Replicable Machine Learning: Theory and Algorithms for Stochastic Convex and Non-Convex Optimization

Raman Arora
Johns Hopkins University

Kaibo Zhang
Johns Hopkins University

Abstract

Replicable algorithms produce identical outputs with high probability when run on independent samples drawn from the same distribution, providing strong reproducibility guarantees for machine learning pipelines. We study replicability in machine learning in Vapnik's general learning setting, which encompasses stochastic optimization over convex and non-convex loss classes, establishing algorithms with near-optimal sample complexity across these settings. For general Lipschitz losses over a bounded parameter space, we show that the exponential mechanism combined with correlated sampling achieves optimal $O(1/\sqrt{n})$ excess risk with ρ-replicability guarantees, but at the cost of exponential runtime. For general Lipschitz losses, the exponential mechanism with correlated sampling achieves optimal $O(1/\sqrt{n})$ excess risk and ρ-replicability, but with exponential runtime. For strongly convex losses over a d-dimensional parameter space, empirical risk minimization (ERM) paired with randomized rounding achieves $\widetilde{O}(\sqrt{d}/(\rho\sqrt{n}))$ excess risk in polynomial time. For general convex losses, regularized ERM yields excess risk of $\widetilde{O}(n^{-1/4})$. We further extend our techniques to overparameterized neural networks in the Neural Tangent Kernel (NTK) regime. Taken together, our results provide evidence for a fundamental computational-statistical tradeoff in replicable learning, whereby optimal replicability requires exponential time while our polynomial-time algorithms incur a modest but provable statistical penalty.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 Introduction

Reproducibility crisis in machine learning is well-documented: small changes in datasets, random seeds, or hyperparameters can yield qualitatively different models (Pineau et al., 2021; Henderson et al., 2018). This instability undermines scientific progress and practical deployment.

Impagliazzo et al. (2022) introduced *replicable learning*: algorithms that produce identical outputs with high probability when run on independent samples from the same distribution using the same random seed. Formally, an algorithm $\mathcal{A}$ is ρ-replicable if for independent samples $S, S' \sim \mathcal{D}^n$ and shared randomness r, $\mathbb{P}_{S,S',r}[\mathcal{A}(S;r) = \mathcal{A}(S';r)] \geq 1-\rho$. Remarkably, replicability *implies* generalization (Impagliazzo et al., 2022): algorithms stable to complete dataset replacement must have small generalization gaps.

We study replicable learning in d-dimensions through the lens of stochastic optimization, i.e., minimizing population risk $L(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$ over some hypothesis class $h \in \mathcal{H}$. This framework unifies PAC learning, regression, and reinforcement learning (Vapnik, 1999; Sridharan et al., 2009). Our contributions are:

- **General non-convex optimization:** Using the exponential mechanism with correlated sampling, we achieve ρ-replicability with optimal

$O(1/\sqrt{n})$ excess risk for arbitrary Lipschitz losses (Section 3.1), though with exponential runtime.

- **Efficient strongly convex optimization:** For μ-strongly convex losses, we combine empirical risk minimization (ERM) with randomized rounding to achieve ρ-replicability and near-optimal $\tilde{O}(\sqrt{d}/(\rho\sqrt{n}))$ excess risk in polynomial time (Section 3.3).

- **General convex optimization:** Using regularized ERM, we obtain efficient replicable algorithms for convex Lipschitz losses, achieving $\tilde{O}(n^{-1/4})$ rates (Section 3.4).

- **Neural networks:** We extend our techniques to overparameterized neural network training in the NTK regime (Jacot et al., 2018; Ji and Telgarsky, 2019), obtaining the first replicability guarantees for this setting (Section 3.5).

Table 1 summarizes our main results across different problem settings, revealing a clear computational-statistical tradeoff landscape. We show that correlated sampling (Bun et al., 2023) achieves optimal rates for general problems but requires exponential time, while rounding-based methods provide polynomial-time efficiency but achieve optimal rates only under additional structural assumptions.

2 Problem Setup and Preliminaries

We adopt the stochastic optimization view of learning (Vapnik, 1999; Sridharan et al., 2009): given samples $z \sim \mathcal{D}$ from an unknown distribution, minimize the population risk $L(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h, z)]$ where $\ell(h, z)$ is a loss function and $h \in \mathcal{H}$ is a hypothesis. This framework unifies PAC learning (0-1 loss), regression (squared loss), and reinforcement learning (cumulative reward).

Definition 2.1 (Replicability (Impagliazzo et al., 2022))**.** *A randomized algorithm $\mathcal{A}$ is ρ-replicable if for every distribution $\mathcal{D}$ and independent samples $S_1, S_2 \sim \mathcal{D}^n$ with shared randomness r, $\mathbb{P}_{S_1,S_2,r}[\mathcal{A}(S_1; r) = \mathcal{A}(S_2; r)] \geq 1 - \rho$.*

Replicability ensures algorithms output identical hypotheses across independent samples from the same distribution. Remarkably, Impagliazzo et al. (2022) showed that replicability *implies generalization* without requiring uniform convergence: if outputs barely depend on the specific sample, empirical risk must concentrate around population risk.

Connections to Stability and Privacy Replicability is equivalent to *total variation (TV) stability* (Bun et al., 2023): output distributions under different samples have small TV distance. This connects replicability to classical stability notions (Bousquet and Elisseeff, 2002), differential privacy (Dwork and Roth, 2014), and pseudodeterminism (Goldreich et al., 2013). Bun et al. (2023) proved that replicability, differential privacy, and perfect generalization are statistically equivalent up to quadratic sample overhead, though they may differ computationally (Kalavasis et al., 2024).

Replicability Mechanisms. Two complementary techniques enforce TV-stability:

1. **Correlated sampling.** Given distributions P_α, P_β with TV distance d_{TV}, we construct coupled samples that agree with probability $\geq 1 - O(d_{\mathrm{TV}})$. Algorithm 1 implements this via rejection sampling from a reference measure Q_0: draw candidates $(\xi_t, y_t) \sim Q_0 \times \mathrm{Unif}[0, c_0]$ and accept ξ_t when $y_t \leq dP_\alpha/dQ_0(\xi_t)$. Here $c_0 \geq dP_\alpha/dQ_0$ for any α. Using the same random seed across samples couples their outputs. From (Bavarian et al., 2016), this approach achieves optimal replicability $\mathbb{P}[A(P_\alpha) \neq A(P_\beta)] \leq 2d_{\mathrm{TV}}/(1 + d_{\mathrm{TV}})$ but requires exponential runtime when acceptance probabilities are small. (Hopkins and Moran, 2025) discusses the number of random bits required in PAC learning. If we only allow polynomial number of random bits, the sample complexity will be exponentially large.

Algorithm 1 Correlated Sampling

Input: Reference measure Q_0, constant c_0, density ratios $\{dP_\alpha/dQ_0\}_\alpha$, sequence $(\xi_t, y_t) \sim Q_0 \times \mathrm{Unif}([0, c_0])$

for $t = 1, 2, \ldots$ **do**
  if $y_t \leq dP_\alpha/dQ_0(\xi_t)$ **then**
    return ξ_t
  end if
end for

Table 1: Comparison of replicable learning algorithms across problem settings. Here $\tilde{O}(\cdot)$ suppresses logarithmic factors in n, d, δ, ρ.

Problem	Excess Risk	Replicability	Runtime
General Lipschitz			
Exponential mechanism	$\tilde{O}(\frac{\sqrt{d^3}}{\rho\sqrt{n}})$	$1-\rho$	$\exp(d)$
Strongly Convex			
ERM + rounding	$\tilde{O}(\frac{d^2}{\rho^2 n})$	$1-\rho$	$\text{poly}(d,n)$
Convex Lipschitz			
RERM + rounding	$\tilde{O}(\frac{\|h^*\|\sqrt{d}}{\sqrt{\rho}n^{\frac{1}{4}}})$	$1-\rho$	$\text{poly}(d,n)$
Neural Networks (NTK)			
RERM + rounding	$\tilde{O}(\frac{\sqrt{d}}{\gamma^3\sqrt{\rho}n^{\frac{1}{4}}})$	$1-\rho$	$\text{poly}(m,d,n)$

2. **Randomized rounding.** For efficiency, we discretize outputs via random grids. Select a random orthogonal basis and partition each axis into intervals of length a. Map each vector x to the center $R(x)$ of its grid cell. This achieves $\mathbb{P}[R(x) \neq R(x')] \leq \sqrt{d}\|x - x'\|/a$ with rounding error $\|R(x) - x\| \leq O(a\sqrt{d})$, running in $\text{poly}(d)$ time. The $\sqrt{d}$ factor loss compared to correlated sampling is the price of polynomial runtime. The ConstructFoams scheme (Kindler et al., 2012) achieves optimal $O(\|x - x'\|/a)$ probability but also requires exponential time.

Related Work Impagliazzo et al. (2022) introduced replicability for discrete problems (statistical queries, heavy hitters, PAC learning), establishing the replicability-generalization connection. Bun et al. (2023) formalized the equivalence to TV-stability and connections to differential privacy. Kalavasis et al. (2024) proved computational separations between replicability and privacy under cryptographic assumptions. Recent work extends replicability to reinforcement learning (Eaton et al., 2023), bandits, clustering, and specific concept classes (Noivirt et al., 2026). Our work provides the first comprehensive treatment of replicable continuous optimization across convex and non-convex settings.

3 Main Results and Techniques

This section presents our main algorithmic contributions for replicable learning across different problem structures. Our development proceeds in two complementary phases. First, in Section 3.1, we develop a general approach based on the exponential mechanism (McSherry and Talwar, 2007) combined with correlated sampling. This technique applies to arbitrary non-convex losses satisfying mild smoothness conditions and achieves statistically optimal $O(1/\sqrt{n})$ excess risk rates. However, as is typical with the exponential mechanism, the approach requires exponential computational cost in the dimension in the worst case.

To address computational efficiency, we turn in Sections 3.2–3.4 to algorithms based on randomized rounding. These methods exploit concentration properties of ERMs in convex problems, achieving polynomial runtime at the cost of slightly relaxed replicability probabilities. Key insight is that when an ERM solution concentrates sharply around the population minimizer, adding controlled noise and discretizing via randomized grids preserves both generalization and replicability.

Our results instantiate a fundamental computational vs. statistical tradeoff: correlated sampling achieves the statistically optimal coupling probability $1-O(\|x-x'\|/a)$ for vectors differing by $\|x-x'\|$ when adding noise of scale a, but requires exponential time; randomized rounding achieves probability $1-O(\sqrt{d}\|x-x'\|/a)$, degraded by a factor of $\sqrt{d}$, but runs in polynomial time. For strongly convex problems (Section 3.3), this degradation is mild as we still obtain near-optimal rates. For general convex problems (Section 3.4), we employ regularized ERM to induce strong convexity, but this introduces a bias-variance tradeoff yielding suboptimal $n^{-1/4}$ rates. Finally, in Section 3.5, we demonstrate that our techniques extend beyond classical con-

vex optimization to modern neural network training. Working in the neural tangent kernel (NTK) regime (Jacot et al., 2018), we show that gradient descent on overparameterized networks reduces to kernel ridge regression, which we can make replicable via our regularization-based approach. We conclude in Section 4 with a discussion and some open questions.

3.1 General Non-Convex Problems

We begin with the most general setting, where we make minimal assumptions on the loss function. Consider a bounded parameter space $\mathcal{H} = \{h : \|h - h_0\|_2 \leq B\} \subseteq \mathbb{R}^d$. Let the loss $\ell(h, z)$ be L-Lipschitz in h and satisfy $0 \leq \ell(h, z) \leq M_0 \ \forall h, z$. Given a sample $S = \{z_1, \ldots, z_n\} \sim \mathcal{D}^n$, define the empirical risk $L_S(h) = \frac{1}{n}\sum_{i=1}^n \ell(h, z_i)$ and population risk $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h, z)]$. Let $h^* = \operatorname{argmin}_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$ denote the population risk minimizer.

The exponential mechanism provides a general-purpose replicable learning algorithm that works for such bounded Lipschitz losses, including non-convex problems. The key insight, building on work in differential privacy (McSherry and Talwar, 2007; Dwork and Roth, 2014), is that uniform convergence guarantees allow us to couple the output distributions across different samples via correlated sampling. Unlike ERM-based approaches that require structural assumptions for concentration, the exponential mechanism only requires that empirical risks concentrate uniformly over the hypothesis class, a property that follows from standard covering number arguments (Shalev-Shwartz and Ben-David, 2014).

