly convex functions, we can leverage the faster convergence of biased SGD algorithms to achieve a tighter second-moment bound that has a better dependence on δ.

4.1 PSGD-R2D: Rewinding with Projected SGD

We first consider projected SGD (PSGD) algorithms, for which we can achieve clean unlearning guarantees with minimal assumptions and reasonable dependence on δ. In particular, in this setting the gradient of the loss function is uniformly bounded over $\mathcal{C}$. This allows us to leverage Hoeffding's inequality for sums of bounded independent random variables to achieve a tail bound that yields a favorable dependence on δ.

Theorem 4.3. *Suppose that the loss function ℓ satisfies Assumptions 4.1, and within the closed, bounded, and convex set $\mathcal{C} \subset \mathbb{R}^d$, the gradient of ℓ is uniformly bounded by some constant $G \geq 0$ such that for all $z \in \mathcal{Z}$ and $\theta \in \mathcal{C}$, $\|\nabla \ell(z; \theta)\| \leq G$. We implement PSGD-R2D (Algorithms 1 and 2) with $\sigma = \frac{\Sigma\sqrt{2\log(1.25/\delta)}}{\varepsilon}$. Then for $0 < \varepsilon \leq 1$ and $0 < \delta \leq \frac{1}{2}$, PSGD-R2D is an $(\varepsilon, 2\delta)$-unlearning algorithm for the following choices of Σ.*

- *For general functions, we have*

$$\Sigma = G\eta\sqrt{\frac{2((1+\eta L)^{2T} - (1+\eta L)^{2K})\log(1/\delta)}{(1+\eta L)^2 - 1}} + \frac{2Gm((1+\eta L)^T - (1+\eta L)^K)}{nL}.$$

- *If ℓ is convex, then for $\eta \leq \frac{2}{L}$ we have*

$$\Sigma = G\eta\sqrt{2(T-K)\log(1/\delta)} + \frac{2G\eta m}{n}(T-K).$$

- *If ℓ is μ-strongly convex (5), then for $\eta \leq \frac{\mu}{L^2}$ and $\gamma = \sqrt{1-\eta\mu}$ we have*

$$\Sigma = G\eta\sqrt{\frac{2(\gamma^{2K} - \gamma^{2T})\log(1/\delta)}{1-\gamma^2}} + \frac{2G\eta m(\gamma^K - \gamma^T)}{n(1-\gamma)}.$$

Proof. See Appendix A.5.

Theorem 4.3 highlights the privacy-utility-complexity tradeoff of PSGD-R2D, through the relationship between Σ, ε and K. We observe that for all settings, Σ decreases to zero as K increases from zero to T, and it decays exponentially for strongly convex functions and linearly for convex functions. For general nonconvex functions, Σ still decreases to zero as K goes to T, but at a slower rate. This reflects the contractive, semi-contractive, and expansive properties of gradient algorithms on strongly

convex, convex, and nonconvex functions respectively. We note that we require η small enough for convex and strongly convex functions in order for contraction to occur on these systems, but no conditions on η for general nonconvex functions, where trajectories diverge exponentially regardless of η.

For strongly convex functions, we achieve K that is *better than sublinear* in T for constant noise; in fact, it converges to a constant for large T, implying that $T - K$, the computational advantage of unlearning, is potentially infinite as T goes to infinity.

Corollary 4.4. *If ℓ is μ-strongly convex, then for any constant Σ we have that K converges to a constant as $T \to \infty$.*

Proof. See Appendix A.5.1.

This implies that for a given level of privacy (ε, δ) and noise, the number of unlearning iterations required is uniformly upper bounded even if the number of training iterations T is arbitrarily large.

4.2 SGD-R2D: Rewinding with SGD

Now we consider rewinding without projection. A key challenge of analyzing SGD unlearning algorithms on an unbounded domain is that the noise and unlearning bias may also be unbounded. We therefore require a reasonable assumption that the second moment of the stochastic gradient is relatively bounded, as in Assumption 4.5.

Assumption 4.5. For all datasets $\tilde{\mathcal{D}} \sim \mathcal{Z}$, let $g_{\mathcal{B}}(\theta)$ represent the gradient estimator constructed from a batch $\mathcal{B} \sim \tilde{\mathcal{D}}$ uniformly sampled with replacement. Then the second moment of the gradient estimator is relatively bounded with constants $B, C \geq 0$ such that for all $\theta \in \mathbb{R}^d$,

$$\mathbb{E}[\|g_{\mathcal{B}}(\theta)\|^2] \leq B\|\nabla \mathcal{L}_{\tilde{\mathcal{D}}}(\theta)\|^2 + C. \tag{3}$$

For finite-sum problems as considered in this work, Assumption 4.5 is satisfied by strongly convex loss functions and functions that satisfy the Polyak–Łojasiewicz (PL) inequality, as established in Khaled and Richtárik [2023], Garrigos and Gower [2024], Gower et al. [2021]. In particular, for batch sampling with replacement, the constant B can be *explicitly* computed from L and the batch size b, and the constant C can be estimated from how close the model is to interpolation, where perfect interpolation (or overparametrization) implies $C = 0$.

The unlearning bias can be bounded by the χ^2-divergence between the empirical distributions of $\mathcal{D}$ and $\mathcal{D}'$, denoted as $P_{\mathcal{D}}$ and $P_{\mathcal{D}'}$ respectively (Lemma A.17 in the Appendix). This conveniently enables the use of existing conditions on the noise like Assumption 4.5.

Finally, when we leverage the above rationale, we note that the accumulated bias and noise effects during training can be upper bounded by the loss at initialization. To achieve a dataset-independent result, we require the following assumption that the loss is finite at initialization for all $z \in \mathcal{Z}$.

Assumption 4.6. For all $z \in \mathcal{Z}$ and $\theta \in \mathbb{R}^d$, $\ell(z; \theta) \geq 0$. Moreover, the loss function is bounded at initialization θ_0 by some constant ℓ_{θ_0}, such that for all $z \in \mathcal{Z}$, we have $\ell(z; \theta_0) \leq \ell_{\theta_0}$.

Assumption 4.6 is satisfied if the underlying support of the data distribution $\mathcal{Z}$ is bounded. With the above stipulations, we are able to achieve SGD-R2D unlearning on unbounded domains as follows.

Theorem 4.7. *Suppose that the loss function ℓ satisfies Assumptions 4.1, 4.5, and 4.6 and we implement SGD-R2D (Algorithms 1 and 2) with $\mathcal{C} = \mathbb{R}^d$ and $\sigma = \frac{\Sigma\sqrt{2\log(1.25/\delta)}}{\varepsilon\delta}$. Then for $0 < \varepsilon \leq 1$ and $\delta > 0$, SGD-R2D is an $(\varepsilon, 2\delta)$-unlearning algorithm for the following Σ.*

- *For general functions, we have*
$$\Sigma = O(((1 + \eta L)^T - (1 + \eta L)^K)(T - K)^{1/2}).$$
- *If ℓ is convex, then for $\eta \leq \frac{2}{L}$ we have*
$$\Sigma = O((T - K)^{3/2}).$$
- *If ℓ is μ-strongly convex (5), then for $\eta \leq \frac{\mu}{L^2}$ and $\gamma = \sqrt{1 - \eta\mu}$ we have*
$$\Sigma = O((\gamma^K - \gamma^T)(T - K)^{1/2}).$$

For all statements, the $O(\cdot)$ notation hides dependencies on B, C, η, L, ℓ_{θ_0}, m, n, and μ.

Proof. See Appendix A.6.

For the unbounded setting, we rely on a weaker first-moment bound on $\|\theta'_T - \theta''_K\|$ with Lemma 4.2, leading to a weaker dependence on δ compared to the projected setting.

4.3 D2D: Descending with SGD

Finally, we prove that SGD-D2D achieves certified unlearning. During training, the biased SGD iterates $\{\theta_t\}$ converge to a neighborhood of $\theta^{*'}$, the global minimum of $\mathcal{L}_{\mathcal{D}'}$, and upon unlearning, the unbiased SGD iterates $\{\theta''_t\}$ converge closer to $\theta^{*'}$. Our approach differs from the original D2D proof, where Lipschitz *continuity* and strong convexity are combined to show that $\theta^{*'}$ is close to the global minimum of $\mathcal{L}_{\mathcal{D}}$, a property that does not hold outside of this highly constrained setting.

Theorem 4.8. *Suppose that ℓ is μ-strongly convex and satisfies Assumptions 4.1, 4.5, and 4.6. Let $\eta \leq \frac{1}{BL}$, $\frac{m}{n} < \frac{1}{6B+1}$, and $T = K + \frac{\log(\ell_{\theta_0}) - \log(\frac{5C}{4B\mu})}{\log(\frac{1}{1-\eta\mu/2})}$. Then for $0 < \varepsilon \leq 1$ and $\delta > 0$, SGD-D2D (Algorithms 3 and 4) is an $(\varepsilon, 2\delta)$-certified unlearning algorithm with $\sigma = \frac{\Sigma}{\varepsilon}\sqrt{\frac{2\log(1.25/\delta)}{\delta}}$, where*

$$\Sigma^2 = \frac{5C}{\mu^2 B}((1 - \frac{\eta\mu}{2})^{2K} + 2(1 - \frac{\eta\mu}{2})^K) + \frac{4L\eta C}{\mu^2}.$$

Proof. See Appendix A.7.

Theorem 4.8 shows that SGD-D2D can achieve certified unlearning as long as the proportion of unlearned data is small enough, which allows the unlearning bias to be "folded into" the standard unbiased SGD analysis, yielding linear convergence to the global minimum. The noise decays *exponentially* with the number of unlearning iterates K. This mirrors the original D2D result in Neel et al. [2021], but the stochastic setting yields an $O(\eta)$ constant term at the end which cannot be eliminated. Moreover, without the bounded gradient/Lipschitz smoothness assumption of Neel et al. [2021], the computation advantage of unlearning, $T - K$, depends on the initialization; in fact, if θ_0 happens to be close enough to the optimum of $\mathcal{L}_{\mathcal{D}'}$, then unlearning may not be advantageous at all. We note that while SGD-D2D can achieve a better dependence on δ than SGD-R2D without projection, it is impossible to prove that it achieves unlearning on nonconvex functions.

5 Experiments

To reflect practical unlearning implementations, we conduct experiments to investigate the impact of noiseless ($\sigma = 0$) rewinding vs. unlearning. We adopt the setup in Mu and Klabjan [2026] applied to noiseless ($\sigma = 0$) PSGD instead of GD. We train a binary classifier with SGD and the cross-entropy loss function, and we unlearn a subset of the data from the trained model. For small-scale experiments, we use the eICU dataset in Pollard et al. [2018], a large multi-center intensive care unit (ICU) tabular dataset, and we train a multilayer perceptron (MLP) with three hidden layers to predict whether patients stay in the hospital for longer or shorter than a week, using their intake data. For large-scale experiments, we consider the Lacuna-100 dataset used in Golatkar et al. [2020], constructed from the VGGFace2 dataset (Cao et al. [2018]) and the MAAD-Face annotations (Terhörst et al. [2020]) as described in Mu and Klabjan [2026], and we train a ResNet-18 model (He et al. [2016]) to perform binary gender classification. In contrast to prior work, our experiments focus on examining the difference between descending vs. rewinding. We considering varying K, or number of unlearning iterations, and we evaluate the effects of D2D and R2D at the same level of computation and batch size. Following the precedent in Zhang et al. [2024], we implement PSGD on a ball of radius R to maintain the strictness of our results, while choosing R large enough to minimize impact on model utility. For additional details, see Appendix B.

We evaluate unlearning efficacy using three metrics. First, we compute the L2 distance in parameter space between unlearning and retraining, where better unlearning minimizes this distance. Second, we consider the model performance on the unlearned dataset $Z = \mathcal{D}_{unlearn}$ before and after unlearning. A decrease in performance suggests that the model is losing information about the unlearned samples. Finally, we apply membership inference attacks (MIAs) which attempt to distinguish between

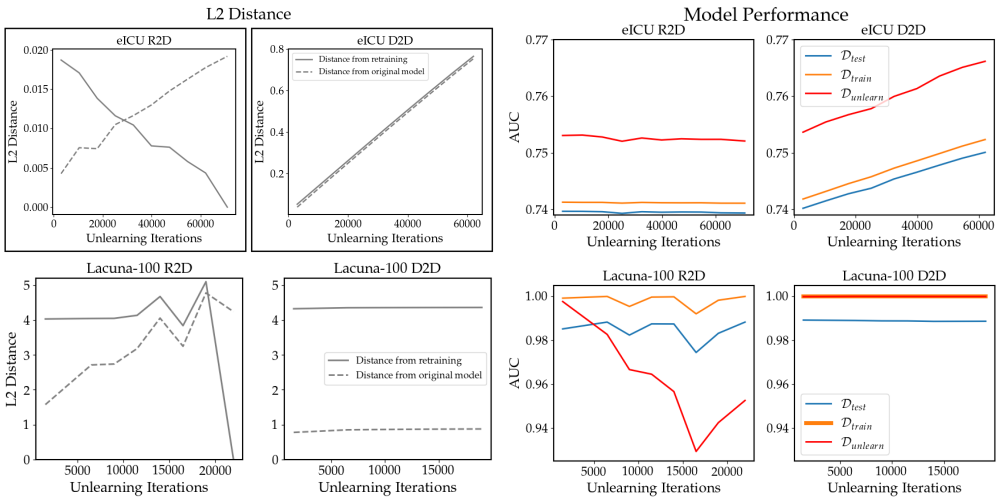


Figure 1: Comparison of unlearning capabilities of noiseless ($\sigma = 0$) PSGD-R2D and PSGD-D2D. The eICU results are in the top row and the Lacuna-100 results are in the bottom row. The left two columns display the L2 distance from unlearning and the original model as the number of unlearning iterations increases. The right two columns display the model performance on the unlearned, retained, and test sets, denoted as $\mathcal{D}_{unlearn}$, $\mathcal{D}_{train}$, and $\mathcal{D}_{test}$ respectively.

