Safe Start: Configuring Optimization Algorithms for Decision-Making under Extreme Risks

Henry Lam, Wasin Meesena
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027
{khl2114,wm2501}@columbia.edu

Abstract

We consider stochastic optimization where the goal is not only to optimize an average-case objective, but also mitigate the occurrence of rare catastrophic events. This problem is motivated from safety-aware decision-making and AI training. We first argue that, in the presence of a simulation model, natural attempts to integrate variance reduction into optimization, even executed in a reasonable adaptive fashion, encounters fundamental challenges in guaranteeing realistic runtime when using common stochastic gradient descent algorithms. This challenge arises from the extreme sensitivity of tail-based objectives with respect to the decision variables, which renders the failure of traditional Lipschitz-based analyses. We offer remedies based on a new notion of *safe start* that allows for efficient finite-time error control, and show how the sampling complexity scales favorably under the combination of safe start and variance reduction. We illustrate our methodologies on examples in portfolio Value-at-Risk and extreme-quantile estimation.

1 Introduction

In many high-stakes problems, it is critical for decision-making to account for the occurrence of rare catastrophic events, in addition to standard average-case performances. For example, in designing human-interacting physical systems such as self-driving vehicles, the objective must prioritize the prevention of road conflicts and fatalities [1, 2]. Similarly, financial portfolio management needs to balance gain with the hedging against extreme losses [3, 4]. In simulation modeling, estimating these catastrophic events, and understanding how they occur, has been a long-standing focus under the umbrella of *rare-event simulation*. Despite many established approaches in this literature, decision optimization with the goal of *preventing* catastrophic events appears to be substantially open and has only very recently started to gather attention [5–8]. On a high level, this work attempts to systematically understand and remedy the challenges in running optimization algorithms for objectives that critically avoid rare events. As we will explain, this involves several novel fundamental issues that deviate from the established literature in the *evaluation* of rare event likelihoods.

We consider the following generic "rare-event optimization" formulation:

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}) = \underbrace{f(\boldsymbol{x})}_{\text{average-case performance}} + \lambda \underbrace{p(\boldsymbol{x})}_{\text{tail risk}}. \tag{1}$$

where $\boldsymbol{x}$ is the decision, f the expectation of some loss function, and p a tail-risk term, i.e., it involves the tail of the underlying randomness, such as the probability of a rare event or value-at-risk at a high level. $\lambda > 0$ is a weighting parameter that balances the expected loss and the tail risk. To put some perspective, first, note that one can typically choose a conservative decision $\boldsymbol{x}$ to almost

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ML×OR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

fully avoid the catastrophic event captured in $p(\boldsymbol{x})$. This, however, would usually be impractical and thus, for most problems, accounting for catastrophic events means a balance between average-case performance and the tail risk, which is exactly what (1) is motivated from. Second, as the weighting parameter λ increases, more weight is put on the tail risk. Typically, λ should be roughly of order $1/p(x^*)$, where x^* is an optimal solution, for formulation (1) to be meaningful. This is because this is the order of λ that would allow the two terms in (1) to be of the same magnitude at x^*; otherwise, it would mean that we either focus on the average-case only, or the target solution is highly conservative. To explain the above further, as well as understanding the choice of λ more concretely, formulation (1) could be viewed in two ways/examples:

Chance-constrained optimization: A natural optimization approach to explicitly avoid target risks is a chance-constrained optimization [9, 10], namely

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \;\text{ subject to }\; p(\boldsymbol{x}) \leq \tilde{\lambda}. \tag{2}$$

where $\tilde{\lambda}$ is often called the tolerance level. If p denotes the probability of an adversarial event, then (2) stipulates that the decision must explicitly control this probability to be within $\tilde{\lambda}$. When $\tilde{\lambda}$ is a low level, e.g., 0.001%, this means $p(x^*)$ must be similarly small and, a heuristic calculation of the Lagrangian reformulation of (2) would reveal that λ acts as the Lagrangian multiplier for the chance constraint in (2), and is of order $1/p(x^*)$.