Our analysis relies on the following uniform convergence result, which can be derived via standard VC-style or Rademacher complexity arguments:

Lemma 3.1 (Uniform Convergence). *If* $n \geq \frac{2M_0^2}{\epsilon^2} \cdot \left(\ln\left(\frac{2}{\delta}\right) + d\ln\left(\frac{8BL}{\epsilon} + 1\right)\right)$, *then with probability at least* $1-\delta$ *over* $S \sim \mathcal{D}^n$, $\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$.

Remark. While we use Lipschitzness to establish uniform convergence, problem-specific structure may yield sharper bounds. For instance, if $\ell(\cdot, z)$ has low local Rademacher complexity near h^*, localization techniques can reduce the dependence on d and B (Bartlett et al., 2005).

3.1.1 Correlated Sampling for the Exponential Mechanism

Given the empirical risk $L_S(h)$ and a parameter $\eta > 0$, the exponential mechanism outputs a random hypothesis h sampled from the probability distribution with density proportional to $\exp(-\eta L_S(h))$. Since we cannot compute the normalization constant in closed form, we implement this via rejection sampling as described in Algo. 2. This algorithm can be seen as a specialization of Algorithm 1

The algorithm requires a reference distribution Q_0 (taken to be uniform over $\mathcal{H}$) and accepts candidates with probability proportional to $\exp(-\eta L_\alpha(\xi_t))$, where L_α represents the loss function for a particular sample. By using the same random seed (sequence of candidate points and thresholds) across different samples, we couple their output distributions. The following result establishes that Algorithm 2 both produces samples from the correct distribution and ensures replicability when loss functions are close.

Algorithm 2 Correlated Sampling for Exponential Mechanism

Input: Uniform probability measure Q_0 over $\mathcal{H}$, loss functions $\{L_\alpha(h)\}_{\alpha\in\mathcal{I}}$, temperature η
Input: Sequence $(\xi_t, y_t) \sim Q_0 \times \mathrm{Unif}([0,1])$ of i.i.d. samples
for $t = 1, 2, 3, \ldots$ **do**
  if $y_t \leq \exp(-\eta L_\alpha(\xi_t))$ **then**
    return ξ_t
  end if
end for

Proposition 3.2. *The output* $A(P_\alpha)$ *of Algorithm 2 satisfies:*

1. $A(P_\alpha)$ *has probability density* $p_\alpha(h) \propto \exp(-\eta L_\alpha(h))$.
2. $\mathbb{P}[A(P_\alpha) \neq A(P_\beta)] \leq \eta \cdot \sup_{h\in\mathcal{H}} |L_\alpha(h) - L_\beta(h)|$.

Proof sketch. Part (1) follows from the rejection sampling principle: the acceptance probability is proportional to the target density. For part (2), we note that the coupling succeeds (outputs agree) unless the acceptance decisions differ, which is when a specific ξ_t is accepted by one loss function L_α, but rejected by the other loss func-

tion L_β. Since the ratio between $\exp(-\eta L_\alpha)$ and $\exp(-\eta L_\beta)$ is very close to 1, this event is unlikely to happen. We upper bound this probability by $1 - \exp(-\eta \cdot \sup_{h\in\mathcal{H}} |L_\alpha(h) - L_\beta(h)|) \leq \eta \cdot \sup_{h\in\mathcal{H}} |L_\alpha(h) - L_\beta(h)|$. □

3.1.2 Replicability and Utility Guarantees

We now establish that the exponential mechanism achieves both strong replicability and optimal utility guarantees. The key is to balance the temperature parameter: larger η increases replicability (by making the distribution more peaked around low-loss regions) but requires more samples to ensure empirical minimizer is near population minimizer.

Proposition 3.3 (Replicability). *Suppose $\delta < \rho/2$ and $n \geq \frac{2M_0^2}{\epsilon^2} \cdot \left(\ln\left(\frac{2}{\delta}\right) + d\ln\left(\frac{8BL}{\epsilon} + 1\right)\right)$. Then, Algorithm 2 with $\eta = \frac{\rho - 2\delta}{2\epsilon}$ and loss function $L(h) = L_S(h)$ is ρ-replicable.*

Proof sketch. By Lemma 3.1, with probability at least $1-2\delta$ over independent samples S, S', we have $\sup_h |L_S(h) - L_{S'}(h)| \leq 2\epsilon$. Conditioning on this event and applying Proposition 3.2, the outputs agree with probability at least $1-\eta\cdot 2\epsilon = 1-(\rho-2\delta)$. By a union bound, the overall replicability is at least $1 - 2\delta - (\rho - 2\delta) = 1 - \rho$. □

For utility, we require a quantitative bound on how well the exponential mechanism minimizes the empirical risk. The following lemma shows that sampling from $\exp(-\eta L_S)$ concentrates mass on near-optimal hypotheses.

Lemma 3.4 (Utility of Exponential Mechanism). *Given any fixed sample S, Algorithm 2 returns a random h_S satisfying $\mathbb{P}\left[L_S(h_S) - \inf_h L_S(h) \geq \alpha\right] \leq e^{-\frac{\eta\alpha}{2}} \max\left\{\left(\frac{4BL}{\alpha}\right)^d, 1\right\}$.*

Proof sketch. The volume of the set $\{h : L_S(h) \leq \inf L_S + \alpha\}$ can be lower-bounded using the Lipschitz constant, yielding a ball of radius $\Omega(\alpha/L)$. The ratio of the measure of this set under $\exp(-\eta L_S)$ to the measure of the entire space is at least $\exp(-\eta\alpha) \cdot (\alpha/(BL))^d$. The result follows by bounding the probability of sampling outside this favorable set. □

Combining the uniform convergence, replicability, and utility guarantees yields our main result for general non-convex problems:

Proposition 3.5 (Excess Risk). *If the sample size $n \geq \frac{2M_0^2}{\epsilon^2} \cdot \left(\ln\left(\frac{2}{\delta}\right) + d\ln\left(\frac{8BL}{\epsilon} + 1\right)\right)$, then with probability at least $1 - 2\delta$ over $S \sim \mathcal{D}^n$, Algorithm 2 with $\eta = \frac{\rho - 2\delta}{2\epsilon}$ and loss function $L(h) = L_S(h)$ returns h_S satisfying $L_\mathcal{D}(h_S) - \inf_h L_\mathcal{D}(h) \leq 2\epsilon\left(1 + \frac{2}{\rho - 2\delta}\left(\ln\left(\frac{1}{\delta}\right) + \left[d \cdot \ln\left(\frac{(\rho-2\delta)BL}{\epsilon \ln(1/\delta)}\right)\right]_+\right)\right)$.*

Proof sketch. By Lemma 3.1, $L_S(h_S) \leq L_\mathcal{D}(h_S) + \epsilon$ and $L_\mathcal{D}(h^*) \leq L_S(h^*) + \epsilon$. From Lemma 3.4 with α chosen to make the tail probability at most δ, we have $L_S(h_S) \leq L_S(h^*) + \alpha$ with high probability. Combining these inequalities and choosing α appropriately yields the result. □

Remark 1 (Excess Risk Rate for General Lipschitz Losses). *Proposition 3.5 yields the $\widetilde{O}(\sqrt{d^3}/(\rho\sqrt{n}))$ rate reported in Table 1. To see this, set $\epsilon = \widetilde{O}(M_0\sqrt{d/n})$, the threshold at which the uniform convergence condition in Lemma 3.1 is met (ignoring logarithmic factors in n, d, and δ). Substituting into the excess risk bound of Proposition 3.5, the dominant term is*

$$L_\mathcal{D}(\mathcal{A}(S)) - L_\mathcal{D}(h^*) \;=\; \widetilde{O}\left(\frac{\epsilon\, d}{\rho}\right) \;=\; \widetilde{O}\left(\frac{M_0\sqrt{d^3}}{\rho\sqrt{n}}\right).$$

This matches the minimax optimal $O(\sqrt{d/n})$ rate up to the d/ρ factor, which reflects the cost of coupling the output distributions across samples via correlated sampling. However, the runtime is exponential in d due to the rejection sampling procedure, as the expected number of steps scales as $e^{\Omega(d)}$ when the acceptance probability varies significantly over $\mathcal{H}$. This motivates our turn to more efficient algorithms for structured problems.

3.2 Efficient Replicability via Randomized Rounding

For convex problems, the empirical risk minimizer concentrates sharply around the population minimizer, opening the door to a different approach: rather than sampling from a global distribution over $\mathcal{H}$, we compute the ERM solution, add controlled noise, and discretize the output via randomized rounding. This strategy achieves polynomial runtime at the cost of slightly weaker replicability

probabilities, specifically, a $\sqrt{d}$ factor degradation compared to optimal correlated sampling.

The randomized rounding scheme works as follows. We first generate a random orthonormal basis by sampling a uniformly random orthogonal matrix from $\mathrm{SO}(d)$, yielding basis vectors $\{v_1, \ldots, v_d\}$. We then partition each coordinate direction (defined by v_i) into intervals of length a, with the partition offset chosen uniformly at random. This creates a random grid tessellation of $\mathbb{R}^d$ with cell side length a. For any vector $x \in \mathbb{R}^d$, we define $R(x)$ to be the center of the grid cell containing x.

Proposition 3.6 (Properties of Randomized Rounding)**.** *Randomized rounding scheme satisfies:*

1. ***Rounding error:*** $\|R(x) - x\|_2 \leq \frac{a}{2}\sqrt{d}$, $\forall x$.
2. ***Replicability:*** *For any two fixed vectors* $x, x' \in \mathbb{R}^d$, $\mathbb{P}[R(x) \neq R(x')] \leq \frac{\sqrt{d}\|x-x'\|_2}{a}$.

Proof sketch. For part (1), the maximum distance from any point to the center of its grid cell is achieved at the corners, where the distance is $(a/2)\sqrt{d}$. For part (2), two points are rounded to different centers only if they lie in different grid cells. The probability of this event is bounded by the probability that at least one of the d coordinate projections (in the random basis) crosses a grid boundary. Each coordinate crosses a boundary with probability at most $\|x - x'\|_2/a\sqrt{d}$, and summing over coordinates yields the result. □

Comparison with correlated sampling. The ConstructFoams rounding scheme of Kindler et al. (2012) achieves the optimal probability bound $\mathbb{P}[R(x) \neq R(x')] \leq c\|x-x'\|_2/a$ without the $\sqrt{d}$ factor, matching correlated sampling. However, like correlated sampling, ConstructFoams requires exponential time in d. Our randomized grid approach sacrifices this $\sqrt{d}$ factor to obtain $\mathrm{poly}(d)$ runtime, making it practical for high-dimensional problems. In the convex settings we consider next, this degradation is acceptable. It introduces an additional $\sqrt{d}$ multiplicative factor in the sample complexity, which is mild compared to d or d^2 dependence already present from uniform convergence.

We now apply this rounding technique to convex optimization problems, beginning with the favorable case of strong convexity.

3.3 Strongly Convex Optimization

We first consider the setting where the loss function exhibits strong convexity, ensuring rapid concentration of the ERM solution. Let the parameter space be $\mathcal{H} = \mathbb{R}^d$ and assume the loss $\ell(h, z)$ is μ-strongly convex and differentiable w.r.t. h. Denote $h_S = \mathrm{argmin}_h L_S(h)$ and $h^* = \mathrm{argmin}_h L_{\mathcal{D}}(h)$. Assume that $\|\nabla_h \ell(h^*, z)\|_2 \leq M$ for all z.

Strong convexity provides powerful concentration: the ERM solution h_S lies close to h^* with high probability, and this proximity persists across independent samples. Following lemma, which follows from standard arguments combining strong convexity with concentration of gradients (Shalev-Shwartz and Ben-David, 2014), quantifies this behavior.

Lemma 3.7 (Concentration of Strongly Convex ERM)**.** *With probability at least* $1-\delta$ *over* $S \sim \mathcal{D}^n$, $\|h_S - h^*\|_2 \leq \frac{M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right)$.

