On Optimization Complexity of Second-Order Certified Unlearning

Anonymous Authors[1]

Abstract

We study machine unlearning: the removal of memorized training data from a trained model. Specifically, we investigate the algorithmic complexity of certified unlearning from an optimization perspective. We formalize the goal of an unlearning algorithm as simultaneously achieving certified unlearning and optimization accuracy. Utilizing the notion of uniformly convex regularizers, we prove new bounds on the distance between initial and unlearned models using a novel substitute for generalization error. Thus we theoretically demonstrate that if the removed data is well-predicted by the unlearned model, the corresponding optimization problem is simple. Furthermore, we develop a new second-order unlearning algorithm with an anisotropic Gaussian mechanism and state-of-the-art global convergence. We prove fast rates for our method in achieving certified unlearning for linear models with quasi-self-concordant losses. As a direct application, our theory covers unlearning for logistic and exponential regressions and shows a provable benefit of utilizing second-order information compared to first-order unlearning methods.

1. Introduction

1.1. Certified Unlearning and Optimization

A crucial requirement for modern AI systems is the ability to *unlearn* the data, that is to allow the data provider (e.g., an individual user or an institutional organization) to select *which data is no longer allowed to use*. To satisfy ethical and legal conditions, such prohibitions must be strictly enforced. At the same time, retraining a large model from scratch to forget a certain small portion of data is often very expensive or even impossible. Therefore, we are interested in efficient practical algorithms that enable removing the data from the model without retraining from scratch. In this work, we study the complexity of the machine unlearning problem through the lens of optimization theory. We develop a new *second-order unlearning algorithm*, equipped with state-of-the-art fast global convergence rates and a strong theoretically certified unlearning guarantee, which ensures trustworthy models.

Let us denote by A_N the entire dataset of size N on which we have trained some machine learning model $\boldsymbol{x}^\star_F$. The goal of unlearning is to forget the subset of the training data A_N from our trained model $\boldsymbol{x}^\star_F$. Denote by $A_n \subset A_N$ the subset of A_N of size n that remains after unlearning, frequently called the *retain set*. Therefore, our goal is to forget the data $A_N \setminus A_n$ of size $m = N - n$, and typically $m \ll N$ (removing a small portion of data).

Before specifying the certified unlearning precisely, first let us fix a *learning algorithm*. In this work, we analyze the unlearning problem from the optimization perspective. Therefore, the process of learning consists of *empirical risk minimization* (ERM). We denote by $F_\psi(\cdot)$ the corresponding ERM objective on the whole dataset A_N that we use to train an initial model, and by $f_\psi(\cdot)$ on the retain data A_n (see definition of optimization formulations in Section 2).

We assume that our initial model $\boldsymbol{x}^\star_F$ is an exact solution to the ERM objective on the entire dataset A_N

$$\boldsymbol{x}^\star_F \quad := \quad \arg\min_{\boldsymbol{x}} F_\psi(\boldsymbol{x}). \tag{1}$$

This assumption follows the previous literature (Sekhari et al., 2021). It provides a strong theoretical baseline, while it can be relaxed. It will also be convenient to define the precise solution on the retain data A_n

$$\boldsymbol{x}^\star_f \quad := \quad \arg\min_{\boldsymbol{x}} f_\psi(\boldsymbol{x}). \tag{2}$$

Knowing $\boldsymbol{x}^\star_f$ provides an ideal solution to the unlearning problem, and we do not assume knowledge of $\boldsymbol{x}^\star_f$.

We define an *unlearning algorithm* as a randomized procedure $\boldsymbol{y} = \mathbb{U}(\boldsymbol{x}, A_N \setminus A_n, A_N)$ that takes as input, correspondingly, a model $\boldsymbol{x}$, the data $A_N \setminus A_n$ that we want to forget, and the whole dataset A_N, or a set of possible statistics from it. It returns a new unlearned model $\boldsymbol{y}$.

We use the following formal definition of *certified unlearning*. Let us fix a desired level $q \in (0, 1)$ and small $\delta > 0$.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

We say that $\mathbb{U}$ is (q, δ)-*unlearning* if for any measurable set Y:

$$\begin{aligned} \mathbb{P}\big(\boldsymbol{y} \in Y\big) &\leq e^q \cdot \mathbb{P}\big(\boldsymbol{y}^\star \in Y\big) + \delta \\ \mathbb{P}\big(\boldsymbol{y}^\star \in Y\big) &\leq e^q \cdot \mathbb{P}\big(\boldsymbol{y} \in Y\big) + \delta \end{aligned} \tag{3}$$

where $\boldsymbol{y} := \mathbb{U}(\boldsymbol{x}_F^\star, A_N \setminus A_n, A_N)$ is a typical use of the unlearning algorithm, starting from the trained model $\boldsymbol{x}_F^\star$ on the entire dataset, and $\boldsymbol{y}^\star := \mathbb{U}(\boldsymbol{x}_f^\star, \varnothing, A_n)$ is the *idealized theoretical use*, as if we train the model from scratch to obtain $\boldsymbol{x}_f^\star$. Note that this definition is aligned with the standard ones from the literature (Koloskova et al., 2025; Ginart et al., 2019; Guo et al., 2020). It is also closely related to the notion of *differentiable privacy* (Dwork & Roth, 2014; Feldman et al., 2022).

The naive mechanism to ensure (3) consists of adding Gaussian noise to the output model of a training procedure (Dwork & Roth, 2014). In principle, *any training algorithm* can be turned into (q, δ)-unlearning, if the variance of the added noise is sufficiently large. However, such noise may erase not only the information to be forgotten, but also useful information about the retained data A_n, yielding a model that is far from the desired solution (2). Hence, in this work, along with the certified unlearning guarantee (3), we require *the output $\boldsymbol{y}$ to be close to the exact minimum*:

$$\mathbb{E}\big[\|\boldsymbol{y} - \boldsymbol{x}_f^\star\|\big] \leq \epsilon, \tag{4}$$

for a desired optimization tolerance $\epsilon > 0$.

