Proceedings of Machine Learning Research vol 336:1–41, 2026 39th Annual Conference on Learning Theory

Clipping the Price of Adaptivity at the Tail

Itai Kreisler KREISLER@MAIL.TAU.AC.IL
Yair Carmon YCARMON@TAUEX.TAU.AC.IL
Oliver Hinder OHINDER@PITT.EDU

Editors: Steve Hanneke and Tor Lattimore

Abstract

Adaptive stochastic convex optimization (SCO) methods face a fundamental "price of adaptivity" barrier: under the standard set of assumptions, they cannot efficiently adapt to large uncertainty in both the initial distance to optimality and the Lipschitz constant. We circumvent this barrier by requiring a small amount of additional structure common to many learning problems. Specifically, we assume that the objective decomposes into a model and a loss function, enabling us to intervene by modifying the model's output before it passes to the loss function. Under this assumption, we design a method that clips the learned model output in tail events where it deviates too much from the output of a fixed reference model. Our method matches the optimal bounds for known-parameter SCO up to logarithmic factors in the uncertainty in the distance and Lipschitz parameters, thus efficiently adapting to large uncertainty in both.

Keywords: Stochastic optimization, Convex optimization, Parameter-free optimization

1. Introduction

This paper studies machine learning objectives composed of a model and a loss function. Specifically, for a model m and a loss function h, the learning problem can be written as

$$\operatorname*{minimize}_x F(x), \quad \text{where} \quad F(x) := \mathbb{E}_{s\sim\mathcal{P}}[h(m(x;s);s)]. \tag{1}$$

Such model–loss decompositions arise broadly in practice, e.g., neural networks with various loss functions. However, optimization theory typically analyzes the more general problem of minimizing a stochastic function f without assuming this internal structure:

$$\operatorname*{minimize}_x F(x), \quad \text{where} \quad F(x) := \mathbb{E}_{s\sim\mathcal{P}}[f(x;s)]. \tag{2}$$

Stochastic convex optimization is a fundamental problem in machine learning and serves as the foundation for classical stochastic gradient methods. While classical methods can attain the optimal optimality gap, they typically hinge on knowing the Lipschitz parameter $L_\star$ and the distance $D_\star$ from the initial point to the optimum. *Parameter-free* methods (e.g., Orabona and Pál, 2016; Cutkosky and Orabona, 2018; Bhaskara et al., 2020; Mhammedi and Koolen, 2020; Jacobsen and Cutkosky, 2022; Carmon and Hinder, 2022; Ivgi et al., 2023; Kreisler et al., 2024) relax the requirement for precise knowledge of the problem parameters, aiming to achieve a near-optimal optimality gap using only loose estimates. Lower bounds (Carmon and Hinder, 2024; Attia and Koren, 2024; Khaled and Jin, 2024) for parameter-free optimization show that parameter-free methods must pay a significant multiplicative price due to uncertainty in $L_\star$ and $D_\star$. For example, Carmon and Hinder (2024) show that if ρ and ℓ are the multiplicative uncertainties in the distance to optimality and Lipschitz constant, respectively, then there is an $\Omega(\min\{\ell,\rho\}\sqrt{\ln(1/\delta)}/\sqrt{T})$ additional multiplicative cost from not knowing problem parameters. This shows that it is difficult to adapt to high uncertainty in both the Lipschitz constant and distance to optimality. However, this lower bound applies to the general stochastic convex structure given in Equation (2) but not the model–loss structure of Equation (1).

Our contributions

We identify that the aforementioned lower bound for parameter-free methods is intimately tied to tail events: rare samples for which the optimized function behaves very differently from the bulk of the distribution. Such rare samples may go unobserved by the optimization algorithm, making it impossible to reliably characterize the tail behavior of the stochastic noise.

However, under the model–loss structure of Equation (1), it is possible to detect and mitigate these rare events, circumventing this lower bound. In particular, under this model–loss structure, our key idea is to clip the output of the model at *inference-time*. This clipping occurs on rare events in which the model provides an extremely large outlier prediction. In these cases, clipping prevents us from paying a high price for a wildly incorrect prediction. We present two methods that use this idea: one that prioritizes computational efficiency and the other that prioritizes sample efficiency. Both methods build on Lawrence et al. (2025), which focused on the standard parameter-free optimization paradigm, i.e., Equation (2).

In the computational setting, we propose a method that performs grid search over hyperparameters, generates candidate weights using a standard optimization algorithm such as SGD, and then selects one of the resulting weights paired with an associated clipping parameter. We show that this method, with probability at least $1-\delta$, achieves an optimality gap of $\widetilde{O}(L_\star D_\star \sqrt{\ln(1/\delta)}/\sqrt{T})$ where T is the number of gradient and function evaluations made by the method, and $\widetilde{O}(\cdot)$ hides poly-logarithmic factors. Notably, our method only incurs a logarithmic cost in uncertainty in the problem parameters compared with the polynomial dependence exhibited in the lower bounds. This optimality gap beats the lower bounds proved in Carmon and Hinder (2024); Attia and Koren (2024); Khaled and Jin (2024), which scale linearly with $\min\{\ell, \rho\}$.

In the sample complexity setting, we propose a method that carefully solves a regularized optimization problem and then applies inference-time clipping to the resulting learned model. This gives a suboptimality gap of $O(L_\star D_\star \sqrt{\ln(1/\delta)}/\sqrt{N})$ for the Lipschitz case, where N is the number of samples made available to the algorithm. This matches lower bounds for parameter-known stochastic convex optimization (Carmon and Hinder, 2024).

Paper outline

Section 2 surveys related work. Section 3 defines the notation used throughout the paper, as well as the instance classes of functions considered in this work. In Section 4 we propose a parameter-free method that achieves a near-optimal optimality gap in the computational complexity setting. Specifically, in Section 4.2 we introduce a model selection algorithm, and in Section 4.5 we show how to use this algorithm to derive a parameter-free optimization method. In Section 5 we propose a parameter-free method in the sample complexity setting that achieves the optimal optimality gap, without requiring any prior lower or upper bounds on the problem parameters $L_\star$ and $D_\star$. Finally, Section 6 discusses which lower bounds continue to hold under the model–loss decomposition assumption defined in Equation (1).

2. Related Work

Parameter-free methods Parameter-free methods aim to achieve a near-optimal optimality gap without requiring precise knowledge of the problem parameters, historically with emphasis on the initial distance to the optimum. Work in this field originates in the setting of online convex optimization (OCO) (Luo and Schapire, 2015; Orabona and Pál, 2016; Cutkosky and Orabona, 2018; Mhammedi and Koolen, 2020; Bhaskara et al., 2020; Jacobsen and Cutkosky, 2022; Zhang et al., 2022). This line of work was subsequently expanded to design parameter-free methods, particularly for the setting of stochastic convex optimization (SCO) (e.g., Orabona, 2014; Orabona and Tommasi, 2017; Chen et al., 2022; Carmon and Hinder, 2022; Ivgi et al., 2023; Kreisler et al., 2024; Liu et al., 2025).

Lower bounds In non-smooth stochastic convex optimization with known Lipschitz constant, $L_\star$, and initial distance to optimality, $D_\star$, classical results show that the expected loss is $\Omega\left(L_\star D_\star/\sqrt{T}\right)$ (Nemirovski and Yudin, 1983; Duchi, 2018). For the high-probability case, Carmon and Hinder (2024) proves the lower bound of $\Omega\left(L_\star D_\star\sqrt{\ln(1/\delta)}/\sqrt{T}\right)$. These lower bounds have matching upper bounds (e.g., Orabona, 2021), and consequently define the optimal optimality gaps for non-smooth stochastic convex optimization.

Recently, Carmon and Hinder (2024) proves that parameter-free methods incur a penalty due to uncertainty in the Lipschitz parameter $L_\star$ and the distance $D_\star$ from the initial point to the optimum. Specifically, they assume that while $L_\star$ and $D_\star$ are unknown, they are bounded such that $L_\star \in [1, \ell]$ and $D_\star \in [1, \rho]$ for some known values ℓ and ρ. They prove that if stochastic gradients are almost surely bounded by $L_\star$, the optimality gap necessarily suffers a multiplicative degradation of at least $\Omega\left(\min\{\ell, \rho\}\sqrt{\ln(1/\delta)}/\sqrt{T}\right)$ relative to the optimal rate. Furthermore, they prove that in the second-moment setting, in which $L_\star^2$ bounds the expectation of the stochastic gradient norm squared, the multiplicative penalty is at least $\Omega\left(\min\{\ell, \rho\}\right)$. This implies that, in heavy-tailed settings, achieving a near-optimal optimality gap is impossible without prior knowledge of at least one of the problem parameters.

Attia and Koren (2024) and Khaled and Jin (2024) also provide lower bounds for parameter-free optimization. Attia and Koren (2024) derives comparable lower bounds to Carmon and Hinder (2024) in the Lipschitz case. Khaled and Jin (2024) provides a slightly weaker lower bound than Attia and Koren (2024), demonstrating that the multiplicative price incurred due to parameter uncertainty cannot be bounded by any polylogarithmic term.

Critically, all these lower bounds focus on the standard stochastic optimization setting, i.e., Equation (2), allowing us to beat them by considering the model–loss structure, i.e., Equation (1). We note that *some* of the lower bounds presented in Carmon and Hinder (2024) do transfer to Equation (1); we discuss this in more detail in Section 6 and its supplement, Appendix D.

Reliable Model Selection Lawrence et al. (2025) demonstrates that the standard holdout model selection approach can fail catastrophically in SCO. To address this, they introduce a general framework for reliable model selection. They also propose an optimization method for the sample complexity setting. However, both methods require a good estimate of the Lipschitz constant (i.e., within a factor of $\sqrt{T}$ of the true value). In this paper, we build upon and extend many of the results and proof techniques introduced by Lawrence et al. (2025), eliminating the need for a good estimate of the Lipschitz constant.

Improper learning In optimization and machine learning theory, one typically fixes a hypothesis class $\mathcal{H}$ (e.g., linear classifiers) as a representation for the target concept class. A learning algorithm is said to be *proper* if it is required to output a hypothesis from $\mathcal{H}$. In contrast, an algorithm is called *improper* if it is allowed to output any predictor, possibly outside $\mathcal{H}$, as long as the predictor attains low prediction error under the data distribution. This relaxation often substantially enlarges the space of candidate predictors and consequently can yield learning algorithms that succeed in settings where proper learning is statistically or computationally infeasible (e.g., Montasser et al., 2019; Hanneke, 2016).

The algorithms we present in this work can be viewed as improper, as they output a weight x and a clipping parameter c for model $m_c(x;\cdot)$, rather than just outputting a weight x for the model $m(x;\cdot)$. Unlike many works in this field that output an ensemble of predictors aggregated via majority vote, our approach outputs a single model, using the initial hypothesis solely to constrain the predictions.

A particularly relevant piece of work in this area is Vovk (2006), who provide an online learning method that exploits a loss–model structure in the least squares setup. However, that work assumes known bounds on the labels.

3. Notation

We include 0 in the set of natural numbers $\mathbb{N}$. For an integer $K \in \mathbb{N}$ we define $[K] \coloneqq \{1, \cdots, K\}$. We denote $\ln_+(x) = \ln(e+x)$. Let $\mathcal{X}, \mathcal{Y}$ denote closed convex sets, let $\mathbb{S}$ denote a sample space, and let $\mathcal{P}$ denote a distribution over $\mathbb{S}$. We denote the number of samples by N, and the number of model and gradient evaluations by T. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be norms on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\|\cdot\|_{\alpha*}$ and $\|\cdot\|_{\beta*}$ be the dual norms of $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, respectively. The corresponding operator norm, induced by the norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, is defined as $\|A\|_{\alpha,\beta} \coloneqq \sup\{\|Ax\|_\beta \mid \|x\|_\alpha \le 1\}$. We denote by $Q_p(\mathbb{D}) \coloneqq \min\{z \in \mathbb{R} : \mathbb{P}_{s\sim\mathbb{D}}[s \le z] \ge p\}$ the p-th quantile of distribution $\mathbb{D}$, where $\mathbb{D}$ is a distribution over $\mathbb{R}$; this definition mimics that of previous work (e.g., Borovkov, 1999, Ch. 3).

Throughout, we assume that the function $f : \mathcal{X} \times \mathbb{S} \to \mathbb{R}$ admits a decomposition into a model and a loss function, i.e., $f(x;s) \coloneqq h(m(x;s);s)$, where $m : \mathcal{X} \times \mathbb{S} \to \mathcal{Y}$ and $h : \mathcal{Y} \times \mathbb{S} \to \mathbb{R}$. For convenience, we treat the functions m, h and f as differentiable. We denote the class of all differentiable functions as $\mathcal{I}_{\text{diff}} \coloneqq \{q \mid \forall s \in \mathbb{S} : q(\cdot;s) \text{ is differentiable}\}$. In appendix K we discuss how to extend our results to non-differentiable functions. We define $F(x) \coloneqq \mathbb{E}_{s\sim\mathcal{P}}[f(x;s)]$. We denote by x_0 the point at which the optimization algorithms are initialized. We define a clipping function into a domain $\mathcal{D} \subseteq \mathcal{Y}$ as a mapping $\text{Clip}_{\mathcal{D}}(\cdot) : \mathcal{Y} \to \mathcal{D}$ such that $\text{Clip}_{\mathcal{D}}(z) = z$ for all $z \in \mathcal{D}$. We allow $\text{Clip}_{\mathcal{D}}(\cdot)$ to be any such mapping; any choice suffices for our theoretical analysis. We define the clipped model as $m_c(x;s) \coloneqq \text{Clip}_{\{u\in\mathcal{Y}:\|u-m(x_0;s)\|_\beta\le c\}}(m(x;s))$, where $c \in \mathbb{R}$ is the clipping parameter. Subsequently, we define the clipped function as $f_c(x;s) \coloneqq h(m_c(x;s);s)$ and $F_c(x) \coloneqq \mathbb{E}_{s\sim\mathcal{P}}[f_c(x;s)]$.

We define $\text{binary-ceil}(x;b) \coloneqq \min\{2^n b \mid n \in \mathbb{N} \text{ and } 2^n b \ge x\mathbb{1}_{\{b\neq 0\}}\}$, i.e., rounding $x \neq 0$ up to the nearest power-of-two multiple of b. For any $p > 1$ and $\bar{b} \ge \underline{b} > 0$, we define $\text{Grid}_p(\underline{b}, \bar{b}) \coloneqq \{\underline{b} \cdot p^n\}_{n=0}^{\lfloor \log_p(\bar{b}/\underline{b})\rfloor} \cup \{\bar{b}\}$, i.e., all the powers of p between the bounds $\underline{b}$ and $\bar{b}$.

In addition, we assume throughout that F attains its minimum, i.e., there exists $x_\star \in \mathcal{X}$ such that $F(x_\star) = \inf_{x\in\mathcal{X}} F(x)$. We denote by $x_\star$ the optimal point closest to the initialization x_0, i.e., $x_\star \coloneqq \arg\min_{x\in\text{optima of } F} \|x_0 - x\|_\alpha$. We define the unknown initial distance to the optimum as $D_\star \coloneqq \|x_0 - x_\star\|_\alpha$, and denote by $\overline{D}_\star \coloneqq \max(D_\star, \underline{\rho})$ the maximum of $D_\star$ and a known lower bound $\underline{\rho}$ on $D_\star$. We define the instance classes of convex stochastic optimization problems whose optimal solution lies within distance at most $\overline{\rho}$ of the initialization as

$$\mathcal{I}_{\text{Dist}}^{\overline{\rho},\|\cdot\|_\alpha} := \{q \in \mathcal{I}_{\text{diff}} \mid \|x_0 - x_\star\|_\alpha \leq \overline{\rho} \text{ and } \mathbb{E}_{s\sim\mathcal{P}} q(\cdot;s) \text{ is convex}\},$$
$$\hat{\mathcal{I}}_{\text{Dist}}^{\overline{\rho},\|\cdot\|_\alpha} := \{q \in \mathcal{I}_{\text{diff}} \mid \|x_0 - x_\star\|_\alpha \leq \overline{\rho} \text{ and } \forall s \in \mathbb{S} : q(\cdot;s) \text{ is convex}\}.$$

Similarly, we define the unknown Lipschitz constant in the context of the Lipschitz case as $L_\star := \sup_{x\in\mathcal{X};s\in\mathbb{S}} \|\nabla_x m(x;s)\|_{\alpha,\beta}$, and in the context of the second-moment-Lipschitz case as $L_\star := \sup_{x\in\mathcal{X}} \sqrt{\mathbb{E}_{s\sim\mathcal{P}} \|\nabla_x m(x;s)\|^2_{\alpha,\beta}}$. We denote $\overline{L}_\star := \max(L_\star, \underline{\ell})$ as the maximum of $L_\star$ and a known lower bound $\underline{\ell}$ on $L_\star$. We then define the instance classes of Lipschitz and second-moment-Lipschitz problems with Lipschitz constant at most $\overline{\ell}$ as

$$\mathcal{I}_{\text{Lip}}^{\overline{\ell},\|\cdot\|} := \{q \in \mathcal{I}_{\text{diff}} \mid \forall x \in \mathcal{X}; s \in \mathbb{S} : \|\nabla_x q(x;s)\| \leq \overline{\ell}\},$$
$$\mathcal{I}_{\text{SM-Lip}}^{\overline{\ell},\|\cdot\|} := \left\{q \in \mathcal{I}_{\text{diff}} \mid \forall x \in \mathcal{X} : \mathbb{E}_{s\sim\mathcal{P}} \|\nabla_x q(x;s)\|^2 \leq \overline{\ell}^2\right\}.$$

We note that any Lipschitz function is also a second-moment-Lipschitz function, i.e., $\mathcal{I}_{\text{Lip}}^{\overline{\ell},\|\cdot\|} \subset \mathcal{I}_{\text{SM-Lip}}^{\overline{\ell},\|\cdot\|}$, and that $\mathcal{I}_{\text{SM-Lip}}^{\overline{\ell},\|\cdot\|}$ also includes problems with heavy-tailed gradient noise.

4. A computationally efficient parameter-free method that exploits model–loss structure

In this section, we propose a parameter-free method that achieves a near-optimal optimality gap. To this end, our method operates under the model–loss decomposition described in Equation (1) and introduces clipping at the model output level.

Our parameter-free method combines grid search, a known optimization method, and a model selection procedure. By performing grid search over possible hyperparameters and applying a known optimization method, such as SGD, at each grid point, we generate a collection of candidate weights, one of which attains a small optimality gap. We then introduce a model selection method that, given this collection of candidates, selects a candidate paired with a clipping parameter such that the selected pair achieves a loss close to the minimum loss among all candidates.

4.1. The prior work that we leverage

Our model selection algorithm uses the model selection framework proposed by Lawrence et al. (2025). Their algorithm takes as input a reference model x_0, a list of candidate models $x_1, \ldots, x_K$, and corresponding confidence interval widths $\tau_1, \ldots, \tau_K$. A key condition is the requirement of Lemma 1, which states that each τ_k bounds the deviation between the validation loss and the population loss of x_k, relative to the reference model x_0. From the list of models $x_0, \ldots, x_K$, their algorithm selects one that balances a low validation loss with a low risk, where the latter is reflected by a small τ. The selected model enjoys a strong guarantee on its performance relative to the other candidates $x_0, \ldots, x_K$; this guarantee is stated explicitly in Lemma 1. Consequently, obtaining confidence widths $\tau_1, \ldots, \tau_K$ that are both valid and sufficiently small is crucial for the effectiveness of the reliable model selection method of Lawrence et al. (2025). For completeness, we present their reliable model selection method in Appendix A as Algorithm 3.