Extreme quantile estimation/regression: Suppose $\xi \in \mathbb{R}$ is a simulable random variable. Then, taking $f(\boldsymbol{x}) = \boldsymbol{x}$ and $p(\boldsymbol{x}) = \mathbb{E}_\xi \left[\xi - \boldsymbol{x}\right]_+$ (where $\boldsymbol{x} \in \mathbb{R}$) in (1) is precisely the quantile estimation problem, for the target level $1/\lambda$. Moreover, this objective itself becomes the conditional value-at-risk. In this formulation, λ is the reciprocal of the tail probability of ξ, which in many cases has the same magnitude as $p(\boldsymbol{x})$ (ignoring logarithmic terms). Moreover, the above can be generalized to quantile regression [11, 12], where $\boldsymbol{x}$ now represents the set of parameters (e.g., linear coefficients) in a quantile regression model $f(\boldsymbol{x}, \boldsymbol{z})$ with covariates $\boldsymbol{z}$, and $p(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{z},\xi} \left[\xi - f(\boldsymbol{x}, \boldsymbol{z})\right]_+$.

Basic Evaluation Challenges and Remedies. Our goal is to obtain a good solution for (1), under the paradigm that the underlying randomness ξ is simulable and, moreover, susceptible to variance reduction. Let us first explain the latter, which relates to the historical focus of the rare-event simulation literature. More concretely, most of the works in this literature study the evaluation of rare-event probabilities or associated risk quantities [13–16]. Suppose $p = P(\xi \in S)$ denotes a rare-event probability of ξ falling into a region S (there is no decision $\boldsymbol{x}$ here to make). Then, with ξ simulable, a natural estimator would be the sample proportion of hits onto S. This, however, is a poor estimator when p is tiny because, as readily intuited, more likely than not we would simply output 0 as the estimate due to 0 hits. Indeed, if we would like to ensure the discrepancy between the estimate and the target probability to be within ϵ, relative to the target probability (instead of the absolute discrepancy), then from the Chebyshev inequality $P(|\hat{p} - p| > \epsilon p) \leq \sigma^2/(\epsilon^2 p^2 n)$ where σ^2 is the per-run variance, we see that the required sample size is $\sigma^2/(\epsilon^2 p^2 \alpha)$ with a $(1-\alpha)$ confidence level. Unfortunately, for naive sample proportion, this translates to $(1-p)/(\epsilon^2 p \alpha)$ which blows up the sample size when p is tiny.

To tackle the above challenges, *variance reduction* methods [17, 18] aim to drive down the variance $\sigma^2 = p(1-p)$ in sample proportion, or crude Monte Carlo, to much smaller. In large deviations settings, this usually means, up to a logarithmic factor, order p^2 instead of p as $p \to 0$, which translates to a sample size that only depends on p logarithmically. These variance reduction methods include importance sampling [13–15], multilevel splitting [19–21], and their variations [22–24].

Curse of Circularity and Adaptivity. The aforementioned evaluation challenges propagate to the optimization problem (1): To solve (1), we first need to have a good estimate of $p(\boldsymbol{x})$ in the objective function. This, however, gives rise to a "curse of circularity" [5, 25], which in turn arises from the double-swordness of variance reduction schemes in general. To explain, most rare-event variance reduction schemes, prominently importance sampling, is known to be highly sensitive to the problem setting, i.e., the scheme needs to be well-designed and well-tuned according to the problem configuration to achieve the variance reduction, otherwise it can perform very poorly. However, by the nature of optimization, this configuration is not known in advance since the decision has not been obtained. That is, on one hand a good optimizer needs a good variance reduction scheme, and on the other hand a good variance reduction scheme requires knowledge about the optimal solution, which leads to a "curse of circularity". To break this curse, [5, 25] use adaptivity, by iteratively updating solutions where at each iteration the variance reduction is applied as if the current solution

is optimal. In particular, stochastic gradient descent (SGD) or stochastic approximation is a natural iterative procedure to apply adaptive variance reduction. [5, 25] show central limit theorems for the updated solutions that reveal, indeed, that such adaptive procedures lead to asymptotic variance that is minimax optimal, i.e., it is on par with the variance where the optimal solution is known, and hence the variance reduction is well-tuned, in advance.