1.2. Contributions

We establish rigorous complexity bounds for optimization algorithms to simultaneously achieve both *certified unlearning* (3) and *optimization* (4) guarantees, for a given triplet (q, δ, ϵ) of parameters specifying the problem. To the best of our knowledge, we are the first to establish fast global convergence of second-order (Newton-type) methods for unlearning, when the distance $\|\boldsymbol{x}_F^\star - \boldsymbol{x}_f^\star\|$ between models can be arbitrarily far and propose a novel *anisotropic Gaussian mechanism*, which is well-suitable for the geometry of the problem. We summarize contributions as follows:

- Under assumption of uniform convexity of a model regularizer (see definition (6); a particular case is ℓ_2-regularization), we show how to relate the distance between models $\boldsymbol{x}_F^\star$ and $\boldsymbol{x}_f^\star$, and the functional residual with a quantity $\pi(\boldsymbol{x})$ representing *prediction error* of a model $\boldsymbol{x}$ on unlearned data (Lemma 2.1). Thus, we show that if the unlearned model generalizes well on the removed data, then the distance is small, and the corresponding optimization problem is simple.
- We developed a *new certified second-order unlearning method* (Algorithm 1). It is based on computing the Hessian of the empirical loss, and using the state-of-the-art globalization of Newton's method for fast convergence even if $\boldsymbol{x}_F^\star$ and $\boldsymbol{x}_f^\star$ are far from each other. To ensure the certified unlearning, we develop a novel mechanism of adding *anisotropic* Normal distribution, which is naturally aligned with the Hessian.
- Under the smoothness condition of *quasi-self-concordance* of the loss (Bach, 2010; Sun & Tran-Dinh, 2019; Karimireddy et al., 2018; Doikov, 2025), we show fast global convergence for our algorithm (Theorem 3.2). To the best of our knowledge, our complexity bound is also new to the optimization literature. A working example that satisfies all our theoretical assumptions is *logistic* or *exponential* regression with linear models, augmented by *any uniformly convex regularizer*. We prove the certified unlearning guarantee for our algorithm in Theorem 3.3.

1.3. Notation

We fix some positive definite symmetric matrix $\boldsymbol{B} \in \mathbb{R}^{d \times d}$, and define the primal-dual pair of generalized Euclidean norms, for any $\boldsymbol{x}, \boldsymbol{s} \in \mathbb{R}^d$:

$$\|\boldsymbol{x}\| := \langle \boldsymbol{B}\boldsymbol{x}, \boldsymbol{x}\rangle^{1/2}, \quad \|\boldsymbol{s}\|_* = \langle \boldsymbol{s}, \boldsymbol{B}^{-1}\boldsymbol{s}\rangle^{1/2}. \tag{5}$$

We use the dual norm to measure the size of the gradients. In the simplest case, we can set $\boldsymbol{B} := \boldsymbol{I}$ (identity matrix), which recovers the standard Euclidean norm. In general, matrix $\boldsymbol{B} = \boldsymbol{B}^\top \succ 0$ allows to better capture the geometry of the problem (see Proposition A.1). We also use $\boldsymbol{B}$ for our novel sampling mechanism that ensures certified unlearning.

We say that a convex, not necessary differentiable, regularizing function $\psi : \operatorname{dom}\psi \to \mathbb{R}$, is *uniformly convex* of degree $p \geq 2$ with constant $\mu > 0$ (see, e.g., Chapter 4.2.2 in (Nesterov, 2018)) if the symmetrized Bregman divergence is bounded as follows, for all $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom}\psi$:

$$\bar{\beta}_\psi(\boldsymbol{x}; \boldsymbol{y}) := \langle \psi'(\boldsymbol{x}) - \psi'(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \geq \mu\|\boldsymbol{x} - \boldsymbol{y}\|^p, \tag{6}$$

where $\psi'(\boldsymbol{x}) \in \partial\psi(\boldsymbol{x})$ is an arbitrary selection of subgradients. Uniformly convex functions of degree $p = 2$ are called *strongly convex*.

2. Optimization Problem

Let us consider the unlearning problem from the optimization perspective. We denote by F the initial objective of *training on the full dataset*, augmenting it with a possible simple[1] regularizer ψ:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \Big[F_\psi(\boldsymbol{x}) = F(\boldsymbol{x}) + \psi(\boldsymbol{x}) \Big]. \tag{7}$$

[1] Namely, we assume that we can efficiently solve a second-order subproblem involving ψ in our algorithm. The main example is ℓ_2-regularization: $\psi(x) = \frac{\mu}{2}\|x\|^2$, which satisfies (6) with $p = 2$. Morevore, we can cover simple constraints in our model.

We denote by $\boldsymbol{x}_F^\star \in \mathbb{R}^d$ a solution to the initial training problem (7), which is available to us. We use $\boldsymbol{x}_F^\star$ as a *starting point* for our unlearning algorithms. For simplicity, we assume that $\boldsymbol{x}_F^\star$ is an exact minimizer to (7), while this assumption can be relaxed, using an approximate solution.

Further, we have the following decomposition of the initial objective, for some $0 \leq \gamma \leq 1$:

$$F(\boldsymbol{x}) \quad = \quad (1-\gamma) f(\boldsymbol{x}) + \gamma u(\boldsymbol{x}), \tag{8}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is the loss on *core data*, that we keep in the dataset, and $u : \mathbb{R}^d \to \mathbb{R}$ is the loss on *forget data*, that we are required to remove from training. We assume that all training components, $F(\cdot)$, $f(\cdot)$, and $u(\cdot)$ are differentiable functions, while the regularizer ψ can be non-differentiable (e.g. a mixture of ℓ_2 and ℓ_1-regularizers, or indicator of convex constraints). Our main optimization objective consists of minimizing the following function on retain data:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \Big[f_\psi(\boldsymbol{x}) \quad = \quad f(\boldsymbol{x}) + \psi(\boldsymbol{x}) \Big]. \tag{9}$$

We denote by $\boldsymbol{x}_f^\star \in \mathbb{R}^d$ a solution to (9), which we want to find. To characterize the global complexity of solving (9), we introduce the quantity:

$$\pi(\boldsymbol{x}) \quad := \quad \|\nabla u(\boldsymbol{x}) + \psi'(\boldsymbol{x})\|_*,$$

which has an interpretation of a *prediction error* of a model $\boldsymbol{x}$ on the forget data ∇u. In what follows, we show that this quantity plays the main role in the *optimization complexity of algorithmic unlearning*. We can relate $\pi(\boldsymbol{x}_f^\star)$ to the distance between solutions of two problems (7) and (9), and $\pi(\boldsymbol{x}_F^\star)$ to the functional residual:

Lemma 2.1. *Let objectives f and F be convex with the relation defined in (8). Then,*

$$\bar{\beta}_\psi(\boldsymbol{x}_F^\star; \boldsymbol{x}_f^\star) \cdot \|\boldsymbol{x}_F^\star - \boldsymbol{x}_f^\star\|^{-1} \quad \leq \quad \gamma \pi(\boldsymbol{x}_f^\star), \tag{10}$$

For uniformly convex regularizers (6), we have

$$\|\boldsymbol{x}_F^\star - \boldsymbol{x}_f^\star\| \quad \leq \quad \left[\tfrac{\gamma}{\mu} \pi(\boldsymbol{x}_f^\star) \right]^{\frac{1}{p-1}} \tag{11}$$

and the bound for the functional residual:

$$f_\psi(\boldsymbol{x}_F^\star) - f_\psi(\boldsymbol{x}_f^\star) \;\leq\; \tfrac{p-1}{p} \left[\tfrac{\gamma}{(1-\gamma)\mu^{1/p}} \pi(\boldsymbol{x}_F^\star) \right]^{\frac{p}{p-1}}. \tag{12}$$

Lemma 2.1 shows that if the unlearned model works well on the removed data (in other words, the removed data does not affect the generalization of the model and falls well within the distribution of the retain data), then the distance between the minimizers is also small. We also conclude that the distance in the left hand side of (11) and the functional residual (12) can be controlled by varying parameter γ and μ. We can make the distance between the minimizers small both by setting $\gamma \to 0$ or by increasing the regularization parameter $\mu > 0$.