Proof sketch. By strong convexity and first-order optimality condition, $\|h_S - h^*\|_2 \leq \frac{1}{\mu}\|\nabla L_S(h^*)\|_2$. Since $\mathbb{E}[\nabla L_S(h^*)] = \nabla L_{\mathcal{D}}(h^*) = 0$ and the gradient is an average of n independent random vectors with ℓ_2 norm bounded by M, standard concentration using McDiarmid's inequality yields the result. □

Given this concentration, our algorithm is straightforward: compute the ERM solution h_S, then apply the randomized rounding scheme from Sec. 3.2. Let $\tilde{h}_S = R(h_S)$ denote the rounded output.

Replicability. The concentration guarantee immediately implies replicability when we choose the grid size appropriately.

Proposition 3.8 (Replicability for Strongly Convex ERM)**.** *Suppose* $\delta < \rho/2$. *If the rounding scheme uses grid size* $a = \frac{2M\sqrt{d}}{(\rho-2\delta)\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right)$, *then* $\tilde{h}_S$ *is* ρ*-replicable.*

Proof. By Lemma 3.7, with probability at least $1-2\delta$ over independent samples S, S', we have

$$\|h_S - h_{S'}\|_2 \leq 2 \cdot \frac{M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right).$$

By Proposition 3.6, conditioning on this event, $\mathbb{P}[\tilde{h}_S \neq \tilde{h}_{S'}] \leq \frac{\sqrt{d}\|h_S - h_{S'}\|_2}{a} \leq \rho - 2\delta$. A union bound yields overall replicability $1 - \rho$. □

Generalization. The rounding error is controlled by the grid size a, which we have chosen inversely proportional to $\sqrt{n}$. This ensures the estimation error and discretization error are balanced.

Proposition 3.9 (Generalization for Strongly Convex ERM)**.** *Suppose $\delta < \rho/2$. With probability at least $1-\delta$ over $S \sim \mathcal{D}^n$, the rounding scheme with $a = \frac{2M\sqrt{d}}{(\rho-2\delta)\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)$ ensures*

$$\|\tilde{h}_S - h^*\|_2 \leq \frac{M(d+1)}{(\rho-2\delta)\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right).$$

Proof. By the triangle inequality, $\|\tilde{h}_S - h^*\|_2 \leq \|h_S - h^*\|_2 + \|\tilde{h}_S - h_S\|_2$. The first term is bounded by Lemma 3.7, and the second by Proposition 3.6(1) using $\|R(x)-x\| \leq a\sqrt{d}/2$. □

Remark 2 (Excess Risk Rate for Strongly Convex ERM)**.** *If we additionally assume the population risk $L_{\mathcal{D}}$ is β-smooth, then $L_{\mathcal{D}}(\tilde{h}_S) - L_{\mathcal{D}}(h^*) \leq \frac{\beta}{2}\|\tilde{h}_S - h^*\|_2^2$, giving excess risk $\widetilde{O}(M^2d^2/(\rho^2\mu^2 n))$, which scales as $\widetilde{O}(d^2/(\rho^2 n))$ after absorbing problem-dependent constants. This matches the non-replicable rate of $O(1/n)$ for strongly convex problems up to the d^2/ρ^2 factor from randomized rounding. The factor d degradation relative to correlated sampling (which would give $\widetilde{O}(d/(\rho^2 n))$) is the price of polynomial-time computation: correlated sampling achieves the optimal coupling probability $1-O(\|x-x'\|/a)$, while randomized rounding achieves $1-O(\sqrt{d}\|x-x'\|/a)$, inflating both the grid size and the rounding error by $\sqrt{d}$. The dependence on ρ reflects the tradeoff between replicability (larger ρ easier) and accuracy.*

3.4 General Convex Optimization

When the loss is convex but not strongly convex, the ERM solution may not concentrate as sharply. For instance, in the Lipschitz-bounded setting, the distance $\|h_S - h_{S'}\|_2$ between ERM solutions on independent samples can be as large as $O(1)$ even as $n \to \infty$ if the problem is poorly conditioned. To address this, we employ *regularized empirical risk minimization* (RERM): we add an ℓ_2 regularization term to the loss, artificially inducing strong convexity (Shalev-Shwartz and Ben-David, 2014).

Let the parameter space be $\mathcal{H} = \mathbb{R}^d$, and assume the loss $\ell(h,z)$ is convex, differentiable, and L-Lipschitz in h. Let $h^* = \operatorname{argmin}_h L_{\mathcal{D}}(h)$ denote the unregularized population risk minimizer. For a regularization parameter $\mu > 0$, define the regularized empirical risk minimizer

$$h_S = \operatorname*{argmin}_h \left[L_S(h) + \frac{\mu}{2}\|h\|_2^2\right].$$

The regularized loss is μ-strongly convex, so Lemma 3.7 applies with gradient bound determined by the Lipschitz constant L.

Lemma 3.10 (Concentration of Regularized ERM)**.** *With probability at least $1-2\delta$ over independent samples $(S, S') \sim \mathcal{D}^{2n}$,*

$$\|h_S - h_{S'}\|_2 \leq \frac{2L}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right).$$

As in the strongly convex case, apply randomized rounding to h_S to obtain the output $\tilde{h}_S = R(h_S)$.

Proposition 3.11 (Replicability for Regularized ERM)**.** *Suppose $\delta < \rho/2$. If the rounding scheme uses grid size $a = \frac{2L\sqrt{d}}{(\rho-2\delta)\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)$, then $\tilde{h}_S$ is ρ-replicable.*

The proof is analogous to Proposition 3.8.

Generalization via uniform stability. To bound the excess risk, we appeal to the theory of uniform stability (Bousquet and Elisseeff, 2002), which provides a powerful tool for analyzing regularized algorithms. An algorithm is *uniformly stable* with rate $\alpha(n)$ if changing a single training example alters the output by at most $\alpha(n)$ in loss.

Definition 3.12 (Uniform Stability)**.** *An algorithm A is called uniformly stable with rate $\alpha(n)$ if for any dataset $S = \{z_1, \ldots, z_n\}$ and any $S^{(i)} = \{z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_n\}$ differing in one example, $\sup_z |\ell(A(S^{(i)}), z) - \ell(A(S), z)| \leq \alpha(n)$.*

It is well-known that μ-regularized ERM for L-Lipschitz losses is uniformly stable with rate $2L^2/(\mu n)$ (Shalev-Shwartz and Ben-David, 2014). This stability directly implies generalization: the expected excess risk satisfies $\mathbb{E}[L_{\mathcal{D}}(h_S)] - L_{\mathcal{D}}(h^*) \leq 2L^2/(\mu n) + (\mu/2)\|h^*\|_2^2$, where the second term reflects the bias from regularization.

For our rounded output $\tilde{h}_S$, the rounding error introduces an additional term bounded by $L \cdot a\sqrt{d}/2$ by Lipschitzness. Combining these observations yields the following result.

Proposition 3.13 (Generalization for Regularized ERM)**.** *Setting $\delta = \rho/4$, the rounding scheme with $a = \frac{2L\sqrt{d}}{(\rho-2\delta)\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)$ returns $\tilde{h}_S$ satisfying $\mathbb{E}_S[L_{\mathcal{D}}(\tilde{h}_S) - L_{\mathcal{D}}(h^*)] \leq \frac{\mu}{2}\|h^*\|_2^2 + \frac{2L^2(d+1)}{\mu\rho\sqrt{n}}\left(1+\sqrt{2\ln\frac{4}{\rho}}\right)$.*

Remark 3 (Excess Risk Rate for General Convex Optimization)**.** *Setting the regularization parameter μ to balance the bias and variance terms in Proposition 3.13, i.e., setting*

$$\mu^2 = \frac{4L^2(d+1)}{\|h^*\|_2^2\rho\sqrt{n}}\left(1+\sqrt{2\ln\frac{4}{\rho}}\right),$$

yields an excess risk of $\tilde{O}(\|h^*\|_2 L\sqrt{d}/(\sqrt{\rho}n^{\frac{1}{4}}))$.

This $n^{-\frac{1}{4}}$ rate is significantly slower than the standard $n^{-\frac{1}{2}}$ rate for non-replicable convex optimization, and slower than the n^{-1} rate we achieved for strongly convex problems. This suboptimality is inherent to the approach. Without strong convexity, we cannot simultaneously achieve sharp concentration (which requires large μ) and low bias (which requires small μ). The noise injection we use here is not well-suited to general convex problems. In contrast, the exponential mechanism of Section 3.1 automatically adapts to the curvature of the loss landscape and achieves the optimal $n^{-\frac{1}{2}}$ rate even for non-convex losses, though at exponential computational cost. This illustrates a fundamental computational statistical tradeoff in replicable learning. Optimal rates seem to require either exponential time (correlated sampling) or problem-specific structure (e.g., strong convexity).

3.5 Application to Neural Networks

We now demonstrate that our techniques extend beyond classical convex optimization to modern neural network training. We work in the *neural tangent kernel* (NTK) regime (Jacot et al., 2018; Du et al., 2019; Allen-Zhu et al., 2019), where gradient descent on an overparameterized network behaves like kernel ridge regression with a fixed kernel determined by the initialization. This reduction allows us to apply our RERM-based replicability approach from Section 3.4.

Problem Setup. Consider a binary classification problem with data (x, y) where $x \in \mathbb{R}^d$ with $\|x\|_2 = 1$ and $y \in \{-1, +1\}$. We train a two-layer ReLU network $f(x; W, \mathbf{a}) = \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s\sigma(\langle w_s, x\rangle)$, where $W \in \mathbb{R}^{m\times d}$ is the weight matrix with rows w_s, $\mathbf{a} \in \{-1,+1\}^m$ is a fixed random sign vector, $\sigma(z) = \max\{0, z\}$ is the ReLU activation, and m is the width. We initialize $w_{s,0} \sim \mathcal{N}(0, I_d)$ and $a_s \sim \mathrm{Unif}(\{-1,+1\})$, and train only the first layer weights W using the logistic loss $\ell(z) = \ln(1+\exp(-z))$. Following Ji and Telgarsky (2019), we assume the data satisfies a margin condition in the infinite-width kernel space:

Assumption 1 (NTK Separability)**.** *There exist $\gamma > 0$ and a mapping $\bar{v} : \mathbb{R}^d \to \mathbb{R}^d$ with $\|\bar{v}(z)\|_2 \leq 1$ such that for any $(x, y) \sim \mathcal{D}$,*

$$y\int_{\mathbb{R}^d}\langle\bar{v}(z), \phi_x(z)\rangle\, d\mu_{\mathcal{N}}(z) \geq \gamma,$$

where $\phi_x(z) = x \cdot \mathbb{1}[\langle z, x\rangle > 0]$ and $\mu_{\mathcal{N}}$ is the standard Gaussian measure on $\mathbb{R}^d$.

This assumption states that the data is linearly separable with margin γ in the feature space defined by the NTK at initialization. Under this condition, Ji and Telgarsky (2019) showed that gradient descent finds a global optimum efficiently when $m = \Omega(\mathrm{poly}(1/\gamma)\log n)$.

3.5.1 NTK Reduction to Kernel Method

The key insight of NTK theory is that for sufficiently large m, the network remains close to initialization throughout training, and its behavior is well-approximated by a linear model in the gradient features $\nabla_W f(x; W_0, \mathbf{a})$. We make this precise through a series of lemmas stated without proof; see Ji and Telgarsky (2019); Allen-Zhu et al. (2019).

Lemma 3.14. *If $m \geq 25\ln(2n/\delta)$, then with probability at least $1-2\delta$ over the random initialization, for all W satisfying $\|w_s - w_{s,0}\|_2 \leq r/\sqrt{m}, \forall s$ and all training examples x_i, $\left|f(x_i; W, \mathbf{a}) - \frac{1}{\sqrt{m}}\sum_{s=1}^m a_s\langle\phi_{x_i}(w_{s,0}), w_s - w_{s,0}\rangle\right| \leq C(m, n, \delta, r)$, where $C(m, n, \delta, r) = \sqrt{2\ln(4n/\delta)} + \sqrt{2/\pi}\cdot r^2/\sqrt{m} + r\sqrt{\ln(n/\delta)/(2m)}$.*

This shows that within a ball of radius $r/\sqrt{m}$ around initialization for any s (which we denote H_r), the network function is approximately a linear function of the weights. We can thus define a convex surrogate loss $\tilde{\ell}(W, z_i) = \ell(\frac{y_i}{\sqrt{m}}\sum_{s=1}^m a_s\langle\phi_{x_i}(w_{s,0}), w_s - w_{s,0}\rangle - C(m, n, \delta, r))$, and minimize the regularized empirical risk

$$W_S = \operatorname*{argmin}_{W \in H_r} \left[\tilde{L}_S(W) + \frac{\mu}{2} \|W - W_0\|_F^2 \right],$$

where $\tilde{L}_S(W) = \frac{1}{n} \sum_{i=1}^n \tilde{\ell}(W, z_i)$.