Safe Start. While central limit theorems and asymptotic variance provide positive guidance on the choice and performance of adaptive variance reduction, such results do not inform the sampling requirements to achieve a good solution. This issue is critically distinct from the evaluation task – In evaluation, the variance of an estimator translates *directly* into the required sample size, since the standard Chebyshev inequality reveals that the latter is proportional to the (squared) relative error. In optimization, the situation is substantially more complicated with the absence of simple guiding inequalities. To this end, our main contribution in this paper is two-fold: First, we show that, in general, embedding adaptive variance reduction into an iterative algorithm like SGD is fundamentally inadequate in terms of sampling complexity, even if the variance reduction is excellent for the evaluation task. Second, we show that, by using *safe start* – initializing the solution to be in a "safe" region where the risk term $p(\boldsymbol{x})$ in (1) is suitably small, we can guarantee a sub-exponential sampling complexity instead of an exploding exponential complexity, as long as we use a good-for-evaluation variance reduction.

2 Problem Setting

We consider (1), but introducing further a "rarity" parameter $\gamma > 0$ that signifies the target rarity level (i.e., $\gamma \to \infty$ means the likelihood of the rare event goes to 0). This rarity parameter is a modeling artifact that has proven useful in the rare-event literature. In particular, in standard large deviations the rare-event probability is often exponentially decaying in a suitable parameter γ. This implies an exponential growth of relative error in γ for naive Monte Carlo. On the other hand, an estimator is deemed (weakly) efficient if its per-run variance grows sub-exponentially in γ. We consider

$$\mathcal{P}(\gamma) := \min_{\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^n} F(\boldsymbol{x}; \gamma) \tag{3}$$

where $F(\boldsymbol{x}; \gamma) = f(\boldsymbol{x}) + \lambda(\gamma) p(\boldsymbol{x})$. Like in the introduction, we denote $\mathbf{x}^*(\gamma)$ as an optimal solution to $\mathcal{P}(\gamma)$. Moreover, for convenience, we denote $p^*(\gamma) := p(\mathbf{x}^*(\gamma))$ which is the tail risk at this optimal solution when the rarity level is γ. We assume an asymptotic characterization of $p^*(\gamma)$ in a large deviation principle (LDP) [26] style:

Assumption 1 (LDP of target risk)**.** *There exists a rate $I > 0$ such that $\lim_{\gamma \to \infty} -\frac{1}{\gamma} \log p^*(\gamma) = I$. Moreover, the weighting parameter $\lambda(\gamma)$ is chosen to match this in exponential scale, in the sense that $\limsup_{\gamma \to \infty} \frac{1}{\gamma} \log \lambda(\gamma) \leq I$.*

That is, the target extreme risk decays at an exponential rate of the rarity level, and it is attained when the weighting parameter grows at the same exponential rate. We impose an additional assumption:

Assumption 2 (Second-moment bounds)**.** *Consider unbiased estimators $G_f(\mathbf{x}, \xi)$ and $G_p(\mathbf{x}, \xi)$ for $\nabla f(\mathbf{x})$ and $\nabla p(\mathbf{x})$. There exist constants $\sigma_f, \sigma_p \geq 0$ such that, for all $\mathbf{x} \in \mathcal{X}$,*

$$\mathbb{E}\Big[\big\|G_f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\big\|^2\Big] \leq \sigma_f^2, \qquad \mathbb{E}\Big[\big\|G_p(\mathbf{x}, \xi) - \nabla p(\mathbf{x})\big\|^2\Big] \leq \sigma_p^2 \|\nabla p(\mathbf{x})\|^2.$$

The second bound in Assumption 2 is equivalent to a strong efficiency property for a gradient estimator in the evaluation setting [27, 28]. When the risk term is extremely small, such a bound can only be obtained via efficient variance reduction. Additional regularity assumptions including convexity properties of the objective function are provided in Appendix A.

Next, we define a *safe* solution as one whose extreme risk is less than the target risk, and that is sufficiently close to the average-case minimum $\tilde{\boldsymbol{x}}^*$.