2.1. Examples and Assumptions

Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^d$ represents the *core data*, and $\boldsymbol{a}_{n+1}, \ldots, \boldsymbol{a}_{n+m} \in \mathbb{R}^d$ is the *forget data*. We denote by $N = n + m$ the initial dataset size, and typically $n \gg m$ (only a very small portion of the data is getting removed).

Example 1 (Linear Models)**.** Let $\ell : \mathbb{R} \to \mathbb{R}$ be a loss function. Consider the objective of training the generalized linear models:

$$F(\boldsymbol{x}) \quad := \quad \tfrac{1}{N} \sum_{i=1}^N \ell(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle),$$

and the corresponding core and forget components:

$$\begin{aligned} f(\boldsymbol{x}) \quad &:= \quad \tfrac{1}{n} \sum_{i=1}^n \ell(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle), \\ u(\boldsymbol{x}) \quad &:= \quad \tfrac{1}{m} \sum_{j=n+1}^{n+m} \ell(\langle \boldsymbol{a}_j, \boldsymbol{x} \rangle). \end{aligned} \tag{13}$$

Then, decomposition (8) holds with $\boxed{\gamma = \tfrac{m}{N}}$. Note that $\gamma \to 0$ when $N \gg m$, which is a common scenario.

We assume that the loss function $\ell : \mathbb{R} \to \mathbb{R}$ is convex, differentiable, and sufficiently smooth. In particular, we use the following notion of smoothness which captures the local geometry of the objective (Bach, 2010):

Assumption 2.2 (Quasi-Self-Concordance)**.** Assume that it holds, for some constant $M \geq 0$:

$$|\ell'''(t)| \quad \leq \quad M \cdot \ell''(t), \qquad t \in \mathbb{R}. \tag{14}$$

Example 2 (Quasi-Self-Concordant Losses)**.** The following functions satisfy assumption (14):

- *Quadratic loss*, $\ell(t) = \frac{1}{2} t^2$. Then $M = 0$.
- *Exponential loss*, $\ell(t) = e^t$. Then $M = 1$.
- *Logistic loss*, for classification of two classes, which can be written as, $\ell(t) = \log(1 + e^t)$. Then $M = 1$.
- The previous example can be generalized to the *multi-class logistic loss*, with the same constant $M = 1$.

The parameter $M \geq 0$ measures how far the loss is from quadratic function. Then, to capture the geometry of the retained model (13), we define

$$\boxed{\boldsymbol{B} \quad := \quad \sum_{i=1}^n \boldsymbol{a}_i \boldsymbol{a}_i^\top,} \tag{15}$$

which defines the global norm (5) and anisotropic Gaussian mechanism that we use in the method. This choice ensures data-agnostic second-order approximation of the objective (Doikov, 2025). For non-linear models, a suitable choice is the Hessian at the initialization: $\boldsymbol{B} := \nabla^2 f(\boldsymbol{x}_0)$, which approximates (15).

3. Algorithm

At each iteration $k \geq 0$ of our algorithm, we use a positive definite matrix $\boldsymbol{H}_k = \boldsymbol{H}_k^\top \succ 0$ that is designed to capture a second-order information about objective and accelerate the global unlearning. Using this matrix, we define the second-order model m_k around $\boldsymbol{x}_k$, with regularizer:

$$m_k(\boldsymbol{x}) := \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} \rangle + \tfrac{1}{2}\langle \boldsymbol{H}_k(\boldsymbol{x} - \boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \psi(\boldsymbol{x}).$$

We discuss how to choose the matrix $\boldsymbol{H}_k$ in the next sections. In our algorithm, we minimize this model for $K \geq 1$ iterations, starting from the minimum $\boldsymbol{x}_F^\star$ of the full model (7). After that, to reach a certified unlearning guarantee, we apply anisotropic Gaussian mechanism to the output.

Algorithm 1 Certified Second-Order Unlearning

Init: Set $\boldsymbol{x}_0 := \boldsymbol{x}_F^\star$. Fix $\boldsymbol{B} \succ 0$, $\sigma > 0$, and $K \geq 1$.
1: **For** $k = 0 \dots K-1$ **do**
2: Compute next step $\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \big[m_k(\boldsymbol{x})\big]$
3: **End for**
4: Sample a normal vector $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{B}^{-1})$
5: **Return** $\boldsymbol{y}_K = \boldsymbol{x}_K + \boldsymbol{\xi}$

For convenience, we denote the norm of the current (sub)gradient at each iterate by $g_k := \|\nabla f(\boldsymbol{x}_k) + \psi'(\boldsymbol{x}_k)\|_*$. We analyze two instances of our algorithm.

3.1. Gradient Method Baseline

This is the first-order baseline of our approach, which selects $\boldsymbol{B} := \boldsymbol{I}$ (isotropic Gaussian noise) and $\boldsymbol{H}_k := \frac{\eta}{2}\boldsymbol{I}$, for a certain step-size parameter $\eta > 0$. In case $\psi(x) \equiv 0$ (no regularization), iterations of our algorithm read as

$$\boldsymbol{x}_{k+1} \quad = \quad \boldsymbol{x}_k - \tfrac{1}{\eta}\nabla f(\boldsymbol{x}_k),$$

which is the standard gradient descent. For general ψ, each iteration can be represented through the prox operator of ψ.