3.5.2 Replicability and Generalization

Since the surrogate loss is Lipschitz and we are minimizing a strongly convex objective, the analysis follows the pattern of Section 3.4. We apply randomized rounding in the md-dimensional space of weight matrices (view W as a vector in $\mathbb{R}^{md}$), getting $\tilde{W}_S = R(W_S)$ after projection back onto H_r.

Lemma 3.15. *With probability at least $1-2\delta$ over independent samples $(S, S') \sim \mathcal{D}^{2n}$,*

$$\|W_S - W_{S'}\|_F \leq \frac{2}{\mu\sqrt{n}} \left(1 + \sqrt{2 \ln \frac{1}{\delta}}\right).$$

Proposition 3.16 (Replicability for NTK)**.** *The rounding scheme with grid size $a = \frac{4\sqrt{md}}{\rho\mu\sqrt{n}} \left(1 + \sqrt{2 \ln \frac{4}{\rho}}\right)$ ensures $\tilde{W}_S$ is ρ-replicable.*

For generalization, we combine the empirical risk bound (following from stability of RERM) with a uniform convergence guarantee adapted to the NTK setting. The details involve controlling the variation of the network function over H_r using Rademacher complexity; we refer to Ji and Telgarsky (2019) for the precise argument.

Theorem 3.17 (Generalization bound)**.** *Let $\epsilon \in (0,1)$. Let $m \geq \left(163\sqrt{\ln \frac{4n}{\delta}} + 39 \ln \frac{1}{\epsilon}\right)^2 / \gamma^4$. Set $r = \frac{12}{\gamma}\sqrt{2 \ln \frac{4n}{\delta}} + \frac{4}{\gamma} \ln \frac{1}{\epsilon}$, $\mu = \sqrt{\frac{4md(1+\sqrt{2\ln(4/\rho)})}{\rho r^2 \sqrt{n}}}$. Then, w.p. at least $1 - 4\delta$, the population risk, $\mathbb{P}_{(x,y)\sim\mathcal{D}}[yf(x; \tilde{W}_S, \mathbf{a}) \leq 0]$ of $\tilde{W}_S$ is bounded by*

$$2\epsilon + 4\sqrt{\frac{mdr^2}{\rho\sqrt{n}}(1 + \sqrt{2\ln(4/\rho)})} + \frac{r}{\sqrt{n}} + 3\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Remark 4 (Excess Risk Rate for Neural Networks in the NTK Regime)**.** *Setting $m = \Theta\big((\ln(4n/\delta) + \ln^2(1/\epsilon))/\gamma^4\big)$ (the minimum width required for the NTK approximation to hold) and the corresponding $r = \widetilde{\Theta}(1/\gamma)$, the bound on the generalization error in Theorem 3.17 scales as $\tilde{O}\left(\epsilon + \frac{\sqrt{d}}{\gamma^3\sqrt{\rho}n^{1/4}} + \frac{1}{\gamma\sqrt{n}}\right)$. To achieve error $O(\epsilon)$, we require $n = \tilde{\Omega}\left(\frac{1}{\gamma^2\epsilon^2} + \frac{d^2}{\rho^2\gamma^{12}\epsilon^4}\right)$ samples. The first term $\tilde{\Omega}(1/\gamma^2\epsilon^2)$ is standard NTK sample complexity without replicability (Ji and Telgarsky, 2019). The second term $\tilde{\Omega}(d^2/(\rho^2\gamma^{12}\epsilon^4))$ reflects the cost of replicability: the dependence on d and ρ comes from the rounding scheme, while the $1/\epsilon^4$ term arises from the bias-variance tradeoff in RERM. Since the second term dominates, the overall population error is $\tilde{O}\left(\frac{\sqrt{d}}{\gamma^3\sqrt{\rho}n^{1/4}}\right)$.*

4 Discussion and Conclusion

We have developed a comprehensive framework for replicable stochastic optimization across convex and non-convex settings. Our exponential mechanism achieves minimax optimal $O(1/\sqrt{n})$ rates for general non-convex problems, demonstrating that replicability does not fundamentally limit statistical efficiency. However, optimal replicability probabilities via correlated sampling require exponential time. Polynomial-time algorithms using randomized rounding incur a $\sqrt{d}$ factor degradation in replicability probability but achieve near-optimal rates when problem structure (strong convexity) is available. For general convex problems, regularization forces a bias-variance tradeoff yielding suboptimal $n^{-1/4}$ rates, highlighting a gap between efficient and exponential-time methods.

Open questions. Several important directions remain: (1) Can efficient algorithms achieve optimal $n^{-1/2}$ rates for general convex replicable learning, or does this require exponential time? (2) Can replicability extend beyond the NTK regime to practical neural network training? (3) What are tight lower bounds for replicable learning as a function of problem structure? (4) How can these algorithms be made practical for large-scale problems?

Conclusion. Our work establishes replicability as a viable stability notion for machine learning, providing clear algorithmic paths forward depending on problem structure and computational constraints. The tools we develop, correlated sampling for statistical optimality, randomized rounding for efficiency, bridge the gap between the reproducibility needed for scientific progress and the inherent instability of stochastic algorithms. By providing formal guarantees that algorithms produce consistent results across independent runs, replicable learning offers a foundation for restoring trust in machine learning through mathematical guarantees enforced by algorithmic design.

Acknowledgement

This work was supported, in part, by NSF CAREER award IIS-1943251.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*. PMLR, 2019.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4), 2005.

Mohammad Bavarian, Badih Ghazi, Elad Haramaty, Pritish Kamath, Ronald L Rivest, and Madhu Sudan. Optimality of correlated sampling strategies. *ArXiv:1612.01041*, 2016.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, 2023.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Cynthia Dwork and Aaron Roth. *The algorithmic foundations of differential privacy*. Now Publishers Inc, 2014.

Eric Eaton, Marcel Hussing, Michael Kearns, and Jessica Sorrell. Replicable reinforcement learning. *Advances in Neural Information Processing Systems*, 36:15172–15185, 2023.

Oded Goldreich, Shafi Goldwasser, and Dana Ron. On the possibilities and limitations of pseudodeterministic algorithms. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 127–138, 2013.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Max Hopkins and Shay Moran. The role of randomness in stability. *arXiv preprint arXiv:2502.08007*, 2025.

Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–967, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Alkis Kalavasis, Amin Karbasi, Grigoris Velegkas, and Felix Zhou. On the computational landscape of replicable learning. *Advances in Neural Information Processing Systems*, 37:105887–105927, 2024.

Guy Kindler, Anup Rao, Ryan O'Donnell, and Avi Wigderson. Spherical cubes: optimal foams from computational hardness amplification. *Communications of the ACM*, 55(10):90–97, 2012.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

Moshe Noivirt, Jessica Sorrell, and Eliad Tsfadia. Computationally efficient replicable learning of parities. *arXiv preprint arXiv:2602.09499*, 2026.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. *Advances in Neural Information Processing Systems*, 21, 2009.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Not Applicable**]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**]
   (b) Complete proofs of all theoretical results. [**Yes**]
   (c) Clear explanations of any assumptions. [**Yes**]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Not Applicable**]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Not Applicable**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Not Applicable**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Not Applicable**]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Not Applicable**]
   (b) The license information of the assets, if applicable. [**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**]
   (d) Information about consent from data providers/curators. [**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

Supplementary Materials

A Detailed Proofs

This appendix contains complete proofs of all results stated in the main paper. We organize the proofs by the section in which the corresponding result appears.

Notation. Throughout the appendix, we use the following notation:

- $L_{\mathcal{D}}(h) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h,z)]$: population risk
- $L_S(h) = \frac{1}{n}\sum_{i=1}^{n}\ell(h,z_i)$: empirical risk on sample $S = \{z_1,\ldots,z_n\}$
- $h^* = \operatorname{argmin}_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$: population risk minimizer
- h_S: the output of an algorithm given dataset S
- $\tilde{h}_S = R(h_S)$: rounded output
- $[x]_+ = \max\{x, 0\}$: positive part

A.1 Missing Proofs in Section 3.1

Proof of Lemma 3.1. Let $\{h_1, h_2, \ldots, h_K\} \subseteq \mathcal{H}$ be an $\frac{\epsilon}{4L}$-net of $\mathcal{H}$, that is, $\forall h \in \mathcal{H}$, there exists $1 \le i \le K$, such that $\|h - h_i\|_2 \le \frac{\epsilon}{4L}$. By a standard packing number argument, we can construct the $\frac{\epsilon}{4L}$-net such that $h_1, h_2, \ldots, h_K$ are $\frac{\epsilon}{4L}$-separated, that is, $\forall i \ne j$, $\|h_i - h_j\|_2 > \frac{\epsilon}{4L}$. Therefore, the balls centered at h_i of radius $\frac{\epsilon}{8L}$ are disjoint balls. We can upper bound K via calculating the total volume of these balls: $K \cdot \left(\frac{\epsilon}{8L}\right)^d \le \left(B + \frac{\epsilon}{8L}\right)^d$. This implies $K \le \left(1 + \frac{8BL}{\epsilon}\right)^d$.

If $n \ge \frac{2M_0^2}{\epsilon^2}\cdot\left(\ln\left(\frac{2}{\delta}\right) + d\ln\left(\frac{8BL}{\epsilon}+1\right)\right)$, $\forall 1 \le i \le K$, Hoeffding's inequality gives $\mathbb{P}_{S\sim\mathcal{D}^n}[|L_S(h_i) - L_{\mathcal{D}}(h_i)| > \frac{\epsilon}{2}] \le 2\exp\left(-\frac{\epsilon^2 n}{2M_0^2}\right) \le \frac{\delta}{K}$. A union bound implies $\mathbb{P}_{S\sim\mathcal{D}^n}[|L_S(h_i) - L_{\mathcal{D}}(h_i)| \le \frac{\epsilon}{2}, \forall 1 \le i \le K] \ge 1 - \delta$.

From the definition of the covering number, $\forall h \in \mathcal{H}$, there exists $1 \le i \le K$, such that $\|h - h_i\|_2 \le \frac{\epsilon}{4L}$. $|L_S(h_i) - L_{\mathcal{D}}(h_i)| \le \frac{\epsilon}{2}$ implies

$$\begin{aligned}|L_S(h) - L_{\mathcal{D}}(h)| &\le |L_S(h) - L_S(h_i)| + |L_S(h_i) - L_{\mathcal{D}}(h_i)| + |L_{\mathcal{D}}(h_i) - L_{\mathcal{D}}(h)| \\ &\le L\cdot\|h - h_i\|_2 + \frac{\epsilon}{2} + L\cdot\|h - h_i\|_2 \le \epsilon.\end{aligned}$$

Therefore, $\mathbb{P}_{S\sim\mathcal{D}^n}[|L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon, \forall h \in \mathcal{H}] \ge 1 - \delta$. □

Proof of Proposition 3.2. In this Proposition, we assume $0 \le L_\alpha(h) \le M_0$. We first prove part 1. Denote

$p_\alpha := \mathbb{P}_{(\xi_t, y_t)\sim Q_0 \times \mathrm{Unif}([0,1])}[y_t \le \exp(-\eta L_\alpha(\xi_t))]$. For any measurable set $A \subseteq \mathcal{H}$,

$$\begin{aligned}\mathbb{P}[A(P_\alpha) \in A] &= \sum_{t=1}^{\infty} [\text{output } \xi_t \text{ and } \xi_t \in A] \\ &= \sum_{t=1}^{\infty} (1-p_\alpha)^{t-1} \cdot \int_A \exp(-\eta L_\alpha(h)) dQ_0(h) \\ &= \frac{1}{p_\alpha} \int_A \exp(-\eta L_\alpha(h)) dQ_0(h).\end{aligned}$$

The density $p_\alpha(h) = \frac{1}{p_\alpha}\exp(-\eta L_\alpha(h)) \propto \exp(-\eta L_\alpha(h))$. For the second part, we first define $m_t = \min\{\exp(-\eta L_\alpha(\xi_t)), \exp(-\eta L_\beta(\xi_t))\}$ and $M_t = \max\{\exp(-\eta L_\alpha(\xi_t)), \exp(-\eta L_\beta(\xi_t))\}$.