Definition 1 (Safe solution)**.** *Under Assumption 1, $\{\mathbf{x}_0(\gamma)\}_{\gamma > 0}$ is a sequence of safe solutions if*

$$\limsup_{\gamma \to \infty} \frac{1}{\gamma} \log \|\mathbf{x}_0(\gamma) - \tilde{\mathbf{x}}^*\| = 0 \quad \textit{and} \quad \limsup_{\gamma \to \infty} \frac{1}{\gamma} \log p\big(\mathbf{x}_0(\gamma)\big) \leq -I.$$

We call these conditions Distance-Safe *and* Tail-Risk-Safe*, respectively.*

We also define our performance criterion, which is based on the attained objective being close to the oracle optimal.

Definition 2 (Rarity-incorporated optimization sample complexity)**.** *For an algorithm $\mathcal{A}$ and initialization $\mathbf{x}_0$, define $T(\varepsilon, \kappa; \gamma, \mathcal{A}, \mathbf{x}_0)$ as the minimum number of stochastic gradient samples required for $\mathcal{A}$, at rarity level γ, to produce an iterate $\hat{\mathbf{x}}$ that, with probability at least $1-\kappa$, satisfying*

$$F(\hat{\mathbf{x}}; \gamma) \;\leq\; (1+\varepsilon) \cdot F^*(\gamma) \text{ where } F^*(\gamma) := \min_{\mathbf{x} \in \mathcal{X}} F(\boldsymbol{x}; \gamma).$$

Note that the complexity $T(\varepsilon, \kappa; \gamma, \mathcal{A}, \mathbf{x}_0)$ in Definition 2 accounts for the number of sampled gradients at each iteration, when they are averaged in mini-batch schemes.

3 Main Results

We first discuss negative results which reveal that, without using safe start, the sampling complexity can depend exponentially in γ even if we have a good variance reduction scheme in hand. In fact, we show the stronger statement that even in the special case where the gradient evaluation is noiseless, unsafe start can still perform poorly.

Theorem 1 (Exponential sample complexity from unsafe starts)**.** *Consider the setup $\mathcal{P}(\gamma) := \min_{x,y \in \mathbb{R}_{\geq 0}} F(x, y; \gamma) \;=\; x^2 + y \;+\; \lambda(\gamma)\, e^{-x-y}$ with rarity level γ and $\lambda(\gamma) = e^{\gamma}$. Denote by $(x^*(\gamma), y^*(\gamma))$ the optimal solution to $\mathcal{P}(\gamma)$ with the exponential decay rate of the target risk I. For gradient descent with a fixed step size, there exists a sequence of* tail-risk-unsafe *but distance-safe starts where the number of iterations required to reach a target accuracy grows exponentially with γ.*

Next, suppose we use safe start. Then the next proposition highlights the need for variance reduction, i.e., Assumption 2. Together with Theorem 1, we see that without either safe start or variance reduction, SGD can perform poorly.

Theorem 2 (Exponential sample complexity from high variance)**.** *Consider the setup $\mathcal{P}(\gamma) := \min_{x \in \mathbb{R}_{\geq 0}} F(x; \gamma) \;=\; x \;+\; \lambda(\gamma)\, e^{-x}$ with risk term $p(x) = e^{-x}$ and $\lambda(\gamma) = e^{\gamma}$. Assume unbiased estimators $G_f(\mathbf{x}, \xi)$ and $G_p(\mathbf{x}, \xi)$ without weak efficiency, i.e., $\mathbb{E}\Big[\big\|G_p(\mathbf{x}, \xi) - \nabla p(\mathbf{x})\big\|^2\Big] \leq \|\nabla p(\mathbf{x})\|$. Then we can construct an instance where there exists a safe starting region S' of size $O(\gamma)$ such that, for any fixed step size $\eta \in \mathbb{R}^+$, there exists an initial point $x_0 \in S'$ for which SGD requires an exponential number of iterations to reach the target accuracy, almost surely.*

With the above, we are now ready to show our remedying positive results: With a safe initialization and variance reduction, SGD with constant or harmonic step sizes exhibit small sample complexity.