Theorem 3.1. *Assume that the second derivative of the loss is bounded: $L \geq \ell''(t), \forall t$. We set $\eta := L\|\boldsymbol{A}\|^2$, where $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is the matrix of the retained data. Then, we achieve $\|\boldsymbol{x}_K - \boldsymbol{x}_f^\star\| \leq \varepsilon$ after the following number of the gradient steps, for $p > 2$:*

$$K \quad = \quad O\Big(\tfrac{L\|\boldsymbol{A}\|^2}{\mu^{2/p}} \Big[\big(\tfrac{1}{\mu\varepsilon^p}\big)^{\frac{p-2}{p}} - \big(\tfrac{\mu^{1/p}}{\gamma\pi(\boldsymbol{x}_F^\star)}\big)^{\frac{p-2}{p-1}} \Big] \Big), \tag{16}$$

and, for $p = 2$ (strongly convex case):

$$K \quad = \quad O\Big(\tfrac{L\|\boldsymbol{A}\|^2}{\mu} \ln \tfrac{\gamma\pi(\boldsymbol{x}_F^\star)}{\mu\varepsilon} \Big). \tag{17}$$

3.2. Newton Method with Gradient Regularization

In our most advanced second-order instance of the algorithm, we choose, as in (Doikov et al., 2024; Doikov, 2025):

$$\boxed{\boldsymbol{H}_k \quad := \quad \nabla^2 f(\boldsymbol{x}_k) + M g_k \boldsymbol{B}} \tag{18}$$

where M is a quasi-self-concordant parameter of the loss, and $\boldsymbol{B} = \sum_{i=1}^n \boldsymbol{a}_i \boldsymbol{a}_i^\top$ (see Proposition A.1). The most important case is when $\psi(\boldsymbol{y}) = \frac{\mu}{2}\|\boldsymbol{y}\|^2$ (strongly-convex regularizer, $p = 2$). Then, each iteration of our method can be written explicitly, as follows, for $k \geq 0$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \Big(\nabla^2 f(\boldsymbol{x}_k) + (M g_k + \mu)\boldsymbol{B}\Big)^{-1} \nabla f_\psi(\boldsymbol{x}_k),$$

and for $M = 0$ this is the classical Newton method as applied to (9). By employing the gradient regularization, we ensure *fast global convergence* of our algorithm, as shown in the following theorem. Note that for a general regularizer ψ, the model $m_k(\boldsymbol{y})$ is strongly convex due (18), and can be solved efficiently by first-order optimization subroutines, without additional data samples.

Theorem 3.2. *For quasi-self-concordant loss and uniformly convex regularizer, we achieve $\|\boldsymbol{x}_K - \boldsymbol{x}_f^\star\| \leq \varepsilon$, after the following number of the Newton steps with (18):*

$$K \quad = \quad O\Big(M \Big[\tfrac{\gamma}{\mu}\pi(\boldsymbol{x}_F^\star)\Big]^{1/p} + \log \tfrac{\gamma\pi(\boldsymbol{x}_F^\star)}{\mu\varepsilon^{p-1}} \Big). \tag{19}$$

We see that, in contrast to the gradient method, the complexity (19) of the Newton method is much better, as $\varepsilon > 0$ enters as an additive logarithmic term. Moreover, the rate of the Newton method does not depend on the size of the input data $\|\boldsymbol{A}\|^2$ as in (16) and (17). The main complexity factor is the first term, which is small when either $\gamma \to 0$ (unlearning a small portion of data) or $\pi(\boldsymbol{x}_F^\star)$ is small (good prediction of the full model on unlearned data).

3.3. Certified Unlearning Guarantee

Applying anisotropic Gaussian mechanism in the output of Algorithm 1, we are able to prove our main result on certified unlearning.

Theorem 3.3. *Let $\boldsymbol{B} \succ 0$ be arbitrary and assume that $\boldsymbol{x}_K$ satisfies the optimization guarantee $\|\boldsymbol{x}_K - \boldsymbol{x}_f^\star\| \leq \varepsilon$. For any $q \in (0,1)$ and $\delta > 0$, set*

$$\sigma \quad := \quad \tfrac{\varepsilon}{q}\max\Big\{1, 2\sqrt{2\ln\tfrac{2}{\delta}}\Big\}. \tag{20}$$

Then, Algorithm 1 ensures certified (q,δ)-unlearning. Moreover, the result satisfies the optimization guarantee:

$$\mathbb{E}\Big[\|\boldsymbol{y}_K - \boldsymbol{x}_f^\star\|\Big] \quad \leq \quad \sqrt{\varepsilon^2 + \sigma^2 d}. \tag{21}$$

Corollary 3.4. *For a given triplet of parameters (q,δ,ϵ), Algorithm 1 achieves (q,δ)-unlearning (3) and ϵ-bound for optimization guarantee (4) in total of*

$$K = O\Big(M \Big[\tfrac{\gamma}{\mu}\pi(\boldsymbol{x}_F^\star)\Big]^{1/p} + \log\Big[\tfrac{\gamma\pi(\boldsymbol{x}_F^\star)}{\mu}\Big(\tfrac{d\ln 1/\delta}{q\epsilon^2}\Big)^{\frac{p-1}{2}}\Big]\Big)$$

iterations (retained data passes). Therefore, we see that all key parameters (q,δ,ϵ) enter under the logarithm, and the efficiency depends mainly on γ (the portion of removed data), $\mu > 0$ (regularization coefficient), and $\pi(\boldsymbol{x}_F^\star)$.

References

Bach, F. Self-concordant analysis for logistic regression. 2010.

Doikov, N. Minimizing quasi-self-concordant functions by gradient regularization of Newton method. *Mathematical Programming*, pp. 1–39, 2025.

Doikov, N., Mishchenko, K., and Nesterov, Y. Super-universal regularized Newton method. *SIAM Journal on Optimization*, 34(1):27–56, 2024.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Feldman, V., McMillan, A., and Talwar, K. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 954–964. IEEE, 2022.

Ginart, A., Guan, M., Valiant, G., and Zou, J. Y. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. In *International Conference on Machine Learning*, pp. 3832–3842. PMLR, 2020.

Karimireddy, S. P., Stich, S. U., and Jaggi, M. Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. *arXiv preprint arXiv:1806.00413*, 2018.

Koloskova, A., Allouah, Y., Jha, A., Guerraoui, R., and Koyejo, S. Certified unlearning for neural networks. In *International Conference on Machine Learning*, pp. 31275–31298. PMLR, 2025.

Nesterov, Y. *Lectures on convex optimization*, volume 137. Springer, 2018.

Nikolov, A., Talwar, K., and Zhang, L. The geometry of differential privacy: the sparse and approximate cases. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 351–360, 2013.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Sun, T. and Tran-Dinh, Q. Generalized self-concordant functions: a recipe for Newton-type methods. *Mathematical Programming*, 178(1-2):145–213, 2019.

A. Appendix

The main consequence of condition (14) that we use is the following bound for the Hessian of the linear models (see Lemma 2 in (Doikov, 2025) for the proof):

Proposition A.1. *Consider generalized linear models (13) with quasi-self-concordant loss (14). Then,*

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})\|_* \quad \leq \quad M\|\boldsymbol{y}-\boldsymbol{x}\|_{\boldsymbol{x}}^2 \cdot \varphi(M\|\boldsymbol{y}-\boldsymbol{x}\|), \tag{22}$$

where $\varphi(t) := \frac{e^t-1-t}{t^2} > 0$ is a convex monotone function, $\|\boldsymbol{y}-\boldsymbol{x}\|_{\boldsymbol{x}} = \langle \nabla^2 f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle^{1/2}$ is the local norm induced by the Hessian of the unlearning problem, and

$$\boldsymbol{B} \quad = \quad \sum_{i=1}^n \boldsymbol{a}_i \boldsymbol{a}_i^\top,$$

as in (15), defines the global norm (5) that we use in the method.