$$1 \ge \frac{m_t}{M_t} = \exp(-\eta \cdot |L_\alpha(\xi_t) - L_\beta(\xi_t)|) \ge \exp\left(-\eta \cdot \sup_{h\in\mathcal{H}} |L_\alpha(h) - L_\beta(h)|\right) \ge 1 - \eta \cdot \sup_{h\in\mathcal{H}} |L_\alpha(h) - L_\beta(h)|.$$

Condition on $\{\xi_t\}_{t\ge 1}$,

$$\begin{aligned}\mathbb{P}\left[A(P_\alpha) = A(P_\beta)\right] &\ge \sum_{t=1}^{\infty} \mathbb{P}[A(P_\alpha) \text{ and } A(P_\beta) \text{ both output } \xi_t] \\ &= \sum_{t=1}^{\infty} \prod_{s=1}^{t-1} \mathbb{P}\left[y_s > M_s\right] \cdot \mathbb{P}\left[y_t \le m_t\right] \\ &= \sum_{t=1}^{\infty} \prod_{s=1}^{t-1} (1 - M_s) \cdot m_t \\ &\ge \left(1 - \eta \cdot \sup_{h\in\mathcal{H}} |L_\alpha(h) - L_\beta(h)|\right) \sum_{t=1}^{\infty} \prod_{s=1}^{t-1} (1-M_s) \cdot M_t \\ &= \left(1 - \eta \cdot \sup_{h\in\mathcal{H}} |L_\alpha(h) - L_\beta(h)|\right) \left(1 - \prod_{t=1}^{\infty} (1 - M_t)\right) \\ &= 1 - \eta \cdot \sup_{h\in\mathcal{H}} |L_\alpha(h) - L_\beta(h)|. \qquad (M_t \ge \exp(-\eta M_0) > 0 \Rightarrow \textstyle\prod_{t=1}^{\infty}(1-M_t) = 0)\end{aligned}$$

Therefore, $\mathbb{P}\left[A(P_\alpha) \ne A(P_\beta)\right] = 1 - \mathbb{P}\left[A(P_\alpha) = A(P_\beta)\right] \le \eta \cdot \sup_{h\in\mathcal{H}} |L_\alpha(h) - L_\beta(h)|$. □

Proof of Proposition 3.3. By Lemma 3.1, with probability at least $1 - 2\delta$ over independent samples S, S', we have $\sup_h |L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon$ and $\sup_h |L_{S'}(h) - L_{\mathcal{D}}(h)| \le \epsilon$. Triangle inequality implies $\sup_h |L_S(h) - L_{S'}(h)| \le 2\epsilon$. Let h_S denote the output of Algorithm 2 with loss function $L_S(h)$.

$$\begin{aligned}\mathbb{P}\left[h_S \ne h_{S'}\right] &= \mathbb{P}\left[h_S \ne h_{S'}, \sup_h |L_S(h) - L_{S'}(h)| > 2\epsilon\right] + \mathbb{P}\left[h_S \ne h_{S'}, \sup_h |L_S(h) - L_{S'}(h)| \le 2\epsilon\right] \\ &\le \mathbb{P}\left[\sup_h |L_S(h) - L_{S'}(h)| > 2\epsilon\right] + \mathbb{P}\left[h_S \ne h_{S'}, \sup_h |L_S(h) - L_{S'}(h)| \le 2\epsilon\right] \\ &\le 2\delta + \mathbb{P}\left[h_S \ne h_{S'} \middle| \sup_h |L_S(h) - L_{S'}(h)| \le 2\epsilon\right] \\ &\le 2\delta + \eta \cdot 2\epsilon = \rho. \qquad \text{(Proposition 3.2)}\end{aligned}$$

□

Proof of Lemma 3.4. Let $h_S^* \in \operatorname*{argmin}_{h\in\mathcal{H}} L_S(h)$ denote the empirical risk minimizer. Denote $H_1 := \{h \in \mathcal{H} | L_S(h) - L_S(h_S^*) \geq \alpha\}$ and $H_2 := \left\{h \in \mathcal{H} \middle| L_S(h) - L_S(h_S^*) \leq \frac{\alpha}{2}\right\}$.

$$\begin{aligned}
\mathbb{P}\left[L_S(h_S) - \inf_h L_S(h) \geq \alpha\right] =& \mathbb{P}\left[h_S \in H_1\right] \\
=& \frac{\int_{H_1} \exp(-\eta L_S(h)) dQ_0(h)}{\int_{\mathcal{H}} \exp(-\eta L_S(h)) dQ_0(h)} \\
\leq& \frac{\int_{H_1} \exp(-\eta L_S(h)) dQ_0(h)}{\int_{H_2} \exp(-\eta L_S(h)) dQ_0(h)} \\
\leq& \exp(-\frac{\eta\alpha}{2}) \frac{Q_0(H_1)}{Q_0(H_2)} \\
\leq& \exp(-\frac{\eta\alpha}{2}) \frac{Q_0(\mathcal{H})}{Q_0(H_2)}.
\end{aligned}$$

Since $L_S(h)$ is L-Lipschitz, $H_3 := \left\{h \in \mathcal{H} \middle| \|h - h_S^*\|_2 \leq \frac{\alpha}{2L}\right\} \subseteq H_2$. It is easy to verify that H_3 contains a ball of radius $\min\{\frac{\alpha}{4L}, B\}$. Therefore,

$$\frac{Q_0(\mathcal{H})}{Q_0(H_2)} \leq \frac{Q_0(\mathcal{H})}{Q_0(H_3)} \leq \frac{B^d}{\min\{\frac{\alpha}{4L}, B\}^d} = \max\left\{\left(\frac{4BL}{\alpha}\right)^d, 1\right\}.$$

$$\mathbb{P}\left[L_S(h_S) - \inf_h L_S(h) \geq \alpha\right] \leq \exp(-\frac{\eta\alpha}{2}) \frac{Q_0(\mathcal{H})}{Q_0(H_2)} \leq \exp(-\frac{\eta\alpha}{2}) \max\left\{\left(\frac{4BL}{\alpha}\right)^d, 1\right\}.$$

□

Proof of Proposition 3.5. Applying $\alpha = \frac{2}{\eta}\left(\ln(1/\delta) + \left[d \cdot \ln\left(\frac{(\rho-2\delta)BL}{\epsilon \ln(1/\delta)}\right)\right]_+\right)$ in Lemma 3.4, we know that

$$\begin{aligned}
\mathbb{P}\left[L_S(h_S) - \inf_h L_S(h) \geq \alpha\right] &\leq \exp\left(-\frac{\eta\alpha}{2}\right) \max\left\{\left(\frac{4BL}{\alpha}\right)^d, 1\right\} \\
&= \max\left\{\exp\left(-\frac{\eta\alpha}{2}\right) \cdot \left(\frac{4BL}{\alpha}\right)^d, \exp\left(-\frac{\eta\alpha}{2}\right)\right\} \\
&\leq \max\left\{\exp\left(-\frac{\eta\alpha}{2}\right) \cdot \left(\frac{4BL}{\frac{2}{\eta}\ln(1/\delta)}\right)^d, \delta\right\} && (\alpha \geq \tfrac{2}{\eta}\ln(1/\delta)) \\
&= \max\left\{\exp\left(-\frac{\eta\alpha}{2}\right) \cdot \left(\frac{(\rho-2\delta)BL}{\epsilon \ln(1/\delta)}\right)^d, \delta\right\} && (\eta = \tfrac{\rho-2\delta}{2\epsilon}) \\
&\leq \delta,
\end{aligned}$$

where we used $\alpha \geq \frac{2}{\eta}\left(\ln(1/\delta) + d \cdot \ln\left(\frac{(\rho-2\delta)BL}{\epsilon \ln(1/\delta)}\right)\right)$ in the last inequality. Therefore, with probability $\geq 1-\delta$, $L_S(h_S) - \inf_h L_S(h) \leq \frac{2}{\eta}\left(\ln(1/\delta) + \left[d \cdot \ln\left(\frac{(\rho-2\delta)BL}{\epsilon \ln(1/\delta)}\right)\right]_+\right)$. Lemma 3.1 gives that with probability $\geq 1-\delta$, $\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$. A union bound implies these two events hold simultaneously with

probability $\geq 1-2\delta$. Assume these two events hold and recall that $h^* \in \operatorname*{argmin}_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$. We have

$$\begin{aligned}
L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \epsilon \\
&\leq \inf_h L_S(h) + \frac{2}{\eta}\left(\ln(1/\delta) + \left[d\cdot\ln\left(\frac{(\rho-2\delta)BL}{\epsilon\ln(1/\delta)}\right)\right]_+\right) + \epsilon \\
&\leq L_S(h^*) + \frac{2}{\eta}\left(\ln(1/\delta) + \left[d\cdot\ln\left(\frac{(\rho-2\delta)BL}{\epsilon\ln(1/\delta)}\right)\right]_+\right) + \epsilon \\
&\leq L_{\mathcal{D}}(h^*) + \frac{2}{\eta}\left(\ln(1/\delta) + \left[d\cdot\ln\left(\frac{(\rho-2\delta)BL}{\epsilon\ln(1/\delta)}\right)\right]_+\right) + 2\epsilon \\
&= L_{\mathcal{D}}(h^*) + 2\epsilon\left(1 + \frac{2}{\rho-2\delta}\left(\ln\left(\frac{1}{\delta}\right) + \left[d\cdot\ln\left(\frac{(\rho-2\delta)BL}{\epsilon\ln(1/\delta)}\right)\right]_+\right)\right).
\end{aligned}$$

□

A.2 Missing Proof in Section 3.2

Proof of Proposition 3.6. For part 1 of this proposition, the maximum distance from any point to the center of its grid cell is achieved at the corners, where the distance is $\frac{a}{2}\sqrt{d}$. Now we look at part 2. Condition on any fixed basis vector v_i. Since the partition offset is chosen uniformly at random, the probability that x and x' cross the grid boundary in the direction of v_i is

$$\begin{aligned}
&\mathbb{P}\left[x \text{ and } x' \text{ cross the grid boundary in the direction of } v_i \middle| v_i\right] \\
=&\max\left\{\frac{|\langle x-x', v_i\rangle|}{a}, 1\right\} \\
\leq&\frac{|\langle x-x', v_i\rangle|}{a}.
\end{aligned}$$

Now we take the expectation over the random basis v_i. Note that the distribution of v_i is uniform on the unit sphere.

$$\begin{aligned}
&\mathbb{P}\left[x \text{ and } x' \text{ cross the grid boundary in the direction of } v_i\right] \\
\leq&\mathbb{E}_{v_i}\frac{|\langle x-x', v_i\rangle|}{a} \\
\leq&\frac{\sqrt{\mathbb{E}_{v_i}\left[\langle x-x', v_i\rangle^2\right]}}{a} \\
=&\frac{\sqrt{(x-x')^\top\mathbb{E}_{v_i}\left[v_iv_i^\top\right](x-x')}}{a} \\
=&\frac{\|x-x'\|_2}{a\sqrt{d}}.
\end{aligned}$$

A union bound implies

$$\mathbb{P}\left[x \text{ and } x' \text{ cross at least one grid boundary}\right] \leq \frac{\sqrt{d}\|x-x'\|_2}{a}.$$

□

A.3 Missing Proofs in Section 3.3

Proof of Lemma 3.7. $h^* = \operatorname{argmin}_h L_{\mathcal{D}}(h)$ implies $\nabla_h L_{\mathcal{D}}(h^*) = \mathbb{E}_{z\sim\mathcal{D}}\left[\nabla_h \ell(h^*, z)\right] = 0$. The empirical approximation of this gradient is $\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_h \ell(h^*, z_i)\right]$. Define $f(z_1, \ldots, z_n) := \left\|\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_h \ell(h^*, z_i)\right]\right\|_2$.