Theorem 3 (Safe-start SGD achieve sub-exponential sample complexity)**.** *Suppose Assumptions 1-5 hold. Consider safe initializations $\{\boldsymbol{x}_0(\gamma)\}_{\gamma > 0}$. Let SGD take either: 1) a constant step size with $\eta \in \big[\frac{1}{10 L_S(\gamma)},\, \frac{1}{L_S(\gamma)}\big]$ where $L_S(\gamma) := L_f + 2L_{p,2}\, F(\boldsymbol{x}_0(\gamma); \gamma)$ (called $\mathcal{A}_{Con}$), or 2) a harmonic step size with $\eta_t = \frac{\alpha(\gamma)}{K(\gamma)+t}$ with $\alpha(\gamma) \in \left[\frac{2}{\mu}, \frac{20}{\mu}\right]$ and $\frac{K(\gamma)}{L_S(\gamma)} \in \left[\frac{40}{\mu}, \frac{400}{\mu}\right]$ (called $\mathcal{A}_{Har}$). Then, both $T(\epsilon, \kappa; \gamma, \mathcal{A}_{Con}, \boldsymbol{x}_0(\gamma))$ and $T(\epsilon, \kappa; \gamma, \mathcal{A}_{Har}, \boldsymbol{x}_0(\gamma))$ grow at a sub-exponential rate in γ.*

Experimental Results. We demonstrate that safe start is critical for rare-event optimization, by reporting two representative tasks. The details of these experiments can be found in Appendix B.

Value at Risk (VaR). Following [29, 30], we estimate the q-VaR of a derivatives portfolio at level $q = 1 - 10^{-5}$ via $\min_{\theta \in \mathbb{R}} \; \theta + \gamma\, \mathbb{E}\big[(L(t, \xi) - \theta)_+\big]$, where the risk term (the hinge loss) is small for large θ (safe start). Importance sampling (IS) is constructed from the quadratic approximation of the loss as in [30]. **Findings.** Without IS, no choice of initialization/step size yields an accurate estimate; with IS, both constant-step and harmonic-step SGDs reach the true VaR (≈ 233) within 10^3–2×10^3 iterations from safe starts (e.g., $\theta_0 \geq 500$). The result is summarized in Table 1.

Extreme Quantile Regression. We learn a high-dimensional conditional q-quantile regressor via $\min_{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{101}} \; \mathbb{E}\Big[\, Q_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{X}) + \gamma \big(Y - Q_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{X})\big)_+ \,\Big]$, where $Q_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{X}) = \tilde{\boldsymbol{\theta}}^\top \tilde{\boldsymbol{X}}$. We embed IS by sampling Y' from $\mathcal{N}(Q_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{X}), 1)$ and reweighting by the likelihood ratio, yielding an unbiased low-variance gradient estimator. **Findings.** Safe starts drive all variants of methods to small errors rapidly; unsafe starts stall due to the sensitivity of the extrem-risk term. See Fig. 1. Importantly, our insight of safe/unsafe starts appears to carry into high-dimensional settings.

References

[1] Zhiyuan Huang, Henry Lam, David J LeBlanc, and Ding Zhao. Accelerated evaluation of automated vehicles using piecewise mixture models. *IEEE Transactions on Intelligent Transportation Systems*, 19(9):2845–2855, 2017.

[2] Matthew O'Kelly, Aman Sinha, Hongseok Namkoong, Russ Tedrake, and John C Duchi. Scalable end-to-end autonomous vehicle testing via rare-event simulation. *Advances in neural information processing systems*, 31, 2018.

[3] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

[4] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015.

[5] Shengyi He, Guangxin Jiang, Henry Lam, and Michael C Fu. Adaptive importance sampling for efficient stochastic root finding and quantile estimation. *Operations Research*, 72(6):2612–2630, 2024.

[6] Shanyin Tong, Anirudh Subramanyam, and Vishwas Rao. Optimization under rare chance constraints. *SIAM Journal on Optimization*, 32(2):930–958, 2022. doi: 10.1137/20M1382490. URL `https://doi.org/10.1137/20M1382490`.