A.1. Proof of Lemma 2.1

The optimality condition for the minimizer $\boldsymbol{x}_F^\star$ of the initial model (7) is

$$\nabla F(\boldsymbol{x}_F^\star) + \psi'(\boldsymbol{x}_F^\star) \quad = \quad 0 \qquad \text{where} \qquad \psi'(\boldsymbol{x}_F^\star) \quad \in \quad \partial\psi(\boldsymbol{x}_F^\star). \tag{23}$$

At the same time, the optimality condition for the minimizer $\boldsymbol{x}_f^\star$ of the unlearned model (9) is

$$\nabla f(\boldsymbol{x}_f^\star) + \psi'(\boldsymbol{x}_f^\star) \quad = \quad 0 \qquad \text{where} \qquad \psi'(\boldsymbol{x}_f^\star) \quad \in \quad \partial\psi(\boldsymbol{x}_f^\star). \tag{24}$$

By convexity of F, we have

$$\begin{aligned}
\bar{\beta}_\psi(\boldsymbol{x}_F^\star; \boldsymbol{x}_f^\star) \quad &= \quad \langle \psi'(\boldsymbol{x}_F^\star) - \psi'(\boldsymbol{x}_f^\star), \boldsymbol{x}_F^\star - \boldsymbol{x}_f^\star \rangle \\
&\leq \quad \langle \nabla F(\boldsymbol{x}_F^\star) + \psi'(\boldsymbol{x}_F^\star) - \nabla F(\boldsymbol{x}_f^\star) - \psi'(\boldsymbol{x}_f^\star), \boldsymbol{x}_F^\star - \boldsymbol{x}_f^\star \rangle \\
&\overset{(23)}{=} \quad \langle \nabla F(\boldsymbol{x}_f^\star) + \psi'(\boldsymbol{x}_f^\star), \boldsymbol{x}_f^\star - \boldsymbol{x}_F^\star \rangle \\
&\overset{(8)}{=} \quad (1-\gamma)\langle \nabla f(\boldsymbol{x}_f^\star) + \psi'(\boldsymbol{x}_f^\star), \boldsymbol{x}_f^\star - \boldsymbol{x}_F^\star \rangle + \gamma \langle \nabla u(\boldsymbol{x}_f^\star) + \psi'(\boldsymbol{x}_f^\star), \boldsymbol{x}_f^\star - \boldsymbol{x}_F^\star \rangle \\
&\overset{(24)}{=} \quad \gamma \langle \nabla u(\boldsymbol{x}_f^\star) + \psi'(\boldsymbol{x}_f^\star), \boldsymbol{x}_f^\star - \boldsymbol{x}_F^\star \rangle \quad \leq \quad \gamma \|\nabla u(\boldsymbol{x}_f^\star) + \psi'(\boldsymbol{x}_f^\star)\|_* \cdot \|\boldsymbol{x}_f^\star - \boldsymbol{x}_F^\star\|.
\end{aligned}$$

Rearranging the terms gives (10).

Then, applying the uniform convexity (6) to (10) provides with (11).

Finally, since $f_\psi(\boldsymbol{x}) = f(\boldsymbol{x}) + \psi(\boldsymbol{x})$ is uniformly convex as a sum of a convex function f and uniformly convex regularizer ψ, by simple integration we obtain, for any $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom}\psi$ and $f'_\psi(\boldsymbol{x}) \in \partial f(\boldsymbol{x})$:

$$f_\psi(\boldsymbol{y}) \quad \geq \quad f_\psi(\boldsymbol{x}) + \langle f'_\psi(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle + \tfrac{\mu}{p}\|\boldsymbol{y}-\boldsymbol{x}\|^p. \tag{25}$$

Minimizing the left and the right hand side independently with respect to y gives:

$$f_\psi(\boldsymbol{x}) - f_\psi(\boldsymbol{x}_f^\star) \quad \leq \quad \tfrac{p-1}{p} \frac{\|f'_\psi(\boldsymbol{x})\|^{\frac{p}{p-1}}}{\mu^{\frac{1}{p-1}}}, \qquad \boldsymbol{x} \in \operatorname{dom}\psi. \tag{26}$$

It remains to substitute $\boldsymbol{x} := \boldsymbol{x}_F^\star$ and notice that, due to (8), $\|f'_\psi(\boldsymbol{x}_F^\star)\|_* = \frac{\gamma}{1-\gamma}\|\nabla u(\boldsymbol{x}_F^\star) + \psi'(\boldsymbol{x}_F^\star)\|_* = \pi(\boldsymbol{x}_F^\star)$, which completes the proof. □

A.2. Proof of Theorem 3.2

Let us denote by $f_k := f_\psi(\boldsymbol{x}_k) - f_\psi(\boldsymbol{x}_f^\star) \geq 0$ the functional residual at iteration $k \geq 0$. Then, for one step of the Newton method with gradient regularization, we have the following progress (see Theorem 3.2 in (Doikov, 2025) for $(*)$), employing additionally uniform convexity of the regularizer:

$$f_k - f_{k+1} \overset{(*)}{\geq} \frac{1}{2M}\left[\frac{g_{k+1}}{g_k}\right]^2 g_k \overset{(26)}{\geq} c \cdot \left[\frac{g_{k+1}}{g_k}\right]^2 f_k^{\frac{p-1}{p}}, \tag{27}$$

where $c := \frac{\mu^{1/p}}{2M}\left[\frac{p}{p-1}\right]^{\frac{p-1}{p}}$.