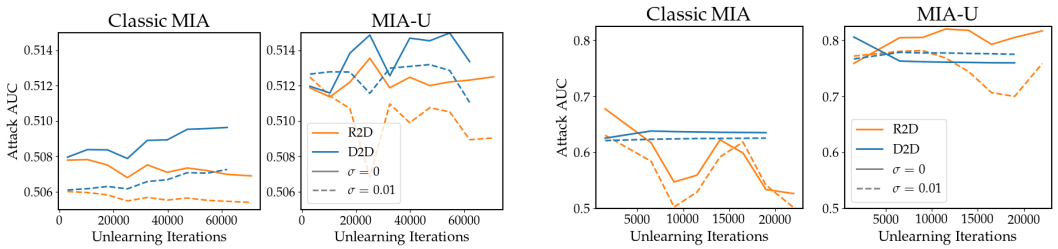
unlearned data and data that has never been in the training dataset. A less successful attack, indicated by a lower Area Under the Receiver Operating Characteristic Curve (AUC), indicates that unlearning is more successful. We utilize both the classic MIA from Shokri et al. [2017] that considers the model outputs after unlearning as well as the more sophisticated unlearning MIA (MIA-U) from Chen et al. [2021] which compares the model outputs before and after unlearning.

Figure 1 compares the first two unlearning metrics, L2 distance and model performance, for PSGD-R2D and PSGD-D2D. Our two experimental settings illustrate varying practical outcomes of the unlearning methods. For eICU, PSGD-R2D has a reliable outcome: the unlearned model moves away from the original model and towards the retrained model. This is reflected in the L2 distance as well as the model performance on $\mathcal{D}_{unlearn}$, which decreases with increased rewinding even as the performance on the retained and test sets remain constant. In contrast, PSGD-D2D causes the model to progress away from both the original and retrained models. The model performance on all data, including $\mathcal{D}_{unlearn}$, improves, suggesting that the optimization algorithm has found a new descent direction on the new loss function $\mathcal{L}_{\mathcal{D}'}$. For Lacuna-100, rewinding still reliably moves the model away from the original trained model, but the L2 distance from the retrained model does not change significantly. This may be because in the stochastic setting, PSGD-R2D is only theoretically guaranteed to reduce the distance in expectation. However, PSGD-R2D does still reduce the model performance on $\mathcal{D}_{unlearn}$ while keeping the performance on other sets constant. In contrast, PSGD-D2D displays very little movement in parameter space or in model performance, suggesting that it is stalled at a stationary point at initialization. Moreover, a large number of descent steps risks overfitting, as shown in the decrease in performance on $\mathcal{D}_{test}$.

Figures 2a and 2b display the results of the MIAs. While the output of the MIAs can be highly nonlinear (especially for non-i.i.d. data), in general more unlearning decreases the attack success, especially for rewinding. Moreover, adding a small amount of noise tends to reduce the success of the attack. Ultimately, we find that descent-based algorithms (including D2D, Chien et al. [2024b], and Koloskova et al. [2023]) can either improve the model performance on all sets, which may obfuscate whether unlearning has occurred, or it may stall at a stationary point. On the other hand, R2D has a more reliable unlearning effect, but may not get any performance boost from unlearning.

6 Conclusion

In this work, we prove certified unlearning guarantees for SGD-R2D and SGD-D2D. Experiments demonstrate the practical effects of each method.



(a) Comparison of MIA success for R2D and D2D on the eICU dataset.

(b) Comparison of MIA success for R2D and D2D on the Lacuna-100 dataset.

Figure 2: Comparison of MIA success for R2D and D2D models across different datasets.

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*, CCS '16, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Ahmad Ajalloeian and Sebastian U. Stich. Analysis of SGD with biased gradient estimators. *CoRR*, abs/2008.00051, 2020. URL https://arxiv.org/abs/2008.00051.

Youssef Allouah, Joshua Kazdan, Rachid Guerraoui, and Sanmi Koyejo. The utility and complexity of in- and out-of-distribution machine unlearning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HVFMooKrHX.

Umit Yigit Basaran, Sk Miraj Ahmed, Amit Roy-Chowdhury, and Basak Guler. A certified unlearning approach without access to source data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=8lt5776GLB.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, 2015. doi: 10.1109/SP.2015.35.

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*, CCS '21, page 896–911, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384544.

Eli Chien, Chao Pan, and Olgica Milenkovic. Certified graph unlearning. In *NeurIPS 2022 Workshop: New Frontiers in Graph Learning*, 2022. URL https://openreview.net/forum?id=wCxlGc9ZCwi.

Eli Chien, Haoyu Peter Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=3LKuC8rbyV.

Eli Chien, Haoyu Peter Wang, Ziang Chen, and Pan Li. Certified machine unlearning via noisy stochastic gradient descent. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=h3k2NXu5bJ.

Rishav Chourasia and Neil Shah. Forget unlearning: Towards true data-deletion in machine learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6028–6073. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chourasia23a.html.

Yury Demidovich, Grigory Malinovsky, Igor Sokolov, and Peter Richtarik. A guide through the zoo of biased SGD. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 23158–23171. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/484d254ff80e99d543159440a06db0de-Paper-Conference.pdf.

Derek Driggs, Jingwei Liang, and Carola-Bibiane SchÃ¶nlieb. On biased stochastic gradient estimation. *Journal of Machine Learning Research*, 23(24):1–43, 2022. URL http://jmlr.org/papers/v23/20-316.html.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, 8 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Yann Fraboni, Martin Van Waerebeke, Kevin Scaman, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. SIFU: Sequential informed federated unlearning for efficient and provable client unlearning in federated optimization. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3457–3465. PMLR, 5 2024. URL https://proceedings.mlr.press/v238/fraboni24a.html.

Guillaume Garrigos and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods, 2024. URL https://arxiv.org/abs/2301.11235.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/cb79f8fa58b91d3af6c9c991f63962d3-Paper.pdf.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Robert Gower, Othmane Sebbouh, and Nicolas Loizou. SGD for structured nonconvex functions: Learning rates, minibatching and interpolation. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1315–1323. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/gower21a.html.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3832–3842. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/guo20c.html.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi Malvajerdi, and Christopher Waites. Adaptive machine unlearning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Goz-qsH1F14.

Robert Hall, Larry Wasserman, and Alessandro Rinaldo. Random differential privacy. *Journal of Privacy and Confidentiality*, 4(2), 3 2013. doi: 10.29012/jpc.v4i2.621. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/621.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Xiaowei Hu, Prashanth L.A., András György, and Csaba Szepesvari. (bandit) convex optimization with biased noisy gradient oracles. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 819–828, Cadiz, Spain, 5 2016. PMLR. URL https://proceedings.mlr.press/v51/hu16b.html.

Yaxi Hu, Bernhard Schölkopf, and Amartya Sanyal. Online learning and unlearning, 2025. URL https://arxiv.org/abs/2505.08557.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2008–2016. PMLR, 2021. URL https://proceedings.mlr.press/v130/izzo21a.html.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In Paolo Frasconi, Niels Landwehr, Giuseppe Manco, and Jilles Vreeken, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham, 2016. Springer International Publishing.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=AU4qHN2VkS. Survey Certification.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17343–17363. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/koloskova23a.html.

Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. Certified unlearning for neural networks. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=3rWQlV3s1I.

Leo Kozachkov, Patrick Wensing, and Jean-Jacques Slotine. Generalization as dynamical robustness–the role of Riemannian contraction in supervised learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=Sb6p5mcefw.

Jiaqi Liu, Jian Lou, Zhan Qin, and Kui Ren. Certified minimax unlearning with generalization rates and deletion capacity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=6H8Md75kAw.

Siqiao Mu and Diego Klabjan. Rewind-to-delete: Certified machine unlearning for nonconvex functions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=FgjcLXIUjr.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 931–962. PMLR, 2021. URL https://proceedings.mlr.press/v132/neel21a.html.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178, September 2018. ISSN 2052-4463.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

Xinbao Qiao, Meng Zhang, Ming Tang, and Ermin Wei. Hessian-free online certified unlearning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=C3TrHWanh5.

Benjamin I. P. Rubinstein and Francesco Aldà. Pain-free random differential privacy with sensitivity sampling. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2950–2959. PMLR, 8 2017. URL https://proceedings.mlr.press/v70/rubinstein17a.html.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=pvCLqcsLJ1N.

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)*, pages 3–18, 2017.

Eduardo D. Sontag. Remarks on input to state stability of perturbed gradient flows, motivated by model-free feedback control learning. *Systems & Control Letters*, 161:105138, 2022. ISSN 0167-6911. doi: https://doi.org/10.1016/j.sysconle.2022.105138. URL https://www.sciencedirect.com/science/article/pii/S0167691122000056.

Vinith Menon Suriyakumar and Ashia Camage Wilson. Algorithms that approximate data removal: New results and limitations. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=G4VOQPYxBsI.

Philipp Terhörst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Maad-face: A massively annotated attribute dataset for face images. *CoRR*, abs/2012.01030, 2020. URL https://arxiv.org/abs/2012.01030.

Yiwen Tu, Pingbang Hu, and Jiaqi W. Ma. A reliable cryptographic framework for empirical machine unlearning evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=TYoYJStuN9.

Enayat Ullah and Raman Arora. From adaptive query release to machine unlearning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34642–34667. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ullah23a.html.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A. Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 4126–4142. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/ullah21a.html.

Xi Wu, Fengan Li, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 1307–1322, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450341974. doi: 10.1145/3035918.3064047. URL https://doi.org/10.1145/3035918.3064047.

Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards certified unlearning for deep neural networks. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=1mf1ISuyS3.

Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private ERM for smooth objectives. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3922–3928, 2017. doi: 10.24963/ijcai.2017/548. URL https://doi.org/10.24963/ijcai.2017/548.

Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramer. Position: Membership inference attacks cannot prove that a model was trained on your data. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 333–345, Los Alamitos, CA, USA, April

2025. IEEE Computer Society. URL https://doi.ieeecomputersociety.org/10.1109/SaTML64287.2025.00025.

A Proofs

A.1 Overview

In Appendix A.2, we list some helper lemmas that are standard results in probability and optimization. In Appendix A.3, we prove the theorems governing the first or second moment sensitivity bounds, which are used for all later proofs. In Appendix A.4, we establish the contractive, semi-contractive, and expansive properties of gradient algorithms on strongly convex, convex, and nonconvex functions respectively. These results, which reflect well-known behaviors of gradient systems, are essential for proving unlearning for PSGD-R2D in Appendix A.5 and SGD-R2D in Appendix A.6. Finally, in Appendix A.7, we prove unlearning for SGD-D2D, under the same unbounded domain conditions of the analysis of SGD-R2D in the previous section.

A.2 Helper Lemmas and Definitions

In addition to general nonconvex functions, we analyze the cases of convex and strongly convex functions, defined below in Definitions A.1 and A.2.

Definition A.1. For all $z \in \mathcal{Z}$, the function $\ell(z;\theta)$ is convex in θ such that for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\ell(z;\theta_1) \geq \ell(z;\theta_2) + \langle \nabla \ell(z;\theta_2), \theta_1 - \theta_2 \rangle. \tag{4}$$

Definition A.2. For all $z \in \mathcal{Z}$, the function $\ell(z;\theta)$ is μ-strongly convex such that for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\ell(z;\theta_1) \geq \ell(z;\theta_2) + \nabla \ell(z;\theta_2)^T(\theta_1 - \theta_2) + \frac{\mu}{2}\|\theta_2 - \theta_1\|^2. \tag{5}$$

The following lemmas are standard results in probability and optimization which we utilize to obtain the results in this paper.