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^n} f(z_1, \ldots, z_n) &= \mathbb{E}_{S\sim\mathcal{D}^n}\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_h \ell(h^*, z_i)\right]\right\|_2 \\
&\leq \sqrt{\mathbb{E}_{S\sim\mathcal{D}^n}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_h \ell(h^*, z_i)\right]\right\|_2^2\right]} \\
&= \sqrt{\frac{1}{n}\mathbb{E}_{z\sim\mathcal{D}}\left[\|\nabla_h \ell(h^*, z)\|_2^2\right]} \leq \frac{M}{\sqrt{n}}.
\end{aligned}$$

From McDiarmid's inequality, $\forall t > 0$,

$$\mathbb{P}\left[f(z_1, \ldots, z_n) - \mathbb{E}_{S\sim\mathcal{D}^n} f(z_1, \ldots, z_n) > t\right] \leq \exp\left(-\frac{t^2 n}{2M^2}\right).$$

Setting $t = M\sqrt{\frac{2\ln\frac{1}{\delta}}{n}}$, we obtain with probability $\geq 1-\delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_h \ell(h^*, z_i)\right]\right\|_2 \leq \frac{M}{\sqrt{n}} + t = \frac{M}{\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right).$$

The strong convexity of $L_S(h)$ implies

$$\begin{aligned}
\|h_S - h^*\|_2 &\leq \frac{1}{\mu}\|\nabla_h L_S(h_S) - \nabla_h L_S(h^*)\|_2 \\
&= \frac{1}{\mu}\|\nabla_h L_S(h^*)\|_2 \\
&= \frac{1}{\mu}\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_h \ell(h^*, z_i)\right]\right\|_2 \\
&\leq \frac{M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right).
\end{aligned}$$

□

Proof of Proposition 3.8. By Lemma 3.7, with probability at least $1-2\delta$ over independent samples S, S', we have $\|h_S - h^*\|_2 \leq \frac{M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right)$ and $\|h_{S'} - h^*\|_2 \leq \frac{M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right)$. Triangle inequality implies $\|h_S - h_{S'}\|_2 \leq \frac{2M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right)$. We apply Proposition 3.6 to ensure replicability.

$$\begin{aligned}
&\mathbb{P}\left[\tilde{h}_S \neq \tilde{h}_{S'}\right] \\
=&\mathbb{P}\left[\tilde{h}_S \neq \tilde{h}_{S'}, \|h_S - h_{S'}\|_2 > \frac{2M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right)\right] + \mathbb{P}\left[\tilde{h}_S \neq \tilde{h}_{S'}, \|h_S - h_{S'}\|_2 \leq \frac{2M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right)\right] \\
\leq&\mathbb{P}\left[\|h_S - h_{S'}\|_2 > \frac{2M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right)\right] + \mathbb{P}\left[\tilde{h}_S \neq \tilde{h}_{S'} \middle| \|h_S - h_{S'}\|_2 \leq \frac{2M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right)\right] \\
\leq&2\delta + \frac{\sqrt{d}}{a}\cdot\frac{2M}{\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{1}{\delta}}\right) = \rho. \qquad \text{(Proposition 3.6)}
\end{aligned}$$

□

Proof of Proposition 3.9. From Proposition 3.6, $\|\tilde{h}_S - h_S\|_2 \leq \frac{a\sqrt{d}}{2} = \frac{Md}{(\rho-2\delta)\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)$. From Lemma 3.7, with probability at least $1-\delta$, $\|h_S - h^*\|_2 \leq \frac{M}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)$. Triangle inequality implies with probability $\geq 1-\delta$,

$$\|\tilde{h}_S - h^*\|_2 \leq \frac{Md}{(\rho-2\delta)\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right) + \frac{M}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right) \leq \frac{M(d+1)}{(\rho-2\delta)\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right). \quad (\rho \leq 1)$$

□

A.4 Missing Proofs in Section 3.4

Proof of Lemma 3.10. Denote $h^*_\mu = \operatorname{argmin}_h \left[L_{\mathcal{D}}(h) + \frac{\mu}{2}\|h\|_2^2\right]$. It implies $\nabla_h L_{\mathcal{D}}(h^*_\mu) + \mu h^*_\mu = 0$. The empirical approximation of this term is $\frac{1}{n}\sum_{i=1}^n \left[\nabla_h \ell(h^*_\mu, z_i) + \mu h^*_\mu\right]$. Define $f(z_1,\ldots,z_n) := \left\|\frac{1}{n}\sum_{i=1}^n \left[\nabla_h \ell(h^*_\mu, z_i) + \mu h^*_\mu\right]\right\|_2$.

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^n} f(z_1,\ldots,z_n) &= \mathbb{E}_{S\sim\mathcal{D}^n}\left\|\frac{1}{n}\sum_{i=1}^n \left[\nabla_h \ell(h^*_\mu, z_i) + \mu h^*_\mu\right]\right\|_2 \\
&\leq \sqrt{\mathbb{E}_{S\sim\mathcal{D}^n}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left[\nabla_h \ell(h^*_\mu, z_i) + \mu h^*_\mu\right]\right\|_2^2\right]} \\
&= \sqrt{\frac{1}{n}\mathbb{E}_{z\sim\mathcal{D}}\left[\|\nabla_h \ell(h^*_\mu, z) + \mu h^*_\mu\|_2^2\right]} \\
&= \sqrt{\frac{1}{n}\mathbb{E}_{z\sim\mathcal{D}}\left[\|\nabla_h \ell(h^*_\mu, z)\|_2^2 - \|\mu h^*_\mu\|_2^2\right]} \leq \frac{L}{\sqrt{n}}.
\end{aligned}$$

From McDiarmid's inequality, $\forall t > 0$,

$$\mathbb{P}\left[f(z_1,\ldots,z_n) - \mathbb{E}_{S\sim\mathcal{D}^n} f(z_1,\ldots,z_n) > t\right] \leq \exp\left(-\frac{t^2 n}{2L^2}\right).$$

Setting $t = L\sqrt{\frac{2\ln\frac{1}{\delta}}{n}}$, we obtain with probability $\geq 1-\delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \left[\nabla_h \ell(h^*_\mu, z_i) + \mu h^*_\mu\right]\right\|_2 \leq \frac{L}{\sqrt{n}} + t = \frac{L}{\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right).$$

The strong convexity of $L_S(h) + \frac{\mu}{2}\|h\|_2^2$ implies

$$\begin{aligned}
\|h_S - h^*_\mu\|_2 &\leq \frac{1}{\mu}\|\nabla_h L_S(h_S) + \mu h_S - \nabla_h L_S(h^*_\mu) - \mu h^*_\mu\|_2 \\
&= \frac{1}{\mu}\|\nabla_h L_S(h^*_\mu) + \mu h^*_\mu\|_2 \\
&= \frac{1}{\mu}\left\|\frac{1}{n}\sum_{i=1}^n \left[\nabla_h \ell(h^*, z_i) + \mu h^*_\mu\right]\right\|_2 \\
&\leq \frac{L}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right).
\end{aligned}$$

Similarly, with probability $\geq 1-\delta$ over S', $\|h_{S'} - h^*_\mu\|_2 \leq \frac{L}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)$. A union bound and triangle inequality imply that with probability $\geq 1-2\delta$,

$$\|h_S - h_{S'}\|_2 \leq \|h_S - h^*_\mu\|_2 + \|h_{S'} - h^*_\mu\|_2 \leq \frac{2L}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right).$$

□

Proof of Proposition 3.11. By Lemma 3.10, with probability at least $1-2\delta$ over independent samples S, S', we have $\|h_S - h_{S'}\|_2 \leq \frac{2L}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)$. We apply Proposition 3.6 to ensure replicability.

$$\begin{aligned}
&\mathbb{P}\left[\tilde{h}_S \neq \tilde{h}_{S'}\right] \\
=&\mathbb{P}\left[\tilde{h}_S \neq \tilde{h}_{S'}, \|h_S - h_{S'}\|_2 > \frac{2L}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)\right] + \mathbb{P}\left[\tilde{h}_S \neq \tilde{h}_{S'}, \|h_S - h_{S'}\|_2 \leq \frac{2L}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)\right] \\
\leq&\mathbb{P}\left[\|h_S - h_{S'}\|_2 > \frac{2L}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)\right] + \mathbb{P}\left[\tilde{h}_S \neq \tilde{h}_{S'} \middle| \|h_S - h_{S'}\|_2 \leq \frac{2L}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)\right] \\
\leq&2\delta + \frac{\sqrt{d}}{a}\cdot\frac{2L}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right) = \rho. \qquad \text{(Proposition 3.6)}
\end{aligned}$$

□

Proof of Proposition 3.13. We first establish uniform stability of regularized ERM (Step 1), then use stability to bound the generalization gap (Step 2), and finally combine with rounding error to obtain the excess risk bound (Step 3).

Step 1: Uniform stability. We show that the μ-regularized ERM algorithm is uniformly stable with rate $\frac{2L^2}{\mu n}$. Recall that for $S = \{z_1, \ldots, z_n\}$ and $S^{(i)} = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n\}$, denote $h_S = \operatorname{argmin}_h \left[L_S(h) + \frac{\mu}{2}\|h\|_2^2\right]$ and $h_{S^{(i)}} = \operatorname{argmin}_h \left[L_{S^{(i)}}(h) + \frac{\mu}{2}\|h\|_2^2\right]$. By the strong convexity of $L_S(h) + \frac{\mu}{2}\|h\|_2^2$ and $L_{S^{(i)}}(h) + \frac{\mu}{2}\|h\|_2^2$, we have

$$L_S(h_{S^{(i)}}) + \frac{\mu}{2}\|h_{S^{(i)}}\|_2^2 \geq L_S(h_S) + \frac{\mu}{2}\|h_S\|_2^2 + \frac{\mu}{2}\|h_{S^{(i)}} - h_S\|_2^2$$

and

$$L_{S^{(i)}}(h_S) + \frac{\mu}{2}\|h_S\|_2^2 \geq L_{S^{(i)}}(h_{S^{(i)}}) + \frac{\mu}{2}\|h_{S^{(i)}}\|_2^2 + \frac{\mu}{2}\|h_S - h_{S^{(i)}}\|_2^2.$$

Adding these two equations, we get

$$\mu\|h_{S^{(i)}} - h_S\|_2^2 \leq \frac{1}{n}\left[\ell(h_{S^{(i)}}, z_i) - \ell(h_S, z_i) + \ell(h_S, z'_i) - \ell(h_{S^{(i)}}, z'_i)\right] \leq \frac{2L}{n}\|h_{S^{(i)}} - h_S\|_2.$$

This implies $\|h_{S^{(i)}} - h_S\|_2 \leq \frac{2L}{\mu n}$. We thus have $\sup_z |\ell(h_{S^{(i)}}, z) - \ell(h_S, z)| \leq L \cdot \|h_{S^{(i)}} - h_S\|_2 \leq \frac{2L^2}{\mu n}$, which is exactly uniform stability.

Step 2: Generalization Gap. The generalization gap of h_S can be bounded as

$$\begin{aligned}
&\mathbb{E}_{S\sim\mathcal{D}^n}\left[L_{\mathcal{D}}(h_S) - L_S(h_S)\right] \\
=&\mathbb{E}_{\{z_1,\ldots,z_n,z'_1,\ldots,z'_n\}\sim\mathcal{D}^{2n}}\left[\frac{1}{n}\sum_{i=1}^n \ell(h_{S^{(i)}}, z_i) - \frac{1}{n}\sum_{i=1}^n \ell(h_S, z_i)\right]
\end{aligned}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\{z_1,\ldots,z_n,z_1',\ldots,z_n'\}\sim\mathcal{D}^{2n}}\left[\ell(h_{S^{(i)}},z_i)-\ell(h_S,z_i)\right]\leq\frac{2L^2}{\mu n}.$$

Step 3: Excess Risk Bound. The excess risk of $\tilde{h}_S$ can be bounded as

$$\begin{aligned}
&\mathbb{E}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(\tilde{h}_S)-L_{\mathcal{D}}(h^*)]\\
=&\mathbb{E}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(\tilde{h}_S)-L_{\mathcal{D}}(h_S)]+\mathbb{E}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(h_S)-L_S(h_S)]+\mathbb{E}_{S\sim\mathcal{D}^n}[L_S(h_S)-L_{\mathcal{D}}(h^*)]\\
\leq&L\|\tilde{h}_S-h_S\|_2+\frac{2L^2}{\mu n}+\mathbb{E}_{S\sim\mathcal{D}^n}\left[L_S(h_S)+\frac{\mu}{2}\|h_S\|_2^2-L_{\mathcal{D}}(h^*)\right]\\
\leq&L\cdot\frac{a}{2}\sqrt{d}+\frac{2L^2}{\mu n}+\mathbb{E}_{S\sim\mathcal{D}^n}\left[L_S(h^*)+\frac{\mu}{2}\|h^*\|_2^2-L_{\mathcal{D}}(h^*)\right] && \text{(Proposition 3.6+definition of RERM)}\\
=&\frac{L^2d}{(\rho-2\delta)\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)+\frac{2L^2}{\mu n}+\frac{\mu}{2}\|h^*\|_2^2\\
=&\frac{\mu}{2}\|h^*\|_2^2+\frac{2L^2}{\mu n}+\frac{2L^2d}{\mu\rho\sqrt{n}}\left(1+\sqrt{2\ln\frac{4}{\rho}}\right) && (\delta=\tfrac{\rho}{4})\\
\leq&\frac{\mu}{2}\|h^*\|_2^2+\frac{2L^2(d+1)}{\mu\rho\sqrt{n}}\left(1+\sqrt{2\ln\frac{4}{\rho}}\right). && (\rho\leq 1)
\end{aligned}$$

□

Remark. The $n^{-1/4}$ rate in Proposition 3.13 is suboptimal compared to the $n^{-1/2}$ rate achieved by the exponential mechanism (Proposition 3.5). This gap arises because Gaussian noise and randomized rounding are not well-adapted to general convex problems. The regularization parameter μ must balance approximation bias ($\mu\|h^*\|_2^2/2$) against the variance of the regularized solution ($\propto 1/\mu$), leading to a bias-variance tradeoff that yields $n^{-1/4}$ rates. The exponential mechanism circumvents this by sampling directly from a distribution adapted to the loss landscape.