Note that due to concavity of $\varphi(t) = t^{1/p}$, we have, for any $a, b > 0$:

$$b^{1/p} \leq a^{1/p} + \tfrac{1}{p}a^{-\frac{p-1}{p}}(b-a) \qquad \Leftrightarrow \qquad a^{1/p} - b^{1/p} \geq \tfrac{1}{p}a^{-\frac{p-1}{p}}(a-b). \tag{28}$$

Therefore, we have

$$f_k^{1/p} - f_{k+1}^{1/p} \overset{(28)}{\geq} \tfrac{1}{p}f_k^{-\frac{p-1}{p}}(f_k - f_{k+1}) \overset{(27)}{\geq} \tfrac{c}{p}\left[\frac{g_{k+1}}{g_k}\right]^2. \tag{29}$$

Telescoping this progress for the first $K \geq 1$ iterations, and using the inequality between arithmetic and geometric means, we get

$$\begin{aligned} f_0^{1/p} - f_K^{1/p} &\overset{(29)}{\geq} \tfrac{cK}{p} \cdot \tfrac{1}{K}\sum_{i=0}^{K-1}\left[\frac{g_{i+1}}{g_i}\right]^2 \geq \tfrac{cK}{p} \cdot \left[\prod_{i=0}^{K-1}\frac{g_{i+1}}{g_i}\right]^{2/K} = \tfrac{cK}{p} \cdot \left[\frac{g_K}{g_0}\right]^{2/K} \\ &= \tfrac{cK}{p} \cdot \exp\left(\tfrac{2}{K}\log\tfrac{g_K}{g_0}\right) \geq \tfrac{cK}{p} \cdot \left(1 + \tfrac{2}{K}\log\tfrac{g_K}{g_0}\right). \end{aligned} \tag{30}$$

Using that $f_K \geq 0$ and rearranging the terms, we obtain

$$K \overset{(30)}{\leq} \tfrac{p}{c}f_0^{1/p} + 2\log\tfrac{g_0}{g_K} = \tfrac{2pM}{\mu^{1/p}}\left[\tfrac{p-1}{p}\right]^{\frac{p-1}{p}} f_0^{1/p} + 2\log\tfrac{g_0}{g_K}. \tag{31}$$

Assuming that $\|\boldsymbol{x}_K - \boldsymbol{x}_f^\star\| \geq \varepsilon$ and using uniform convexity, we obtain the lower bound on the current (sub)gradient norm:

$$g_K = \|f'_\psi(\boldsymbol{x}_K)\|_* \overset{(26),(25)}{\geq} \frac{\mu}{(p-1)^{\frac{p-1}{p}}}\|\boldsymbol{x}_K - \boldsymbol{x}_f^\star\|^{p-1} \geq \frac{\mu}{(p-1)^{\frac{p-1}{p}}}\varepsilon^{p-1}.$$

Substituting this bound into (31), and using the bound on the initial functional residual:

$$f_0 = f_\psi(\boldsymbol{x}_F^\star) - f_\psi(\boldsymbol{x}_f^\star) \overset{(12)}{\leq} \tfrac{p-1}{p}\left[\tfrac{\gamma}{(1-\gamma)\mu^{1/p}}\pi(\boldsymbol{x}_F^\star)\right]^{\frac{p}{p-1}} \tag{32}$$

completes the proof. □

A.3. Proof of Theorem 3.1

Since for analyzing the gradient method, we assume that the loss function has bounded second derivative: $\ell''(t) \leq L, \forall t$, we conclude that the smooth part f of the objective in (9) has Lipschitz continuous gradient with constant $L_f := L\|\boldsymbol{A}\|^2$, where $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is the matrix composed by the retaining data $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{R}^d$. In the gradient method, we use the classical choice of the stepsize parameter (see, e.g. (Nesterov, 2018)), as

$$\boxed{\eta := L_f = L\|\boldsymbol{A}\|^2.} \tag{33}$$

Optimality condition for one method step, taking into account the regularizer ψ, is

$$\nabla f(\boldsymbol{x}_k) + \eta(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) + \psi'(\boldsymbol{x}_{k+1}) = 0, \qquad \psi'(\boldsymbol{x}_{k+1}) \in \partial\psi(\boldsymbol{x}_{k+1}). \tag{34}$$

Therefore, using Lipschitzness of the graduent, we conclude that

$$L_f\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| \geq \|\nabla f(\boldsymbol{x}_{k+1}) - \nabla f(\boldsymbol{x}_k)\| \overset{(34),(33)}{=} \|f'_\psi(\boldsymbol{x}_{k+1}) + L_f(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\|.$$

Taking square of both sides and rearranging the terms, we obtain

$$\langle f'_\psi(\boldsymbol{x}_{k+1}), \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle \quad \geq \quad \tfrac{1}{2L_f}\|f'_\psi(\boldsymbol{x}_{k+1})\|^2 \quad = \quad \tfrac{1}{2L_f} g_{k+1}^2. \tag{35}$$

Then, for the functional residual $f_k := f_\psi(\boldsymbol{x}_k) - f_\psi(\boldsymbol{x}_f^\star) \geq 0$ and employing the uniform convexity of the regularizer, we obtain the recurrence:

$$f_k - f_{k+1} \quad \geq \quad \langle f'_\psi(\boldsymbol{x}_{k+1}), \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle \overset{(35)}{\geq} \tfrac{1}{2L_f} g_{k+1}^2 \overset{(26)}{\geq} c \cdot f_{k+1}^{\frac{2(p-1)}{p}} \quad = \quad c \cdot f_{k+1}^{\alpha}, \tag{36}$$

where $c := \left(\frac{p}{p-1}\right)^{\frac{2(p-1)}{p}} \frac{\mu^{2/p}}{2L_f}$ and $\alpha := \frac{2(p-1)}{p} \in [1, 2)$.

Note that the function $\varphi(t) = t^{\alpha-1}$ is concave. Thus, for any $a, b > 0$:

$$b^{\alpha-1} \quad \leq \quad a^{\alpha-1} + (\alpha-1)a^{\alpha-2}(b-a) \qquad \Leftrightarrow \qquad \tfrac{1}{\alpha-1}\left[a^{\alpha-1} - b^{\alpha-1}\right] \quad \geq \quad a^{\alpha-2}(a-b), \tag{37}$$

where we treat the left hand side of the last expression as the limit when $\alpha \to 1$:

$$\lim_{\alpha\to 1} \tfrac{1}{\alpha-1}\left[a^{\alpha-1} - b^{\alpha-1}\right] \quad = \quad \log \tfrac{b}{a}.$$

Hence, we obtain

$$\tfrac{1}{\alpha-1}\left[\tfrac{1}{f_{k+1}^{\alpha-1}} - \tfrac{1}{f_k^{\alpha-1}}\right] \quad = \quad \tfrac{1}{f_{k+1}^{\alpha-1} f_k^{\alpha-1}(\alpha-1)}\left[f_k^{\alpha-1} - f_{k+1}^{\alpha-1}\right] \overset{(37)}{\geq} \tfrac{f_k - f_{k+1}}{f_{k+1}^{\alpha-1} f_k} \overset{(36)}{\geq} c \cdot \tfrac{f_{k+1}}{f_k}. \tag{38}$$