Lemma A.3. *Let $\mathcal{C} \subset \mathbb{R}^d$ be a nonempty, closed, and convex set. For each $\theta \in \mathbb{R}^d$, let $\Pi_{\mathcal{C}}(\theta)$ denote the (unique) projection of θ onto $\mathcal{C}$ such that $\Pi_{\mathcal{C}}(x) = \arg\min_{x' \in \mathcal{C}} \|x - x'\|$.. Then the projection is nonexpansive such that for all $\theta, \theta' \in \mathbb{R}^d$,*

$$\|\Pi_{\mathcal{C}}(\theta) - \Pi_{\mathcal{C}}(\theta')\| \leq \|\theta - \theta'\|. \tag{6}$$

Lemma A.4. *For constants $a \geq 0$, $b \geq 0$, we have*

$$\sqrt{a+b} \leq \sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)},$$
$$\sqrt{a} + \sqrt{b} + \sqrt{c} \leq \sqrt{3(a+b+c)}.$$

Proof. We can use the AM-GM inequality as follows.

$$(\sqrt{a} + \sqrt{b})^2 = a + 2\sqrt{ab} + b \leq 2(a+b).$$

□

Lemma A.5. *Suppose X, Y are independent random variables over the same domain and $\mathbb{E}[X] = 0$. Then*

$$\mathbb{E}[\|X - Y\|^2] = \mathbb{E}[\|X\|^2] + \mathbb{E}[\|Y\|^2].$$

Lemma A.6. *Suppose for all $z \in \mathcal{D}$, $\ell(z;\theta)$ is L-Lipschitz smooth for all $\theta \in \mathbb{R}^d$. Then $\mathcal{L}_{\mathcal{D}}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(z;\theta)$ is also L-Lipschitz smooth in θ.*

Lemma A.7. *(Tower property of expectation) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\mathcal{F}_t\}_{t \in T}$ a filtration with $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$ for $s \leq t$. If the random variable X is integrable ($\mathbb{E}[|X|] < \infty$), then for all $s \leq t$:*

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{F}_t] \mid \mathcal{F}_s] = \mathbb{E}[X \mid \mathcal{F}_s] \quad \text{a.s.}$$

The following lemmas pertain to functions that satisfy the Polyak–Łojasiewicz (PL) inequality. PL functions can be nonconvex, and μ-strongly convex functions are also μ-PL.

Definition A.8. Let $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and denote $\mathcal{L}^\star := \inf_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta)$. We say that $\mathcal{L}$ satisfies the *Polyak–Łojasiewicz (PL) inequality* with parameter $\mu > 0$ if

$$\frac{1}{2}\|\nabla \mathcal{L}(\theta)\|^2 \;\geq\; \mu\big(\mathcal{L}(\theta) - \mathcal{L}^\star\big) \qquad \text{for all } \theta \in \mathbb{R}^d. \tag{7}$$

Lemma A.9. *(Karimi et al. [2016]) Suppose a function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ satisfies the PL inequality (Definition A.8) with parameter $\mu > 0$. Let $\mathcal{X} = \arg\min_{\theta \in \Theta} \mathcal{L}(\theta)$ represent the solution set of $\mathcal{L}$, and let $\theta^* = Proj_{\mathcal{X}}(\theta) = \arg\min_{x \in \mathcal{X}} \|x - \theta\|$. Then $\mathcal{L}$ satisfies the quadratic growth condition such that*

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \geq \frac{\mu}{2}\|\theta^* - \theta\|^2 \qquad \text{for all } \theta \in \mathbb{R}^d. \tag{8}$$

Lemma A.10. *Let $\mathcal{L}(\theta)$ be a L-Lipschitz smooth, μ-PL function. Then $\mu \leq L$.*

Proof. Let θ^* represent the global minima of $\mathcal{L}$. Then by Lipschitz smoothness and the fact that $\nabla\mathcal{L}(\theta^*) = 0$, we have for any $\theta \in \mathbb{R}^d$,

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \leq \frac{L}{2}\|\theta - \theta^*\|^2.$$

Then by quadratic growth (8),

$$\begin{aligned} \frac{\mu}{2}\|\theta - \theta^*\|^2 &\leq \frac{L}{2}\|\theta - \theta^*\|^2, \\ \mu &\leq L. \end{aligned}$$

□

A.3 Proof of Lemma 4.2

We prove the first part of Lemma 4.2. Then the second and third part follow through a Markov bound.

Lemma A.11. *Let x and y be random variables over some domain Ω, and suppose*

$$\mathbb{P}[\|x - y\| \geq \Sigma] \leq \delta.$$

Let ξ, ξ' be draws from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Then $X = x + \xi$, $Y = y + \xi'$ are $(\varepsilon, 2\delta)$-indistinguishable if

$$\sigma = \frac{\Sigma\sqrt{2\log(1.25/\delta)}}{\varepsilon}.$$

Proof. Let E represent the event that $\|x - y\| \leq \Sigma$, such that $\mathbb{P}[E] > 1 - \delta$ and $\mathbb{P}[\neg E] \leq \delta$. Then when E occurs, the standard Gaussian mechanism holds with sensitivity Σ, and we have for $\sigma = \frac{\Sigma\sqrt{2\log(1.25/\delta)}}{\varepsilon}$,

$$\mathbb{P}[X \in S|E] \leq e^{\varepsilon}\mathbb{P}[Y \in S|E] + \delta.$$

Without loss of generality, we have for $S \subset \Omega$,

$$\begin{aligned} \mathbb{P}[X \in S] =& \mathbb{P}[X \in S \mid E]\mathbb{P}[E] + \mathbb{P}[X \in S \mid \neg E]\mathbb{P}[\neg E] \\ \leq& \mathbb{P}[X \in S \mid E]\mathbb{P}[E] + \delta \\ \leq& (e^{\varepsilon}\mathbb{P}[Y \in S \mid E] + \delta)\mathbb{P}[E] + \delta \\ \leq& e^{\varepsilon}\mathbb{P}[Y \in S \mid E]\mathbb{P}[E] + \delta + \delta \\ =& e^{\varepsilon}\mathbb{P}[E \mid Y \in S]\mathbb{P}[Y \in S] + 2\delta \\ \leq& e^{\varepsilon}\mathbb{P}[Y \in S] + 2\delta. \end{aligned}$$

□

We note that independence is not required for Lemma A.11 to hold, as the bounds are true for arbitrary joint distribution between x, y, X and Y.

Lemma A.12. *Let x and y be random variables over some domain Ω, and suppose*

$$\mathbb{E}[\|x - y\|^2] \leq \Sigma^2.$$

Let ξ, ξ' be draws from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Then $X = x + \xi$, $Y = y + \xi'$ are $(\varepsilon, 2\delta)$-indistinguishable if

$$\sigma = \frac{\Sigma}{\varepsilon}\sqrt{\frac{2\log(1.25/\delta)}{\delta}}.$$

Proof. By Markov's inequality, for any $t > 0$ we have the tail bound

$$\mathbb{P}[\|x - y\| > t] \leq \frac{\Sigma^2}{t^2},$$

so for $0 < \delta < 1$,

$$\mathbb{P}[\|x - y\| > \frac{\Sigma}{\sqrt{\delta}}] \leq \delta.$$

□

Lemma A.13. *Let x and y be random variables over some domain Ω, and suppose*

$$\mathbb{E}[\|x - y\|] \leq \Sigma.$$

Let ξ, ξ' be draws from the Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Then $X = x + \xi$, $Y = y + \xi'$ are $(\varepsilon, 2\delta)$-indistinguishable if

$$\sigma = \frac{\Sigma\sqrt{2\log(1.25/\delta)}}{\varepsilon\delta}.$$

Proof. By Markov's inequality, for any $t > 0$ we obtain the tail bound

$$\mathbb{P}[\|x - y\| > t] \leq \frac{\Sigma}{t},$$

so for $0 < \delta < 1$,

$$\mathbb{P}[\|x - y\| > \frac{\Sigma}{\delta}] \leq \delta.$$

□

A.4 Contraction of gradient systems

Lemma A.14. *Suppose ℓ satisfies Assumption 4.1. For some data sample $z \sim \mathcal{Z}$, we define the gradient descent map as*

$$T_\eta^z(\theta) = \theta - \eta\nabla\ell(z; \theta).$$

1. *For general L-smooth function ℓ, we have for all $z \in \mathcal{Z}$*
$$\|T_\eta^z(\theta) - T_\eta^z(\theta')\| \leq (1 + \eta L)\|\theta - \theta'\|. \tag{9}$$

2. *If ℓ is convex and $\eta \leq \frac{2}{L}$, we have that the map is nonexpansive such that for all $z \in \mathcal{Z}$*
$$\|T_\eta^z(\theta) - T_\eta^z(\theta')\| \leq \|\theta - \theta'\|. \tag{10}$$

3. *If ℓ is μ-strongly convex and $\eta \leq \frac{\mu}{L^2}$, we have that the map is contractive such that for all $z \in \mathcal{Z}$*
$$\|T_\eta^z(\theta) - T_\eta^z(\theta')\|^2 \leq (1 - \mu\eta)\|\theta - \theta'\|^2. \tag{11}$$

4. *Finally, let $\tilde{\mathcal{D}} \sim \mathcal{Z}$ represent a dataset of arbitrary size, and define the map*
$$T_\eta^{\tilde{\mathcal{D}}}(\theta) = \theta - \eta\frac{1}{|\tilde{\mathcal{D}}|}\sum_{z\in\tilde{\mathcal{D}}}\nabla\ell(z; \theta).$$
If there exists $\gamma > 0$ such that $\|T_\eta^z(\theta) - T_\eta^z(\theta')\| \leq \gamma\|\theta - \theta'\|$ for all $z \in \mathcal{Z}$, then for all $\tilde{\mathcal{D}}$ we have
$$\|T_\eta^{\tilde{\mathcal{D}}}(\theta) - T_\eta^{\tilde{\mathcal{D}}}(\theta')\| \leq \gamma\|\theta - \theta'\|. \tag{12}$$

Proof. To prove the first part (9), we have

$$\begin{aligned}\|T_\eta^z(\theta) - T_\eta(\theta')\| =& \|\theta - \eta\nabla\ell(z; \theta) - \theta' + \eta\nabla\ell(z; \theta')\| \\ \leq& \|\theta - \theta'\| + \|\eta\nabla\ell(z; \theta) - \eta\nabla\ell(z; \theta')\| \\ \overset{\text{Assumption 4.1}}{\leq}& \|\theta - \theta'\| + \eta L\|\theta - \theta'\|.\end{aligned}$$

To prove the second part (10), we use co-coercivity of convex and L-smooth functions such that for $\theta, \theta' \in \Theta$

$$\frac{1}{L}\|\nabla \ell(z;\theta) - \nabla \ell(z;\theta')\|^2 \leq \langle \nabla \ell(z;\theta) - \nabla \ell(z;\theta'), \theta - \theta' \rangle.$$

We have

$$\begin{aligned}
\|\theta - \eta\nabla \ell(z;\theta) - \theta' + \eta \nabla \ell(z;\theta')\|^2 =& \|\theta - \theta'\|^2 - 2\eta\langle \theta - \theta', \nabla \ell(z;\theta) - \nabla \ell(z;\theta')\rangle + \eta^2 \|\nabla \ell(z;\theta) - \nabla \ell(z;\theta')\|^2 \\
\leq& \|\theta - \theta'\|^2 - 2\eta \frac{1}{L}\|\nabla \ell(z;\theta) - \nabla \ell(z;\theta')\|^2 + \eta^2 \|\nabla \ell(z;\theta) - \nabla \ell(z;\theta')\|^2 \\
=& \|\theta - \theta'\|^2 - (2\frac{\eta}{L} - \eta^2)\|\nabla \ell(z;\theta) - \nabla \ell(z;\theta')\|^2 \\
\leq& \|\theta - \theta'\|^2.
\end{aligned}$$

To prove the third part (11), if we know ℓ is μ-strongly convex, we know that

$$\langle \theta - \theta', \nabla \ell(z;\theta) - \nabla \ell(z;\theta') \rangle \geq \mu \|\theta - \theta'\|^2.$$

We have

$$\begin{aligned}
\|\theta - \eta\nabla \ell(z;\theta) - \theta' + \eta \nabla \ell(z;\theta')\|^2 \leq& \|\theta - \theta'\|^2 - 2\eta\mu\|\theta - \theta'\|^2 + \eta^2 L^2 \|\theta - \theta'\|^2 \\
\leq& (1 - 2\eta\mu + \eta^2 L^2)\|\theta - \theta'\|^2 \\
\overset{\eta \leq \frac{\mu}{L^2}}{\leq}& (1 - \eta\mu)\|\theta - \theta'\|^2.
\end{aligned}$$

Finally, to prove the last part (12), we have

$$\begin{aligned}
\|T_\eta^{\tilde{\mathcal{D}}}(\theta) - T_\eta^{\tilde{\mathcal{D}}}(\theta')\| =& \|\theta - \eta \frac{1}{|\tilde{\mathcal{D}}|}\sum_{z \in \tilde{\mathcal{D}}} \nabla \ell(z;\theta) - \theta' + \eta \frac{1}{|\tilde{\mathcal{D}}|}\sum_{z \in \tilde{\mathcal{D}}} \nabla \ell(z;\theta')\| \\
=& \|\frac{1}{|\tilde{\mathcal{D}}|}\sum_{z \in \tilde{\mathcal{D}}} [\theta - \nabla \ell(z;\theta) - \theta' + \nabla \ell(z;\theta')]\| \\
\leq& \frac{1}{|\tilde{\mathcal{D}}|}\sum_{z \in \tilde{\mathcal{D}}} \|\theta - \nabla \ell(z;\theta) - \theta' + \nabla \ell(z;\theta')\| \\
=& \frac{1}{|\tilde{\mathcal{D}}|}\sum_{z \in \tilde{\mathcal{D}}} \|T_\eta^z(\theta) - T_\eta^z(\theta')\| \leq \gamma \|\theta - \theta'\|.
\end{aligned}$$

□

Lemma A.15. *Suppose there exists $\gamma > 0$ such that $\|T_\eta^z(\theta) - T_\eta^z(\theta')\| \leq \gamma\|\theta - \theta'\|$ for all $z \in \mathcal{Z}$. Suppose w_t, w_t' represent PSGD iterates (2) on the same loss function $\mathcal{L}_\mathcal{D}$, and $\|w_0 - w_0'\| \leq D$. Then there exists a coupling of $\{w_t\}_{t\geq 0}$, $\{w_t'\}_{t \geq 0}$ such that for all $t \geq 0$*

$$\|w_t - w_t'\| \leq \gamma^t D.$$

Moreover, the result also applies for SGD without projection.