Lemma 1 (Lemma 1 of Lawrence et al. (2025)) *Let $\hat{\delta} \in (0,1)$. Let K be the number of models, and N be the validation sample set size. Let $\tau_1, \ldots, \tau_K$ be non-negative scalars such that with probability $1-\hat{\delta}$ for all $k \in [K]$ we have $\left| F(x_k) - F(x_0) - \frac{1}{N}\sum_{i=1}^{N}(f(x_k; s_i) - f(x_0; s_i)) \right| \le \tau_k$, and let $\tau_0 = 0$. Then, with probability $1-\hat{\delta}$, $F(x_{reliable}) \le \min_{k\in[K]\cup\{0\}}[F(x_k) + 2\tau_k]$, where $x_{reliable}$ is the result of Algorithm 3.*

4.2. Our model selection algorithm

We present a model selection method Algorithm 1, and show in Proposition 3 and corollaries in Section 4.5 that combining a grid search with our algorithm yields a result that, with high probability, has a near-optimal suboptimality.

The purpose of Algorithm 1 is to select a weight x paired with a clipping parameter c such that the loss $F_c(x)$ is small relative to the losses of all candidate weights. The algorithm receives a set of weight candidates $\{x_g\}_{g\in\mathcal{G}}$, possibly generated using some optimization method, such as grid search over SGD step sizes. For each candidate x_g, it considers a set of clipping parameters $\mathcal{C}_g$, where $\mathcal{C}_g$ is defined as the n largest elements of the set $\{\|m(x_g; s_i) - m(x_0; s_i)\|_\beta\}_{i=1}^{N}$. For each pair of weight and clipping parameter, the algorithm computes a confidence interval width τ. Finally, Algorithm 1 applies Algorithm 3 to select the pair (x, c) that minimizes the validation loss augmented by the risk term τ. The resulting weight and clipping parameter are guaranteed to achieve a loss close to that of the best candidate weight.

The key ingredient of the algorithm is the introduction of clipping at the model output at inference-time. This clipping enables the use of Algorithm 3 to choose a weight and clipping parameter with sufficiently small suboptimality. In contrast, attempting to achieve the same suboptimality by using grid search alone, without clipping the model output, is doomed to fail, as it contradicts Carmon and Hinder (2024, Theorem 3).

Introducing clipping at the model output allows us to characterize a confidence interval width τ. For each pair (x, c), we compute the empirical mean of the difference between the clipped loss at (x, c) and the loss at the initial weight x_0, and denote it by $u(x,c) \coloneqq \frac{1}{N}\sum_{j=1}^{N}(f_c(x; s_j) - f(x_0; s_j))$. Similarly, we compute the empirical variance $v(x,c) \coloneqq \frac{1}{N-1}\sum_{j=1}^{N}((f_c(x; s_j) - f(x_0; s_j)) - u(x,c))^2$. To account for the union bound over candidate pairs, we define $a(x,c) \coloneqq |\mathcal{G}|\log_2^2\left(\frac{2c}{\underline{\ell}\|x-x_0\|_\alpha}\right)$ if $x \ne x_0$, and $a(x,c) \coloneqq |\mathcal{G}|$ for $x = x_0$. Finally, the confidence interval width is defined as

$$\tau(x,c) \coloneqq \sqrt{\frac{2v(x,c)\cdot\ln(16a(x,c)/\delta)}{N}} + c\frac{14\ln(16a(x,c)/\delta)}{3(N-1)}. \tag{3}$$

To ensure that the set $\mathcal{C}_g$ contains a sufficiently small clipping parameter–specifically, one no larger than binary-ceil$\left(\sqrt{\frac{N}{\ln(4/\delta)}}L_\star\|x_g - x_0\|_\alpha; \underline{\ell}\|x_g - x_0\|_\alpha\right)$, while keeping $\mathcal{C}_g$ small, we set $n = \lceil 4\ln(4/\delta)\rceil$.

Algorithm 1 invokes Algorithm 3. Therefore, we apply Lemma 1 to guarantee that the result of Algorithm 1 achieves the desired small loss. To this end, we first show in the lemma below that the requirement of Lemma 1 is satisfied. We present the proof of this lemma in Section F.3.

Lemma 2 *Let $N \ge 2$ be the number of validation samples. For some $n \in \mathbb{N}$, let $\{(x_{g_i}, c_i)\}_{i\in[n]}$ be a set of candidate pairs such that, for all $i \in [n]$ we have $x_{g_i} \in \{x_g\}_{g\in\mathcal{G}}$ and $c_i \in \left\{2^k\underline{\ell}\|x_{g_i} - x_0\|_\alpha \mid k \in \mathbb{N}\right\}$.*

Algorithm 1: Model selection with model output clipping

Input: Initialization x_0, candidates $\{x_g\}_{g\in\mathcal{G}}$, number of validation set samples $N \geq 2$, samples $s_1, \dots, s_N \overset{\text{iid}}{\sim} \mathcal{P}$, a lower bound $\underline{\ell}$ on the Lipschitz constant, and a confidence level δ.

Set $n = \lceil 4 \ln(4/\delta) \rceil$.

for $g \in \mathcal{G}$ **do**

  Let $\mathcal{C}_g$ be the n largest elements in $\{\text{binary-ceil}(\|m(x_g; s_i) - m(x_0; s_i)\|_\beta; \underline{\ell}\|x_g - x_0\|_\alpha)\}_{i=1}^N$.

  for $c_{g,i} \in \mathcal{C}_g$ **do**

    Define the clipped model: $m_{c_{g,i}}(x_g; s) \coloneqq \text{Clip}_{\{u:\|u-m(x_0;s)\|_\beta \leq c_{g,i}\}}(m(x_g; s))$.

    Set $\tau_{g,i} = \tau(x_g, c_{g,i})$, where $\tau(\cdot, \cdot)$ is as defined in Equation (3).

  end

end

Run Algorithm 3 with sample set of $\{s_i\}_{i\in[N]}$, with the initial weight of x_0 as the reference model, $\{(x_g, c_{g,i})\}_{(g,i)\in\mathcal{G}\times[n]}$ as the additional candidates, and $\{\tau_{g,i}\}_{(g,i)\in\mathcal{G}\times[n]}$ as the confidence intervals.

return the resulting x, c of the model selection algorithm.

For each $i \in [n]$, define $\tau_i \coloneqq \tau(x_{g_i}, c_i)$. The requirement of Lemma 1 holds for $\hat{\delta} = \delta/2$, i.e., with probability $1 - \hat{\delta}$ for all $i \in [n]$ we have $\left| F_{c_i}(x_{g_i}) - F(x_0) - \frac{1}{N} \sum_{j=1}^N (f_{c_i}(x_{g_i}; s_j) - f(x_0; s_j)) \right| \leq \tau_i$.

We now state a high-probability guarantee on the loss of the output of Algorithm 1. In Section 4.5, we combine this guarantee with the guarantee for clipped SGD to derive a near-optimal optimality gap.

Proposition 3 *Let $\delta \in (0, 1)$, and let the number of validation samples be $N \geq 4\ln(4/\delta)$. Let the candidates $\{x_g\}_{g\in\mathcal{G}}$ be independent of the validation samples $\{s_i\}_{i\in[N]}$. Let $m \in \mathcal{I}_{\text{SM-Lip}}^{\infty,\|\cdot\|_{\alpha,\beta}}$ and $h \in \mathcal{I}_{\text{Lip}}^{1,\|\cdot\|_{\beta*}}$. If x, c are the output of Algorithm 1, then with probability at least $1 - \delta$ we have*

$$F_c(x) \leq \min_{g\in\mathcal{G}\cup\{0\}} \left[F(x_g) + O\left(\overline{L}_\star \|x_g - x_0\|_\alpha \sqrt{\frac{\ln_+(\tilde{a}/\delta)\ln_+(\tilde{a})}{N}} \right) \right],$$

where $\tilde{a} = |\mathcal{G}| \ln_+\left(\sqrt{N} \cdot \overline{L}_\star / \underline{\ell}\right)$.

4.3. Warm-up: the Lipschitz case

The proof of the following proposition is simpler than the proof of Proposition 3, and can be seen as a warm-up for the latter. Before proving Proposition 3, we first consider the simpler setting in which the model m is Lipschitz, i.e., $m \in \mathcal{I}_{\text{Lip}}^{\infty,\|\cdot\|_{\alpha,\beta}}$.

Proposition 4 *Let $\delta \in (0, 1)$, and let the number of samples be $N \geq 2$. Let the candidates $\{x_g\}_{g\in\mathcal{G}}$ be independent of the validation samples $\{s_i\}_{i\in[N]}$. Let $m \in \mathcal{I}_{\text{Lip}}^{\infty,\|\cdot\|_{\alpha,\beta}}$ and $h \in \mathcal{I}_{\text{Lip}}^{1,\|\cdot\|_{\beta*}}$. If x, c are* *the result of Algorithm 1, where we use* $n = 1$*, then with probability at least* $1 - \delta$ *we have*

$$F_c(x) \le \min_{g \in \mathcal{G} \cup \{0\}} \left[F(x_g) + O\left(\overline{L}_\star \|x_g - x_0\|_\alpha \sqrt{\frac{\ln_+(\tilde{a}/\delta)}{N}} \right) \right],$$

where $\tilde{a} = |\mathcal{G}| \ln_+(\overline{L}_\star / \underline{\ell})$.

Proof For all $g \in \mathcal{G}$, let $c_{g,0} \coloneqq \text{binary-ceil}(L_\star \|x_g - x_0\|_\alpha; \underline{\ell}\|x_g - x_0\|_\alpha)$. This value is an intermediate quantity through which we relate $F_{c_{g,0}}(\cdot)$ to $F(\cdot)$ without affecting the algorithm. In this proof, we consider an augmented candidate set $\{(x_g, c_{g,i})\}_{(g,i) \in \mathcal{G} \times ([n] \cup \{0\})}$ as the set of candidates passed to Algorithm 3. For candidate $(x_g, c_{g,0})$, $\tau_{g,0}$ is calculated in the same way as for the other candidates, i.e., $\tau_{g,0} = \tau(x_g, c_{g,0})$. We define $c_{0,0} \coloneqq 0$ and $\tau_{0,0} \coloneqq 0$ for convenience, noting that $m_c(x_0; s) = m(x_0; s)$ and $f_c(x_0; s) = f(x_0; s)$ for every $c \ge 0$ and $s \in \mathbb{S}$. We first show that Algorithm 3 returns the same output for both the original set of candidates and the modified set of candidates. Because $m(\cdot, s)$ is $L_\star$-Lipschitz, Lemma 19 states that for any $g \in \mathcal{G}$ and sample $s \in \mathbb{S}$ we have that

$$\|m(x_g; s) - m(x_0; s)\|_\beta \le L_\star \|x_g - x_0\|_\alpha \le c_{g,0}. \tag{4}$$

Thus, for any $g \in \mathcal{G}$ we have $c_{g,0} \ge c_{g,1}$, and consequently $\tau_{g,0} \ge \tau_{g,1}$. Moreover, the validation losses under the clipping values $c_{g,0}$ and $c_{g,1}$ coincide: $\frac{1}{N}\sum_{j=1}^N f_{c_{g,0}}(x_g; s_j) = \frac{1}{N}\sum_{j=1}^N f_{c_{g,1}}(x_g; s_j)$. Algorithm 3 first constructs a reliable set $\mathcal{F}$ of candidates that minimize the sum of validation loss and the risk term τ, and then selects from $\mathcal{F}$ the candidate with minimal validation loss. Therefore, for every $g \in \mathcal{G}$, the candidate $(x_g, c_{g,1})$ dominates $(x_g, c_{g,0})$ in the sense that it has no larger validation loss and a smaller (or equal) risk term. It follows that the candidates $\{(x_g, c_{g,0})\}_{g \in \mathcal{G}}$ do not affect Algorithm 3. Hence, the algorithm returns the same output for the original and the modified candidate sets.

We now bound $F_{c_{g,0}}(x_g) + \tau_{g,0}$ for any $g \in \mathcal{G}$. As a direct result of Equation (4), we have $F_{c_{g,0}}(x_g) = F(x_g)$. In addition, because $h(\cdot; s)$ is 1-Lipschitz, $v(x_g, c_{g,0}) \le O\left(c_{g,0}^2\right) = O\left(\left(\overline{L}_\star \|x_g - x_0\|_\alpha\right)^2\right)$. Therefore, $\tau_{g,0} \le O\left(\overline{L}_\star \|x_g - x_0\|_\alpha \sqrt{\frac{\ln_+(\tilde{a}/\delta)}{N}}\right)$. Thus, by combining the previous equation with the fact that $F_{c_{g,0}}(x_g) = F(x_g)$, we obtain

$$F_{c_{g,0}}(x_g) + 2\tau_{g,0} \le F(x_g) + O\left(\overline{L}_\star \|x_g - x_0\|_\alpha \sqrt{\frac{\ln_+(\tilde{a}/\delta)}{N}} \right). \tag{5}$$

Finally, as Lemma 2 shows that the requirements of Lemma 1 hold for $\delta/2$, Lemma 1 guarantees that if x, c is the result of Algorithm 1, then with probability at least $1 - \delta/2$,

$$F_c(x) \le \min_{\substack{i \in [n] \cup \{0\} \\ g \in \mathcal{G} \cup \{0\}}} \left[F_{c_{g,i}}(x_g) + 2\tau_{g,i} \right].$$

Hence, as the additional candidates $\{(x_g, c_{g,0})\}_{g \in \mathcal{G} \cup \{0\}}$ do not affect the algorithm, Equation (5) implies Proposition 4. ■

4.4. Proof outline for Proposition 3

The proof for the case $m \in \mathcal{I}_{\text{SM-Lip}}^{\infty,\|\cdot\|_{\alpha,\beta}}$ follows steps similar to the proof for the case $m \in \mathcal{I}_{\text{Lip}}^{\infty,\|\cdot\|_{\alpha,\beta}}$. However, some changes are needed to adapt the proof to the case $m \in \mathcal{I}_{\text{SM-Lip}}^{\infty,\|\cdot\|_{\alpha,\beta}}$. We now present an outline of the proof.

Unlike the Lipschitz case, in the second-moment-Lipschitz case the quantity $\|m_{c_i}(x_{g_i}; s) - m(x_0; s)\|_\beta$ may be unbounded with non-negligible probability. Thus, for any choice of $c_{g,0}$ we cannot guarantee that $c_{g,0}$ bounds $\|m_{c_i}(x_{g_i}; s) - m(x_0; s)\|_\beta$. Therefore, unlike what we showed for the Lipschitz case, we cannot conclude that $F_{c_{g,0}}(x_g) = F(x_g)$ in the second-moment-Lipschitz case. Instead, to balance the trade-off between a small value of $\left|F_{c_{g,0}}(x_g) - F(x_g)\right|$ and a small risk term $\tau_{g,0}$, we choose $c_{g,0} \coloneqq \text{binary-ceil}\left(\sqrt{\frac{N}{\ln(4/\delta)}} L_\star \|x_g - x_0\|_\alpha; \underline{\ell}\|x_g - x_0\|_\alpha\right)$. Leveraging the facts that $h(\cdot; s)$ is 1-Lipschitz, that m is $L_\star$-second-moment-Lipschitz, and our choice of $c_{g,0}$, we obtain that $|F_{c_{g,0}}(x_g) - F(x_g)| \le L_\star \|x_g - x_0\|_\alpha \sqrt{\frac{\ln(4/\delta)}{N}}$, which is sufficiently small.

By using Theorem 10 of Maurer and Pontil (2009) (see Theorem 15), with high probability we bound the empirical variance $\sqrt{v(x_g, c_{g,0})}$ using the true variance and the clipping parameter. Thus, our choice of $c_{g,0}$ results in a sufficiently small risk term: $\tau_{g,0} \le O\left(\overline{L}_\star \|x_g - x_0\|_\alpha \sqrt{\frac{\ln(16\tilde{a}/\delta)}{N}} \sqrt{\frac{\ln(16\tilde{a}/\delta)}{\ln(4/\delta)}}\right)$.

The above suffices to prove that Algorithm 3 selects a good candidate from $\{(x_g, c_{g,i})\}_{(g,i)\in\mathcal{G}\times([n]\cup\{0\})}$, similarly to how it was proven in the Lipschitz case. All that is left is to show that adding the candidate $(x_g, c_{g,0})$ does not affect Algorithm 3. We show that since $n = \lceil 4\ln(4/\delta)\rceil$ is sufficiently large, $c_{g,0} \ge c_{g,n}$. This results in the validation loss for both clipping parameters $c_{g,0}$ and $c_{g,n}$ being very close:

$$\left|\frac{1}{N}\sum_{i=1}^{N} f_{c_{g,0}}(x_g; s_i) - \frac{1}{N}\sum_{i=1}^{N} f_{c_{g,n}}(x_g; s_i)\right| \le 2c_{g,0}\frac{n-1}{N} \le 16\overline{L}_\star \|x_g - x_0\|_\alpha \sqrt{\frac{\ln(4/\delta)}{N}}.$$

As a consequence, we artificially increase the value of $\tau_{g,0}$ by $16\overline{L}_\star \|x_g - x_0\|_\alpha \sqrt{\frac{\ln(4/\delta)}{N}}$, maintaining a small enough risk term $\tau_{g,0}$. Thus, we obtain that $\frac{1}{N}\sum_{i=1}^{N} f_{c_{g,n}}(x_g; s_i) + \tau_{g,c_{g,n}} \le \frac{1}{N}\sum_{i=1}^{N} f_{c_{g,0}}(x_g; s_i) + \tau_{g,c_{g,0}}$. This means two things: (i) The safe set $\mathcal{F}$ calculated in Algorithm 3 remains unchanged, except that $(x_g, c_{g,0})$ may also be included in $\mathcal{F}$, even though we added $(x_g, c_{g,0})$ as a candidate. (ii) If the candidate $(x_g, c_{g,0})$ is in $\mathcal{F}$, then $(x_g, c_{g,n})$ is also in $\mathcal{F}$. Now, as we can generalize Lemma 1 to hold for any candidate in $\mathcal{F}$ and not only to the result of Algorithm 3 (see Lemma 9), we can assume that the additional candidate $(x_g, c_{g,0})$ does not affect Algorithm 3.

Finally, by combining all the above results as in the Lipschitz case, we conclude Proposition 3. We present the full proof in Appendix H.

4.5. A parameter-free guarantee for our model selection method

To obtain a final optimality gap, we combine Algorithm 1 together with grid search and a base optimization algorithm. Let $\mathcal{G}$ be the set of hyperparameters over which we perform grid search, and $\{x_g\}_{g\in\mathcal{G}}$ be the solutions obtained by optimizing the objective for each $g \in \mathcal{G}$. We reuse the same N samples across all $|\mathcal{G}|$ runs of the optimization algorithm, yielding a total optimization complexity of $T_{\text{opt}} = \Theta\left(|\mathcal{G}|N\right)$.

Algorithm 1 uses additional N validation samples. However, Algorithm 1 does not calculate gradients, and thus we must quantify the evaluation complexity of Algorithm 1. For Algorithm 1, we consider all operations aside from the model evaluation as negligible. We thus take $T_{\text{select}} = \Theta\left(|\mathcal{G}|N\right)$ as the evaluation complexity of Algorithm 1; we further justify taking $T_{\text{select}} = \Theta\left(|\mathcal{G}|N\right)$ in Appendix B. Finally, we take $T = T_{\text{opt}} + T_{\text{select}} = \Theta\left(|\mathcal{G}|N\right)$ as the total evaluation complexity of the combined procedure.