Telescoping this inequality for the first $K \geq 1$ iterations, and using the inequality between arithmetic and geometric means, we get

$$\begin{aligned}
\tfrac{1}{\alpha-1}\left[\tfrac{1}{f_K^{\alpha-1}} - \tfrac{1}{f_0^{\alpha-1}}\right] \overset{(38)}{\geq} cK \cdot \tfrac{1}{K}\sum_{i=0}^{K-1} \tfrac{f_{i+1}}{f_i} \quad &\geq \quad cK \cdot \left[\prod_{i=0}^{K-1} \tfrac{f_{i+1}}{f_i}\right]^{1/K} \quad = \quad cK \cdot \left[\tfrac{f_K}{f_0}\right]^{1/K} \\
&= \quad cK \cdot \exp\left(\tfrac{1}{K}\log\tfrac{f_K}{f_0}\right) \quad \geq \quad cK \cdot \left(1 + \tfrac{1}{K}\log\tfrac{f_K}{f_0}\right).
\end{aligned}$$

Rearranging the terms, we have

$$\begin{aligned}
K \quad &\leq \quad \tfrac{1}{c(\alpha-1)}\left[\tfrac{1}{f_K^{\alpha-1}} - \tfrac{1}{f_0^{\alpha-1}}\right] + \log\tfrac{f_0}{f_K} \\
&= \quad 2p \cdot \left(\tfrac{p-1}{p}\right)^{\frac{2(p-1)}{p}} \cdot \tfrac{L\|\boldsymbol{A}\|^2}{\mu^{2/p}} \cdot \tfrac{1}{p-2}\left[\left(\tfrac{1}{f_K}\right)^{\frac{p-2}{p}} - \left(\tfrac{1}{f_0}\right)^{\frac{p-2}{p}}\right] + \log\tfrac{f_0}{f_K}.
\end{aligned} \tag{39}$$

It remains to use the upper bound (32) on the initial functional residual f_0, and the lower bound on f_K, assuming that $\|\boldsymbol{x}_K - \boldsymbol{x}_f^\star\| \geq \varepsilon$:

$$f_K \quad = \quad f_\psi(\boldsymbol{x}_K) - f_\psi(\boldsymbol{x}_f^\star) \overset{(25)}{\geq} \tfrac{\mu}{p}\|\boldsymbol{x}_K - \boldsymbol{x}_f^\star\|^p \quad \geq \quad \tfrac{\mu}{p}\varepsilon^p.$$

Substituting these estimates into (39) completes the proof. □

A.4. Proof of Theorem 3.3

Our proof follows the standard reasoning used in Gaussian mechanism for differential privacy and unlearning (see, e.g., (Nikolov et al., 2013; Dwork & Roth, 2014)). Since we analyze it from the optimization perspective, and, additionally, in Algorithm 1 we employ *anisotropic Normal distribution*, which is better suitable for the second-order geometry, we provide the full proof for completeness of our presentation.

Let $\boldsymbol{y}_K = \boldsymbol{x}_K + \boldsymbol{\xi}$ be the output of Algorithm 1 starting from $\boldsymbol{x}_0 := \boldsymbol{x}_F^\star$, where $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{B}^{-1})$ is the Normal noise.

At the same time, note that if we run Algorithm 1 from $\boldsymbol{x}_0 := \boldsymbol{x}_f^\star$, which satisfies the optimality condition:

$$\nabla f(\boldsymbol{x}_f^\star) + \psi'(\boldsymbol{x}_f^\star) \quad = \quad 0, \qquad \psi'(\boldsymbol{x}_f^\star) \in \partial\psi(\boldsymbol{x}_f^\star), \tag{40}$$

then all iterates are the same: $\boldsymbol{x}_K = \boldsymbol{x}_{K-1} = \ldots = \boldsymbol{x}_0 = \boldsymbol{x}_f^\star$ (so $\boldsymbol{x}_f^\star$ is a fixed point of the iterates). Indeed, due to $\boldsymbol{H}_k \succ 0$, every next iterate $\boldsymbol{x}_{k+1}$ is a unique solution of the following equation:

$$\nabla f(\boldsymbol{x}_k) + \boldsymbol{H}_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k) + \psi'(\boldsymbol{x}_{k+1}) \;\; = \;\; 0, \qquad \psi'(\boldsymbol{x}_{k+1}) \;\in\; \partial\psi(\boldsymbol{x}_{k+1}), \tag{41}$$

and it is easy to see that if $\boldsymbol{x}_k = \boldsymbol{x}_f^\star$, which satisfies (40), then $\boldsymbol{x}_{k+1} := \boldsymbol{x}_f^\star$ satisfies (41). Hence, running Algorithm 1 from $\boldsymbol{x}_0 := \boldsymbol{x}_f^\star$ for any number of iterations $K \geq 0$, we always have as the result $\boldsymbol{x}_K = \boldsymbol{x}_f^\star$. Let us denote the output of Algorithm 1 in this case by $\boldsymbol{y}^\star := \boldsymbol{x}_f^\star + \boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{B}^{-1})$.

Therefore, to show certified (q, δ)-unlearning, by definition, we need to show for any measurable $Y \subseteq \mathbb{R}^n$ that

$$\begin{aligned} \mathbb{P}\big(\boldsymbol{y}_K \in Y\big) \;\; &\leq \;\; e^q \cdot \mathbb{P}\big(\boldsymbol{y}^\star \in Y\big) + \delta, \\ \mathbb{P}\big(\boldsymbol{y}^\star \in Y\big) \;\; &\leq \;\; e^q \cdot \mathbb{P}\big(\boldsymbol{y}_K \in Y\big) + \delta. \end{aligned} \tag{42}$$

To establish (42), we follow the reasoning from (Nikolov et al., 2013), extending it to our case. We denote $\boldsymbol{v} := \boldsymbol{x}_K - \boldsymbol{x}_f^\star$, and by our optimization guarantee, we have

$$\|\boldsymbol{v}\| \;\; := \;\; \langle \boldsymbol{B}\boldsymbol{v}, \boldsymbol{v}\rangle^{1/2} \;\; \leq \;\; \varepsilon. \tag{43}$$

Let $p(\boldsymbol{\xi})$ be the probability density function of $\mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{B}^{-1})$:

$$p(\boldsymbol{\xi}) \;\; \propto \;\; \exp\Big(-\tfrac{1}{2\sigma^2}\langle \boldsymbol{B}\boldsymbol{\xi}, \boldsymbol{\xi}\rangle\Big) \;\; = \;\; \exp\Big(-\tfrac{1}{2\sigma^2}\|\boldsymbol{\xi}\|^2\Big),$$

and consider the quantity:

$$\begin{aligned} D_{\boldsymbol{v}}(\boldsymbol{\xi}) \;\; &:= \;\; \ln \tfrac{p(\boldsymbol{\xi})}{p(\boldsymbol{\xi}+\boldsymbol{v})} \;\; = \;\; \ln p(\boldsymbol{\xi}) - \ln p(\boldsymbol{\xi}+\boldsymbol{v}) \;\; = \;\; \tfrac{1}{2\sigma^2}\|\boldsymbol{\xi}+\boldsymbol{v}\|^2 - \tfrac{1}{2\sigma^2}\|\boldsymbol{\xi}\|^2 \\ &= \;\; \tfrac{1}{2\sigma^2}\|\boldsymbol{v}\|^2 + \tfrac{1}{\sigma^2}\langle \boldsymbol{B}\boldsymbol{v}, \boldsymbol{\xi}\rangle \;\overset{(43)}{\leq}\; \tfrac{\varepsilon^2}{2\sigma^2} + \tfrac{1}{\sigma^2}\langle \boldsymbol{B}\boldsymbol{v}, \boldsymbol{\xi}\rangle. \end{aligned} \tag{44}$$