Proof. We can couple $\{w_t\}_{t\geq 0}$ and $\{w_t'\}_{t\geq 0}$ by choosing the same sampled mini-batch $\mathcal{B}_t$ at every step, such that we have

$$\begin{aligned}
\|w_t - w_t'\| =& \|\Pi_\mathcal{C}(w_{t-1} - \eta g_{\mathcal{B}_t}(w_{t-1})) - \Pi_\mathcal{C}(w_{t-1}' - \eta g_{\mathcal{B}_t}(w_{t-1}'))\| \\
\overset{(6)}{\leq}& \|w_{t-1} - \eta g_{\mathcal{B}_t}(w_{t-1}) - w_{t-1}' + \eta g_{\mathcal{B}_t}(w_{t-1}')\| \\
=& \|w_{t-1} - \eta \frac{1}{|\mathcal{B}_t|}\sum_{z \in \mathcal{B}_t} \nabla \ell(z; w_{t-1}) - w_{t-1}' + \eta \frac{1}{|\mathcal{B}_t|}\sum_{z \in \mathcal{B}_t} \nabla \ell(z; w_{t-1})\| \\
\leq& \gamma \|\theta_{T-K+t-1}' - \theta_{t-1}''\|.
\end{aligned}$$

□

A.5 PSGD-R2D

In this section, we prove Theorem 4.3, which establishes certified unlearning for PSGD-R2D. We define the iterates $\{\theta_t\}_{t=0}^T$, $\{\theta'_t\}_{t=0}^T$, $\{\theta''_t\}_{t=0}^K$ as follows.

- $\{\theta_t\}_{t=0}^T$ represents the PSGD (2) learning iterates on $\mathcal{L}_{\mathcal{D}}$, starting from θ_0, where $\mathcal{B}_t \sim \mathcal{D}$ and θ_t is updated as follows,

$$\theta_t = \Pi_{\mathcal{C}}(\theta_{t-1} - \eta g_{\mathcal{B}_t}(\theta_{t-1})). \tag{13}$$

- $\{\theta'_t\}_{t=0}^T$ represents the learning iterates on $\mathcal{L}_{\mathcal{D}'}$, where $\theta'_0 = \theta_0$, $\mathcal{B}'_t \sim \mathcal{D}'$, and θ'_t is updated as follows,

$$\theta'_t = \Pi_{\mathcal{C}}(\theta'_{t-1} - \eta g_{\mathcal{B}'_t}(\theta'_{t-1})). \tag{14}$$

- $\{\theta''_t\}_{t=0}^K$ represents the unlearning iterates on $\mathcal{L}_{\mathcal{D}'}$, where $\theta''_0 = \theta_{T-K}$, $\mathcal{B}'_t \sim \mathcal{D}'$, and θ''_t is updated as follows,

$$\theta''_t = \Pi_{\mathcal{C}}(\theta''_{t-1} - \eta g_{\mathcal{B}'_t}(\theta''_{t-1})). \tag{15}$$

To prove Theorem 4.3, we derive a tail bound in Theorem A.16 by considering the accumulated disturbances as a sum of independent bounded random variables, which produces a tail bound via Hoeffding's inequality of the form

$$\mathbb{P}[\|\theta'_T - \theta''_K\| \geq \Sigma] \leq \delta.$$

This can be combined with Lemma 4.2 to yield the (ε, δ) guarantee. We can determine Σ for contracting, semi-contracting, and expanding systems, which correspond to strongly convex, convex, and nonconvex loss functions (Lemma A.14).

Theorem A.16. *Suppose the loss function ℓ satisfies Assumption 4.1 and within the closed, bounded, and convex set $\mathcal{C}$, the gradient of ℓ is uniformly bounded by some constant $G \geq 0$ such that for all $z \in \mathcal{Z}$ and $\theta \in \mathcal{C}$, $\|\nabla \ell(z;\theta)\| \leq G$. Moreover, suppose for all $z \in \mathcal{Z}$ and for any $\theta, \theta' \in \mathbb{R}^d$, the gradient descent map $T_\eta^z(\theta) = \theta - \eta \nabla \ell(z;\theta)$ satisfies the following property for some $\gamma > 0$,*

$$\|T_\eta^z(\theta) - T_\eta^z(\theta')\| \leq \gamma \|\theta - \theta'\|. \tag{16}$$

Let θ_t, θ'_t and θ''_t denote the training, retraining, and unlearning iterates as defined in (13), (14) and (15) respectively. If $\gamma \neq 1$, there exists a coupling of θ_t, θ'_t and θ''_t such that

$$\mathbb{P}\Big[\|\theta'_T - \theta''_K\| \geq G\eta \sqrt{\frac{2(\gamma^{2K} - \gamma^{2T})\log(1/\delta)}{1-\gamma^2}} + \frac{2G\eta m(\gamma^K - \gamma^T)}{n(1-\gamma)}\Big] \leq \delta,$$

and if $\gamma = 1$, then we have

$$\mathbb{P}\Big[\|\theta'_T - \theta''_K\| \geq G\eta\sqrt{2(T-K)\log(1/\delta)} + \frac{2G\eta m}{n}(T-K)\Big] \leq \delta,$$

where the probability is taken with respect to the resulting joint distribution of θ_t, θ'_t and θ''_t.

Proof. We describe our coupling of $\{\theta_t\}_{t=0}^T$ and $\{\theta'_t\}_{t=0}^T$. At each step t, we sample b data samples uniformly *with replacement* to form batches. Let $\mathcal{B}_t = z_1, z_2, ..., z_b$ and $\mathcal{B}'_t = z'_1, z'_2, ..., z'_b$ represent batches sampled at time t from $\mathcal{D}$ and $\mathcal{D}'$ respectively. Because the batches are sampled *with replacement*, we can treat each data sample as i.i.d. to one another (and drawing a particular sample does not impact the distribution of the next). Moreover, at each time step t, we can choose a favorable coupling of $\mathcal{B}_t$ and $\mathcal{B}'_t$ such that they contain the exact same samples except for when $\mathcal{B}_t$ contains samples from the unlearned set. Specifically, for $z_i \in \mathcal{B}_t$, if $z_i \in \mathcal{D}'$, then $z'_i = z_i$. If $z_i \notin \mathcal{D}'$, then z'_i

is simply some other point sampled uniformly from $\mathcal{D}'$. With this coupling, we have

$$\begin{aligned}
\|\theta_t - \theta'_t\| =& \|\Pi_{\mathcal{C}}(\theta_{t-1} - \eta g_{\mathcal{B}_t}(\theta_{t-1})) - \Pi_{\mathcal{C}}(\theta'_{t-1} - \eta g_{\mathcal{B}'_t}(\theta'_{t-1}))\| \\
\overset{(6)}{\leq}& \|\theta_{t-1} - \eta g_{\mathcal{B}_t}(\theta_{t-1}) - \theta'_{t-1} + \eta g_{\mathcal{B}'_t}(\theta'_{t-1})\| \\
=& \|\theta_{t-1} - \eta \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(z_i; \theta_{t-1}) - \theta'_{t-1} + \eta \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(z'_i; \theta'_{t-1})\| \\
\leq& \|\theta_{t-1} - \eta \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(z'_i; \theta_{t-1}) - \theta'_{t-1} + \eta \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(z'_i; \theta'_{t-1})\| + \frac{\eta}{b}\|\sum_{i=1}^{b}(\nabla \ell(z_i; \theta_{t-1}) - \nabla \ell(z'_i; \theta_{t-1}))\| \\
\overset{(16) \text{ and Lemma A.14}}{\leq}& \gamma \|\theta_{t-1} - \theta'_{t-1}\| + \frac{\eta}{b}\sum_{i=1}^{b} \|\nabla \ell(z_i; \theta_{t-1}) - \nabla \ell(z'_i; \theta_{t-1})\|
\end{aligned}$$

Let $d_{i,t} = \nabla \ell(z_i; \theta_{t-1}) - \nabla \ell(z'_i; \theta_{t-1})$. Then the recursive relationship evaluates to

$$\|\theta_t - \theta'_t\| \leq \frac{\eta}{b}\sum_{\tau=1}^{t} \gamma^{t-\tau} \sum_{i=1}^{b} \|d_{i,\tau}\|.$$

Under the coupling described above, we have that $d_{i,t} = 0$ with probability $\frac{n-m}{n}$. When z_i belongs to the unlearned set, which occurs with probability $\frac{m}{n}$, we can bound $d_{i.t}$ as

$$\|d_{i,t}\| \leq 2G,$$

as the gradient is uniformly bounded. We can therefore define the independent binomial variables $B_t = Binom(b, \frac{m}{n})$, coupled to the random batch sampling such that

$$\|\theta_t - \theta'_t\| \leq \frac{\eta}{b}\sum_{\tau=1}^{t} \gamma^{t-\tau} \sum_{i=1}^{b} \|d_{i,\tau}\| \leq \frac{2G\eta}{b}\sum_{\tau=1}^{t} \gamma^{t-\tau} B_\tau.$$

Since $0 \leq \gamma^{t-\tau} B_t \leq \gamma^{t-\tau} b$ uniformly, we can apply Hoeffding's inequality to the sum S_t defined as follows

$$\begin{aligned}
S_t =& \sum_{\tau=1}^{t} \gamma^{t-\tau} B_\tau, \\
\mathbb{E}[S_t] =& \sum_{\tau=1}^{t} \gamma^{t-\tau} \mathbb{E}[B_\tau] = \sum_{\tau=1}^{t} \gamma^{t-\tau} \frac{bm}{n}
\end{aligned}$$

and we have the following tail bound for all $\Delta > 0$

$$\mathbb{P}[S_t \geq \Delta + \mathbb{E}[S_t]] \leq \exp(-\frac{2\Delta^2}{\sum_{\tau=1}^{t} \gamma^{2(t-\tau)} b^2})$$

Therefore, we also have for some $0 < \delta \leq 1$

$$\begin{aligned}
\mathbb{P}[S_t \geq& \sqrt{\frac{1}{2}\sum_{\tau=1}^{t} \gamma^{2(t-\tau)} b^2 \log(1/\delta)} + \mathbb{E}[S_t]] \leq \delta \\
\mathbb{P}[S_t \geq& \sqrt{\frac{1}{2}\sum_{\tau=1}^{t} \gamma^{2(t-\tau)} b^2 \log(1/\delta)} + \frac{bm}{n}\sum_{\tau=1}^{t} \gamma^{t-\tau}] \leq \delta,
\end{aligned}$$

From Lemma A.15, we have that we can choose a coupling of $\{\theta'_t\}_{t=T-K}^{T}$, $\{\theta''_t\}_{t=0}^{K}$, such that

$$\begin{aligned}
\|\theta'_T - \theta''_K\| \leq& \gamma^K \|\theta'_{T-K} - \theta''_0\|, \\
=& \gamma^K \|\theta'_{T-K} - \theta_{T-K}\|, \\
\leq& \gamma^K \frac{2G\eta}{b} S_{T-K}.
\end{aligned}$$

Therefore, we can use the above bound to derive a tail bound for $\|\theta'_T - \theta''_K\|$

$$\mathbb{P}\Big[\|\theta'_T - \theta''_K\| \geq \frac{2G\eta\gamma^K}{b}\Big(\sqrt{\frac{1}{2}\sum_{\tau=1}^{T-K}\gamma^{2(T-K-\tau)}b^2\log(1/\delta)} + \frac{bm}{n}\sum_{\tau=1}^{T-K}\gamma^{T-K-\tau}\Big)\Big] \leq \delta,$$

$$\mathbb{P}\Big[\|\theta'_T - \theta''_K\| \geq 2G\eta\gamma^K\Big(\sqrt{\frac{1}{2}\sum_{\tau=1}^{T-K}\gamma^{2(T-K-\tau)}\log(1/\delta)} + \frac{m}{n}\sum_{\tau=1}^{T-K}\gamma^{T-K-\tau}\Big)\Big] \leq \delta,$$

For $\gamma \neq 1$, we have

$$\mathbb{P}\Big[\|\theta'_T - \theta''_K\| \geq G\eta\gamma^K\sqrt{\frac{2(1-\gamma^{2(T-K)})\log(1/\delta)}{1-\gamma^2}} + \frac{2G\eta m(\gamma^K - \gamma^T)}{n(1-\gamma)}\Big] \leq \delta$$

$$\mathbb{P}\Big[\|\theta'_T - \theta''_K\| \geq G\eta\sqrt{\frac{2(\gamma^{2K} - \gamma^{2T})\log(1/\delta)}{1-\gamma^2}} + \frac{2G\eta m(\gamma^K - \gamma^T)}{n(1-\gamma)}\Big] \leq \delta$$

and for $\gamma = 1$, we have

$$\mathbb{P}\Big[\|\theta'_T - \theta''_K\| \geq G\eta\sqrt{2(T-K)\log(1/\delta)} + \frac{2G\eta m}{n}(T-K)\Big] \leq \delta.$$

□

Theorem A.16 can be combined with Lemma A.14 and Lemma 4.2/Lemma A.11 to yield the result in Theorem 4.3.