We note that Algorithm 1 requires a sample size $N \geq 4\ln(4/\delta)$ to satisfy the requirements of Proposition 3. If N is below this threshold, we default to outputting the initial weight x_0. This is justified when $N \leq \alpha$ (for some $\alpha > 0$), since:

$$F(x_0) - F(x_\star) = \mathbb{E}_{s\sim\mathcal{P}}[f(x_0; s) - f(x_\star; s)] \leq L_\star\|x_0 - x_\star\|_\alpha \leq L_\star\|x_0 - x_\star\|_\alpha\sqrt{\alpha/N}. \quad (6)$$

We now informally state the optimality gap obtained by combining Algorithm 1 together with clipped SGD.

Corollary 5 *Let $\|\cdot\|_\alpha = \|\cdot\|_2$ be the Euclidean norm. Let $\delta \in (0, 1/2)$, $m \in \mathcal{I}_{\text{SM-Lip}}^{\overline{\ell},\|\cdot\|_{2,\beta}}$, $h \in \mathcal{I}_{\text{Lip}}^{1,\|\cdot\|_{\beta*}}$, and $f \in \mathcal{I}_{\text{Dist}}^{\overline{\rho},\|\cdot\|_2}$. By grid searching over $\text{Grid}_2\left(\underline{\ell}, \overline{\ell}\right) \times \text{Grid}_2\left(\underline{\rho}, \overline{\rho}\right)$, running clipped SGD (see Carmon and Hinder (2024)) with suitable parameters for each grid point, using the same N samples in each run, and then applying Algorithm 1, we obtain x, c such that, with probability at least $1 - 2\delta$,*

$$F_c(x) - F(x_\star) \leq O\left(\overline{L}_\star\overline{D}_\star\sqrt{\frac{\ln_+(\tilde{a}/\delta)\ln_+(\tilde{a})}{N}}\right) = O\left(\overline{L}_\star\overline{D}_\star\sqrt{\frac{\ln_+(\tilde{a}/\delta)\ln_+(\tilde{a})\ln_+\left(\overline{\ell}/\underline{\ell}\right)\ln_+\left(\overline{\rho}/\underline{\rho}\right)}{T}}\right),$$

where $\tilde{a} = \ln_+\left(\overline{\ell}/\underline{\ell}\right)\ln_+\left(\overline{\rho}/\underline{\rho}\right)\ln_+\left(\sqrt{N}\cdot\overline{L}_\star/\underline{\ell}\right)$.

Corollary 5 follows almost immediately from Proposition 3 and the guarantee for clipped SGD (Carmon and Hinder, 2024, Proposition 1a). We prove Corollary 5 in Appendix C. Additionally, in Appendix C, we derive slightly better optimality gaps for the Lipschitz setting.

5. A parameter-free method with the best possible sample complexity

This section focuses on producing a parameter-free method that matches the sample complexity of optimization with known parameters. Sample complexity refers to the number of function samples that the algorithm uses, and can be contrasted with stochastic gradient evaluation complexity (the subject of the previous section), which focuses on the number of stochastic gradient evaluations. The sample complexity perspective allows algorithms to perform operations that may be computationally infeasible in practice, such as computing an exact minimizer of the empirical objective over the N samples. In other words, sample-efficient algorithms make more efficient use of small quantities of data, whereas computationally efficient algorithms make more efficient use of compute.

We introduce Algorithm 2 and analyze its optimality gap in Theorem 8 under the sample complexity setting. Our algorithm achieves the best possible optimality gap without requiring *any* prior knowledge of either the Lipschitz constant $L_\star$ or the initial distance to the optimum $D_\star$.

Algorithm 2 first uses N samples from $\mathcal{P}$ to estimate a lower bound $\hat{L}$ on the Lipschitz constant $L_\star$. It then draws $2N$ additional samples from $\mathcal{P}$, conditioned on $m(\cdot; s)$ being $\hat{L}$-Lipschitz. Following Lawrence et al. (2025), the algorithm finds a weight with a small optimality gap under this conditioned

Algorithm 2: Optimization with model output clipping

Input: Initialization x_0, number of samples N, constants $\delta, \phi, \psi_\alpha$, and optimization algorithm **A** mapping a ball radius and N samples to an approximate minimizer of F in the ball.

if $N \leq 7$ *or* $\ln(1/\delta) > \frac{N}{2e}$ **then Output:** $x_0, 0$.

Compute the maximum Lipschitz constant $\hat{L}$ of $\{m(\cdot; s_i)\}_{i=1}^N$.

Compute $\lambda = 9 \cdot 66\phi \frac{\hat{L}\sqrt{\ln \frac{\psi_\alpha}{\delta}}}{\sqrt{N}}$.

$k \leftarrow 0$.

for $j = 1, \ldots, 5N$ **do**

  Draw $s \sim \mathcal{P}$, **if** $m(\cdot; s)$ *is* $\hat{L}$*-Lipschitz* **then** $k \leftarrow k+1$, $s_{N+k} \leftarrow s$.

end

if $k < 2N$ **then Output:** FAILURE.

Choose $\hat{x}_\lambda \in \arg\min_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=N+1}^{2N} f(x; s_i) + \lambda \|x - x_0\|_\alpha$.

Set $x_{\text{perfect}} \leftarrow \mathbf{A}\Big(\hat{L}, 3\|\hat{x}_\lambda - x_0\|_\alpha; s_{2N+1}, \ldots, s_{3N}\Big)$.

Output: $x_{\text{perfect}}, \hat{L}\|x_{\text{perfect}} - x_0\|_\alpha$.

distribution. Finally, it outputs this weight together with a clipping parameter to guarantee a small optimality gap under the original distribution $\mathcal{P}$.

For the algorithm to work properly, we make two assumptions. First, given a radius $\hat{R}$ and the Lipschitz constant $\hat{L}$, the optimization algorithm **A** must be able to achieve a small optimality gap.

Assumption 6 *There exist constants $\phi \geq 1$ and $\psi_\alpha \geq 1$ such that for all $\hat{L} > 0$, $f \in \mathcal{I}_{Dist}^{\infty, \|\cdot\|_\alpha}$, $\hat{R} \geq 0$, $\delta \in \left(0, \frac{1}{5}\right)$, and any distribution $\tilde{\mathcal{P}}$ over $\mathbb{S}$ such that $f(\cdot; s)$ is $\hat{L}$-Lipschitz a.s., **A** satisfies*

$$\mathbb{P}_{s_1, \ldots, s_N \overset{iid}{\sim} \tilde{\mathcal{P}}} \left[\hat{F}\Big(\boldsymbol{A}\Big(\hat{L}, \hat{R}; s_1, \ldots, s_N\Big)\Big) - \min_{x \in \mathcal{X}: \|x - x_0\|_\alpha \leq \hat{R}} \hat{F}(x) \leq \phi \frac{\hat{L}\hat{R}\sqrt{\ln \frac{\psi_\alpha}{\delta}}}{\sqrt{N}} \right] \geq 1 - \delta, \text{ and}$$

$$\mathbb{P}_{s_1, \ldots, s_N \overset{iid}{\sim} \tilde{\mathcal{P}}} \left[\boldsymbol{A}\Big(\hat{L}, \hat{R}; s_1, \ldots, s_N\Big) \in \Big\{ x \in \mathcal{X} \mid \|x - x_0\|_\alpha \leq \hat{R} \Big\} \right] = 1,$$

where $\hat{F}(x) \coloneqq \int_{s \sim \tilde{\mathcal{P}}} f(x; s)$.

Second, similarly to Lawrence et al. (2025), we assume that the empirical mean of the stochastic gradients $\nabla f(x; s)$ is a good approximation to the true gradient $\nabla F(x)$.

Assumption 7 *There exist $\phi \geq 1$ and $\psi_\alpha \geq 1$ such that for any $f \in \mathcal{I}_{diff}$, any $\delta \in \left(0, \frac{1}{5}\right)$, any $x \in \mathcal{X}$, any $\hat{L} \geq 0$, and any distribution $\tilde{\mathcal{P}}$ over $\mathbb{S}$ such that $f(\cdot; s)$ is $\hat{L}$-Lipschitz a.s., we have*

$$\mathbb{P}_{s_1, \ldots, s_N \overset{iid}{\sim} \tilde{\mathcal{P}}} \left[\left\| \nabla \mathbb{E}_{s \sim \tilde{\mathcal{P}}} f(x; s) - \frac{1}{N} \sum_{i=1}^N \nabla f(x; s_i) \right\|_{\alpha *} > 594 \cdot \frac{\phi}{2} \cdot \frac{\hat{L}\sqrt{\ln \frac{\psi_\alpha}{\delta}}}{\sqrt{N}} \right] \leq \delta.$$

Assumption 6 is known to hold for many optimization algorithms with $\phi = O(1)$. Specifically, if $\|\cdot\|_\alpha = \|\cdot\|_1$, then Assumption 6 holds for $\psi_\alpha = d$ with entropic mirror descent (i.e., mirror descent with KL divergence) (Beck and Teboulle, 2003; Nemirovski and Yudin, 1983) with adaptive step sizes (Orabona, 2021). Additionally, for the Euclidean norm $\|\cdot\|_\alpha = \|\cdot\|_2$, the assumption holds $\psi_\alpha = 1$ with ADAPTIVE SGD (e.g., Gupta et al., 2017).

Similarly, Assumption 7 holds with $\phi = O(1)$ for $\|\cdot\|_\alpha = \|\cdot\|_1$ and $\|\cdot\|_\alpha = \|\cdot\|_2$, with $\psi_\alpha = d$, and $\psi_\alpha = 1$, respectively. We provide the proof that Assumption 7 holds in these cases in Appendix G.

When both assumptions hold, we get the following optimality gap guarantee. For the Euclidean norm $\|\cdot\|_\alpha = \|\cdot\|_2$, and when the optimization algorithm **A** is ADAPTIVE SGD, this matches the best possible sample complexity guarantee (Carmon and Hinder, 2024, Proposition 1b).

Theorem 8 *Let $\delta \in \left(0, \frac{1}{5}\right)$, $m \in \mathcal{I}_{\mathit{Lip}}^{\infty,\|\cdot\|_{\alpha,\beta}}$, $h \in \mathcal{I}_{\mathit{Lip}}^{1,\|\cdot\|_{\beta*}}$ and $f \in \hat{\mathcal{I}}_{\mathit{Dist}}^{\infty,\|\cdot\|_\alpha}$. If Assumptions 6 and 7 hold with ϕ and ψ_α, and $x_{\mathit{perfect}}, \hat{L}\|x_{\mathit{perfect}} - x_0\|_\alpha$ are the outputs of Algorithm 2, then with probability at least $1 - 5\delta$ we have*

$$F_{\hat{L}\|x_{\mathit{perfect}} - x_0\|_\alpha}(x_{\mathit{perfect}}) \le F(x_\star) + O\left(\phi \frac{L_\star \|x_\star - x_0\|_\alpha \sqrt{\ln \frac{\psi_\alpha}{\delta}}}{\sqrt{N}}\right).$$

Proof outline for Theorem 8 (full proof in Appendix I). We define $L(s)$ as the Lipschitz constant of $m(\cdot; s)$, and $L(\mathcal{P})$ as the distribution over the Lipschitz constants resulting from the distribution $\mathcal{P}$. We prove that with high probability $\hat{L} = \max_{i\in[N]} L(s_i)$ is greater than the quantile $Q_{1-\ln(1/\delta)/N}(L(\mathcal{P}))$. This means that for most samples $s \sim \mathcal{P}$, $m(\cdot; s)$ is $\hat{L}$-Lipschitz. As a consequence, due to the model clipping, the samples $s \in \mathbb{S}$ for which $m(\cdot; s)$ is not $\hat{L}$-Lipschitz increase the optimality gap by at most $\left(\hat{L}\|x_{\text{perfect}} - x_0\|_\alpha + L_\star\|x_\star - x_0\|_\alpha\right)\frac{\ln(1/\delta)}{N}$.

We further show that among $5N$ samples drawn from $\mathcal{P}$, at least $2N$ have Lipschitz constant below $Q_{1-\ln(1/\delta)/N}(L(\mathcal{P}))$ with high probability, implying that Algorithm 2 does not output failure.

Define $\hat{F}(x) := \mathbb{E}_{s\sim\mathcal{P}}\left[f(x; s) \mid L(s) \le \hat{L}\right]$. We denote $\hat{R} := 3\|\hat{x}_\lambda - x_0\|_\alpha$. Additionally, for any $\theta \ge 0$, we define $\hat{x}_\theta^\star \in \arg\min_{x\in\mathcal{X}}\left\{\hat{F}(x) + \theta\|x - x_0\|_\alpha\right\}$. Using Lemma 2 of Lawrence et al. (2025), we show that there exists $\lambda_{\hat{R}} \in [\lambda/3, 3\lambda]$ such that $\|\hat{x}_{\lambda_{\hat{R}}}^\star - x_0\|_\alpha \in \left[\frac{\hat{R}}{33}, \hat{R}\right]$. Now, similarly to Lawrence et al. (2025), we obtain

$$\min_{x\in\mathcal{X}:\|x-x_0\|_\alpha\le\hat{R}} \hat{F}(x) \le \hat{F}\left(\hat{x}_{\lambda_{\hat{R}}}^\star\right) \le \hat{F}(x_\star) + \lambda_{\hat{R}}\left(\|x_\star - x_0\|_\alpha - \|\hat{x}_{\lambda_{\hat{R}}}^\star - x_0\|_\alpha\right). \tag{7}$$

In the case that $\hat{L}\hat{R} \le O(L_\star\|x_\star - x_0\|_\alpha)$, i.e. $\hat{L}\hat{R}$ is small enough relative to $O(L_\star\|x_\star - x_0\|_\alpha)$, we continue in a similar manner to Lawrence et al. (2025) and use Assumption 6 to prove the final optimality gap. However, $\hat{L}\hat{R}$ can be significantly larger than $O(L_\star\|x_\star - x_0\|_\alpha)$. In this case, we exploit the negative term $-\lambda_{\hat{R}}\|\hat{x}_{\lambda_{\hat{R}}}^\star - x_0\|_\alpha$ on the right-hand side of Equation (7) to cancel the positive terms in the bound, and thus, with high probability $F_{\hat{L}\|x_{\text{perfect}} - x_0\|_\alpha}\left(x_{\text{perfect}}\right) \le F(x_\star)$. As a result, we obtain the required optimality gap for all cases.

6. Lower bounds for the model–loss setup

Our algorithms succeed in achieving a near-optimal optimality gap of $\widetilde{O}\left(L_\star D_\star\sqrt{\ln(1/\delta)}/\sqrt{N}\right)$, surpassing the lower bounds in the case of uncertainty in both the Lipschitz constant and the distance from the initial point to the optimum. We achieve this because we assume that the optimized function can be decomposed into a model and a loss, and introduce a method to clip the model output. Therefore, at first glance, it is not clear that any of the known lower bounds apply.

However, in some cases, we can decompose a function f into a model m and a loss function h such that all the complexity of f is inherited by h. In particular, consider lower-bound constructions in which each function f_i is $L_\star$-Lipschitz for a known constant $L_\star$. For each such f_i, we decompose f_i into $m_i(x; s) := xL_\star$ and $h_i(x; s) := f_i(x/L_\star; s)$. Under this decomposition, each m_i is $L_\star$-Lipschitz, and each h_i is 1-Lipschitz. Moreover, having the ability to modify the output of $m_i(x; s)$ is equivalent to modifying the input x. Consequently, any lower-bound argument that applies to the original functions f_i continues to apply in the model–loss setting. Therefore, the lower bound $\Omega\left(L_\star D_\star \sqrt{\ln(1/\delta)}/\sqrt{N}\right)$ established in Carmon and Hinder (2024, Proposition 1b) for the case where $L_\star$ and $D_\star$ are known also holds in our setting. Hence, the optimality gaps established in this work are still near-optimal even in the model–loss setting.

In contrast, not all lower bounds carry over to the model–loss setting. For instance, the lower bounds established in Carmon and Hinder (2024, Theorem 3) no longer apply, as we obtain strictly better optimality gaps in this work. In Carmon and Hinder (2024, Theorem 3), the Lipschitz constant $L_\star$ is unknown, and only coarse lower and upper bounds on $L_\star$ are assumed. Their construction involves two types of functions: one with a uniformly small Lipschitz constant, and another that has a small Lipschitz constant for most samples but a large Lipschitz constant on a rare subset of samples. Consequently, the reduction argument from the previous paragraph does not apply to this setting. In our framework, however, the ability to clip the model output limits the loss contribution of rare high-Lipschitz samples. This mechanism prevents the adversarial behavior exploited in Carmon and Hinder (2024, Theorem 3), enabling us to achieve near-optimal optimality gaps despite the unknown Lipschitz constant.

In Appendix D, we formally state the lower bound $\Omega\left(L_\star D_\star \sqrt{\ln(1/\delta)}/\sqrt{N}\right)$ that carries over to the model–loss setting.

Acknowledgments

This work was supported by the NSF-BSF program, under NSF grant #2239527 and BSF grant #2022663. YC acknowledges support from the Israeli Science Foundation (ISF) grant no. 2486/21.

References

Amit Attia and Tomer Koren. How free is parameter-free stochastic optimization? In *International Conference on Machine Learning (ICML)*, 2024.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Online learning with imperfect hints. In *International Conference on Machine Learning (ICML)*, 2020.

A.A. Borovkov. *Probability Theory*. CRC Press, 1999.

Yair Carmon and Oliver Hinder. Making SGD parameter-free. In *Conference on Learning Theory (COLT)*, 2022.

Yair Carmon and Oliver Hinder. The price of adaptivity in stochastic convex optimization. In *Conference on Learning Theory (COLT)*, 2024.

Keyi Chen, John Langford, and Francesco Orabona. Better parameter-free stochastic optimization with ODE updates for coin-betting. In *AAAI Conference on Artificial Intelligence*, 2022.

Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Conference on Learning Theory (COLT)*, 2018.

John C Duchi. Introductory lectures on stochastic optimization. *The Mathematics of Data*, 25: 99–186, 2018.

Vineet Gupta, Tomer Koren, and Yoram Singer. A unified approach to adaptive regularization in online and stochastic optimization. *arXiv:1706.06569*, 2017.

Steve Hanneke. The optimal sample complexity of pac learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. We refer to the latest arXiv version: https://arxiv.org/abs/1808.03204v8.

Maor Ivgi, Oliver Hinder, and Yair Carmon. DoG is SGD's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning (ICML)*, 2023. We refer to the latest arXiv version: https://arxiv.org/abs/2302.12022.

Andrew Jacobsen and Ashok Cutkosky. Parameter-free mirror descent. In *Conference on Learning Theory (COLT)*, 2022.

Ahmed Khaled and Chi Jin. Tuning-free stochastic optimization. In *International Conference on Machine Learning (ICML)*, 2024.

Itai Kreisler, Maor Ivgi, Oliver Hinder, and Yair Carmon. Accelerated parameter-free stochastic optimization. In *Conference on Learning Theory (COLT)*, 2024.

Jared Lawrence, Ari Kalinsky, Hannah Bradfield, Yair Carmon, and Oliver Hinder. The sample complexity of parameter-free stochastic convex optimization. *arXiv:2506.11336*, 2025.