A.5 Missing Proofs in Section 3.5

We first restate a few results in Ji and Telgarsky (2019).

Given the initialization $(W_0,\mathbf{a})$, for any $1\leq s\leq m$, define $\bar{u}_s:=\frac{1}{\sqrt{m}}a_s\bar{v}(w_{s,0})$, where $\bar{v}$ is given by Assumption 1. Collect $\bar{u}_s$ into the rows of the matrix $\overline{U}\in\mathbb{R}^{m\times d}$. It holds that $\|\bar{u}_s\|_2\leq 1/\sqrt{m}$ and $\|\overline{U}\|_F\leq 1$.

Lemma A.1 (Lemma 2.3 in Ji and Telgarsky (2019))**.** *Under Assumption 1, given any $\delta\in(0,1)$, with probability at least $1-\delta$ over the random initialization, it holds simultaneously for all $1\leq i\leq n$ that*

$$y_i\left\langle\nabla_W f(x_i;W_0,\mathbf{a}),\overline{U}\right\rangle\geq\gamma-\sqrt{\frac{2\ln(n/\delta)}{m}}.$$

For any W, and any $\epsilon>0$, and any $1\leq i\leq n$, define

$$\alpha_i(W,\epsilon)=\frac{1}{m}\sum_{s=1}^{m}\mathbb{1}\left[|\langle w_s,x_i\rangle|\leq\epsilon\right].$$

Lemma A.2 (Lemma 2.4 in Ji and Telgarsky (2019))**.** *For any $\epsilon>0$, with probability at least $1-\delta$ over the random initialization, it holds simultaneously for all $1\leq i\leq n$ that*

$$\alpha_i(W_0,\epsilon)\leq\sqrt{\frac{2}{\pi}}\epsilon+\sqrt{\frac{\ln(n/\delta)}{2m}}.$$

The next lemma controls the output of the network at initialization.

Lemma A.3 (Lemma 2.5 in Ji and Telgarsky (2019)). *Given any $\delta \in (0,1)$, if $m \geq 25\ln(2n/\delta)$, then with probability at least $1-\delta$ over the random initialization, it holds simultaneously for all $1 \leq i \leq n$ that $|f(x_i; W_0, \mathbf{a})| \leq \sqrt{2\ln(4n/\delta)}$.*

Now we can prove Lemma 3.14.

Proof of Lemma 3.14. Combining Lemma A.3 and Lemma A.2 with $\epsilon = r/\sqrt{m}$, we know that with probability at least $1-2\delta$ over the random initialization, it holds simultaneously for all $1 \leq i \leq n$ that $|f(x_i; W_0, \mathbf{a})| \leq \sqrt{2\ln(4n/\delta)}$ and $\alpha_i(W_0, r/\sqrt{m}) \leq \sqrt{2/\pi} \cdot r/\sqrt{m} + \sqrt{\frac{\ln(n/\delta)}{2m}}$. For all W satisfying $\|w_s - w_{s,0}\|_2 \leq r/\sqrt{m}, \forall s$,

$$
\begin{aligned}
&\left| f(x_i; W, \mathbf{a}) - f(x_i; W_0, \mathbf{a}) - \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \langle \phi_{x_i}(w_{s,0}), w_s - w_{s,0} \rangle \right| \\
=&\left| \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma(\langle w_s, x_i \rangle) - \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \sigma(\langle w_{s,0}, x_i \rangle) - \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \mathbb{1}[\langle w_{s,0}, x_i \rangle > 0] \cdot \langle w_s - w_{s,0}, x_i \rangle \right| \\
=&\left| \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \left(\sigma(\langle w_s, x_i \rangle) - \sigma(\langle w_{s,0}, x_i \rangle) - \mathbb{1}[\langle w_{s,0}, x_i \rangle > 0] \cdot \langle w_s - w_{s,0}, x_i \rangle \right) \right|. \qquad (1)
\end{aligned}
$$

We consider two different types of s.

First, if $|\langle w_{s,0}, x_i \rangle| > r/\sqrt{m}$, combining with $|\langle w_s - w_{s,0}, x_i \rangle| \leq \|w_s - w_{s,0}\|_2 \leq r/\sqrt{m}$, we know that $\langle w_s, x_i \rangle$ and $\langle w_{s,0}, x_i \rangle$ have the same sign, and thus

$$
\sigma(\langle w_s, x_i \rangle) - \sigma(\langle w_{s,0}, x_i \rangle) - \mathbb{1}[\langle w_{s,0}, x_i \rangle > 0] \cdot \langle w_s - w_{s,0}, x_i \rangle = 0.
$$

Second, if $|\langle w_{s,0}, x_i \rangle| \leq r/\sqrt{m}$, then

$$
|\sigma(\langle w_s, x_i \rangle) - \sigma(\langle w_{s,0}, x_i \rangle) - \mathbb{1}[\langle w_{s,0}, x_i \rangle > 0] \cdot \langle w_s - w_{s,0}, x_i \rangle| \leq |\langle w_s - w_{s,0}, x_i \rangle| \leq r/\sqrt{m}.
$$

We can use the above discussion to simplify equation 1.

$$
\begin{aligned}
&\left| f(x_i; W, \mathbf{a}) - f(x_i; W_0, \mathbf{a}) - \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \langle \phi_{x_i}(w_{s,0}), w_s - w_{s,0} \rangle \right| \\
=&\left| \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \left(\sigma(\langle w_s, x_i \rangle) - \sigma(\langle w_{s,0}, x_i \rangle) - \mathbb{1}[\langle w_{s,0}, x_i \rangle > 0] \cdot \langle w_s - w_{s,0}, x_i \rangle \right) \right| \\
=&\left| \frac{1}{\sqrt{m}} \sum_{s:|\langle w_{s,0}, x_i \rangle| \leq r/\sqrt{m}} a_s \left(\sigma(\langle w_s, x_i \rangle) - \sigma(\langle w_{s,0}, x_i \rangle) - \mathbb{1}[\langle w_{s,0}, x_i \rangle > 0] \cdot \langle w_s - w_{s,0}, x_i \rangle \right) \right| \\
\leq&\frac{1}{\sqrt{m}} \sum_{s:|\langle w_{s,0}, x_i \rangle| \leq r/\sqrt{m}} |\sigma(\langle w_s, x_i \rangle) - \sigma(\langle w_{s,0}, x_i \rangle) - \mathbb{1}[\langle w_{s,0}, x_i \rangle > 0] \cdot \langle w_s - w_{s,0}, x_i \rangle| \\
\leq&\frac{1}{\sqrt{m}} \cdot m \cdot \alpha_i(W_0, r/\sqrt{m}) \cdot r/\sqrt{m} \\
\leq&\sqrt{2/\pi} \cdot r^2/\sqrt{m} + r\sqrt{\ln(n/\delta)/(2m)}.
\end{aligned}
$$

Triangle inequality implies

$$
\left| f(x_i; W, \mathbf{a}) - \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \langle \phi_{x_i}(w_{s,0}), w_s - w_{s,0} \rangle \right|
$$

$$\leq \left| f(x_i; W, \mathbf{a}) - f(x_i; W_0, \mathbf{a}) - \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \langle \phi_{x_i}(w_{s,0}), w_s - w_{s,0} \rangle \right| + |f(x_i; W_0, \mathbf{a})|$$
$$\leq \sqrt{2\ln(4n/\delta)} + \sqrt{2/\pi} \cdot r^2/\sqrt{m} + r\sqrt{\ln(n/\delta)/(2m)} = C(m, n, \delta, r).$$

□

From Lemma 3.14, we can show that the original logistic loss is upper bounded by the convex surrogate loss. With probability at least $1 - 2\delta$ over the random initialization,

$$\begin{aligned}
\ell(W, z_i) &= \ell(y_i f(x_i; W, \mathbf{a})) \\
&= \ell\left(y_i \cdot \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \langle \phi_{x_i}(w_{s,0}), w_s - w_{s,0} \rangle + y_i \cdot \left(f(x_i; W, \mathbf{a}) - \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \langle \phi_{x_i}(w_{s,0}), w_s - w_{s,0} \rangle \right) \right) \\
&\leq \ell\left(\frac{y_i}{\sqrt{m}} \sum_{s=1}^{m} a_s \langle \phi_{x_i}(w_{s,0}), w_s - w_{s,0} \rangle - C(m, n, \delta, r) \right) = \tilde{\ell}(W, z_i). \qquad (2)
\end{aligned}$$

Proof of Lemma 3.15. Denote $W_\mu^* = \operatorname{argmin}_{W \in H_r} \left[\tilde{L}_{\mathcal{D}}(W) + \frac{\mu}{2} \|W - W_0\|_F^2 \right]$. The gradient of the objective at W_μ^* equals $\nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*) + \mu \left(W_\mu^* - W_0 \right)$. The empirical approximation of this term is $\frac{1}{n} \sum_{i=1}^{n} \left[\nabla_W \tilde{\ell}(W_\mu^*, z_i) + \mu \left(W_\mu^* - W_0 \right) \right]$. Define this approximation error as

$$\begin{aligned}
f(z_1, \ldots, z_n) :&= \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\nabla_W \tilde{\ell}(W_\mu^*, z_i) + \mu \left(W_\mu^* - W_0 \right) \right] - \left[\nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*) + \mu \left(W_\mu^* - W_0 \right) \right] \right\|_F \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_W \tilde{\ell}(W_\mu^*, z_i) - \nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*) \right\|_F .
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{S \sim \mathcal{D}^n} f(z_1, \ldots, z_n) &= \mathbb{E}_{S \sim \mathcal{D}^n} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_W \tilde{\ell}(W_\mu^*, z_i) - \nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*) \right\|_F \\
&\leq \sqrt{\mathbb{E}_{S \sim \mathcal{D}^n} \left[\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_W \tilde{\ell}(W_\mu^*, z_i) - \nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*) \right\|_F^2 \right]} \\
&= \sqrt{\frac{1}{n} \mathbb{E}_{z \sim \mathcal{D}} \left[\|\nabla_W \tilde{\ell}(W_\mu^*, z) - \nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*)\|_F^2 \right]} \\
&= \sqrt{\frac{1}{n} \mathbb{E}_{z \sim \mathcal{D}} \left[\|\nabla_W \tilde{\ell}(W_\mu^*, z)\|_F^2 - \|\nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*)\|_F^2 \right]} \\
&\leq \sqrt{\frac{1}{n} \mathbb{E}_{z \sim \mathcal{D}} \left[\|\nabla_W \tilde{\ell}(W_\mu^*, z)\|_F^2 \right]} \\
&\leq \frac{1}{\sqrt{n}}. \qquad (\|\nabla_W \tilde{\ell}(W_\mu^*, z)\|_F \leq 1)
\end{aligned}$$