A.5.1 Proof of Corollary 4.4

Proof. For fixed Σ, we have

$$\begin{aligned}
\Sigma =& G\eta\sqrt{\frac{2(\gamma^{2K} - \gamma^{2T})\log(1/\delta)}{1-\gamma^2}} + \frac{2G\eta m(\gamma^K - \gamma^T)}{n(1-\gamma)} \\
=& G\eta\gamma^K\sqrt{2\sum_{\tau=1}^{T-K}\gamma^{2(T-K-t)}\log(1/\delta)} + \frac{2G\eta m(\gamma^K - \gamma^T)}{n(1-\gamma)} \\
\leq& G\eta\gamma^K\sum_{\tau=1}^{T-K}\gamma^{T-K-t}\sqrt{2\log(1/\delta)} + \frac{2G\eta m(\gamma^K - \gamma^T)}{n(1-\gamma)} \\
=& G\eta(\frac{\gamma^K - \gamma^T}{1-\gamma})\sqrt{2\log(1/\delta)} + \frac{2G\eta m(\gamma^K - \gamma^T)}{n(1-\gamma)} \\
\leq& (\frac{\gamma^K - \gamma^T}{1-\gamma})(G\eta\sqrt{2\log(1/\delta)} + \frac{2G\eta m}{n}) \\
\gamma^K - \gamma^T \geq& \frac{\Sigma(1-\gamma)}{G\eta\sqrt{2\log(1/\delta)} + \frac{2G\eta m}{n}} \\
K\log\gamma \geq& \log(\frac{\Sigma(1-\gamma)}{G\eta\sqrt{2\log(1/\delta)} + \frac{2G\eta m}{n}} + \gamma^T) \\
K \leq& \frac{1}{\log\gamma}\log(\frac{\Sigma(1-\gamma)}{G\eta\sqrt{2\log(1/\delta)} + \frac{2G\eta m}{n}} + \gamma^T) \leq \frac{1}{\log\gamma}\log(\frac{\Sigma(1-\gamma)}{G\eta\sqrt{2\log(1/\delta)} + \frac{2G\eta m}{n}}),
\end{aligned}$$

where the last step uses the fact that $\frac{\Sigma(1-\gamma)}{G\eta\sqrt{2\log(1/\delta)}+\frac{2G\eta m}{n}} \leq \gamma^K - \gamma^T \leq 1 - \gamma^T$. For small T, K increases at an approximately linear rate, but as T grows to infinity, K converges to a constant value. □

A.6 SGD-R2D

In this section, we prove Theorem 4.7, which establishes certified unlearning for SGD-R2D without projection or bounded domain. We establish preliminaries and present Lemmas A.22 and A.18. We then use those results to prove Theorem A.20, from which Theorem 4.7 directly follows. In Appendix A.7.1, we prove Lemma A.22. In Appendix A.6.2, we prove Lemma A.6.2.

We define the iterates $\{\theta_t\}_{t=0}^T$, $\{\theta'_t\}_{t=0}^T$, $\{\theta''_t\}_{t=0}^K$ as follows.

- $\{\theta_t\}_{t=0}^T$ represents the SGD (1) learning iterates on $\mathcal{L}_\mathcal{D}$, starting from θ_0, where $\mathcal{B}_t \sim \mathcal{D}$ and θ_t is updated as follows,
$$\theta_t = \theta_{t-1} - \eta g_{\mathcal{B}_t}(\theta_{t-1}). \tag{17}$$
- $\{\theta'_t\}_{t=0}^T$ represents the SGD learning iterates on $\mathcal{L}_{\mathcal{D}'}$, where $\theta'_0 = \theta_0$, $\mathcal{B}'_t \sim \mathcal{D}'$ and θ'_t is updated as follows,
$$\theta'_t = \theta'_{t-1} - \eta g_{\mathcal{B}'_t}(\theta'_{t-1}). \tag{18}$$
- $\{\theta''_t\}_{t=0}^K$ represents the SGD unlearning iterates on $\mathcal{L}_{\mathcal{D}'}$, where $\theta''_0 = \theta_{T-K}$, $\mathcal{B}'_t \sim \mathcal{D}'$, and θ''_t is updated as follows,
$$\theta''_t = \theta''_{t-1} - \eta g_{\mathcal{B}'_t}(\theta''_{t-1}). \tag{19}$$

To prove Theorem 4.7, we need to determine Σ such that $\mathbb{E}[\|\theta'_t - \theta''_K\|] \leq \Sigma$. This result can be combined with Lemma 4.2 to yield the (ε, δ) guarantee in Theorem 4.7. We can determine Σ for contracting, semi-contracting, and expanding systems, which correspond to strongly convex, convex, and nonconvex loss functions (Lemma A.14). This yields our final noise guarantees in Theorem 4.7.

As discussed in the main body, we require Lemma A.17, which bounds the unlearning bias in terms of the second moment of the stochastic noise.

Lemma A.17. *For all $\theta \in \mathbb{R}^d$, the unlearning bias is bounded as follows,*
$$\|\nabla \mathcal{L}_\mathcal{D}(\theta) - \nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 \leq \frac{m}{n-m} \mathbb{E}_{z\sim\mathcal{D}}[\|\nabla \ell(\theta; z)\|^2]. \tag{20}$$

Proof. See Appendix A.6.1.

In addition, we require Lemma A.18, a classic result which states that assuming the noise is relatively bounded (Assumption 4.5), the gradient norms of the *unbiased* SGD iterates are bounded. Lemma A.17 and A.18 allow us to isolate and capture the accumulated disturbances from noise and bias, which are upper bounded by a linear term in T and the loss at initialization. This can be combined with Assumption 4.6 to achieve a dataset-independent bound.

Lemma A.18. *(Convergence of SGD) Let θ_t represent SGD iterates (1) on a loss function $\mathcal{L}_\mathcal{D}$. Suppose Assumptions 4.1 and 4.5 are satisfied. Then for $\theta_{t+1} = \theta_t - \eta g_{\mathcal{B}_t}(\theta_t)$ and $\eta \leq \frac{1}{LB}$, we have*
$$\mathbb{E}[\sum_{t=0}^{T-1} \|\nabla \mathcal{L}_\mathcal{D}(\theta_t)\|^2] \leq \frac{2}{\eta}(\mathcal{L}_\mathcal{D}(\theta_0) - \mathcal{L}^*_\mathcal{D}) + L\eta CT.$$

Proof. See Appendix A.6.2.

We use the above results to prove the following Lemma A.19 that bounds the expected distance between θ'_t and θ_t.

Lemma A.19. *Suppose that the loss function ℓ satisfies Assumption 4.1, 4.5, and 4.6. Moreover, suppose for all $z \in \mathcal{Z}$ and for any $\theta, \theta' \in \mathbb{R}^d$, the gradient descent map $T^z_\eta(\theta) = \theta - \eta \nabla \ell(z; \theta)$ satisfies the following property for some $\gamma > 0$,*
$$\|T^z_\eta(\theta) - T^z_\eta(\theta')\| \leq \gamma \|\theta - \theta'\|. \tag{21}$$
Let θ_t and θ'_t denote the training and retraining from scratch iterates as defined in (17) and (18) respectively. there exists a coupling of the iterates $\{\theta_t\}$, $\{\theta'_t\}$ such that for $t \geq 0$ and $\gamma \neq 1$, we have
$$\mathbb{E}[\|\theta_t - \theta'_t\|] \leq \eta \frac{1-\gamma^t}{1-\gamma} \left(3B\left(\frac{2}{\eta}\ell_{\theta_0} + L\eta Ct\right)(\frac{3n-m}{n-m}) + 6(\frac{4n-3m}{n-m})C\right)^{1/2}.$$
If $\gamma = 1$, we have
$$\mathbb{E}[\|\theta_t - \theta'_t\|] \leq \eta t \left(3B\left(\frac{2}{\eta}\ell_{\theta_0} + L\eta Ct\right)(\frac{3n-m}{n-m}) + 6(\frac{4n-3m}{n-m})C\right)^{1/2}.$$

Proof. See Appendix A.6.3.

Lemma A.19 allows us to prove Theorem A.20, which can be combined with Lemma A.14 and Lemma 4.2 to yield the result in Theorem 4.7.

Theorem A.20. *Suppose that the loss function ℓ satisfies Assumptions 4.1, 4.5, and 4.6. Moreover, suppose for all $z \in \mathcal{Z}$ and for any $\theta, \theta' \in \mathbb{R}^d$, the gradient descent map $T_\eta^z(\theta) = \theta - \eta\nabla\ell(z;\theta)$ satisfies the property that for some $\gamma > 0$ $\|T_\eta^z(\theta) - T_\eta^z(\theta')\| \leq \gamma\|\theta - \theta'\|$. Let θ_t, θ_t' and θ_t'' denote the training, retraining, and unlearning iterates as defined in (17), (18) and (19) respectively. If $\gamma \neq 1$, there exists a coupling of θ_t, θ_t' and θ_t'' such that*

$$\mathbb{E}[\|\theta_K'' - \theta_T'\|] \leq \eta\frac{\gamma^K - \gamma^T}{1-\gamma}\left(3B\left(\frac{2}{\eta}\ell_{\theta_0} + L\eta C(T-K)\right)(\frac{3n-m}{n-m}) + 6(\frac{4n-3m}{n-m})C\right)^{1/2},$$

and if $\gamma = 1$, then we have

$$\mathbb{E}[\|\theta_K'' - \theta_T'\|] \leq \eta(T-K)\left(3B\left(\frac{2}{\eta}\ell_{\theta_0} + L\eta C(T-K)\right)(\frac{3n-m}{n-m}) + 6(\frac{4n-3m}{n-m})C\right)^{1/2},$$

where the expectation is taken with respect to the resulting joint distribution of θ_t, θ_t' and θ_t''.

Proof. Lemma A.19 shows that the expected distance between θ_t, θ_t', is bounded as a function of t. So for $\gamma \neq 1$, we have

$$\mathbb{E}[\|\theta_{T-K} - \theta_{T-K}'\|] \leq \eta\frac{1-\gamma^{T-K}}{1-\gamma}\left(3B\left(\frac{2}{\eta}\ell_{\theta_0} + L\eta C(T-K)\right)(\frac{3n-m}{n-m}) + 6(\frac{4n-3m}{n-m})C\right)^{1/2}.$$

From Lemma A.15, we have that we can choose a coupling of $\{\theta_t'\}_{t=T-K}^T$, $\{\theta_t''\}_{t=0}^K$, such that

$$\begin{aligned}
\|\theta_T' - \theta_K''\| &\leq \gamma^K\|\theta_{T-K}' - \theta_0''\|, \\
\mathbb{E}[\|\theta_T' - \theta_K''\|] &\leq \gamma^K\mathbb{E}[\|\theta_{T-K}' - \theta_0''\|], \\
&\leq \eta\frac{\gamma^K - \gamma^T}{1-\gamma}\left(3B\left(\frac{2}{\eta}\ell_{\theta_0} + L\eta C(T-K)\right)(\frac{3n-m}{n-m}) + 6(\frac{4n-3m}{n-m})C\right)^{1/2}.
\end{aligned}$$

The same approach can be applied to the $\gamma = 1$ case, finishing our proof.

□

A.6.1 Proof of Lemma A.17

To bound the unlearning bias, we first require the following result from Polyanskiy and Wu [2025] (Theorem 7.26 and Example 7.4), which follows from the Radon-Nikodym theorem and Cauchy-Schwarz inequality.

Lemma A.21. *Let P and Q be probability measures on a measurable space $\mathcal{X}$ such that $P \ll Q$. Then for any function $f : \mathcal{X} \to \mathbb{R}$, we have*

$$\|\mathbb{E}_P[f] - \mathbb{E}_Q[f]\|^2 \;\leq\; \chi^2(P\|Q) \;\cdot\; \mathrm{Var}_Q(f),$$

where

$$\chi^2(P\|Q) = \int (w(x) - 1)^2 \; dQ(x)$$

and $w(x) = \frac{dP}{dQ}(x)$ denotes the Radon-Nikodym derivative.

Now we can proceed with the proof of Lemma A.17.