Tuo Liu, El Mehdi Saad, Wojciech Kotłowski, and Francesco Orabona. Dual averaging converges for nonconvex smooth stochastic optimization. *arXiv:2505.21394*, 2025.

Haipeng Luo and Robert E Schapire. Achieving all with no parameters: AdaNormalHedge. In *Conference on Learning Theory (COLT)*, 2015.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory (COLT)*, 2009.

Zakaria Mhammedi and Wouter M Koolen. Lipschitz and comparator-norm adaptivity in online learning. In *Conference on Learning Theory (COLT)*, 2020.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory (COLT)*, 2019.

Jean-Jacques Moreau. Proximity and duality in a hilbertian space. *Bulletin of the Mathematical Society of France*, 93:273–299, 1965.

Arkadi Nemirovski and David Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Francesco Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Francesco Orabona. A modern introduction to online learning. *arXiv:1912.13213*, 2021.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Francesco Orabona and Tatiana Tommasi. Training deep networks without learning rates through coin betting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton University Press, 1970.

Vladimir Vovk. On-line regression competitive with reproducing kernel hilbert spaces. In *International Conference on Theory and Applications of Models of Computation*, pages 452–463. Springer, 2006.

Zhiyu Zhang, Ashok Cutkosky, and Ioannis Paschalidis. PDE-based optimal strategy for unconstrained online learning. In *International Conference on Machine Learning (ICML)*, 2022.

Contents

Appendix A. Reliable model selection method

Algorithm 3: Reliable model selection method

Input: A function f, a scalar $\gamma \in [1, \infty)$, candidate solutions $x_0, x_1, \ldots, x_K$, samples $s_1, s_2, \ldots, s_N$, and scalars $\tau_1, \ldots, \tau_K$.

$\tau_0 := 0$

$\theta = \min_{k \in [K] \cup \{0\}} \left(\gamma \tau_k + \frac{1}{N} \sum_{i=1}^{N} f(x_k; s_i) \right)$.

$\mathcal{F} = \left\{ k \in [K] \cup \{0\} \mid \tau_k + \frac{1}{N} \sum_{i=1}^{N} f(x_k; s_i) \leq \theta \right\}$.

$k_{\text{reliable}} = \arg\min_{k \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} f(x_k; s_i)$.

return $x_{k_{\text{reliable}}}$.

Throughout this paper, for simplicity, we use $\gamma = 1$. However, the result can easily be expanded to any fixed $\gamma \in [1, \infty)$.

A.1. Generalization of Lemma 1

Lemma 1 provides a high-probability guarantee on how much worse the result of Algorithm 3 can be relative to the other candidates. We generalize the guarantee to all candidates in $\mathcal{F}$.

Lemma 9 *Let $\hat{\delta} \in (0, 1)$. Let K be the number of models, and N be the validation sample set size. Let $\tau_1, \ldots, \tau_K$ be non-negative scalars such that with probability $1 - \hat{\delta}$ for all $k \in [K]$ we have $\left| F(x_k) - F(x_0) - \frac{1}{N} \sum_{i=1}^{N} (f(x_k; s_i) - f(x_0; s_i)) \right| \leq \tau_k$. Let $\gamma \in [1, \infty)$, and let $\mathcal{F}$ be the set $\mathcal{F}$ calculated by Algorithm 3 with γ. Then, with probability $1 - \hat{\delta}$, for any $k_{\mathcal{F}} \in \mathcal{F}$ we have*

$$F(x_{k_{\mathcal{F}}}) \leq \min_{k \in [K] \cup \{0\}} (F(x_k) + (1 + \gamma)\tau_k).$$

The proof of Lemma 9 is the same as the proof of Lemma 1 presented in Lawrence et al. (2025). For completeness, we reproduce that proof here.

Proof Let $k_{\mathcal{F}}$ be any element of $\mathcal{F}$. Define

$$\bar{F}(x_k) = \frac{1}{N} \sum_{i=1}^{N} f(x_k; s_i).$$

For all $k \in [K]$ we have

$$\begin{aligned} 0 &\overset{(i)}{\leq} \theta - \bar{F}(x_{k_{\mathcal{F}}}) - \tau_{k_{\mathcal{F}}} \overset{(ii)}{\leq} \bar{F}(x_k) + \gamma \tau_k - \bar{F}(x_{k_{\mathcal{F}}}) - \tau_{k_{\mathcal{F}}} \\ &= \gamma \tau_k + \bar{F}(x_k) - \bar{F}(x_0) + \bar{F}(x_0) - \bar{F}(x_{k_{\mathcal{F}}}) - \tau_{k_{\mathcal{F}}} \overset{(iii)}{\leq} (1 + \gamma)\tau_k + F(x_k) - F(x_{k_{\mathcal{F}}}), \end{aligned}$$

where (i) follows from $k_{\mathcal{F}} \in \mathcal{F}$, (ii) follows from the definition of θ, and (iii) follows from the fact that $\left| F(x_j) - F(x_0) - \left(\bar{F}(x_j) - \bar{F}(x_0) \right) \right| \leq \tau_j$ for all $j \in [K]$. ∎

Appendix B. Discussion about the evaluation complexity

Below, we quantify the evaluation complexity of Algorithm 1.

- In the case that $m \in \mathcal{I}_{\text{Lip}}^{\infty,\|\cdot\|_{\alpha,\beta}}$ we only need to include a single value in $\mathcal{C}_g$. In this case, the total evaluation complexity in Algorithm 1 is $T_{\text{select}} = \Theta\left(|\mathcal{G}|N\right)$.
- If the grid search is also done on the Lipschitz constant, and one of the generated x_g is associated with L_g that exceeds $L_\star$ by at most a constant factor, then we can modify $\mathcal{C}_g$ to contain only a single clipping parameter without changing the proof of Proposition 3. We can set $\mathcal{C}_g = \{L_g\|x_g - x_0\|_\alpha\}$ in the context that $m \in \mathcal{I}_{\text{Lip}}^{\infty,\|\cdot\|_{\alpha,\beta}}$, and $\mathcal{C}_g = \left\{\sqrt{\frac{N}{\ln(4/\delta)}} L_g\|x_g - x_0\|_\alpha\right\}$ in the context that $m \in \mathcal{I}_{\text{SM-Lip}}^{\infty,\|\cdot\|_{\alpha,\beta}}$. With this modification, the evaluation complexity of Algorithm 1 is $T_{\text{select}} = \Theta\left(|\mathcal{G}|N\right)$.
- Lastly, not all operations take the same time. Computing gradients and evaluating the output of the model m are considered time-consuming operations. In contrast, obtaining the $|\mathcal{C}_g|$ values needed to form $\mathcal{C}_g$ and evaluating the loss function h take relatively insignificant time. Therefore, if we consider only the operations of evaluating the output of the model m, the total evaluation complexity in Algorithm 1 is $T_{\text{select}} = \Theta\left(|\mathcal{G}|N\right)$. If we do not ignore the remaining operations, then obtaining the values for $\mathcal{C}_g$ takes $\Theta\left(|\mathcal{G}|N \log|\mathcal{C}_g|\right)$. Similarly, evaluating all $\tau_{g,i}$ takes $\Theta\left(|\mathcal{G}|\left(N + |\mathcal{C}_g|^2\right)\right)$ since we do not need to recompute everything from scratch for each clipping parameter in $\mathcal{C}_g$. Moreover, Proposition 3 continues to hold even if, for each $g \in \mathcal{G}$, $\mathcal{C}_g$ contains only the n-th largest value of the *multiset* $\{\text{binary-ceil}(\|m(x_g; s_i) - m(x_0; s_i)\|_\beta; \underline{\ell}\|x_g - x_0\|_\alpha)\}_{i=1}^N$; in this case, there is only one confidence interval τ_g per grid point, reducing the total time required to compute the confidence intervals to $\Theta\left(|\mathcal{G}|N\right)$.

For the reasons above, we take $T_{\text{select}} = \Theta\left(|\mathcal{G}|N\right)$ as the evaluation complexity of Algorithm 1.

Appendix C. Additional upper bounds

Corollary 5 *Let $\|\cdot\|_\alpha = \|\cdot\|_2$ be the Euclidean norm. Let $\delta \in (0, 1/2)$, $m \in \mathcal{I}_{\textit{SM-Lip}}^{\overline{\ell},\|\cdot\|_{2,\beta}}$, $h \in \mathcal{I}_{\textit{Lip}}^{1,\|\cdot\|_{\beta*}}$, and $f \in \mathcal{I}_{\textit{Dist}}^{\overline{\rho},\|\cdot\|_2}$. By grid searching over Grid$_2\left(\underline{\ell}, \overline{\ell}\right) \times$ Grid$_2\left(\underline{\rho}, \overline{\rho}\right)$, running clipped SGD (see Carmon and Hinder (2024)) with suitable parameters for each grid point, using the same N samples in each run, and then applying Algorithm 1, we obtain x, c such that, with probability at least $1 - 2\delta$,*

$$F_c(x) - F(x_\star) \le O\left(\overline{L}_\star\overline{D}_\star\sqrt{\frac{\ln_+(\tilde{a}/\delta)\ln_+(\tilde{a})}{N}}\right) = O\left(\overline{L}_\star\overline{D}_\star\sqrt{\frac{\ln_+(\tilde{a}/\delta)\ln_+(\tilde{a})\ln_+\left(\overline{\ell}/\underline{\ell}\right)\ln_+\left(\overline{\rho}/\underline{\rho}\right)}{T}}\right),$$

where $\tilde{a} = \ln_+\left(\overline{\ell}/\underline{\ell}\right)\ln_+\left(\overline{\rho}/\underline{\rho}\right)\ln_+\left(\sqrt{N}\cdot\overline{L}_\star/\underline{\ell}\right)$.

Proof If $N < 4\ln(4/\delta)$, we return $x_0, 0$ instead of performing a grid search. In this case, applying Equation (6) gives

$$F_0(x_0) - F(x_\star) = F(x_0) - F(x_\star) \le O\left(\overline{L}_\star\overline{D}_\star\sqrt{\frac{\ln_+(1/\delta)}{N}}\right),$$

which is the desired guarantee.

Now suppose that $N \geq 4\ln(4/\delta)$. Let the grid be $\mathcal{G} = \mathrm{Grid}_2\big(\underline{\ell}, \overline{\ell}\big) \times \mathrm{Grid}_2\big(\underline{\rho}, \overline{\rho}\big)$. For every $(L, D) \in \mathcal{G}$, let $x_{L,D}$ be the weight obtained by running clipped SGD for N steps with step size

$$\eta = \frac{D}{L\sqrt{N}},$$

gradient clipping threshold

$$G_{\mathrm{clip}} = \frac{L\sqrt{N}}{\sqrt{\log(1/\delta)}},$$

Euclidean projection onto

$$\mathcal{X} \cap \{x \mid \|x - x_0\|_2 \leq D\},$$

and the same N samples across all grid points.

By the definition of $\mathcal{G}$, there exists a grid point $(L, D) \in \mathcal{G}$ such that

$$\overline{L}_\star \leq L \leq 2\overline{L}_\star \qquad \text{and} \qquad \overline{D}_\star \leq D \leq 2\overline{D}_\star.$$

For this grid point, Carmon and Hinder (2024, Proposition 1a) implies[1] that, with probability at least $1 - \delta$,

$$F(x_{L,D}) - F(x_\star) \leq O\left(\overline{L}_\star \overline{D}_\star \sqrt{\frac{\ln_+(1/\delta)}{N}}\right).$$

Let x, c be the result obtained by applying Algorithm 1 to the candidates $\{x_{L,D}\}_{(L,D)\in\mathcal{G}}$ using N fresh validation samples. Applying Proposition 3 and taking a union bound over the clipped-SGD guarantee and the validation guarantee yields the desired bound on $F_c(x) - F(x_\star)$. ∎

We can also use Algorithm 1 with Carmon and Hinder (2022) to obtain a slightly different optimality gap in the Lipschitz case. Compared to the previous bound, this gap depends less on the upper and lower bounds on the Lipschitz constant and distance to the optima $\underline{\ell}, \overline{\ell}, \underline{\rho}, \overline{\rho}$, but depends more on δ.

Corollary 10 *Let $\|\cdot\|_\alpha = \|\cdot\|_2$ be the Euclidean norm. Let $m \in \mathcal{I}_{\mathit{Lip}}^{\overline{\ell}, \|\cdot\|_{2,\beta}}$, $h \in \mathcal{I}_{\mathit{Lip}}^{1, \|\cdot\|_{\beta*}}$ and $f \in \mathcal{I}_{\mathit{Dist}}^{\overline{\rho}, \|\cdot\|_2}$. Let $\delta \in (0, 1/2)$. Let the grid be $\mathcal{G} = \mathit{Grid}_{\sqrt{N}}\big(\underline{\ell}, \overline{\ell}\big)$. For every $L \in \mathcal{G}$, let x_L be the resulting weight of running Carmon and Hinder (2022) with $\eta_\varepsilon = \underline{\rho}/L$, and the same N samples across all runs. If $N \geq 2$, let x, c be the result obtained by applying Algorithm 1, with $n = 1$ and using additional N validation samples, on the candidates $\{x_g\}_{g\in\mathcal{G}}$; otherwise, let x, c be $x_0, 0$. With probability at least $1 - 2\delta$, we have*

$$\begin{aligned} F_c(x) - F(x_\star) &\leq O\left(\overline{L}_\star \overline{D}_\star \frac{\ln_+\left(\tilde{b}/\delta\right)\sqrt{\ln_+(\tilde{a}/\delta)} + \ln_+^2\left(\tilde{b}/\delta\right)}{\sqrt{N}}\right) \\ &= O\left(\overline{L}_\star \overline{D}_\star \frac{\left(\ln_+\left(\tilde{b}/\delta\right)\sqrt{\ln_+(\tilde{a}/\delta)} + \ln_+^2\left(\tilde{b}/\delta\right)\right)\sqrt{\log_{\sqrt{N}}\left(e + \overline{\ell}/\underline{\ell}\right)}}{\sqrt{T}}\right), \end{aligned}$$

1. While Carmon and Hinder (2024) assume throughout that $f(\cdot; s)$ is convex for all $s \in \mathbb{S}$, the proof of Proposition 1a only uses that $F(\cdot)$ is convex.

where $\tilde{a} = \log_{\sqrt{N}}\left(e + \overline{\ell}/\underline{\ell}\right) \ln_+\left(\overline{L}_\star/\underline{\ell}\right)$ *and* $\tilde{b} = \ln_+\left(N\overline{D}_\star/\underline{\rho}\right)$.

Proof If $N < 2$, then simply applying Equation (6) gives the desired guarantee

$$F_0(x_0) - F(x_\star) = F(x_0) - F(x_\star) \le O\left(\overline{L}_\star \overline{D}_\star \sqrt{\frac{\ln_+(1/\delta)}{N}}\right).$$

From the definition of $\mathcal{G}$, there exists a point in the grid $L \in \mathcal{G}$ such that $\overline{L}_\star \le L \le \sqrt{N}\overline{L}_\star$. For this point, Carmon and Hinder (2024, Theorem 2) proved that, with probability at least $1 - \delta$,

$$F(x_L) - F(x_\star) \le O\left(\overline{L}_\star \overline{D}_\star \frac{\ln_+^2\left(\tilde{b}/\delta\right)}{\sqrt{N}}\right).$$

It additionally proves that $\|x_L - x_0\| \le O\left(\overline{D}_\star \ln_+\left(\tilde{b}/\delta\right)\right)$. Now, by simply applying Proposition 3 and using the union bound, we obtain the desired guarantee on $F_c(x) - F(x_\star)$. ■

Appendix D. Formally stated lower bound

Although our method clips the model output to achieve improved optimality gaps, we establish lower bounds that apply more broadly to any method that modifies the model output. Specifically, we consider algorithms that, instead of outputting a weight $x \in \mathcal{X}$, output a function $q^m : \mathbb{S} \to \mathcal{Y}$ that represents a generic modification to the model output. In the finite case, we restrict the function to be of the form $q^m(s) := U(m(x_1; s), \ldots, m(x_k; s))$ for some $k \in \mathbb{N}$, where $x_1, \ldots, x_k \in \mathcal{X}$ and $U : \mathcal{Y}^k \to \mathcal{Y}$ are output by the optimization algorithm. More generally, we only require that the output function $q^m : \mathbb{S} \to \mathcal{Y}$ is in the set

$$\mathbb{Q}_m := \{q \in \mathbb{S} \to \mathcal{Y} \mid \forall s_1, s_2 \in \mathbb{S} : \text{if } \forall x \in \mathcal{X} : m(x; s_1) = m(x; s_2) \text{ then } q(s_1) = q(s_2)\}. \tag{8}$$

D.1. Formal setup

We now formally define the basic building blocks of this section: stochastic optimization problems, algorithms, error metrics, and minimax rates. We define them similarly to Carmon and Hinder (2024).

Stochastic optimization under model–loss decomposition (SOML) problems. A SOML problem instance is a tuple $(m, h, f, \mathcal{P})$ containing a distribution $\mathcal{P}$ over $\mathbb{S}$, a sample objective $f : \mathcal{X} \times \mathbb{S} \to \mathbb{R}$, and functions $m : \mathcal{X} \times \mathbb{S} \to \mathcal{Y}$ and $h : \mathcal{Y} \times \mathbb{S} \to \mathbb{R}$ such that $f(x; s) := h(m(x; s); s)$. Let $\mathcal{I}_{\text{SOML}}$ denote the class of all SOML problem instances. We consider two fundamental classes of convex functions with a minimizer at most D away from the initial weight x_0. The first class contains problems for which m is L-Lipschitz and h is 1-Lipschitz.

$$\mathcal{I}^{L,D}_{\text{SOML-Lip}} := \left\{(m, h, f, \mathcal{P}) \in \mathcal{I}_{\text{SOML}} \mid m \in \mathcal{I}^{L,\|\cdot\|_{\alpha,\beta}}_{\text{Lip}} \text{ and } h \in \mathcal{I}^{1,\|\cdot\|_{\beta*}}_{\text{Lip}}, \text{ and } f \in \hat{\mathcal{I}}^{D,\|\cdot\|_\alpha}_{\text{Dist}}\right\}.$$

The second class contains problems for which m is L-second-moment-Lipschitz and h is 1-Lipschitz,

$$\mathcal{I}^{L,D}_{\text{SOML-sm-Lip}} := \left\{(m, h, f, \mathcal{P}) \in \mathcal{I}_{\text{SOML}} \mid m \in \mathcal{I}^{L,\|\cdot\|_{\alpha,\beta}}_{\text{SM-Lip}} \text{ and } h \in \mathcal{I}^{1,\|\cdot\|_{\beta*}}_{\text{Lip}}, \text{ and } f \in \hat{\mathcal{I}}^{D,\|\cdot\|_\alpha}_{\text{Dist}}\right\}.$$

Optimization algorithms. We define a SOML algorithm as an algorithm that has unrestricted access to the functions m, h and f, but observes $\mathcal{P}$ only through samples $s_1, \cdots, s_N \overset{\text{iid}}{\sim} \mathcal{P}$, and outputs $q_N^m : \mathbb{S} \to \mathcal{Y}$ such that $q_N^m \in \mathbb{Q}_m$. Such algorithms are allowed to perform computationally infeasible operations, such as finding the Lipschitz constant of $m(\cdot, s)$ for some $s \in \mathbb{S}$. Thus, the unrestricted access to the functions m, h and f allows the algorithm to access the ith-order derivative for any i. We write $\mathcal{A}_{\text{SOML}}$ for the set of all SOML algorithms.