From McDiarmid's inequality, $\forall t > 0$,

$$\mathbb{P}\left[f(z_1, \ldots, z_n) - \mathbb{E}_{S \sim \mathcal{D}^n} f(z_1, \ldots, z_n) > t \right] \leq \exp\left(-\frac{t^2 n}{2} \right).$$

Setting $t = \sqrt{\frac{2\ln\frac{1}{\delta}}{n}}$, we obtain with probability $\geq 1-\delta$,

$$\left\|\nabla_W \tilde{L}_S(W_\mu^*) - \nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*)\right\|_F \leq \frac{1}{\sqrt{n}} + t = \frac{1}{\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right).$$

The strong convexity of $\tilde{L}_S(W) + \frac{\mu}{2}\|W - W_0\|_F^2$ implies

$$\tilde{L}_S(W_\mu^*) + \frac{\mu}{2}\|W_\mu^* - W_0\|_F^2 \geq \tilde{L}_S(W_S) + \frac{\mu}{2}\|W_S - W_0\|_F^2 + \frac{\mu}{2}\|W_\mu^* - W_S\|_F^2$$

and

$$\begin{aligned}&\tilde{L}_S(W_S) + \frac{\mu}{2}\|W_S - W_0\|_F^2\\ \geq&\tilde{L}_S(W_\mu^*) + \frac{\mu}{2}\|W_\mu^* - W_0\|_F^2 + \langle \nabla_W \tilde{L}_S(W_\mu^*) + \mu\left(W_\mu^* - W_0\right), W_S - W_\mu^*\rangle + \frac{\mu}{2}\|W_\mu^* - W_S\|_F^2.\end{aligned}$$

Adding these two inequalities, we get

$$\begin{aligned}\mu\|W_\mu^* - W_S\|_F^2 &\leq \langle \nabla_W \tilde{L}_S(W_\mu^*) + \mu\left(W_\mu^* - W_0\right), W_\mu^* - W_S\rangle\\ &= \langle \nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*) + \mu\left(W_\mu^* - W_0\right), W_\mu^* - W_S\rangle + \langle \nabla_W \tilde{L}_S(W_\mu^*) - \nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*), W_\mu^* - W_S\rangle\\ &\leq 0 + \langle \nabla_W \tilde{L}_S(W_\mu^*) - \nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*), W_\mu^* - W_S\rangle\\ &\qquad (W_\mu^* \text{ minimizes } \tilde{L}_{\mathcal{D}}(W) + \tfrac{\mu}{2}\|W - W_0\|_F^2 \text{ and } W_S - W_\mu^* \text{ is a feasible direction})\\ &\leq \|\nabla_W \tilde{L}_S(W_\mu^*) - \nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*)\|_F \cdot \|W_\mu^* - W_S\|_F.\end{aligned}$$

This implies with probability at least $1-\delta$,

$$\|W_\mu^* - W_S\|_2 \leq \frac{1}{\mu}\|\nabla_W \tilde{L}_S(W_\mu^*) - \nabla_W \tilde{L}_{\mathcal{D}}(W_\mu^*)\|_2 \leq \frac{1}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right).$$

Similarly, with probability at least $1-\delta$ over S',

$$\|W_\mu^* - W_{S'}\|_2 \leq \frac{1}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right).$$

A union bound and triangle inequality imply that with probability $\geq 1-2\delta$,

$$\|W_S - W_{S'}\|_2 \leq \|W_\mu^* - W_S\|_2 + \|W_\mu^* - W_{S'}\|_2 \leq \frac{2}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right).$$

□

Proof of Proposition 3.16. By Lemma 3.15, with probability at least $1-2\delta$ over independent samples S, S', we have $\|W_S - W_{S'}\|_F \leq \frac{2}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)$. We apply Proposition 3.6 to ensure replicability. Select $\delta = \rho/4$.

$$\begin{aligned}&\mathbb{P}\left[\tilde{W}_S \neq \tilde{W}_{S'}\right]\\ =&\mathbb{P}\left[\tilde{W}_S \neq \tilde{W}_{S'}, \|W_S - W_{S'}\|_F > \frac{2}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)\right] + \mathbb{P}\left[\tilde{W}_S \neq \tilde{W}_{S'}, \|W_S - W_{S'}\|_F \leq \frac{2}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)\right]\\ \leq&\mathbb{P}\left[\|W_S - W_{S'}\|_F > \frac{2}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)\right] + \mathbb{P}\left[\tilde{W}_S \neq \tilde{W}_{S'}\middle| \|W_S - W_{S'}\|_F \leq \frac{2}{\mu\sqrt{n}}\left(1+\sqrt{2\ln\frac{1}{\delta}}\right)\right]\end{aligned}$$

$$\leq 2\delta + \frac{\sqrt{md}}{a} \cdot \frac{2}{\mu\sqrt{n}} \left(1 + \sqrt{2\ln\frac{1}{\delta}}\right) \qquad \text{(Proposition 3.6)}$$
$$= \frac{\rho}{2} + \frac{\sqrt{md}}{a} \cdot \frac{2}{\mu\sqrt{n}} \left(1 + \sqrt{2\ln\frac{4}{\rho}}\right) = \rho.$$

□

To derive the generalization guarantee, we utilize the uniform convergence provided in Ji and Telgarsky (2019). They defined $\mathcal{W}_\rho := \left\{W \in \mathbb{R}^{m\times d} \middle| \|w_s - w_{s,0}\|_2 \leq \rho \text{ for any } 1 \leq s \leq m\right\}$. Equation (B.1) in Ji and Telgarsky (2019) states with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$, for any $W \in \mathcal{W}_\rho$,

$$\mathbb{E}_{z\sim\mathcal{D}}[-\ell'(yf(x;W,\mathbf{a}))] - \frac{1}{n}\sum_{i=1}^{n}[-\ell'(y_i f(x_i;W,\mathbf{a}))] \leq \frac{\rho\sqrt{m}}{2\sqrt{n}} + 3\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{3}$$

Here they used $-\ell'$ as the surrogate loss of the 0-1 loss. This surrogate loss satisfies $0 \leq -\ell'(z) \leq 1$ and $-\ell'(z) \leq \ell(z)$. We will apply $\rho = r/\sqrt{m}$ in our proof.

Proof of Theorem 3.17. We first upper bound the empirical risk.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\ell(y_i f(x_i;\tilde{W}_S,\mathbf{a})) &\leq \frac{1}{n}\sum_{i=1}^{n}\left[\ell(y_i f(x_i;W_S,\mathbf{a})) + \left|f(x_i;\tilde{W}_S,\mathbf{a}) - f(x_i;W_S,\mathbf{a})\right|\right] && (\ell \text{ is 1-Lip})\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\ell(y_i f(x_i;W_S,\mathbf{a})) + \left\|\tilde{W}_S - W_S\right\|_F && (f \text{ is 1-Lip})\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\ell(y_i f(x_i;W_S,\mathbf{a})) + \frac{a\sqrt{md}}{2} && \text{(Proposition 3.6)}\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\tilde{l}(W_S,z_i) + \frac{a\sqrt{md}}{2} && \text{(from equation 2 with probability } \geq 1-2\delta)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\tilde{l}(W_S,z_i) + \frac{\mu}{2}\|W_S - W_0\|_F^2 + \frac{a\sqrt{md}}{2}\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\tilde{l}(W_0 + r\overline{U},z_i) + \frac{\mu}{2}\|W_0 + r\overline{U} - W_0\|_F^2 + \frac{a\sqrt{md}}{2} && \text{(definition of RERM)}\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\tilde{l}(W_0 + r\overline{U},z_i) + \frac{\mu r^2}{2} + \frac{a\sqrt{md}}{2} && (4)
\end{aligned}$$

Now we upper bound the RHS of equation (4). From Lemma A.1, with probability $\geq 1-\delta$, for any $1 \leq i \leq n$,

$$\begin{aligned}
\tilde{l}(W_0 + r\overline{U},z_i) &= \ell\left(\frac{y_i}{\sqrt{m}}\sum_{s=1}^{m} a_s\langle\phi_{x_i}(w_{s,0}), r\bar{u}_s\rangle - C(m,n,\delta,r)\right)\\
&\leq \ell\left(r\left(\gamma - \sqrt{\frac{2\ln(n/\delta)}{m}}\right) - \sqrt{2\ln(4n/\delta)} - \sqrt{\frac{2}{\pi}}\cdot\frac{r^2}{\sqrt{m}} - r\sqrt{\frac{\ln(n/\delta)}{2m}}\right) && \text{(Lemma A.1)}\\
&\leq \exp\left(-r\left(\gamma - \sqrt{\frac{2\ln(n/\delta)}{m}}\right) + \sqrt{2\ln(4n/\delta)} + \sqrt{\frac{2}{\pi}}\cdot\frac{r^2}{\sqrt{m}} + r\sqrt{\frac{\ln(n/\delta)}{2m}}\right)\\
& && (\ell(z) \leq \exp(-z))
\end{aligned}$$

$$\leq \exp\left(-\frac{r\gamma}{2} + \sqrt{2\ln(4n/\delta)} + \sqrt{\frac{2}{\pi}} \cdot \frac{r^2}{\sqrt{m}} + r\sqrt{\frac{\ln(n/\delta)}{2m}}\right)$$

(given condition implies $m \geq \frac{8\ln(n/\delta)}{\gamma^2}$)

$$\leq \exp\left(-\frac{r\gamma}{4}\right). \quad \text{(given conditions imply } \max\left\{\sqrt{2\ln(4n/\delta)}, \sqrt{\tfrac{2}{\pi}} \cdot \tfrac{r^2}{\sqrt{m}}, r\sqrt{\tfrac{\ln(n/\delta)}{2m}}\right\} \leq \tfrac{r\gamma}{12})$$

Combining this inequality with equation (4), we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \ell(y_i f(x_i; \tilde{W}_S, \mathbf{a})) &\leq \frac{1}{n}\sum_{i=1}^{n} \tilde{l}(W_0 + r\overline{U}, z_i) + \frac{\mu r^2}{2} + \frac{a\sqrt{md}}{2} \\
&\leq \exp\left(-\frac{r\gamma}{4}\right) + \frac{\mu r^2}{2} + \frac{a\sqrt{md}}{2} \\
&\leq \epsilon + \frac{\mu r^2}{2} + \frac{a\sqrt{md}}{2} && \text{(given condition implies } r \geq \tfrac{4}{\gamma}\ln\tfrac{1}{\epsilon}) \\
&= \epsilon + \frac{\mu r^2}{2} + \frac{2md}{\rho\mu\sqrt{n}}\left(1 + \sqrt{2\ln\frac{4}{\rho}}\right) && (a = \tfrac{4\sqrt{md}}{\rho\mu\sqrt{n}}\left(1 + \sqrt{2\ln\tfrac{4}{\rho}}\right)) \\
&= \epsilon + 2\sqrt{\frac{mdr^2}{\rho\sqrt{n}}(1 + \sqrt{2\ln(4/\rho)})}. && \text{(plug in } \mu)
\end{aligned}$$

Since $-\ell'(z) \leq \ell(z)$, we get

$$\frac{1}{n}\sum_{i=1}^{n}[-\ell'(y_i f(x_i; \tilde{W}_S, \mathbf{a}))] \leq \epsilon + 2\sqrt{\frac{mdr^2}{\rho\sqrt{n}}(1 + \sqrt{2\ln(4/\rho)})}.$$

From equation 3, with probability $\geq 1 - \delta$,

$$\mathbb{E}_{z\sim\mathcal{D}}[-\ell'(yf(x; \tilde{W}_S, \mathbf{a}))] \leq \epsilon + 2\sqrt{\frac{mdr^2}{\rho\sqrt{n}}(1 + \sqrt{2\ln(4/\rho)})} + \frac{r}{2\sqrt{n}} + 3\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Since the 0-1 loss $\ell_{0-1}(z) = \mathbb{1}[z \leq 0]$ is upper bounded by $-2\ell'(z)$, we conclude that

$$\mathbb{P}_{(x,y)\sim\mathcal{D}}[yf(x; \tilde{W}_S, \mathbf{a}) \leq 0] \leq 2\epsilon + 4\sqrt{\frac{mdr^2}{\rho\sqrt{n}}(1 + \sqrt{2\ln(4/\rho)})} + \frac{r}{\sqrt{n}} + 3\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

□