Proof. In the following, we provide an explicit proof that does not require any prior knowledge for easy understanding. However, we note that the result ultimately follows from defining the empirical distributions of $\mathcal{D}$ and $\mathcal{D}'$ as $P_{\mathcal{D}}$ and $P_{\mathcal{D}'}$, and applying Lemma A.21.

Define $w(x)$ as follows

$$w(x) = \begin{cases} \frac{n}{n-m} & x \in \mathcal{D}', \\ 0 & x \in \mathcal{D} \backslash \mathcal{D}'. \end{cases} \tag{22}$$

Then we can write $\nabla \mathcal{L}_{\mathcal{D}'}(\theta)$ in terms of $w(z_i)$ for all $z_i \in \mathcal{D}'$ as follows,

$$\nabla \mathcal{L}_{\mathcal{D}'}(\theta) = \frac{1}{n-m} \sum_{i=1}^{n-m} \nabla \ell(\theta; z_i) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta; z_i) w(z_i).$$

Then for $\nabla \mathcal{L}_{\mathcal{D}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta; z_i)$, we have

$$\begin{aligned} \|\nabla \mathcal{L}_{\mathcal{D}}(\theta) - \nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 =& \|\frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta; z_i) - \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(\theta; z_i) w(z_i)\|^2, \\ =& \frac{1}{n^2} \|\sum_{i=1}^{n} \nabla \ell(\theta; z_i)(1 - w(z_i))\|^2. \end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \|\nabla \mathcal{L}_{\mathcal{D}}(\theta) - \nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 \leq& \frac{1}{n^2} \sum_{i=1}^{n} \|\nabla \ell(\theta; z_i)\|^2 \sum_{i=1}^{n} (1 - w(z_i))^2, \\ =& \frac{1}{n} \sum_{i=1}^{n} \|\nabla \ell(\theta; z_i)\|^2 \cdot \frac{1}{n} (m + (n-m)(1 - \frac{n}{n-m})^2), \\ =& \mathbb{E}_{z \sim \mathcal{D}}[\|\nabla \ell(\theta; z_i)\|^2] \frac{m}{n-m}. \end{aligned}$$

The theory underlying Lemma A.17 is that we use the second moment bound to write the difference in terms of the χ^2-divergence between $P_{\mathcal{D}}$ and $P_{\mathcal{D}'}$, where $w(x)$ represents a discrete version of the Radon Nikodym derivative $\frac{dP_{\mathcal{D}'}}{dP_{\mathcal{D}}}$. The χ^2-divergence then evaluates to $\frac{m}{n-m}$.

□

A.6.2 Proof of Lemma A.18

Proof. By Lipschitz smoothness (Lemma A.6), we have

$$\begin{aligned} \mathcal{L}_{\mathcal{D}}(\theta_{t+1}) - \mathcal{L}_{\mathcal{D}}(\theta_t) \leq& \nabla \mathcal{L}_{\mathcal{D}}(\theta_t)^T (-\eta \nabla \mathcal{L}_{\mathcal{D}}(\theta_t) + \eta \xi_t) + \frac{L\eta^2}{2} \|g(\theta_t)\|^2 \\ \mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_{t+1})] - \mathcal{L}_{\mathcal{D}}(\theta_t) \leq& \nabla \mathcal{L}_{\mathcal{D}}(\theta_t)^T (-\eta \nabla \mathcal{L}_{\mathcal{D}}(\theta_t)) + \frac{L\eta^2}{2} \mathbb{E}[\|g(\theta_t)\|^2] \\ =& -\eta \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|^2 + \frac{L\eta^2}{2} \mathbb{E}[\|g(\theta_t)\|^2]. \end{aligned}$$

By Assumption 4.5 and $\eta \leq \frac{1}{LB}$, we have

$$\begin{aligned} \mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_{t+1})] - \mathcal{L}_{\mathcal{D}}(\theta_t) \leq& -\eta \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|^2 + \frac{L\eta^2}{2} (B \|\nabla \mathcal{L}(\theta_t)\|^2 + C) \\ \leq& -\frac{\eta}{2} \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|^2 + \frac{L\eta^2}{2} C \\ \frac{\eta}{2} \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|^2 \leq& \mathcal{L}_{\mathcal{D}}(\theta_t) - \mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_{t+1})] + \frac{L\eta^2 C}{2} \\ \frac{\eta}{2} \mathbb{E}[\|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|^2] \leq& \mathbb{E}[\mathbb{E}[\mathcal{L}_{\mathcal{D}}(\theta_t)] - \mathbb{E}[\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_{t+1})] + \frac{L\eta^2 C}{2} \\ \frac{\eta}{2} \mathbb{E}[\sum_{t=0}^{T-1} \|\nabla \mathcal{L}_{\mathcal{D}}(\theta_t)\|^2] \leq& \mathcal{L}_{\mathcal{D}}(\theta_0) - \mathcal{L}_{\mathcal{D}}^* + \frac{L\eta^2 C}{2} T. \end{aligned}$$

□

A.6.3 Proof of Lemma A.19

Proof. We can decompose $\|\theta_t - \theta'_t\|$ as follows,

$$\mathbb{E}[\|\theta_t - \theta'_t\|] \leq \underbrace{\|\theta_{t-1} - \eta\nabla\mathcal{L}_{\mathcal{D}'}(\theta_{t-1}) - \theta'_{t-1} + \eta\nabla\mathcal{L}_{\mathcal{D}'}(\theta'_{t-1})\|}_{(1)} + \eta\underbrace{\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta_{t-1}) - \nabla\mathcal{L}_{\mathcal{D}}(\theta_{t-1})\|}_{(2)} + \eta\underbrace{\mathbb{E}[\|g_{\mathcal{D}'}(\theta'_{t-1}) - \nabla\mathcal{L}_{\mathcal{D}'}(\theta'_{t-1})\| + \|g_{\mathcal{D}}(\theta_{t-1}) - \nabla\mathcal{L}_{\mathcal{D}}(\theta_{t-1})\|]}_{(3)},$$

where (2) represents the unlearning bias and (3) represents the noise. By (21) and (12) of Lemma A.14, we have

$$(1) \leq \gamma\|\theta_{t-1} - \theta'_{t-1}\|.$$

By our bias bound (Lemma A.17) and relative noise bound assumption (Assumption 4.5), we have

$$(2) \leq (\frac{m}{n-m}\mathbb{E}_{z\sim\mathcal{D}}[\|\nabla\ell(z;\theta_{t-1})\|^2])^{1/2} \leq (\frac{m}{n-m}(B\|\nabla\mathcal{L}_{\mathcal{D}}(\theta_{t-1})\|^2 + C))^{1/2}.$$

We can also use Assumption 4.5 to bound the noise terms in (3) as follows,

$$(3) \leq (B\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta'_{t-1})\|^2 + C)^{1/2} + (B\|\nabla\mathcal{L}_{\mathcal{D}}(\theta_{t-1})\|^2 + C)^{1/2}.$$

Combining the above bounds and utilizing the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and Lemma A.4, we have

$$\begin{aligned}
\mathbb{E}[\|\theta_t - \theta'_t\|] \leq& \gamma\|\theta_{t-1} - \theta'_{t-1}\| + \eta\sqrt{B}\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta'_{t-1})\| + \eta\sqrt{2(\frac{n}{n-m})B}\|\nabla\mathcal{L}_{\mathcal{D}}(\theta_{t-1})\| + \eta\sqrt{2(\frac{4n-3m}{n-m})C} \\
\mathbb{E}[\|\theta_t - \theta'_t\|] \leq& \eta\mathbb{E}\left[\sum_{\tau=1}^{t}\gamma^{t-\tau}\left(\sqrt{B}\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta'_{\tau-1})\| + \sqrt{2(\frac{n}{n-m})B}\|\nabla\mathcal{L}_{\mathcal{D}}(\theta_{\tau-1})\| + \sqrt{2(\frac{4n-3m}{n-m})C}\right)\right] \\
\overset{Cauchy-Schwarz}{\leq}& \eta\sqrt{\sum_{\tau=1}^{t}\gamma^{2(t-\tau)}}\left(\sqrt{B\sum_{\tau=1}^{t}\mathbb{E}[\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta'_{\tau-1})\|^2]} + \sqrt{2(\frac{n}{n-m})B\sum_{\tau=1}^{t}\mathbb{E}[\|\nabla\mathcal{L}_{\mathcal{D}}(\theta_{\tau-1})\|^2]}\right) \\
&+ \eta\sqrt{2(\frac{4n-3m}{n-m})C}\sum_{\tau=1}^{t}\gamma^{t-\tau}.
\end{aligned}$$

We observe that we can simplify the geometric sums such that

$$\sqrt{\sum_{\tau=1}^{t}\gamma^{2(t-\tau)}} \leq \sum_{\tau=1}^{t}\gamma^{t-\tau} = \sum_{\tau=0}^{t-1}\gamma^{\tau}.$$

This is equal to $\frac{1-\gamma^t}{1-\gamma}$ when $\gamma \neq 1$ and t when $\gamma = 1$. Plugging this back into the original equation yields

$$\begin{aligned}
\mathbb{E}[\|\theta_t - \theta'_t\|] \leq & \eta \sum_{\tau=0}^{t-1} \gamma^\tau \left(\sqrt{B \sum_{\tau=1}^{t} \mathbb{E}[\|\nabla \mathcal{L}_{\mathcal{D}'}(\theta'_{\tau-1})\|^2]} + \sqrt{2(\frac{n}{n-m}) B \sum_{\tau=1}^{t} \mathbb{E}[\|\nabla \mathcal{L}_{\mathcal{D}}(\theta_{\tau-1})\|^2]} + \sqrt{2(\frac{4n-3m}{n-m})C} \right), \\
\overset{\text{Lemma A.18}}{\leq} & \eta \sum_{\tau=0}^{t-1} \gamma^\tau \left(\sqrt{B \left(\frac{2}{\eta} (\mathcal{L}_{\mathcal{D}'}(\theta_0) - \mathcal{L}^*_{\mathcal{D}'}) + L\eta Ct \right)} + \sqrt{2(\frac{n}{n-m}) B \left(\frac{2}{\eta} (\mathcal{L}_{\mathcal{D}}(\theta_0) - \mathcal{L}^*_{\mathcal{D}}) + L\eta Ct \right)} \right. \\
& \left. + \sqrt{2(\frac{4n-3m}{n-m})C} \right) \\
\overset{\text{Assumption 4.6}}{\leq} & \eta \sum_{\tau=0}^{t-1} \gamma^\tau \left(\sqrt{B \left(\frac{2}{\eta} \ell_{\theta_0} + L\eta Ct \right)} + \sqrt{2(\frac{n}{n-m}) B \left(\frac{2}{\eta} \ell_{\theta_0} + L\eta Ct \right)} + \sqrt{2(\frac{4n-3m}{n-m})C} \right) \\
\overset{\text{Lemma A.4}}{\leq} & \eta \sum_{\tau=0}^{t-1} \gamma^\tau \left(3 \left(B \left(\frac{2}{\eta} \ell_{\theta_0} + L\eta Ct \right) (1 + 2(\frac{n}{n-m})) + 2(\frac{4n-3m}{n-m})C \right) \right)^{1/2} \\
= & \eta \sum_{\tau=0}^{t-1} \gamma^\tau \left(3 \left(B \left(\frac{2}{\eta} \ell_{\theta_0} + L\eta Ct \right) (\frac{3n-m}{n-m}) + 2(\frac{4n-3m}{n-m})C \right) \right)^{1/2} \\
= & \eta \sum_{\tau=0}^{t-1} \gamma^\tau \left(3B \left(\frac{2}{\eta} \ell_{\theta_0} + L\eta Ct \right) (\frac{3n-m}{n-m}) + 6(\frac{4n-3m}{n-m})C \right)^{1/2}.
\end{aligned}$$

□

A.7 SGD-D2D

In this section, we prove Theorem 4.8, which establishes certified unlearning for SGD-D2D on strongly convex functions. We first establish preliminaries and provide an overall proof sketch. In Appendix A.7.1 we prove Lemma A.22, and in Appendix A.7.2 we prove Lemma A.23, which are necessary for achieving the final result. Finally, in Appendix A.7.3, we finish the proof of Theorem 4.8.

We define the iterates $\{\theta_t\}_{t=0}^T$, $\{\theta'_t\}_{t=0}^T$, $\{\theta''_t\}_{t=0}^K$ as follows.