Error metrics. For an algorithm alg, SOML instance $(m, h, f, \mathcal{P})$, and budget N, we let q_N^m denote the algorithm's output given N samples. We define the high-probability error at confidence level δ as

$$\mathsf{Err}_N^\delta(\mathsf{alg}, (m, h, f, \mathcal{P})) = Q_{1-\delta}(\mathbb{E}_{s\sim\mathcal{P}} h(q_N^m(s); s)) - \inf_{x_\star \in \mathcal{X}} \mathbb{E}_{s\sim\mathcal{P}} f(x_\star; s),$$

where $Q_p(Y) \coloneqq \min\{y : \mathbb{P}[Y \le y] \ge p\}$ denotes the p-th quantile of the random variable Y, similar to the definition of $Q_p(\cdot)$ over distributions.

Minimax error. Given an instance class $\mathcal{I}$ and an algorithm alg, we overload the notation above to denote worst-case error over the class,

$$\mathsf{Err}_N^\delta(\mathsf{alg}, \mathcal{I}) \coloneqq \sup_{(m,h,f,\mathcal{P})\in\mathcal{I}} \mathsf{Err}_N^\delta(\mathsf{alg}, (m, h, f, \mathcal{P})).$$

Given an algorithm class $\mathcal{A}$, we further overload our notation to express minimax optimal worst-case error

$$\mathsf{Err}_N^\delta(\mathcal{A}, \mathcal{I}) \coloneqq \inf_{\mathsf{alg}\in\mathcal{A}} \mathsf{Err}_N^\delta(\mathsf{alg}, \mathcal{I}) = \inf_{\mathsf{alg}\in\mathcal{A}} \sup_{(m,h,f,\mathcal{P})\in\mathcal{I}} \mathsf{Err}_N^\delta(\mathsf{alg}, (m, h, f, \mathcal{P})).$$

D.2. Lower bound statement

We now formally establish the lower bound. The proof of the bound is a reduction to the lower bounds established in Carmon and Hinder (2024, Proposition 1b and Theorem 2).

Proposition 11 *For any $L, D > 0$, $\delta \in \left(0, \frac{1}{2}\right)$, and any number of samples $N \ge 1$, we have:*

$$\mathsf{Err}_N^\delta\left(\mathcal{A}_{SOML}, \mathcal{I}_{SOML\text{-}sm\text{-}Lip}^{L,D}\right) \ge \mathsf{Err}_N^\delta\left(\mathcal{A}_{SOML}, \mathcal{I}_{SOML\text{-}Lip}^{L,D}\right) \ge \Omega\left(LD \min\left\{\sqrt{\ln_+(1/\delta)}/\sqrt{N}, 1\right\}\right).$$

Proof The proof mostly follows from the proof of Carmon and Hinder (2024, Proposition 1b). In the proof of Carmon and Hinder (2024, Proposition 1b), they let $\mathcal{X} = \mathbb{R}, \mathbb{S} = \{0, 1\}$ and the optimized function $f : \mathcal{X} \times \mathbb{S} \to \mathbb{R}$ be

$$f(x; s) = \begin{cases} L|x| & \text{s=0} \\ L|x - D| & \text{s=1.} \end{cases}$$

They consider the distributions $P_v \coloneqq \text{Bernoulli}\left(\frac{1+(2v-1)\varepsilon}{2}\right)$ for $v \in \{0, 1\}$ and some $\varepsilon \in [0, \frac{1}{2}]$.

We set $\mathcal{Y} = \mathbb{R}$ and decompose f into a model m and a loss function h as follows:

$$m(x; s) = Lx,$$

and

$$h(x;s) = \begin{cases} |x| & \text{s=0} \\ |x - LD| & \text{s=1.} \end{cases}$$

For this decomposition it is easy to see that $f(x;s) = h(m(x;s);s)$, that m is L-Lipschitz, h is 1-Lipschitz, and that $f \in \hat{\mathcal{I}}^{D,|\cdot|}_{\text{Dist}}$.

Let q_N^m be the output of the optimization algorithm after interacting with m, h and f through N samples. As $m(x;s)$ depends only on x and does not depend on s, $q_N^m(s)$ does not depend on s. As such, we can choose $x_N = q_N^m(\cdot)/L$. Therefore, $f(x_N;s) = h(q_N^m(s);s)$. Thus, from here we can use the same proof as in the proof of Carmon and Hinder (2024, Proposition 1b). Equivalently, any optimization algorithm used for optimizing the problems described here can be used to optimize the problems described in the proof of Carmon and Hinder (2024, Proposition 1b), and vice versa.

We note that $|\cdot|$ is not differentiable at 0. Using a smoothing argument, the lower bound can also be extended to the differentiable setting. For example, we can replace each occurrence of $|\cdot|$ in the construction of $h(\cdot;\cdot)$ with

$$|y|_\gamma := \begin{cases} |y| & \text{if } y \geq \gamma \text{ or } y \leq -\gamma \\ \frac{y^2}{2\gamma} + \frac{\gamma}{2} & \text{if } y \in [-\gamma, \gamma] \end{cases},$$

for some $\gamma > 0$. The modified function is differentiable and differs from the original function by at most $\frac{\gamma}{2}$. By choosing γ much smaller than $\frac{LD}{2}\min\left\{\sqrt{\ln_+(1/\delta)}/\sqrt{N}, 1\right\}$, we obtain that optimizing the modified function is almost equivalent to optimizing the original function. Thus, the lower bound also holds for differentiable functions. ∎

Appendix E. Well-known results

This section collects several well-known probabilistic and optimization results used throughout the paper.

E.1. Chebyshev's inequalities

Theorem 12 (Chebyshev's inequality) *Let X be a random variable with finite variance σ^2. For any $k > 0$, we have*

$$\mathbb{P}[|X - \mathbb{E}[X]| > k\sigma] \leq \frac{1}{k^2}.$$

Theorem 13 (Exponential Chebyshev's inequality) *Let X be a random variable. For any $t, \varepsilon > 0$ we have*

$$\mathbb{P}[X \geq \varepsilon] \leq e^{-t\varepsilon}\mathbb{E}\left[e^{tX}\right].$$

E.2. Concentration bounds

Define the sample variance as

$$V_n(\boldsymbol{X}) \coloneqq \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \frac{1}{n} \sum_{j=1}^{n} X_j \right)^2 .$$

Theorem 14 (Theorem 4 of Maurer and Pontil (2009)) *Let $c > 0$, $n \geq 2$, and $X, X_1, \ldots, X_n$ be i.i.d. random variables with values in $[-c, c]$. Define $\boldsymbol{X} = (X_1, \ldots, X_n)$. Then, for any $\delta > 0$, with probability of at least $1 - \delta$ we have*

$$\mathbb{E}[X] - \frac{1}{n} \sum_{i=1}^{n} X_i \leq \sqrt{\frac{2V_n(\boldsymbol{X}) \ln(2/\delta)}{n}} + c \frac{14 \ln(2/\delta)}{3(n-1)}.$$

Note that by a union bound theorem 14 immediately implies that for any $\delta > 0$, with probability of at least $1 - \delta$ we have

$$\left| \mathbb{E}[X] - \frac{1}{n} \sum_{i=1}^{n} X_i \right| \leq \sqrt{\frac{2V_n(\boldsymbol{X}) \ln(4/\delta)}{n}} + c \frac{14 \ln(4/\delta)}{3(n-1)}.$$

Theorem 15 (Theorem 10 of Maurer and Pontil (2009)) *Let $c > 0$, $n \geq 2$, and $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables with values in $[-c, c]$. Then, for any $\delta > 0$, we have*

$$\mathbb{P}\left[\sqrt{\mathbb{E}_{\boldsymbol{X}}[V_n(\boldsymbol{X})]} > \sqrt{V_n(\boldsymbol{X})} + c \sqrt{\frac{8 \ln(1/\delta)}{n-1}} \right] \leq \delta, \quad \textit{and}$$
$$\mathbb{P}\left[\sqrt{V_n(\boldsymbol{X})} > \sqrt{\mathbb{E}_{\boldsymbol{X}}[V_n(\boldsymbol{X})]} + c \sqrt{\frac{8 \ln(1/\delta)}{n-1}} \right] \leq \delta.$$

Definition 16 (From Pinelis (1994)) *A function $\Psi : \mathcal{X} \mapsto \mathbb{R}$ is called $(2, D)$-smooth for some $D > 0$ if, for all $x, v \in \mathcal{X}$, we have*

$$\Psi(0) = 0,$$
$$|\Psi(x+v) - \Psi(x)| \leq \|v\|, \qquad \textit{and}$$
$$\Psi^2(x+v) - 2\Psi^2(x) + \Psi^2(x-v) \leq 2D^2 \|v\|^2.$$

In particular, for the Euclidean norm, the function $\Psi(x) = \|x\|_2$ is $(2, 1)$-smooth.

Theorem 17 (Corollary 10a of Howard et al. (2020)) *Consider a martingale $(Y_t)_{t \in \mathbb{N}}$ taking values in a separable Banach space $(\mathcal{X}, \|\cdot\|)$. Let the function $\Psi : \mathcal{X} \mapsto \mathbb{R}$ be $(2, D)$-smooth, and define $D_\star \coloneqq 1 \vee D$. Suppose $\|Y_t - Y_{t-1}\| \leq c_t$ a.s. for all $t \in \mathbb{N}$ for some constants $(c_t)_{t \in \mathbb{N}}$, and let $V_t \coloneqq \sum_{i=1}^{t} c_i^2$. Then, for any $x, m > 0$, we have*

$$\mathbb{P}\Big[\exists t \in \mathbb{N} : \Psi(Y_t) \geq x + \frac{x}{2m} \cdot (V_t - m) \Big] \leq 2 \exp\left(-\frac{x^2}{2D_\star^2 m} \right).$$

E.3. Bounds on regularized solutions

Lemma 18 (Lemma 2 of Lawrence et al. (2025)) *For any $\theta \geq 0$, define*

$$\hat{x}^\star_\theta \in \arg\min_{x \in \mathcal{X}} \Big\{ \hat{F}(x) + \theta \|x - x_0\|_\alpha \Big\}.$$

If Assumption 7 holds, then

$$\mathbb{P}\Big(\|\hat{x}^\star_{3\lambda} - x_0\|_\alpha \leq 3\|\hat{x}_\lambda - x_0\|_\alpha \leq 33\|\hat{x}^\star_{\lambda/3} - x_0\|_\alpha \Big) \geq 1 - 2\delta,$$

where $\lambda = 9 \cdot 66\phi \frac{\hat{L}\sqrt{\ln \frac{\psi_\alpha}{\delta}}}{\sqrt{N}}$.

Appendix F. Proofs of lemmas

F.1. Bounding distances between outputs of a Lipschitz function

We now bound $\|m(x; s) - m(y; s)\|_\beta$ under the assumption $m \in \mathcal{I}_{\text{Lip}}^{L_\star, \|\cdot\|_{\alpha,\beta}}$.

Lemma 19 *If $m \in \mathcal{I}_{Lip}^{L_\star, \|\cdot\|_{\alpha,\beta}}$, then for all $x, y \in \mathcal{X}$ and $s \in \mathbb{S}$ we have*

$$\|m(x; s) - m(y; s)\|_\beta \leq L_\star \|x - y\|_\alpha.$$

Proof For all $x, y \in \mathcal{X}$ define the line $\theta_{x,y}(a) := (1 - a) \cdot y + a \cdot x$. We conclude that for all $x, y \in \mathcal{X}$ and $s \in \mathbb{S}$ we have

$$\begin{aligned}
\|m(x; s) - m(y; s)\|_\beta &= \|\int_y^x \nabla_x m(x; s) dx\|_\beta \\
&= \|\int_0^1 \nabla_x m(\theta_{x,y}(a); s) \nabla_a \theta_{x,y}(a) da\|_\beta \\
&\overset{(1)}{\leq} \int_0^1 \|\nabla_x m(\theta_{x,y}(a); s) \nabla_a \theta_{x,y}(a)\|_\beta da \\
&\leq \int_0^1 \|\nabla_x m(\theta_{x,y}(a); s)\|_{\alpha,\beta} \|\nabla_a \theta_{x,y}(a)\|_\alpha da \\
&= \|x - y\|_\alpha \int_0^1 \|\nabla_x m(\theta_{x,y}(a); s)\|_{\alpha,\beta} da,
\end{aligned}$$

where (1) follows from the triangle inequality for integrals. Therefore, since $m \in \mathcal{I}_{\text{Lip}}^{L_\star, \|\cdot\|_{\alpha,\beta}}$, we conclude that for all $x, y \in \mathcal{X}$ and $s \in \mathbb{S}$ we have

$$\|m(x; s) - m(y; s)\|_\beta \leq L_\star \|x - y\|_\alpha.$$

■

F.2. Bounding distances between outputs of a second-moment-Lipschitz function

Analogously to Section F.1, we derive a similar result for the second-moment Lipschitz case.

Lemma 20 *If $m \in \mathcal{I}_{\text{SM-Lip}}^{L_\star, \|\cdot\|_{\alpha,\beta}}$ then for all $x, y \in \mathcal{X}$ we have*

$$\mathbb{E}_{s\sim\mathcal{P}}\left[\|m(x;s) - m(y;s)\|_\beta^2\right] \le L_\star^2 \|x-y\|_\alpha^2.$$

Proof For all $x, y \in \mathcal{X}$ define the line $\theta_{x,y}(a) \coloneqq (1-a)\cdot y + a \cdot x$. Thus, for all $x, y \in \mathcal{X}$ we have

$$\begin{aligned}
\mathbb{E}_{s\sim\mathcal{P}}\left[\|m(x;s) - m(y;s)\|_\beta^2\right] &= \mathbb{E}_{s\sim\mathcal{P}}\left[\|\int_y^x \nabla_x m(x;s)dx\|_\beta^2\right] \\
&= \mathbb{E}_{s\sim\mathcal{P}}\left[\|\int_0^1 \nabla_x m(\theta_{x,y}(a);s)\nabla_a \theta_{x,y}(a)da\|_\beta^2\right] \\
&\overset{(1)}{\le} \mathbb{E}_{s\sim\mathcal{P}}\left[\left(\int_0^1 \|\nabla_x m(\theta_{x,y}(a);s)\nabla_a \theta_{x,y}(a)\|_\beta da\right)^2\right] \\
&\le \mathbb{E}_{s\sim\mathcal{P}}\left[\left(\int_0^1 \|\nabla_x m(\theta_{x,y}(a);s)\|_{\alpha,\beta}\|\nabla_a \theta_{x,y}(a)\|_\alpha da\right)^2\right] \\
&\overset{(2)}{\le} \mathbb{E}_{s\sim\mathcal{P}}\left[\int_0^1 \|\nabla_a \theta_{x,y}(a)\|_\alpha^2 da \int_0^1 \|\nabla_x m(\theta_{x,y}(a);s)\|_{\alpha,\beta}^2 da\right] \\
&= \|x-y\|_\alpha^2 \int_0^1 \mathbb{E}_{s\sim\mathcal{P}}\left[\|\nabla_x m(\theta_{x,y}(a);s)\|_{\alpha,\beta}^2\right] da,
\end{aligned}$$

where (1) follows from the triangle inequality for integrals, and (2) follows from the Cauchy–Schwarz inequality. Therefore, because $m \in \mathcal{I}_{\text{SM-Lip}}^{L_\star, \|\cdot\|_{\alpha,\beta}}$, for all $x, y \in \mathcal{X}$ we have

$$\mathbb{E}_{s\sim\mathcal{P}}\left[\|m(x;s) - m(y;s)\|_\beta^2\right] \le L_\star^2 \|x-y\|_\alpha^2.$$

■

F.3. The risk term τ fulfills the requirement of Lemma 1

Lemma 2 *Let $N \ge 2$ be the number of validation samples. For some $n \in \mathbb{N}$, let $\{(x_{g_i}, c_i)\}_{i\in[n]}$ be a set of candidate pairs such that, for all $i \in [n]$ we have $x_{g_i} \in \{x_g\}_{g\in\mathcal{G}}$ and $c_i \in \left\{2^k \underline{\ell}\|x_{g_i} - x_0\|_\alpha \mid k \in \mathbb{N}\right\}$. For each $i \in [n]$, define $\tau_i \coloneqq \tau(x_{g_i}, c_i)$. The requirement of Lemma 1 holds for $\hat{\delta} = \delta/2$, i.e., with probability $1-\hat{\delta}$ for all $i \in [n]$ we have $\left|F_{c_i}(x_{g_i}) - F(x_0) - \frac{1}{N}\sum_{j=1}^N (f_{c_i}(x_{g_i};s_j) - f(x_0;s_j))\right| \le \tau_i$.*

Proof For any $k \in \mathbb{N}$ and $g \in \mathcal{G}$, define

$$\begin{aligned}
\bar{c}_{g,k} &\coloneqq 2^k \underline{\ell} \|x_g - x_0\|_\alpha, \\
\bar{u}_{g,k} &\coloneqq \frac{1}{N} \sum_{j=1}^{N} \big(f_{\bar{c}_{g,k}}(x_g; s_j) - f(x_0; s_j)\big), \\
\bar{v}_{g,k} &\coloneqq \frac{1}{N-1} \sum_{j=1}^{N} \Big(\big(f_{\bar{c}_{g,k}}(x_g; s_j) - f(x_0; s_j)\big) - \bar{u}_{g,k}\Big)^2, \\
\tilde{a}_k &\coloneqq (k+1)^2 |\mathcal{G}|, \quad \text{and} \\
\bar{\tau}_{g,k} &\coloneqq \sqrt{\frac{2\bar{v}_{g,k} \ln(16\tilde{a}_k/\delta)}{N}} + \bar{c}_{g,k} \frac{14 \ln(16\tilde{a}_k/\delta)}{3(N-1)}.
\end{aligned}$$

Note that for every $g \in \mathcal{G}$ such that $x_g \neq x_0$,

$$\tilde{a}_k = |\mathcal{G}| \log_2^2 \left(\frac{2^{k+1} \underline{\ell} \|x_g - x_0\|_\alpha}{\underline{\ell} \|x_g - x_0\|_\alpha} \right).$$

Additionally, for every $k \in \mathbb{N}$ and $g \in \mathcal{G}$ such that $x_g = x_0$, we have $\bar{\tau}_{g,k} = 0$.