Note that $t := \frac{1}{\sigma^2}\langle \boldsymbol{B}\boldsymbol{v}, \boldsymbol{\xi}\rangle \sim \mathcal{N}\big(0, \frac{\|\boldsymbol{v}\|^2}{\sigma^2}\big)$ is a univariate Normal variable. The classic Chernoff bound ensures that

$$\mathbb{P}\big(|t| \geq \alpha\big) \;\; = \;\; \mathbb{P}\big(\tfrac{1}{\sigma^2}|\langle \boldsymbol{B}\boldsymbol{v}, \boldsymbol{\xi}\rangle| \geq \alpha\big) \;\; \leq \;\; 2\exp\Big(-\tfrac{\alpha^2\sigma^2}{2\|\boldsymbol{v}\|^2}\Big) \;\overset{(43)}{\leq}\; \delta \;\; := \;\; 2\exp\Big(-\tfrac{\alpha^2\sigma^2}{2\varepsilon^2}\Big). \tag{45}$$

Hence, with probability that is greater than $1 - \delta$, we have

$$|D_{\boldsymbol{v}}(\boldsymbol{\xi})| \;\overset{(44)}{\leq}\; \tfrac{\varepsilon^2}{2\sigma^2} + \tfrac{1}{\sigma^2}|\langle \boldsymbol{B}\boldsymbol{v}, \boldsymbol{\xi}\rangle| \;\overset{(45)}{<}\; \tfrac{\varepsilon^2}{2\sigma^2} + \alpha \;\overset{(45)}{=}\; \tfrac{\varepsilon^2}{2\sigma^2} + \tfrac{\varepsilon}{\sigma}\sqrt{2\ln\tfrac{2}{\delta}} \;\; \leq \;\; q, \tag{46}$$

where the last inequality is satisfied for any given $q \in (0, 1)$, and for a corresponding sufficiently small σ. Namely, by the condition of the theorem, we have chosen:

$$\sigma \;\; := \;\; \tfrac{\varepsilon}{q}\max\Big\{1, 2\sqrt{2\ln\tfrac{2}{\delta}}\Big\},$$

which ensures (46).

Now, to justify the unlearning bounds (42), we consider the set

$$S \;\; := \;\; \Big\{\boldsymbol{\xi} \;:\; D_{\boldsymbol{v}}(\boldsymbol{\xi}) < q\Big\}.$$

Notice that for $\boldsymbol{\xi} \in S$, we have

$$p(\boldsymbol{\xi}) \;\; \leq \;\; e^q \cdot p(\boldsymbol{\xi} + \boldsymbol{v}), \tag{47}$$

while by the previous observations, the measure of the complement is small:

$$\overline{S} \;\; = \;\; \Big\{\boldsymbol{\xi} \;:\; D_{\boldsymbol{v}}(\boldsymbol{\xi}) \geq q\Big\}, \qquad \mathbb{P}(\overline{S}) \;\; \leq \;\; \delta. \tag{48}$$

Hence,

$$\mathbb{P}(\boldsymbol{y}_K \in Y) \;=\; \mathbb{P}(\xi \in Y - \boldsymbol{x}_K) \;=\; \int\limits_{Y-\boldsymbol{x}_K} p(\boldsymbol{\xi})d\boldsymbol{\xi} \;=\; \int\limits_{S\cap(Y-\boldsymbol{x}_K)} p(\boldsymbol{\xi})d\boldsymbol{\xi} \;+\; \int\limits_{\overline{S}\cap(Y-\boldsymbol{x}_K)} p(\boldsymbol{\xi})d\boldsymbol{\xi}$$

$$\overset{(48)}{\leq} \int\limits_{S\cap(Y-\boldsymbol{x}_K)} p(\boldsymbol{\xi})d\boldsymbol{\xi} \;+\; \delta \;\overset{(47)}{\leq}\; e^q \cdot \int\limits_{S\cap(Y-\boldsymbol{x}_K)} p(\boldsymbol{\xi}+\boldsymbol{v})d\boldsymbol{\xi} \;+\; \delta$$

$$\leq\; e^q \cdot \int\limits_{Y-\boldsymbol{x}_K} p(\boldsymbol{\xi}+\boldsymbol{v})d\boldsymbol{\xi} \;+\; \delta \;=\; e^q \cdot \mathbb{P}\big(\xi \in Y - \boldsymbol{x}_f^\star\big) \;+\; \delta \;=\; e^q \cdot \mathbb{P}\big(\boldsymbol{y}^\star \in Y\big) \;+\; \delta,$$

which is the first inequality (42). The proof of the second inequality in (42) is identical due to symmetry.

Finally, to ensure the optimization guarantee (21), we observe that

$$\mathbb{E}\big[\|\boldsymbol{y}_K - \boldsymbol{x}_f^\star\|^2\big] \;=\; \mathbb{E}\big[\|\boldsymbol{v}+\boldsymbol{\xi}\|^2\big] \;=\; \mathbb{E}\big[\|\boldsymbol{v}\|^2 + 2\langle \boldsymbol{B}\boldsymbol{v}, \boldsymbol{\xi}\rangle + \|\boldsymbol{\xi}\|^2\big]$$

$$=\; \|\boldsymbol{v}\|^2 + \mathbb{E}\big[\|\boldsymbol{\xi}\|^2\big] \;\overset{(43)}{\leq}\; \varepsilon^2 + \mathbb{E}\big[\|\boldsymbol{\xi}\|^2\big] \;\leq\; \varepsilon^2 + \sigma^2 d,$$

where in the last bound we used the equivalent representation $\boldsymbol{\xi} = \sigma \boldsymbol{B}^{-1/2}\boldsymbol{z}$ with the standard normal vector $\boldsymbol{z} \in \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, which leads to the mean of the standard $\chi^2(d)$ distribution:

$$\mathbb{E}\big[\|\boldsymbol{\xi}\|^2\big] \;=\; \sigma^2 \mathbb{E}\big[\|\boldsymbol{z}\|_2^2\big] \;=\; \sigma^2 d.$$

It remains to use Jensen's inequality for concave function $\sqrt{\cdot}$, in order to obtain (21), which completes the proof. □