- $\{\theta_t\}_{t=0}^T$ represents the SGD (1) learning iterates on $\mathcal{L}_{\mathcal{D}}$, starting from θ_0, where $\mathcal{B}_t \sim \mathcal{D}$ and θ_t is updated as follows,
$$\theta_t = \theta_{t-1} - \eta g_{\mathcal{B}_t}(\theta_{t-1}). \tag{23}$$
- $\{\theta'_t\}_{t=0}^T$ represents the SGD learning iterates on $\mathcal{L}_{\mathcal{D}'}$, where $\theta'_0 = \theta_0$, $\mathcal{B}'_t \sim \mathcal{D}'$, and θ'_t is updated as follows,
$$\theta'_t = \theta'_{t-1} - \eta g_{\mathcal{B}'_t}(\theta'_{t-1}). \tag{24}$$
- $\{\theta''_t\}_{t=0}^K$ represents the SGD unlearning iterates on $\mathcal{L}_{\mathcal{D}'}$, where $\theta''_0 = \theta_T$, $\mathcal{B}'_t \sim \mathcal{D}'$, and θ''_t is updated as follows,
$$\theta''_t = \theta''_{t-1} - \eta g_{\mathcal{B}'_t}(\theta''_{t-1}). \tag{25}$$

The first step of the proof is to show that during training the biased SGD iterates $\{\theta_t\}_{t=0}^T$ will contract to be within some neighborhood of $\theta^{*'}$, the optimum of $\mathcal{L}_{\mathcal{D}'}$ (Lemma A.23). This result replicates the general result in Ajalloeian and Stich [2020] showing linear convergence of biased SGD as long as the following conditions hold: i) the loss function is PL and Lipschitz smooth, ii) the noise satisfies a (relative) bound as in Assumption 4.5, and iii) there exists constants $D \geq 0$, $0 \leq M < 1$, such that the bias at time step t, denoted as d_t, is relatively bounded as follows,

$$\|d_t\|^2 \leq M \|\nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 + D. \tag{26}$$

We show that we satisfy the bias bound condition (26) if the proportion of unlearned data is small enough in Lemma A.22.

Lemma A.22. *Suppose that Assumption 4.5 holds and $\frac{m}{n} \leq \frac{1}{6B+1}$. Then for all $\theta \in \mathbb{R}^d$, the unlearning bias is bounded as follows,*

$$\|\nabla \mathcal{L}_{\mathcal{D}}(\theta) - \nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 \leq \frac{1}{2}\|\nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 + \frac{C}{4B}. \tag{27}$$

Proof. See Appendix A.7.1.

Lemma A.23. *Consider the SGD algorithm on $\mathcal{L}_{\mathcal{D}}$ as defined in (23), and let $\mathcal{L}^*_{\mathcal{D}'}$ represent the minimum value of $\mathcal{L}_{\mathcal{D}'}$. Suppose that $\eta \leq \frac{1}{BL}$ and $\frac{m}{n} < \frac{1}{6B+1}$. Then we have*

$$\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_T)] - \mathcal{L}^*_{\mathcal{D}'} \leq (1 - \frac{\eta\mu}{2})^T(\mathcal{L}_{\mathcal{D}'}(\theta_0) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{C}{4B\mu} + \frac{L\eta C}{\mu},$$

where the expectation is taken with respect to the underlying randomization of $\{\theta_t\}_{t=0}^T$.

Proof. See Appendix A.7.2.

The second step is to show that the *unbiased* SGD iterates $\{\theta'_t\}_{t=0}^T$ also converge close to $\theta^{*'}$ through the classic convergence analysis. Finally, the last step is to show that during unlearning, the *unbiased* SGD iterates $\{\theta''_t\}_{t=0}^K$ will converge closer $\theta^{*'}$, closing the gap up to some neighborhood determined by the stochasticity. By tracking the progress of the loss value $\mathcal{L}_{\mathcal{D}'}$ and leveraging the quadratic growth condition (8), we achieve a second-moment bound on $\|\theta'_T - \theta''_K\|$, which can be combined with Lemma 4.2 to yield (ε, δ)-indistinguishability. In Appendix A.7.3, we carry out these proof components to achieve the result in Theorem 4.8. This approach is unique to strongly convex functions and cannot be achieved for convex and nonconvex functions.

A.7.1 Proof of Lemma A.22

Proof. Combining Lemma A.17 with Assumption 4.5, we have

$$\begin{aligned}
\|\nabla \mathcal{L}_{\mathcal{D}}(\theta) - \nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 \leq & \frac{m}{n-m}(B\|\nabla \mathcal{L}_{\mathcal{D}}(\theta)\|^2 + C), \\
= & \frac{m}{n-m}(B\|\nabla \mathcal{L}_{\mathcal{D}}(\theta) - \nabla \mathcal{L}_{\mathcal{D}'}(\theta) + \nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 + C), \\
\overset{\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2}{\leq} & \frac{m}{n-m}(2B\|\nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 + 2B\|\nabla \mathcal{L}_{\mathcal{D}}(\theta) - \nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 + C) \\
(1 - \frac{2Bm}{n-m})\|\nabla \mathcal{L}_{\mathcal{D}}(\theta) - \nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 \leq & \frac{2Bm}{n-m}\|\nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 + \frac{Cm}{n-m} \\
\|\nabla \mathcal{L}_{\mathcal{D}}(\theta) - \nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 \leq & \frac{2Bm}{n-m-2Bm}\|\nabla \mathcal{L}_{\mathcal{D}'}(\theta)\|^2 + \frac{Cm}{n-m-2Bm},
\end{aligned}$$

where in the last step, the inequality is maintained after dividing both sides by $1 - \frac{2Bm}{n-m}$, which is positive since $\frac{m}{n} < \frac{1}{6B+1} \leq \frac{1}{2B+1}$. We note that $\frac{2Bm}{n-m-2Bm} \leq \frac{1}{2}$, which can be used to simplify to the final result.

□

A.7.2 Proof of Lemma A.23

Proof. We define the bias term d_t and zero-mean noise term ξ_t as

$$\begin{aligned}
d_t &= \nabla \mathcal{L}_{\mathcal{D}'}(\theta_t) - \nabla \mathcal{L}_{\mathcal{D}}(\theta_t) \\
\xi_t &= \nabla \mathcal{L}_{\mathcal{D}}(\theta_t) - g(\theta_t).
\end{aligned}$$

Let $\{\mathcal{F}_t\}_{t\geq 0} = \{\sigma(\xi_0, ..., \xi_{t-1})\}_{t\geq 0}$ denote the natural filtration adapted to ξ_t such that θ_t is $\mathcal{F}_t$-measurable and we have

$$\mathbb{E}[\xi_t \,|\, \mathcal{F}_t] = \mathbb{E}[\nabla \mathcal{L}_{\mathcal{D}}(\theta_t) - g(\theta_t) \,|\, \mathcal{F}_t] = 0.$$

By Lipschitz smoothness (Lemma A.6), we have

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}'}(\theta_{t+1}) - \mathcal{L}_{\mathcal{D}'}(\theta_t) \leq & \nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)^T(-\eta\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + \eta d_t + \eta\xi_t) + \frac{L\eta^2}{2}\|g(\theta_t)\|^2, \\
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_{t+1}) \,|\, \mathcal{F}_t] - \mathcal{L}_{\mathcal{D}'}(\theta_t) \leq & \nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)^T(-\eta\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + \eta d_t) + \frac{L\eta^2}{2}\mathbb{E}[\|g(\theta_t)\|^2 \,|\, \mathcal{F}_t], \\
\overset{\text{Assumption 4.5}}{\leq} & \nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)^T(-\eta\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + \eta d_t) + \frac{L\eta^2}{2}(B\|\nabla\mathcal{L}_{\mathcal{D}}(\theta_t)\|^2 + C), \\
= & \nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)^T(-\eta\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + \eta d_t) + \frac{L\eta^2}{2}(B\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + d_t\|^2 + C), \\
= & -\eta\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)^T(\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + d_t) + \frac{L\eta^2 B}{2}\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + d_t\|^2 + \frac{L\eta^2 C}{2}, \\
\overset{\eta\leq\frac{1}{BL}}{\leq} & -\eta\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)^T(\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + d_t) + \frac{\eta}{2}\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + d_t\|^2 + \frac{L\eta^2 C}{2}.
\end{aligned}$$

By the fact that $\|a+b\|^2 - 2a^T(a+b) = -\|a\|^2 + \|b\|^2$, the bias cross term $\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)^T d_t$ can get "folded" into the norm squared of d_t, which is bounded away with Lemma A.22. We have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_{t+1}) \,|\, \mathcal{F}_t] - \mathcal{L}_{\mathcal{D}'}(\theta_t) \leq & \frac{\eta}{2}(-2\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)^T(\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + d_t) + \|\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t) + d_t\|^2) + \frac{L\eta^2 C}{2}, \\
= & -\frac{\eta}{2}\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)\|^2 + \frac{\eta}{2}\|d_t\|^2 + \frac{L\eta^2 C}{2}.
\end{aligned}$$

Assuming that $\frac{m}{n} \leq \frac{1}{1+6B}$, we can bound the bias term by Lemma A.22 as follows,

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_{t+1}) \,|\, \mathcal{F}_t] - \mathcal{L}_{\mathcal{D}'}(\theta_t) \overset{\text{Lemma A.22}}{\leq} & -\frac{\eta}{2}\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)\|^2 + \frac{\eta}{2}(\frac{1}{2}\|\mathcal{L}_{\mathcal{D}'}(\theta)\|^2 + \frac{C}{4B}) + \frac{L\eta^2 C}{2}, \\
\leq & -\frac{\eta}{4}\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)\|^2 + \frac{\eta C}{8B} + \frac{L\eta^2 C}{2}.
\end{aligned}$$

We can use the Polyak–Łojasiewicz (PL) property of strongly convex functions (Assumption A.2 and Definition A.8) to bound in terms of $\mathcal{L}^*_{\mathcal{D}'} = \inf_{\theta\in\mathbb{R}^d}\mathcal{L}_{\mathcal{D}'}(\theta)$.

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_{t+1}) \,|\, \mathcal{F}_t] - \mathcal{L}_{\mathcal{D}'}(\theta_t) \overset{\text{Assumption A.2}}{\leq} & -\frac{\eta}{2}\mu(\mathcal{L}_{\mathcal{D}'}(\theta_t) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{\eta C}{8B} + \frac{L\eta^2 C}{2}, \\
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_{t+1}) \,|\, \mathcal{F}_t] - \mathcal{L}^*_{\mathcal{D}'} \leq & (1 - \frac{\eta\mu}{2})(\mathcal{L}_{\mathcal{D}'}(\theta_t) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{\eta C}{8B} + \frac{L\eta^2 C}{2}, \\
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_t)] - \mathcal{L}^*_{\mathcal{D}'} \overset{\text{Lemma A.7}}{\leq} & (1 - \frac{\eta\mu}{2})^t(\mathcal{L}_{\mathcal{D}'}(\theta_0) - \mathcal{L}^*_{\mathcal{D}'}) + \eta(\frac{C}{8B} + \frac{L\eta C}{2})\sum_{i=0}^{t-1}(1 - \frac{\eta\mu}{2})^i, \\
\leq & (1 - \frac{\eta\mu}{2})^t(\mathcal{L}_{\mathcal{D}'}(\theta_0) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{2}{\mu}(\frac{C}{8B} + \frac{L\eta C}{2}), \\
\leq & (1 - \frac{\eta\mu}{2})^t(\mathcal{L}_{\mathcal{D}'}(\theta_0) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{C}{4B\mu} + \frac{L\eta C}{\mu},
\end{aligned}$$

where in the second to last step we upper bound the geometric series $\sum_{i=0}^{t-1}(1 - \frac{\eta\mu}{2})^i$. □

A.7.3 Proof of Theorem 4.8

Proof. We can analyze the linear convergence of θ'_t on $\mathcal{L}_{\mathcal{D}'}$ as follows,

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}'}(\theta'_{t+1}) - \mathcal{L}_{\mathcal{D}'}(\theta'_t) \leq & \nabla\mathcal{L}_{\mathcal{D}'}(\theta'_t)^T(-\eta g_{\mathcal{B}'}(\theta'_t)) + \frac{L\eta^2}{2}\|g_{\mathcal{B}'}(\theta'_t)\|^2 \\
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_{t+1}) \,|\, \mathcal{F}_t] - \mathcal{L}_{\mathcal{D}'}(\theta_t) \leq & \nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)^T(-\eta\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)) + \frac{L\eta^2}{2}\mathbb{E}[\|g_{\mathcal{B}'}(\theta_t)\|^2 \,|\, \mathcal{F}_t]. \\
\overset{\text{Assumption 4.5}}{\leq} & \nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)^T(-\eta\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)) + \frac{L\eta^2}{2}(B\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta)\|^2 + C) \\
= & -\eta\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta_t)\|^2 + \frac{L\eta^2 B}{2}\|\nabla\mathcal{L}_{\mathcal{D}'}(\theta)\|^2 + \frac{L\eta^2 C}{2}.
\end{aligned}$$