For any $g \in \mathcal{G}$, $k \in \mathbb{N}$, and any sample $s \in \mathbb{S}$,

$$|f_{\bar{c}_{g,k}}(x_g; s) - f(x_0; s)| \leq \|m_{\bar{c}_{g,k}}(x_g; s) - m(x_0; s)\|_\beta \leq \bar{c}_{g,k},$$

where the first inequality is because the loss function is 1-Lipschitz, and the second inequality follows from the definition of the clipped model. As a result, by applying Theorem 4 of Maurer and Pontil (2009) (see Theorem 14), we obtain that for any $g \in \mathcal{G}$ and $k \in \mathbb{N}$, with probability at least $1 - \delta/\big(4k^2|\mathcal{G}|\big)$

$$\begin{aligned}
&\left| F_{\bar{c}_{g,k}}(x_g) - F(x_0) - \frac{1}{N} \sum_{j=1}^{N} \big(f_{\bar{c}_{g,k}}(x_g; s_j) - f(x_0; s_j)\big) \right| \\
&\qquad \leq \sqrt{\frac{2\bar{v}_{g,k} \ln(16\tilde{a}_k/\delta)}{N}} + \bar{c}_{g,k} \frac{14 \ln(16\tilde{a}_k/\delta)}{3(N-1)} = \bar{\tau}_{g,k}.
\end{aligned}$$

Thus, by using the union bound and since $\sum_{k \in \mathbb{N}} 1/(k+1)^2 \leq 2$, with probability at least $1 - \delta/2$, for any $g \in \mathcal{G}$ and $k \in \mathbb{N}$,

$$\left| F_{\bar{c}_{g,k}}(x_g) - F(x_0) - \frac{1}{N} \sum_{j=1}^{N} \big(f_{\bar{c}_{g,k}}(x_g; s_j) - f(x_0; s_j)\big) \right| \leq \bar{\tau}_{g,k},$$

Finally, since for all $i \in [n]$ we have that $g_i \in \mathcal{G}$ and $c_i \in \{\bar{c}_{g_i,k} \mid k \in \mathbb{N}\}$, the same bound holds for every c_i with probability at least $1 - \delta/2$:

$$\left| F_{c_i}(x_{g_i}) - F(x_0) - \frac{1}{N} \sum_{j=1}^{N} (f_{c_i}(x_{g_i}; s_j) - f(x_0; s_j)) \right| \leq \tau_i,$$

i.e. the requirement of Lemma 1 holds for $\delta/2$. ■

F.4. The sample maximum exceeds a high quantile

Lemma 21 *Let $n \geq 1$ and $\delta \in \left(0, \frac{1}{e}\right)$ such that $\ln(1/\delta) < n$. Define $p = 1 - \frac{\ln(1/\delta)}{n}$. For any $z \geq n$ and i.i.d. random variables $X_1, \dots, X_z \sim \mathcal{P}$, we have*

$$\mathbb{P}[\max(X_1, \dots, X_z) < Q_p(\mathcal{P})] < \delta.$$

Proof Let $p = 1 - \frac{\ln(1/\delta)}{n}$. For $\max(X_1, \dots, X_z) < Q_p(\mathcal{P})$ to hold, every $X_1, \dots, X_z$ must be less than $Q_p(\mathcal{P})$. The probability of that is less than p^z. A well-known inequality states that for any $k > 0$ we have that $\ln(k+1) - \ln(k) = \int_k^{k+1} \frac{1}{t}\, dt \geq \frac{1}{k+1}$. Thus,

$$\begin{aligned}
\ln\left(\frac{n}{n - \ln(1/\delta)}\right) &= (\ln(n) - \ln(n-1)) + \cdots + (\ln(\lceil n - \ln(1/\delta) \rceil) - \ln(n - \ln(1/\delta))) \\
&\geq \frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{\lceil n - \ln(1/\delta) \rceil} \\
&\geq \frac{\ln(1/\delta)}{n}.
\end{aligned}$$

Therefore, since $z \geq n$,

$$z \geq \ln(1/\delta) / \ln\left(\frac{n}{n - \ln(1/\delta)}\right).$$

As a consequence,

$$z \ln\left(\frac{n}{n - \ln(1/\delta)}\right) \geq \ln(1/\delta).$$

Thus,

$$\left(1 - \frac{\ln(1/\delta)}{n}\right)^z \leq \delta,$$

Finally, this means that

$$\mathbb{P}[\max(X_1, \dots, X_z) < Q_p(\mathcal{P})] < p^z \leq \delta.$$

■

F.5. How many samples are needed to obtain N good samples

Lemma 22 *Let $a, M, k, N \geq 0$ such that $N - a \cdot e > 0$, and for $X \sim \mathcal{D}$ we have $\mathbb{P}[X > M] \leq \frac{a}{N}$. If we draw at least*

$$k \cdot N + \max\left(1, \frac{\ln(1/\delta) + a \cdot e \cdot k}{N - a \cdot e} \cdot N\right)$$

random variables i.i.d. from $\mathcal{D}$, then with probability at least $1 - \delta$, at least $k \cdot N$ of the drawn variables are at most M.

Proof We define the random variable Z_i as

$$Z_i = \begin{cases} 1 & \text{if } X_i > M \\ 0 & \text{otherwise} \end{cases}.$$

We have

$$\mathbb{P}[Z_i = 1] = \mathbb{P}[X_i > M] \leq \frac{a}{N}. \tag{9}$$

Let $r \geq 1$ be the number of additional samples we draw in addition to the $k \cdot N$ we want. Using the exponential Chebyshev's inequality (see Theorem 13), we obtain, for all $t > 0$,

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^{k\cdot N+r} Z_i \geq r\right] &\leq e^{-tr}\mathbb{E}\Big[e^{t\sum_{i=1}^{k\cdot N+r} Z_i}\Big] \\
&\overset{(1)}{=} e^{-tr} \prod_{i=1}^{k\cdot N+r} \mathbb{E}\big[e^{tZ_i}\big] \\
&\leq e^{-tr} \prod_{i=1}^{k\cdot N+r} \big(1 \cdot e^0 + \mathbb{P}[Z_i = 1]e^t\big) \\
&\overset{(2)}{\leq} e^{-tr}\left(1 + \frac{a \cdot e^t}{N}\right)^{k\cdot N+r} \\
&\leq e^{-tr+a\cdot e^t(k\cdot N+r)/N},
\end{aligned}$$

where (1) is because Z_i are independent of each other, and (2) is by Equation (9). Setting $t = 1$ yields

$$\mathbb{P}\left[\sum_{i=1}^{k\cdot N+r} Z_i \geq r\right] \leq e^{-r+a\cdot e(k+r/N)}.$$

We obtain

$$e^{-r+a\cdot e(k+r/N)} \leq \delta$$

if and only if

$$r(1 - a \cdot e/N) \geq \ln(1/\delta) + a \cdot e \cdot k.$$

Therefore, if $a < N/e$ and

$$r \geq \max\left(1, \frac{\ln(1/\delta) + a \cdot e \cdot k}{N - a \cdot e} \cdot N\right),$$

we obtain

$$\mathbb{P}\left[\sum_{i=1}^{k\cdot N+r} Z_i \geq r\right] \leq \delta.$$

This implies that, with probability at least $1 - \delta$, fewer than r samples exceed M, and hence at least $k \cdot N$ samples are at most M. ■

F.6. Empirical mean approximates the expectation

Lemma 23 *Let $\|\cdot\|$ be a $(2, D)$-smooth (Definition 16) norm for some $D \geq 1$. Let $V_1, \ldots, V_n$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$. Define $v := \mathbb{E}[V_i]$. If $\|V_i - v\| \leq C$ almost surely for some constant C, then for all $\delta \in \left(0, \frac{2}{e}\right)$,*

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} V_i - v\right\| \geq \frac{DC\sqrt{2\ln(2/\delta)}}{\sqrt{n}}\right] \leq \delta.$$

Proof For any $k \in \mathbb{N}$ define $Y_k := \frac{1}{n}\sum_{i=1}^{k}(V_i - v)$. Denote $Y_0 = 0$ and $\Psi(\cdot) := \|\cdot\|$. We have $\|Y_i - Y_{i-1}\| \leq C/n$ for all $i \in \mathbb{N}$, and $\{Y_i\}_{i\in\mathbb{N}}$ is a martingale. Thus, from Corollary 10a of Howard et al. (2020) (see Theorem 17), for all $x, z > 0$,

$$\mathbb{P}\left[\exists t \in \mathbb{N} : \Psi(Y_t) \geq x + \frac{D^2 x}{2z} \cdot \left(\frac{tC^2}{n^2} - z\right)\right] \leq 2\exp\left(-\frac{x^2}{2D^2 z}\right).$$

Therefore, for $x = DC\sqrt{2\ln(2/\delta)}/\sqrt{n}$ and $z = C^2/n$, we have

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} V_i - v\right\| \geq \frac{DC\sqrt{2\ln(2/\delta)}}{\sqrt{n}}\right] \leq 2\exp(-\ln(2/\delta)) = \delta.$$

■

Appendix G. The mean of stochastic gradients approximates the true gradient

Lemma 24 *Let $d \in \mathbb{N}$ and let $\mathcal{X} \subseteq \mathbb{R}^d$.*

1. *In the case that $\|\cdot\|_\alpha = \|\cdot\|_2$, Assumption 7 holds for $\psi_\alpha = 1$ and any $\phi \geq 1$.*
2. *In the case that $\|\cdot\|_\alpha = \|\cdot\|_1$, Assumption 7 holds for $\psi_\alpha = d$ and any $\phi \geq 1$.*

We first prove Lemma 24 in the case that $\|\cdot\|_\alpha = \|\cdot\|_2$.

Proof In the case that $\|\cdot\|_\alpha = \|\cdot\|_2$ then we also have that $\|\cdot\|_{\alpha*} = \|\cdot\|_2$. We note that $\|\cdot\|_2$ is $(2, 1)$-smooth. Therefore, as $f(\cdot, s)$ is $\hat{L}$-Lipschitz for every s sampled from $\tilde{\mathcal{P}}$, a direct result of Lemma 23 is that

$$\mathbb{P}_{s_1,\ldots,s_N \overset{\text{iid}}{\sim} \tilde{\mathcal{P}}}\left[\left\|\nabla \mathbb{E}_{s\sim\tilde{\mathcal{P}}} f(x; s) - \frac{1}{N}\sum_{i=1}^{N} \nabla f(x; s_i)\right\|_2 \geq \frac{\hat{L}\sqrt{8\ln(2/\delta)}}{\sqrt{N}}\right] \leq \delta.$$

Thus, Assumption 7 holds for $\psi_\alpha = 1$ and any $\phi \geq 1$. ■

We now prove Lemma 24 in the case that $\|\cdot\|_\alpha = \|\cdot\|_1$.

Proof In the case that $\|\cdot\|_\alpha = \|\cdot\|_1$ then we also have that $\|\cdot\|_{\alpha*} = \|\cdot\|_\infty$. Therefore, following the result of the case that $\|\cdot\|_\alpha = \|\cdot\|_2$, for every coordinate $j \in [d]$,

$$\mathbb{P}_{s_1,\ldots,s_N \overset{\text{iid}}{\sim} \tilde{\mathcal{P}}}\left[\left|\left[\nabla \mathbb{E}_{s\sim\tilde{\mathcal{P}}} f(x; s) - \frac{1}{N}\sum_{i=1}^{N} \nabla f(x; s_i)\right]_j\right| \geq \frac{\hat{L}\sqrt{8\ln(2d/\delta)}}{\sqrt{N}}\right] \leq \frac{\delta}{d}.$$

Thus, by a union bound we obtain that

$$\mathop{\mathbb{P}}_{s_1,\ldots,s_N \overset{\text{iid}}{\sim} \tilde{\mathcal{P}}}\left[\left\|\nabla \mathbb{E}_{s\sim\tilde{\mathcal{P}}} f(x;s) - \frac{1}{N}\sum_{i=1}^{N} \nabla f(x;s_i)\right\|_\infty \geq \frac{\hat{L}\sqrt{8\ln(2d/\delta)}}{\sqrt{N}}\right] \leq \delta.$$

As a consequence, Assumption 7 holds for $\psi_\alpha = d$ and any $\phi \geq 1$. ■

Appendix H. Proof of Proposition 3 for the case that m is second-moment-Lipschitz

The following theorem is a slight generalization of Proposition 3; the choice $\hat{a} = 1$ recovers Proposition 3.

Proposition 25 *Let $m \in \mathcal{I}^{\infty,\|\cdot\|_{\alpha,\beta}}_{\text{SM-Lip}}$ and $h \in \mathcal{I}^{1,\|\cdot\|_{\beta*}}_{\text{Lip}}$. Let $\hat{a} \in \left[1, |\mathcal{G}|\log_2^2\left(2\sqrt{N}\cdot \overline{\ell}/\underline{\ell}\right)\right]$, where $\overline{\ell}$ is an upper bound on the Lipschitz constant. Let the number of samples be $N \geq 4\ln(4\hat{a}/\delta)$. If x, c are the result of Algorithm 1, where we use $n = \lceil 4\ln(4\hat{a}/\delta)\rceil$, then with probability at least $1-\delta$ we have*

$$F_c(x) \leq \min_{g\in\mathcal{G}\cup\{0\}}\left[F(x_g) + O\left(\overline{L}_\star\|x_g - x_0\|_\alpha\sqrt{\frac{\ln_+(\tilde{a}/\delta)}{N}}\sqrt{\frac{\ln_+(\tilde{a}/\delta)}{\ln_+(\hat{a}/\delta)}}\right)\right],$$

where $\tilde{a} = \max\left(\hat{a}, |\mathcal{G}|\ln_+\left(\sqrt{N}L_\star/\underline{\ell}\right)\right)$.

Proof Let $\hat{a} \in \left[1, |\mathcal{G}|\log_2^2\left(2\sqrt{N}\cdot\overline{\ell}/\underline{\ell}\right)\right]$. Let n in Algorithm 1 be $n = 4\ln(4\hat{a}/\delta)$. Finally, define $\tilde{a} = \max\left(\hat{a}, |\mathcal{G}|\log_2^2\left(2\sqrt{N}\cdot\overline{L}_\star/\underline{\ell}\right)\right)$.

First, as $m \in \mathcal{I}^{L_\star,\|\cdot\|_{\alpha,\beta}}_{\text{SM-Lip}}$, Lemma 20 states that for all $g \in \mathcal{G}$,

$$\mathbb{E}_{s\sim\mathcal{P}}\left[\|m(x_g;s) - m(x_0;s)\|_\beta^2\right] \leq L_\star^2\|x_g - x_0\|_\alpha^2. \tag{10}$$

Let

$$g_\star = \mathop{\arg\min}_{g\in\mathcal{G}\cup\{0\}} F(x_g) + 94\overline{L}_\star\|x_g - x_0\|_\alpha\sqrt{\frac{\ln(16\tilde{a}/\delta)}{N}}\sqrt{\frac{\ln(16\tilde{a}/\delta)}{\ln(4\hat{a}/\delta)}}. \tag{11}$$

Define $c_{g_\star,0} :=$ binary-ceil$\left(\sqrt{\frac{N}{\ln(4\hat{a}/\delta)}}L_\star\|x_{g_\star} - x_0\|_\alpha; \underline{\ell}\|x_{g_\star} - x_0\|_\alpha\right)$. For candidate $(x_{g_\star}, c_{g_\star,0})$ we define $\tau_{g_\star,0}$ later in the proof. In the proof we use $\{(x_g, c_{g,i})\}_{(g,i)\in\mathcal{G}\times[n]} \cup \{(x_{g_\star}, c_{g_\star,0})\}$ as the set of candidates passed to Algorithm 3. We show that, with high probability, we can assume that Algorithm 3 returns the same output under both the original set of candidates and the modified set of candidates.

We define the random variable Z_i as

$$Z_i = \begin{cases} 1 & \text{if binary-ceil}(\|m(x_{g_\star};s_i) - m(x_0;s_i)\|_\beta; \underline{\ell}\|x_{g_\star} - x_0\|_\alpha) > c_{g_\star,0} \\ 0 & \text{otherwise} \end{cases}.$$

We now find an upper bound on $\mathbb{P}[Z_i = 1]$. First, we bound the variance $\mathrm{Var}_{s\sim\mathcal{P}}[\|m(x_{g_\star}; s) - m(x_0; s)\|_\beta]$:

$$\mathrm{Var}_{s\sim\mathcal{P}}[\|m(x_{g_\star}; s) - m(x_0; s)\|_\beta] \le \mathbb{E}_{s\sim\mathcal{P}}\left[\|m(x_{g_\star}; s) - m(x_0; s)\|_\beta^2\right] \le (L_\star \|x_{g_\star} - x_0\|_\alpha)^2,$$

where the last inequality is the result of Equation (10). Therefore,

$$\begin{aligned}\mathbb{P}[Z_i = 1] &= \mathbb{P}[\text{binary-ceil}(\|m(x_{g_\star}; s_i) - m(x_0; s_i)\|_\beta; \underline{\ell}\|x_{g_\star} - x_0\|_\alpha) > c_{g_\star,0}] \\ &\le \mathbb{P}\left[\|m(x_{g_\star}; s_i) - m(x_0; s_i)\|_\beta > \sqrt{\frac{N}{\ln(4\hat{a}/\delta)}} L_\star \|x_{g_\star} - x_0\|_\alpha\right] \\ &\le \frac{\ln(4\hat{a}/\delta)}{N}, \end{aligned} \tag{12}$$

where the last inequality is a result of Chebyshev's inequality (see Theorem 12).

As a result, using the exponential Chebyshev's inequality (see Theorem 13), we obtain, for all $t > 0$ and $r > 0$,

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^N Z_i \ge r\right] &\le e^{-tr}\mathbb{E}\left[e^{t\sum_{i=1}^N Z_i}\right] \\ &\overset{(1)}{=} e^{-tr}\prod_{i=1}^N \mathbb{E}\left[e^{tZ_i}\right] \\ &\le e^{-tr}\prod_{i=1}^N \left(1\cdot e^0 + \mathbb{P}[Z_i = 1]e^t\right) \\ &\overset{(2)}{\le} e^{-tr}\left(1 + \frac{e^t \cdot \ln(4\hat{a}/\delta)}{N}\right)^N \\ &\le e^{-tr + e^t\cdot\ln(4\hat{a}/\delta)}, \end{aligned}$$

where (1) is because Z_i are independent of each other, and (2) is by Equation (12). Thus, by setting $t = 1$ and $r = 4\ln(4\hat{a}/\delta)$ we obtain that

$$\mathbb{P}\left[\sum_{i=1}^N Z_i \ge 4\ln(4\hat{a}/\delta)\right] \le e^{-\ln(4\hat{a}/\delta)} \le \frac{\delta}{4\hat{a}} \le \frac{\delta}{4}.$$

This implies that with probability at least $1 - \frac{\delta}{4}$ there are fewer than $4\ln(4\hat{a}/\delta)$ samples s_i for which $m(x_{g_\star}; s_i)$ is clipped when $c_{g_\star,0}$ is used as the clipping threshold. As a result, with probability at least $1 - \frac{\delta}{4}$, there exists $i \in [n]$ for which $c_{g_\star,0} \ge c_{g_\star,i}$. Let $i_\star = \arg\max_{i\in[n]; c_{g_\star,i} \le c_{g_\star,0}} c_{g_\star,i}$. The validation loss for both clipping parameters $c_{g_\star,0}$ and $c_{g_\star,i}$ is very close:

$$\begin{aligned}\left|\frac{1}{N}\sum_{j=1}^N f_{c_{g_\star,0}}(x_{g_\star}; s_j) - \frac{1}{N}\sum_{j=1}^N f_{c_{g_\star,i_\star}}(x_{g_\star}; s_j)\right| &\le \frac{1}{N}\sum_{j=1}^N \|m_{c_{g_\star,0}}(x_{g_\star}; s_j) - m_{c_{g_\star,i_\star}}(x_{g_\star}; s_j)\|_\beta \\ &\le 8c_{g_\star,0}\frac{\ln(4\hat{a}/\delta)}{N}, \end{aligned}$$

where the first inequality is because $h(\cdot;s)$ is 1-Lipschitz. Thus, from the definition of $c_{g_\star,0}$, we get

$$\left|\frac{1}{N}\sum_{j=1}^{N} f_{c_{g_\star,0}}(x_{g_\star};s_j) - \frac{1}{N}\sum_{j=1}^{N} f_{c_{g_\star,i_\star}}(x_{g_\star};s_j)\right| \le 16\overline{L}_\star \|x_{g_\star} - x_0\|_\alpha \sqrt{\frac{\ln(4\hat{a}/\delta)}{N}}. \tag{13}$$

For candidate $(x_{g_\star}, c_{g_\star,0})$, we define $\tau_{g_\star,0}$ as follows:

$$\tau_{g_\star,0} \coloneqq \max\left(\tau(x_{g_\star}, c_{g_\star,0}), \tau_{g_\star,i_\star} + 16\overline{L}_\star \|x_{g_\star} - x_0\|_\alpha \sqrt{\frac{\ln(4\hat{a}/\delta)}{N}}\right).$$

Therefore, from Equation (13),

$$\frac{1}{N}\sum_{j=1}^{N} f_{c_{g_\star,i_\star}}(x_{g_\star};s_j) + \tau_{g_\star,i_\star} \le \frac{1}{N}\sum_{j=1}^{N} f_{c_{g_\star,0}}(x_{g_\star};s_j) + \tau_{g_\star,0}.$$

As a consequence, (i) the θ calculated in Algorithm 3 is not affected by the candidate $(x_{g_\star}, c_{g_\star,0})$, and (ii) for $\mathcal{F}$ calculated by Algorithm 3, if $(x_{g_\star}, c_{g_\star,0}) \in \mathcal{F}$ then also $(x_{g_\star}, c_{g_\star,i_\star}) \in \mathcal{F}$. Therefore, by using Lemma 9 instead of Lemma 1, we can assume that if Algorithm 3 should have returned the candidate $(x_{g_\star}, c_{g_\star,0})$ it instead returned the candidate $(x_{g_\star}, c_{g_\star,i_\star})$. This means that with probability at least $1 - \frac{\delta}{4}$ Algorithm 3 is not affected by the candidate $(x_{g_\star}, c_{g_\star,0})$, i.e. we can assume that Algorithm 3 returns the same output for both the original set of candidates and the modified set of candidates.

We now move to show that $(x_{g_\star}, c_{g_\star,0})$ is a good candidate. For $g_\star$ and $c_{g_\star,0}$, we cannot show that $F_{c_{g_\star,0}}(x_{g_\star}) = F(x_{g_\star})$. Therefore, we bound $|F_{c_{g_\star,0}}(x_{g_\star}) - F(x_{g_\star})|$ instead. For any $c \ge 0$, $x \in \mathcal{X}$ and $s \in \mathbb{S}$ such that $\|m(x;s) - m(x_0;s)\|_\beta > c$, we have

$$\begin{aligned}
\|m_c(x;s) - m(x;s)\|_\beta &\le \|m(x;s) - m(x_0;s)\|_\beta + \|m_c(x;s) - m(x_0;s)\|_\beta \\
&\le \|m(x;s) - m(x_0;s)\|_\beta + c \\
&\le 2\|m(x;s) - m(x_0;s)\|_\beta.
\end{aligned}$$

As a consequence, if $c_{g_\star,0} > 0$ then

$$\begin{aligned}
|F_{c_{g_\star,0}}(x_{g_\star}) - F(x_{g_\star})| &= \mathbb{E}_{s\sim\mathcal{P}}\big[|f_{c_{g_\star,0}}(x_{g_\star};s) - f(x_{g_\star};s)|\big] \\
&\le \mathbb{E}_{s\sim\mathcal{P}}\big[\|m_{c_{g_\star,0}}(x_{g_\star};s) - m(x_{g_\star};s)\|_\beta\big] \\
&= \mathbb{E}_{s\sim\mathcal{P}}\Big[\|m_{c_{g_\star,0}}(x_{g_\star};s) - m(x_{g_\star};s)\|_\beta \mathbb{1}_{\left\{\|m(x_{g_\star};s)-m(x_0;s)\|_\beta > c_{g_\star,0}\right\}}\Big] \\
&\le 2\mathbb{E}_{s\sim\mathcal{P}}\Big[\|m(x_{g_\star};s) - m(x_0;s)\|_\beta \mathbb{1}_{\left\{\|m(x_{g_\star};s)-m(x_0;s)\|_\beta > c_{g_\star,0}\right\}}\Big] \\
&\le \frac{2}{c_{g_\star,0}}\mathbb{E}_{s\sim\mathcal{P}}\big[\|m(x_{g_\star};s) - m(x_0;s)\|_\beta^2\big],
\end{aligned} \tag{14}$$

where the first inequality is because $h(\cdot;s)$ is 1-Lipschitz. Additionally, if $c_{g_\star,0} = 0$ then $x_{g_\star} = x_0$, and thus

$$|F_{c_{g_\star,0}}(x_{g_\star}) - F(x_{g_\star})| = |F_{c_{g_\star,0}}(x_0) - F(x_0)| = 0$$

Therefore, by combining Equations (10) and (14) and the previous equation, we get

$$|F_{c_{g_\star,0}}(x_{g_\star}) - F(x_{g_\star})| \le 2L_\star \|x_{g_\star} - x_0\|_\alpha \sqrt{\frac{\ln(4\hat{a}/\delta)}{N}}. \tag{15}$$

Thus, the clipped function is, in expectation, almost equivalent to the non-clipped function.

We now want to bound $\tau_{g_\star,0}$, similarly to what we have done for the Lipschitz case. To do this, we need to bound the sample variance $v(g_\star, c_{g_\star,0})$. We first start by bounding the variance, for any $c \ge 0$

$$\begin{aligned}\mathbb{E}_{s\sim\mathcal{P}}\left[|f_c(x_{g_\star}; s) - f(x_0; s)|^2\right] &\le \mathbb{E}_{s\sim\mathcal{P}}\left[\|m_c(x_{g_\star}; s) - m(x_0; s)\|_\beta^2\right] \\ &\le \mathbb{E}_{s\sim\mathcal{P}}\left[\|m(x_{g_\star}; s) - m(x_0; s)\|_\beta^2\right] \\ &\le L_\star^2 \|x_{g_\star} - x_0\|_\alpha^2,\end{aligned}$$

where the first inequality is because $h(\cdot; s)$ is 1-Lipschitz, and the final inequality is from using Equation (10) again. Now, using Theorem 10 of Maurer and Pontil (2009) (see Theorem 15), with probability at least $1 - \delta/4$, the sample variance is bounded by

$$\sqrt{v(g_\star, c_{g_\star,0})} \le L_\star \|x_{g_\star} - x_0\|_\alpha + c_{g_\star,0}\sqrt{\frac{8\ln(4/\delta)}{N-1}}, \qquad \text{and}$$
$$\sqrt{v(g_\star, c_{g_\star,i_\star})} \le L_\star \|x_{g_\star} - x_0\|_\alpha + c_{g_\star,i_\star}\sqrt{\frac{8\ln(4/\delta)}{N-1}} \le L_\star \|x_{g_\star} - x_0\|_\alpha + c_{g_\star,0}\sqrt{\frac{8\ln(4/\delta)}{N-1}}.$$

Therefore,

$$\begin{aligned}\tau_{g_\star,0} &= \max\left(\tau(x_{g_\star}, c_{g_\star,0}), \tau_{g_\star,i_\star} + 16\overline{L}_\star \|x_{g_\star} - x_0\|_\alpha \sqrt{\frac{\ln(4\hat{a}/\delta)}{N}}\right) \\ &\le (16+2)\overline{L}_\star \|x_{g_\star} - x_0\|_\alpha \sqrt{\frac{\ln(16\tilde{a}/\delta)}{N}} + c_{g_\star,0}\sqrt{\frac{8\ln(4/\delta)}{N-1}}\sqrt{\frac{2\ln(16\tilde{a}/\delta)}{N}} + c_{g_\star,0}\frac{14\ln(16\tilde{a}/\delta)}{3(N-1)} \\ &\le 18\overline{L}_\star \|x_{g_\star} - x_0\|_\alpha \sqrt{\frac{\ln(16\tilde{a}/\delta)}{N}} + 8\overline{L}_\star \|x_{g_\star} - x_0\|_\alpha \sqrt{\frac{\ln(16\tilde{a}/\delta)}{N}}\sqrt{\frac{N}{N-1}} \\ &\quad + 10\overline{L}_\star \|x_{g_\star} - x_0\|_\alpha \sqrt{\frac{\ln(16\tilde{a}/\delta)}{N}}\sqrt{\frac{\ln(16\tilde{a}/\delta)}{\ln(4\hat{a}/\delta)}}\frac{N}{N-1}.\end{aligned}$$

As a result, since $N \ge 2$,

$$\tau_{g_\star,0} \le 54\overline{L}_\star \|x_{g_\star} - x_0\|_\alpha \sqrt{\frac{\ln(16\tilde{a}/\delta)}{N}}\sqrt{\frac{\ln(16\tilde{a}/\delta)}{\ln(4\hat{a}/\delta)}}. \tag{16}$$

Thus, by using Equations (15) and (16),

$$F_{c_{g_\star,0}}(x_{g_\star}) + 2\tau_{g_\star,0} \le F(x_{g_\star}) + 2\cdot 55\overline{L}_\star \|x_{g_\star} - x_0\|_\alpha \sqrt{\frac{\ln(16\tilde{a}/\delta)}{N}}\sqrt{\frac{\ln(16\tilde{a}/\delta)}{\ln(4\hat{a}/\delta)}}. \tag{17}$$

Finally, as Lemma 2 shows that the requirement of Lemma 1 holds for $\delta/2$, Lemmas 1 and 9 guarantee that if x, c is the result of Algorithm 1, then with probability at least $1 - \delta/2$,

$$F_c(x) \le \min_{\substack{g \in \mathcal{G} \cup \{0\} \\ i \in [n] \cup \{0\}}} \left[F_{c_{g,i}}(x_g) + 2\tau_{g,i} \right].$$

Now, because with probability at least $1 - \delta/4$ the additional candidate $(g_\star, c_{g_\star,0})$ does not affect the algorithm, with probability at least $1 - \delta$ we have

$$F_c(x) \le \min_{\substack{g \in \mathcal{G} \cup \{0\} \\ i \in [n]}} \left[F_{c_{g,i}}(x_g) + 2\tau_{g,i} \right].$$

Hence, by the definition of $g_\star$ in Equation (11) and by Equation (17), with probability at least $1 - \delta$,

$$F_c(x) \le \min_{g \in \mathcal{G} \cup \{0\}} \left[F(x_g) + O\left(\overline{L}_\star \|x_g - x_0\|_\alpha \sqrt{\frac{\ln_+(\tilde{a}/\delta)}{N}} \sqrt{\frac{\ln_+(\tilde{a}/\delta)}{\ln_+(\hat{a}/\delta)}} \right) \right].$$

■

Appendix I. Proof of Theorem 8

Theorem 8 *Let* $\delta \in \left(0, \frac{1}{5}\right)$, $m \in \mathcal{I}_{Lip}^{\infty, \|\cdot\|_{\alpha,\beta}}$, $h \in \mathcal{I}_{Lip}^{1, \|\cdot\|_{\beta*}}$ *and* $f \in \hat{\mathcal{I}}_{Dist}^{\infty, \|\cdot\|_\alpha}$. *If Assumptions 6 and 7 hold with* ϕ *and* ψ_α, *and* $x_{perfect}, \hat{L}\|x_{perfect} - x_0\|_\alpha$ *are the outputs of Algorithm 2, then with probability at least* $1 - 5\delta$ *we have*

$$F_{\hat{L}\|x_{perfect} - x_0\|_\alpha}(x_{perfect}) \le F(x_\star) + O\left(\phi \frac{L_\star \|x_\star - x_0\|_\alpha \sqrt{\ln \frac{\psi_\alpha}{\delta}}}{\sqrt{N}} \right).$$

Proof First, we consider the case where $N \le 7$ or $\ln(1/\delta) > \frac{N}{2e}$. In both cases, we have $\ln(\psi_\alpha/\delta) > \frac{N}{2e}$; if $N \le 7$, this follows from $\psi_\alpha \ge 1$ and $\delta < \frac{1}{5}$; otherwise it follows directly from $\psi_\alpha \ge 1$. Therefore,

$$F_0(x_0) = F(x_0) \le F(x_\star) + L_\star \|x_\star - x_0\|_\alpha \le F(x_\star) + O\left(\frac{L_\star \|x_\star - x_0\|_\alpha \sqrt{\ln \frac{\psi_\alpha}{\delta}}}{\sqrt{N}} \right).$$

In these cases, the number of samples used is 0.

Now, we move to prove the theorem for the case that $\ln(\psi_\alpha/\delta) \le \frac{N}{2e}$. Define

$$\hat{F}(x) := \mathbb{E}_{s \sim \mathcal{P}}\left[f(x; s) \mid L(s) \le \hat{L} \right].$$

For any $\theta \ge 0$ define

$$\hat{x}_\theta \in \arg\min_{x \in \mathcal{X}} \left\{ \frac{1}{N} \sum_{i=N+1}^{2N} f(x; s_i) + \theta \|x - x_0\|_\alpha \right\} \text{ and } \hat{x}_\theta^\star \in \arg\min_{x \in \mathcal{X}} \left\{ \hat{F}(x) + \theta \|x - x_0\|_\alpha \right\}.$$

Also, define $\hat{R} \coloneqq 3\|\hat{x}_\lambda - x_0\|_\alpha$. As we assume that Assumption 7 holds, the conditions of Lemma 2 of Lawrence et al. (2025) are satisfied. Thus, from Lemma 2 of Lawrence et al. (2025) (see Lemma 18), with probability at least $1-2\delta$,

$$\|\hat{x}^\star_{3\lambda} - x_0\|_\alpha \le 3\|\hat{x}_\lambda - x_0\|_\alpha \le 33\|\hat{x}^\star_{\lambda/3} - x_0\|_\alpha. \tag{18}$$

If $\|\hat{x}^\star_{\lambda/3} - x_0\|_\alpha \ge \hat{R}$ then, because of Equation (18), there exists $\lambda_{\hat{R}}$ such that $\lambda_{\hat{R}} \in [\lambda/3, 3\lambda]$ and $\|\hat{x}^\star_{\lambda_{\hat{R}}} - x_0\|_\alpha = \hat{R}$ and

$$\min_{x\in\mathcal{X}:\|x-x_0\|_\alpha\le\hat{R}} \hat{F}(x) \le \hat{F}\left(\hat{x}^\star_{\lambda_{\hat{R}}}\right) \le \hat{F}(x_\star) + \lambda_{\hat{R}}\left(\|x_\star - x_0\|_\alpha - \|\hat{x}^\star_{\lambda_{\hat{R}}} - x_0\|_\alpha\right). \tag{19}$$

Otherwise, because of Equation (18), we have that $\|\hat{x}^\star_{\lambda/3} - x_0\|_\alpha \in \left[\frac{\hat{R}}{33}, \hat{R}\right]$, and therefore

$$\min_{x\in\mathcal{X}:\|x-x_0\|_\alpha\le\hat{R}} \hat{F}(x) \le \hat{F}\left(\hat{x}^\star_{\lambda/3}\right) \le \hat{F}(x_\star) + \frac{\lambda}{3}\left(\|x_\star - x_0\|_\alpha - \|\hat{x}^\star_{\lambda/3} - x_0\|_\alpha\right). \tag{20}$$

If $\hat{R} \le 66\|x_\star - x_0\|_\alpha$ then Equations (19) and (20) yields

$$\min_{x\in\mathcal{X}:\|x-x_0\|_\alpha\le\hat{R}} \hat{F}(x) \le \hat{F}(x_\star) + 3\lambda\|x_\star - x_0\|_\alpha$$

Therefore, by Assumption 6, with probability at least $1-3\delta$,

$$\begin{aligned}\hat{F}\left(x_{\text{perfect}}\right) &\le \hat{F}(x_\star) + 3\lambda\|x_\star - x_0\|_\alpha + \phi\frac{\hat{L}\hat{R}\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}} \\ &\le \hat{F}(x_\star) + O\left(\phi\frac{\hat{L}\|x_\star - x_0\|_\alpha\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}}\right).\end{aligned} \tag{21}$$

Otherwise, if $\hat{R} \ge 66\|x_\star - x_0\|_\alpha$, we obtain from Equations (19) and (20) that

$$\min_{x\in\mathcal{X}:\|x-x_0\|_\alpha\le\hat{R}} \hat{F}(x) \le \hat{F}(x_\star) - \frac{\lambda\hat{R}}{3\cdot 66}.$$

Therefore, by Assumption 6, with probability at least $1-3\delta$,

$$\begin{aligned}\hat{F}\left(x_{\text{perfect}}\right) &\le \hat{F}(x_\star) - \frac{\lambda\hat{R}}{3\cdot 66} + \phi\frac{\hat{L}\hat{R}\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}} \\ &\le \hat{F}(x_\star) - 2\frac{\hat{L}\hat{R}\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}}.\end{aligned} \tag{22}$$

Now that we have bounded $\hat{F}\left(x_{\text{perfect}}\right)$, we use this to bound $F_{\hat{L}\|x_{\text{perfect}}-x_0\|_\alpha}\left(x_{\text{perfect}}\right)$. Define $L(s)$ as the Lipschitz constant of $m(\cdot; s)$, and $L(\mathcal{P})$ as the distribution over the Lipschitz constants resulting from the distribution $\mathcal{P}$. Lemma 21 states that with probability at least $1-\delta$,

$$\hat{L} = \max_{i\in[N]} L(s_i) \ge Q_{1-\frac{\ln(1/\delta)}{N}}(L(\mathcal{P})). \tag{23}$$

Thus, for $p \coloneqq \mathbb{P}_{s\sim\mathcal{P}}\left[L(s) > \hat{L}\right]$, we have $p \in \left[0, \frac{\ln(1/\delta)}{N}\right]$. Therefore,

$$\begin{aligned}
&F_{\hat{L}\|x_{\text{perfect}}-x_0\|_\alpha}\left(x_{\text{perfect}}\right) \\
&\quad= (1-p)\mathbb{E}_{s\sim\mathcal{P}}\left[f\left(x_{\text{perfect}}; s\right) \mid L(s) \le \hat{L}\right] \\
&\qquad\quad + p\mathbb{E}_{s\sim\mathcal{P}}\left[h\left(m_{\hat{L}\|x_{\text{perfect}}-x_0\|_\alpha}\left(x_{\text{perfect}}; s\right); s\right) \mid L(s) > \hat{L}\right] \\
&\quad\le (1-p)\mathbb{E}_{s\sim\mathcal{P}}\left[f\left(x_{\text{perfect}}; s\right) \mid L(s) \le \hat{L}\right] \\
&\qquad\quad + p\mathbb{E}_{s\sim\mathcal{P}}\left[f(x_\star; s) + \hat{L}\|x_{\text{perfect}} - x_0\|_\alpha + L_\star\|x_\star - x_0\|_\alpha \mid L(s) > \hat{L}\right] \\
&\quad= F(x_\star) + (1-p)\left(\hat{F}\left(x_{\text{perfect}}\right) - \hat{F}(x_\star)\right) + p\left(\hat{L}\|x_{\text{perfect}} - x_0\|_\alpha + L_\star\|x_\star - x_0\|_\alpha\right), \quad (24)
\end{aligned}$$

where the inequality is because h is 1-Lipschitz and m is $L_\star$-Lipschitz.