Let $\eta \leq \frac{1}{BL}$, then we have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta'_{t+1}) \,|\, \mathcal{F}_t] - \mathcal{L}_{\mathcal{D}'}(\theta'_t) \leq& -\frac{\eta}{2}\|\nabla \mathcal{L}_{\mathcal{D}'}(\theta'_t)\|^2 + \frac{L\eta^2 C}{2}, \\
\overset{\text{Definition A.8}}{\leq}& -\eta\mu(\mathcal{L}_{\mathcal{D}'}(\theta'_t) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{L\eta^2 C}{2}, \\
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta'_t)] - \mathcal{L}^*_{\mathcal{D}'} \leq& (1-\eta\mu)^t(\mathcal{L}_{\mathcal{D}'}(\theta'_0) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{L\eta^2 C}{2}\sum_{i=0}^{t-1}(1-\eta\mu)^i, \\
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta'_t)] - \mathcal{L}^*_{\mathcal{D}'} \leq& (1-\eta\mu)^t(\mathcal{L}_{\mathcal{D}'}(\theta'_0) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{L\eta C}{2\mu}, \\
\leq& (1-\frac{\eta\mu}{2})^t(\mathcal{L}_{\mathcal{D}'}(\theta'_0) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{L\eta C}{2\mu}, \\
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta'_T)] - \mathcal{L}^*_{\mathcal{D}'} \overset{\text{Assumption 4.6}}{\leq}& (1-\frac{\eta\mu}{2})^T \ell_{\theta_0} + \frac{L\eta C}{2\mu}.
\end{aligned}$$

Now we analyze the linear convergence of θ''_t on $\mathcal{L}_{\mathcal{D}'}$, leading to a similar result depending on $\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta''_0)] - \mathcal{L}^*_{\mathcal{D}'}$. This allows us to plug in the results from Lemma A.23. We have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta''_K)] - \mathcal{L}^*_{\mathcal{D}'} \leq& (1-\eta\mu)^K(\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta''_0)] - \mathcal{L}^*_{\mathcal{D}'}) + \frac{L\eta C}{2\mu}, \\
=& (1-\eta\mu)^K(\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta_T)] - \mathcal{L}^*_{\mathcal{D}'}) + \frac{L\eta C}{2\mu}, \\
\overset{\text{Lemma A.23}}{\leq}& (1-\eta\mu)^K((1-\frac{\eta\mu}{2})^T(\mathcal{L}_{\mathcal{D}'}(\theta_0) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{C}{4B\mu} + \frac{L\eta C}{\mu}) + \frac{L\eta C}{2\mu}, \\
\overset{\eta \leq \frac{1}{BL}}{\leq}& (1-\frac{\eta\mu}{2})^K((1-\frac{\eta\mu}{2})^T(\mathcal{L}_{\mathcal{D}'}(\theta_0) - \mathcal{L}^*_{\mathcal{D}'}) + \frac{5C}{4B\mu}) + \frac{L\eta C}{2\mu}, \\
\overset{\text{Assumption 4.6}}{\leq}& (1-\frac{\eta\mu}{2})^K((1-\frac{\eta\mu}{2})^T \ell_{\theta_0} + \frac{5C}{4B\mu}) + \frac{L\eta C}{2\mu}.
\end{aligned}$$

Let

$$T = K + \frac{\log(\ell_{\theta_0}) - \log(\frac{5C}{4B\mu})}{\log(\frac{1}{1-\eta\mu/2})},$$

then we have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta''_K)] - \mathcal{L}^*_{\mathcal{D}'} \leq& (\frac{5C}{4B\mu})((1-\frac{\eta\mu}{2})^{2K} + (1-\frac{\eta\mu}{2})^K) + \frac{L\eta C}{2\mu}, \\
\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta'_T)] - \mathcal{L}^*_{\mathcal{D}'} \leq& (\frac{5C}{4B\mu})(1-\frac{\eta\mu}{2})^K + \frac{L\eta C}{2\mu}.
\end{aligned}$$

By quadratic growth (Lemma A.9), we have

$$\begin{aligned}
\mathbb{E}[\|\theta''_K - \theta'_T\|^2] \leq& 2\mathbb{E}[\|\theta''_K - \theta^{*'}\|^2] + 2\mathbb{E}[\|\theta'_T - \theta^{*'}\|^2], \\
\leq& \frac{4}{\mu}(\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta''_K)] - \mathcal{L}_{\mathcal{D}'}) + \frac{4}{\mu}(\mathbb{E}[\mathcal{L}_{\mathcal{D}'}(\theta'_T)] - \mathcal{L}_{\mathcal{D}'}), \\
\leq& \frac{4}{\mu}((\frac{5C}{4B\mu})((1-\frac{\eta\mu}{2})^{2K} + 2(1-\frac{\eta\mu}{2})^K) + \frac{L\eta C}{\mu}),
\end{aligned}$$

where the expectation is taken with respect to the random implementations of the unlearning algorithm producing θ''_K and the learning algorithm θ'_T. Let

$$\Sigma^2 = \frac{5C}{B\mu^2}((1-\frac{\eta\mu}{2})^{2K} + 2(1-\frac{\eta\mu}{2})^K) + \frac{4L\eta C}{\mu^2}.$$

Then by Lemma 4.2, we obtain $(\varepsilon, 2\delta)$-certified unlearning if the noise is scaled as

$$\sigma = \frac{\Sigma}{\varepsilon}\sqrt{\frac{2\log(1.25/\delta)}{\delta}}.$$

□

B Experiments

B.1 Implementation Details

We follow the experimental setup for unlearning detailed in Section 4.1 and Appendix B of Mu and Klabjan [2026], including dataset preparation, model architecture, unlearning procedure, and MIA implementations. We implement PSGD by projecting iterates onto a ball of radius R centered on the origin. We use the hyperparameters listed in Table 2. Code is open-sourced at the anonymous GitHub repository `https://anonymous.4open.science/r/r2d2-3753/`.

All experiments were run using PyTorch 2.5.0 and CUDA 12.1, on an Intel(R) Core(TM) i7-6850K CPU (3.60GHz) with an NVIDIA GeForce GTX 1080 GPU (8 GB VRAM) or on an Intel(R) Xeon(R) Silver 4208 CPU (2.10GHz) with an NVIDIA RTX A6000 GPU (48 GB).

Table 2: R2D Experiment parameters for the eICU and Lacuna-100 datasets.

Parameter	**eICU and MLP**	**Lacuna-100 and ResNet-18**
Size of training dataset n	94449	32000
Number of users	119282	100
Percent data unlearned	$\sim 1\%$	$\sim 2\%$
Number of model parameters d	136386	11160258
Batch size	64	64
L	0.059955	
G	0.820322	
η	0.001	0.01
Number of training epochs	48	43
R	10	50

B.2 Hyperparameters Selection

We conduct additional experiments to ensure appropriate choices of projection radius R and batch size b. For R, following the precedent in Zhang et al. [2024], we implement projected SGD on a ball with radius R large enough to minimize impact on model utility. We assess this by examining the training loss and error for varying choices of R. For these experiments, we train the model (with batch size $b = 64$) and perform model selection of the parameters with lowest validation loss. As shown in Table 3, the choices of $R = 10$ and $R = 50$ for eICU and Lacuna-100 are have minimal impact on model performance.

For the batch size b, we desire a batch size that is small enough to highlight the effects of stochasticity but large enough so that training is stable enough to yield good model performance. We consider the training and test loss for varying choices of b, where for each the model is trained with the same number of *iterations*. Tables 4 and 5 demonstrate the impact of varying batch sizes on the performance on the validation and training sets.

Table 3: Choice of projection radius R vs. model performance for eICU (left) and Lacuna-100 (right).

eICU Dataset		
R	Train Error	Train Loss
1	0.403646	0.679236
2	0.394975	0.647507
5	0.313513	0.588481
10	0.308939	0.581918
15	0.308907	0.581422
20	0.308907	0.581422

Lacuna-100 Dataset		
R	Train Error	Train Loss
20	0.502000	0.693194
30	0.059719	0.148735
40	0.003563	0.013695
50	0.020094	0.058069
60	0.012969	0.045947
70	0.012969	0.045947

Table 4: Choice of batch size b vs. model performance for the eICU dataset.

	Batch Size	Train Loss	Train Error	Validation Loss	Validation Error
5	8	0.416120	0.250000	0.583045	0.308373
4	16	0.513511	0.250000	0.582881	0.308373
3	32	0.493724	0.250000	0.581867	0.307314
2	64	0.551778	0.296875	0.581444	0.306467
1	128	0.557833	0.265625	0.581424	0.306425
0	256	0.588656	0.313111	0.588421	0.309813

Table 5: Choice of batch size b vs. model performance for the Lacuna-100 dataset.

	Batch Size	Train Loss	Train Error	Validation Loss	Validation Error
5	8	0.315790	0.250000	0.244619	0.096125
4	16	0.098530	0.000000	0.184725	0.074375
3	32	0.020336	0.000000	0.201231	0.056375
2	64	0.010567	0.000000	0.220710	0.049125
1	128	0.000255	0.000000	0.342793	0.051500
0	256	0.126300	0.048500	0.230133	0.093125

C Additional Discussion of Related Work

Certified Unlearning. Beyond the general settings considered in this work, certified unlearning algorithms have also been developed for specific settings, such as linear and logistic models in Guo et al. [2020], Izzo et al. [2021], graph neural networks in Chien et al. [2022], minimax problems in Liu et al. [2023], federated learning in Fraboni et al. [2024], adaptive unlearning requests in Gupta et al. [2021], Chourasia and Shah [2023], and online learning in Hu et al. [2025]. These works do not consider general nonconvex settings.

While working on this manuscript, we became aware of Ullah et al. [2021] and Ullah and Arora [2023], works with similarities that are worth clarifying. This work treats certified unlearning algorithms, where the output of unlearning and retraining are each governed by some probability distributions, and these two distributions are (ϵ, δ)-indistinguishable. While this may seem similar to the definition of TV-stability in Ullah et al. [2021] and Ullah and Arora [2023], these works actually treat *exact* unlearning algorithms – i.e., unlearning is exactly equal to retraining from scratch – and they leverage TV-stability in the analysis to achieve this goal. This is reflected in the unlearning algorithm in Ullah et al. [2021], where the parameters, batch samples, and noise are saved at each step, and upon unlearning they revert to the exact step where the deleted sample was used, and recompute additional descent steps. Their computational advantage holds in expectation, with respect to the probability of recomputation, which grows with the number of training iterates T . In contrast, the number of rewinding steps in K R2D does not depend on the exact sample deleted but rather on the computation budget. Therefore, the computational advantage is deterministic, and the model parameters at only one step need to be saved.

Differential Privacy. The concept of certified unlearning is directly built on the existing mathematical framework of differential privacy (DP) developed in Dwork et al. [2006]. DP formalizes privacy guarantees by leveraging noise injection to limit the impact of the inclusion or exclusion of any one data sample on the output of the algorithm. The level of privacy is controlled by the parameters ε and δ, as stated in the definition below.

Definition C.1. (Dwork and Roth [2014]) A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $S \subset Range(\mathcal{M})$ and for all adjacent datasets $\mathcal{D}, \mathcal{D}' \in \mathbb{N}^{|\mathcal{X}|}$

$$\mathbb{P}[\mathcal{M}(\mathcal{D}) \in S] \leq e^{\varepsilon}\mathbb{P}[\mathcal{M}(\mathcal{D}') \in S] + \delta.$$

A common DP technique is the Gaussian mechanism, which scales Gaussian noise to the *deterministic* global sensitivity, a worst-case bound on how much the algorithm output shifts when the underlying

dataset is changed (Dwork and Roth [2014]). This technique has been applied to privatizing deterministic functions such as dataset queries or full-batch gradient descent, as well as *randomized* algorithms such as SGD as in Abadi et al. [2016], Wu et al. [2017], Zhang et al. [2017]. In particular, to bound the sensitivity, the well-known DP-SGD algorithm from Abadi et al. [2016] clips the gradient at each step to minimize the impact of any single data sample. Alternatively, the black-box SGD algorithm from Wu et al. [2017] requires a uniformly bounded gradient and strongly convex loss function.

Biased SGD. Our work relies on interpreting the certified unlearning problem as a form of biased or disturbed gradient descent. There are two complementary viewpoints in the literature. First, several works analyze biased SGD from a classic optimization perspective to achieve convergence to a minimum or stationary point. Typically, these works establish minimal viable assumptions on the bias or noise to achieve convergence. For example, Demidovich et al. [2023] summarizes biased SGD results for nonconvex and strongly convex (or PL) loss functions. In addition, Hu et al. [2016] considers bandit convex optimization with absolutely bounded bias and on a bounded convex domain, and Driggs et al. [2022] considers biased SGD schemes that satisfy a specific "memory-biased" property, like SARAH and SVRG. In our work, we utilize the groundwork laid in Ajalloeian and Stich [2020] to show unlearning for D2D on strongly convex functions.

The second viewpoint of biased SGD is through the lens of nonlinear contraction theory as established in Kozachkov et al. [2023] Sontag [2022], characterizing the bias at every step as disturbances to a noisy gradient system. The trajectory stability, which determines how far a gradient trajectory will diverge when disturbed, can be used to bound the distance between SGD training on two different loss functions. We leverage well-known properties of these systems to analyze stochastic R2D unlearning.

D Limitations

There are several limitations of our work. First, the reliance on the Gaussian mechanism requires the noise to scale with the dimension of the parameter space, which can cause the noise to become prohibitively large for large models. Second, the results for unconstrained SGD require knowledge of the interpolation constant, which may not be easily determined if the model is not overparametrized. The same can be said of constants such as the Lipschitz smoothness constant, which may need to be estimated. However, the fundamental principle of indistinguishability via Gaussian mechanism still holds, even if the degree of certification is less precise.