Now we continue to bound $F_{\hat{L}\|x_{\text{perfect}}-x_0\|_\alpha}\left(x_{\text{perfect}}\right)$. First, let us consider the case that $\hat{R} \le 66\|x_\star - x_0\|_\alpha$. We have

$$\begin{aligned}
&F_{\hat{L}\|x_{\text{perfect}}-x_0\|_\alpha}\left(x_{\text{perfect}}\right) \\
&\quad\overset{(i)}{\le} F(x_\star) + O\left(\phi\frac{\hat{L}\|x_\star - x_0\|_\alpha\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}}\right) + \frac{\left(\hat{L}\|x_{\text{perfect}} - x_0\|_\alpha + L_\star\|x_\star - x_0\|_\alpha\right)\ln(1/\delta)}{N} \\
&\quad\overset{(ii)}{\le} F(x_\star) + O\left(\phi\frac{L_\star\|x_\star - x_0\|_\alpha\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}}\right), \quad (25)
\end{aligned}$$

where (i) is from Equations (21) and (24), and (ii) is from $\|x_{\text{perfect}} - x_0\|_\alpha \le \hat{R}$, $\hat{L} \le L_\star$, and from the fact that for every $v \in [0,1]$ we have $v \le \sqrt{v}$.

Now, let us consider the case that $\hat{R} \ge 66\|x_\star - x_0\|_\alpha$. Thus, from Equations (22) and (24) and from $\|x_{\text{perfect}} - x_0\|_\alpha \le \hat{R}$,

$$\begin{aligned}
F_{\hat{L}\|x_{\text{perfect}}-x_0\|_\alpha}\left(x_{\text{perfect}}\right) &\overset{(i)}{\le} F(x_\star) - 2\frac{\hat{L}\hat{R}\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}} + \frac{\left(\hat{L}\|x_{\text{perfect}} - x_0\|_\alpha + L_\star\|x_\star - x_0\|_\alpha\right)\ln(1/\delta)}{N} \\
&\overset{(ii)}{\le} F(x_\star) - 2\frac{\hat{L}\hat{R}\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}} + \left(\hat{L}\|x_{\text{perfect}} - x_0\|_\alpha + L_\star\|x_\star - x_0\|_\alpha\right)\frac{\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}} \\
&\overset{(iii)}{\le} F(x_\star) + O\left(\frac{L_\star\|x_\star - x_0\|_\alpha\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}}\right), \quad (26)
\end{aligned}$$

where (i) is from Equations (22) and (24), (ii) is from the fact that for every $v \in [0,1]$ we have $v \le \sqrt{v}$, and (iii) is from $\|x_{\text{perfect}} - x_0\|_\alpha \le \hat{R}$, and $\hat{L} \le L_\star$.

Therefore, Equations (25) and (26) establish that in all cases, with probability at least $1 - 3\delta$, we have

$$F_{\hat{L}\|x_{\text{perfect}}-x_0\|_\alpha}\left(x_{\text{perfect}}\right) \le F(x_\star) + O\left(\phi\frac{L_\star\|x_\star - x_0\|_\alpha\sqrt{\ln\frac{\psi_\alpha}{\delta}}}{\sqrt{N}}\right).$$

Now, Equation (23) shows that $\hat{L}$ exceeds the $1 - \frac{\ln(1/\delta)}{N}$ quantile of $L(\mathcal{P})$. As

$$2N + \frac{\ln(1/\delta) + 2e\ln(1/\delta)}{N - e\ln(1/\delta)} \cdot N \le 2N + \frac{\ln(1/\delta) + 2e\ln(1/\delta)}{e\ln(1/\delta)} \cdot N \le 5N,$$

Lemma 22 shows that among $5N$ samples, with probability at least $1-\delta$, at least $2N$ lie below this quantile. Hence, with high probability, Algorithm 2 does not return FAILURE after the loop, and $6N$ samples suffice.

Finally, by using the union bound, we get the desired results with probability at least $1 - 5\delta$. ■

Appendix J. Discussion of Lipschitz properties in the model–loss decomposition setting

In this section, we discuss the relationship between the Lipschitz constants of the model m, the loss h, and the composed objective function f.

First, we note that the dual norm $\|\cdot\|_{\alpha*}$ can be viewed as the operator norm induced by the norms $\|\cdot\|_{\alpha}$ and $|\cdot|$; both are defined as $\|v\|_{\alpha*} := \sup\{|\langle v, x\rangle| \mid \|x\|_{\alpha} \le 1\}$. Similarly, the dual norm $\|\cdot\|_{\beta*}$ can be viewed as the operator norm induced by the norms $\|\cdot\|_{\beta}$ and $|\cdot|$. Therefore, immediately from the sub-multiplicativity of operator norms, we obtain that

$$\|\nabla f(x;s)\|_{\alpha*} = \|\nabla h(m(x;s);s)\nabla m(x;s)\|_{\alpha*} \le \|\nabla h(m(x;s);s)\|_{\beta*} \cdot \|\nabla m(x;s)\|_{\alpha,\beta} \tag{27}$$

For completeness, we provide a short proof:

Proof For any $v \in \mathcal{X}$ such that $\|v\|_{\alpha} \le 1$, if $\nabla m(x;s)v = 0$ then

$$|\nabla h(m(x;s);s)\nabla m(x;s)v| = 0 \le \|\nabla h(m(x;s);s)\|_{\beta*} \cdot \|\nabla m(x;s)\|_{\alpha,\beta}.$$

Otherwise,

$$\begin{aligned}|\nabla h(m(x;s);s)\nabla m(x;s)v| &= |\nabla h(m(x;s);s)\frac{\nabla m(x;s)v}{\|\nabla m(x;s)v\|_{\beta}}| \cdot \|\nabla m(x;s)v\|_{\beta} \\ &\le \|\nabla h(m(x;s);s)\|_{\beta*} \cdot \|\nabla m(x;s)\|_{\alpha,\beta},\end{aligned}$$

where the inequality follows from the definitions of the dual norm $\|\cdot\|_{\beta*}$ and the operator norm $\|\cdot\|_{\alpha,\beta}$. Thus, from the definition of the dual norm $\|\cdot\|_{\alpha*}$, we obtain that Equation (27) holds. ■

From Equation (27), it immediately follows that if the model m is L-Lipschitz and the loss function h is 1-Lipschitz, then the objective function f is L-Lipschitz. Formally:

Lemma 26 *Let $L > 0$.*

1. *If $m \in \mathcal{I}_{\text{Lip}}^{L,\|\cdot\|_{\alpha,\beta}}$ and $h \in \mathcal{I}_{\text{Lip}}^{1,\|\cdot\|_{\beta*}}$, then $f \in \mathcal{I}_{\text{Lip}}^{L,\|\cdot\|_{\alpha*}}$.*

2. *If $m \in \mathcal{I}_{\text{SM-Lip}}^{L,\|\cdot\|_{\alpha,\beta}}$ and $h \in \mathcal{I}_{\text{Lip}}^{1,\|\cdot\|_{\beta*}}$, then $f \in \mathcal{I}_{\text{SM-Lip}}^{L,\|\cdot\|_{\alpha*}}$.*

Proof

Proof of 26.1 For every $x \in \mathcal{X}$ and $s \in \mathbb{S}$, from Equation (27) we have

$$\|\nabla f(x; s)\|_{\alpha*} \leq \|\nabla h(m(x; s); s)\|_{\beta*} \cdot \|\nabla m(x; s)\|_{\alpha,\beta} \leq L.$$

Thus, $f \in \mathcal{I}_{\text{Lip}}^{L,\|\cdot\|_{\alpha*}}$.

Proof of 26.2 For every $x \in \mathcal{X}$, from Equation (27) we have

$$\mathbb{E}_{s\sim\mathcal{P}}\|\nabla f(x; s)\|_{\alpha*}^2 \leq \mathbb{E}_{s\sim\mathcal{P}}[\|\nabla h(m(x; s); s)\|_{\beta*}^2 \cdot \|\nabla m(x; s)\|_{\alpha,\beta}^2] \leq \mathbb{E}_{s\sim\mathcal{P}}[\|\nabla m(x; s)\|_{\alpha,\beta}^2] \leq L^2.$$

Thus, $f \in \mathcal{I}_{\text{SM-Lip}}^{L,\|\cdot\|_{\alpha*}}$. ■

However, it is possible to construct a model m and a loss function h such that the Lipschitz constant of f is much smaller than the product of the Lipschitz constants of m and h. In what follows, we present examples in which the Lipschitz constant of f is close to this product, showing that the bound above can be tight.

J.1. Lipschitz constant under the absolute loss

In this subsection, we discuss the Lipschitz constant of a model with the absolute loss. Let $\mathcal{Y} \subseteq \mathbb{R}$, and let the loss function be the absolute loss: $h(y; s) := |y - q(s)|$, where $q : \mathbb{S} \mapsto \mathbb{R}$ is some function. It is easy to see that for every $y \in \mathcal{Y}$ and $s \in \mathbb{S}$ such that $y \neq q(s)$, we have $\nabla h(y; s) \in \{-1, 1\}$. Therefore, for every $x \in \mathcal{X}$ and $s \in \mathbb{S}$ such that $m(x; s) \neq q(s)$, we have $\|\nabla f(x; s)\|_{\alpha*} = \|\nabla h(m(x; s); s)\nabla m(x; s)\|_{\alpha*} = \|\nabla m(x; s)\|_{\alpha*}$. Thus, for many models, the Lipschitz constant of f coincides with that of m.

J.2. Linear model with cross-entropy loss

In this section, we discuss the Lipschitz constant of a linear model with cross-entropy loss. Let $d_1, d_2 \in \mathbb{N}$, $\mathcal{X} \subseteq \mathbb{R}^{d_1}$, and $\mathcal{Y} \subseteq \mathbb{R}^{d_2}$. We consider the Euclidean norm $\|\cdot\|_\alpha = \|\cdot\|_\beta = \|\cdot\|_2$ on $\mathcal{X}$ and on $\mathcal{Y}$. Define the softmax function by

$$p_i(y) := \frac{e^{y_i}}{\sum_{j=1}^{d_2} e^{y_j}} \qquad \boldsymbol{p}(y) := (p_1(y), \ldots, p_{d_2}(y)).$$

With a slight abuse of notation, let s be a vector of size $d_1 + 1$, where s_0 denotes the class label and $s_{1:d_1}$ the remaining features.

Proposition 27 *Let the sample set $\mathbb{S}$ be a subset of $\{s \in \mathbb{R}^{d_1+1} \mid s_0 \in [d_2]\}$. Let m be the linear model $m(x; s) := mat(x)s_{1:d_1}$, where $mat(x)_{i,j} := x_{j+(i-1)d_1}$ is the matricization of x. Define the cross-entropy loss h as $h(y; s) := -\log(p_{s_0}(y))$. Consequently, the objective function f is defined as $f(x; s) := -\log(p_{s_0}(mat(x)s_{1:d_1}))$.*

Equation (27) is tight, i.e., for all $x \in \mathcal{X}$ and $s \in \mathbb{S}$,

$$\|\nabla f(x; s)\|_2 = \|\nabla h(m(x; s); s)\|_2 \cdot \|\nabla m(x; s)\|_{2,2}.$$

Furthermore, $h \in \mathcal{I}_{Lip}^{\sqrt{2},\|\cdot\|_{\beta*}}$, and for $x = 0$ and all $s \in \mathbb{S}$ we have $\|\nabla h(m(0; s); s)\|_2 = \sqrt{1 - \frac{1}{d_2}}$.

Proof Formally, for every $i \in [d_2]$ and $j \in [d_1 d_2]$, if $j \in [(i-1) \cdot d_1 + 1, i \cdot d_1]$ we have $[\nabla m(x;s)]_{i,j} = s_{j-(i-1)d_1}$, and otherwise $[\nabla m(x;s)]_{i,j} = 0$. Informally, every row i of $\nabla m(x;s)$ contains the sample $s_{1:d_1}$ starting at the $(i-1) \cdot d_1 + 1$ element,

$$\nabla m(x;s) = \begin{bmatrix} s_1 & \dots & s_{d_1} & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & s_1 & \dots & s_{d_1} & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & & & & & & & \ddots & & & & & & & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & s_1 & \dots & s_{d_1} & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & s_1 & \dots & s_{d_1} \end{bmatrix}.$$

Moreover, the gradient of the loss satisfies $\nabla h(y;s) = \boldsymbol{p}(y) - \boldsymbol{e}_{s_0}$, where $\boldsymbol{e}_{s_0}$ is the vector with a one in the s_0-th position and zeros everywhere else.

Since the operator norm $\|\cdot\|_{2,2}$ is the maximum singular value, we obtain $\|\nabla m(x;s)\|_{2,2} = \|s_{1:d_1}\|_2$. Therefore, by the chain rule,

$$\begin{aligned} \|\nabla f(x;s)\|_2 &= \|\nabla h(m(x;s);s) \nabla m(x;s)\|_2 \\ &= \sqrt{\sum_{i=1}^{d_2} \sum_{j=1}^{d_1} [\nabla h(m(x;s);s)]_i^2 s_j^2} \\ &= \sqrt{\sum_{i=1}^{d_2} [\nabla h(m(x;s);s)]_i^2 \sum_{j=1}^{d_1} s_j^2} \\ &= \|\nabla h(m(x;s);s)\|_2 \cdot \|s_{1:d_1}\|_2. \end{aligned}$$

Thus, for all $x \in \mathcal{X}$ and $s \in \mathbb{S}$ we have

$$\|\nabla f(x;s)\|_2 = \|\nabla h(m(x;s);s)\|_2 \cdot \|\nabla m(x;s)\|_{2,2}.$$

In addition, for all $x \in \mathcal{X}$ and $s \in \mathbb{S}$, we have

$$\|\nabla h(m(x;s);s)\|_2^2 = 1 - 2p_{s_0}(m(x;s)) + \sum_{i=1}^{d_2} p_i^2(m(x;s)) \le 1 + \sum_{i=1}^{d_2} p_i(m(x;s)) = 2.$$

Thus, $\|\nabla h(m(x;s);s)\|_2 \le \sqrt{2}$.

Moreover, for $x = 0$ and all $s \in \mathbb{S}$ we have

$$\|\nabla h(m(0;s);s)\|_2^2 = 1 - 2p_{s_0}(m(0;s)) + \sum_{i=1}^{d_2} p_i^2(m(0;s)) = 1 - \frac{2}{d_2} + \sum_{i=1}^{d_2} \frac{1}{d_2^2}.$$

Thus, $\|\nabla h(m(0;s);s)\|_2 = \sqrt{1 - \frac{1}{d_2}}$. ■

This shows that, for linear models with cross-entropy loss, the upper bound in Equation (27) is tight, i.e., for any point $x \in \mathcal{X}$ and sample $s \in \mathbb{S}$ the gradient norm $\|\nabla f(x;s)\|_2$ is equal to $\|\nabla h(m(x;s);s)\|_2 \cdot \|\nabla m(x;s)\|_{2,2}$. Therefore, depending on the choice of $\mathcal{X}$, $\mathcal{Y}$, and $\mathbb{P}$, the Lipschitz constant of the composed objective f can be close to, or even equal to, the product of the Lipschitz constants of the model m and the loss h.

Appendix K. Extending the results to non-differentiable functions

In this paper, we do not impose convexity requirements on the model and loss functions, but only on the objective function. Therefore, subgradients do not necessarily exist at all points. Instead, for a function q, we use the Clarke generalized gradient (Clarke, 1975):

$$\partial_c q(x) \coloneqq \mathrm{co}\left\{ v \mid \exists \{x_i\}_{i\in\mathbb{N}} \subset \Omega_q : x = \lim_{i\to\infty} x_i \text{ and } v = \lim_{i\to\infty} \nabla q(x_i) \right\},$$

where co is the convex hull, and Ω_q is the set of points where q is differentiable. We note that if q is convex and locally Lipschitz, then $\partial_c q(x)$ coincides with the set of subgradients $\partial q(x)$ for every x in any open subset of the domain (Clarke, 1975, Proposition 1.2); see also Rockafellar (1970, Theorem 25.6).

We define $\mathcal{I}_{\text{cont}}$ as the class of functions q such that: (i) $q(\cdot; s)$ is differentiable almost everywhere for any sample s; (ii) the Clarke gradient $\partial_c q(x; s)$ is nonempty for every point x and sample s; and (iii) for every pair of points x, y, sample s, and every continuously differentiable curve $\gamma : [0, 1] \to \mathrm{dom}(q(\cdot; s))$ satisfying $\gamma(0) = y$ and $\gamma(1) = x$, there exists a selection $\nabla_x q(\gamma(a); s) \in \partial_c q(\gamma(a); s)$ such that

$$q(x; s) = q(y; s) + \int_0^1 \nabla_x q(\gamma(a); s) \nabla_a \gamma(a)\, da.$$

We further assume that the model–loss decomposition satisfies the chain rule; i.e., for all $x \in \mathcal{X}$ and $s \in \mathbb{S}$,

$$\partial_c f(x; s) \subseteq \overline{\mathrm{co}}\{vu \mid v \in \partial_c h(m(x; s); s) \text{ and } u \in \partial_c m(x; s)\},$$

where $\overline{\mathrm{co}}$ denotes the closed convex hull. The purpose of this requirement is to ensure that if m is L-Lipschitz and h is 1-Lipschitz, then f is also L-Lipschitz, as in the differentiable case (see Lemma 26).

We are now able to replace the use of $\mathcal{I}_{\text{diff}}$ with $\mathcal{I}_{\text{cont}}$. For the classes of convex stochastic optimization problems, the definitions remain almost the same:

$$\mathcal{I}_{\text{Dist}}^{\overline{\rho},\|\cdot\|_\alpha} := \{q \in \mathcal{I}_{\text{cont}} \mid \|x_0 - x_\star\|_\alpha \le \overline{\rho} \text{ and } \mathbb{E}_{s\sim\mathcal{P}} q(\cdot; s) \text{ is convex}\},$$
$$\hat{\mathcal{I}}_{\text{Dist}}^{\overline{\rho},\|\cdot\|_\alpha} := \{q \in \mathcal{I}_{\text{cont}} \mid \|x_0 - x_\star\|_\alpha \le \overline{\rho} \text{ and } \forall s \in \mathbb{S} : q(\cdot; s) \text{ is convex}\}.$$

To account for the entire Clarke generalized gradient, we modify the definitions of the classes of Lipschitz and second-moment-Lipschitz problems:

$$\mathcal{I}_{\text{Lip}}^{\overline{\ell},\|\cdot\|} := \left\{q \in \mathcal{I}_{\text{cont}} \mid \forall x \in \mathcal{X}; s \in \mathbb{S}; g \in \partial_c q(x; s) : \|g\| \le \overline{\ell}\right\},$$
$$\mathcal{I}_{\text{SM-Lip}}^{\overline{\ell},\|\cdot\|} := \left\{q \in \mathcal{I}_{\text{cont}} \mid \forall x \in \mathcal{X} : \mathbb{E}_{s\sim\mathcal{P}}\left[\sup_{g\in\partial_c q(x;s)} \|g\|^2\right] \le \overline{\ell}^2\right\}.$$

Except for the proof of Theorem 8, all other proofs remain valid under the modified definitions of the problem classes, provided that $\nabla q(x)$ is interpreted as an arbitrary element of $\partial_c q(x)$ at the point x. Theorem 8, however, requires assuming that f is differentiable, because its proof uses Lemma 2 of Lawrence et al. (2025), which relies on this differentiability assumption. Theorem 8 can be extended to the non-differentiable setting via smoothing arguments, e.g., using the Moreau envelope (Moreau, 1965).