[7] Jose Blanchet, Joost Jorritsma, and Bert Zwart. Optimization under rare events: scaling laws for linear chance-constrained programs. *arXiv preprint arXiv:2407.11825*, 2024.

[8] Anand Deo and Karthyek Murthy. Achieving efficiency in black-box simulation of distribution tails with self-structuring importance samplers. *Operations Research*, 73(1):325–343, 2025.

[9] Aharon Ben-Tal, Arkadi Nemirovski, and Laurent El Ghaoui. Robust optimization. 2009.

[10] Giuseppe Calafiore and Marco C Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical programming*, 102(1):25–46, 2005.

[11] Victor Chernozhukov, Iván Fernández-Val, and Tetsuya Kaji. Extremal quantile regression. *Handbook of Quantile Regression*, pages 333–362, 2017.

[12] Olivier C Pasche and Sebastian Engelke. Neural networks for extreme quantile regression with an application to forecasting of flood risk. *The Annals of Applied Statistics*, 18(4):2818–2839, 2024.

[13] James Antonio Bucklew and J Bucklew. *Introduction to rare event simulation*, volume 5. Springer, 2004.

[14] Sandeep Juneja and Perwez Shahabuddin. Rare-event simulation techniques: An introduction and recent advances. *Handbooks in operations research and management science*, 13:291–350, 2006.

[15] Jose Blanchet and Henry Lam. State-dependent importance sampling for rare-event simulation: An overview and recent advances. *Surveys in Operations Research and Management Science*, 17(1):38–59, 2012.

[16] Reuven Y Rubinstein and Dirk P Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media, 2004.

[17] Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer, 2007.

[18] Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer, 2004.

[19] Paul Glasserman, Philip Heidelberger, Perwez Shahabuddin, and Tim Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47(4):585–600, 1999.

[20] Thomas Dean and Paul Dupuis. Splitting for rare event simulation: A large deviation approach to design and analysis. *Stochastic processes and their applications*, 119(2):562–587, 2009.

[21] Manuel Villén-Altamirano and José Villén-Altamirano. The rare event simulation method restart: efficiency analysis and guidelines for its application. In *Network performance engineering: a handbook on convergent multi-service networks and next generation internet*, pages 509–547. Springer, 2011.

[22] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

[23] Reuven Y Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.

[24] Zdravko I Botev, Pierre L'Ecuyer, and Bruno Tuffin. Markov chain importance sampling with applications to rare event probability estimation. *Statistics and Computing*, 23(2):271–285, 2013.

[25] Liviu Aolaritei, Bart PG Van Parys, Henry Lam, and Michael I Jordan. Stochastic optimization with optimal importance sampling. *arXiv preprint arXiv:2504.03560*, 2025.

[26] Amir Dembo. *Large deviations techniques and applications*. Springer, 2009.

[27] Yuanlu Bai, Shengyi He, Henry Lam, Guangxin Jiang, and Michael C. Fu. Importance sampling for rare-event gradient estimation. In *2022 Winter Simulation Conference (WSC)*, pages 3063–3074, 2022. doi: 10.1109/WSC57314.2022.10015239.

[28] Pierre L'ecuyer, Jose H Blanchet, Bruno Tuffin, and Peter W Glynn. Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 20(1):1–41, 2010.

[29] Shengyi He, Guangxin Jiang, Henry Lam, and Michael C. Fu. Adaptive importance sampling for efficient stochastic root finding and quantile estimation. *Operations Research*, 0(0):null, 0. doi: 10.1287/opre.2023.2484. URL `https://doi.org/10.1287/opre.2023.2484`.

[30] Paul Glasserman, Philip Heidelberger, and Perwez Shahabuddin. Variance reduction techniques for estimating value-at-risk. *Management Science*, 46(10):1349–1364, 2000.

[31] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.

[34] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[35] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[36] Stefan Webb, Tom Rainforth, Yee Whye Teh, and M Pawan Kumar. A statistical approach to assessing neural network robustness. *arXiv preprint arXiv:1811.07209*, 2018.

[37] Yuanlu Bai, Zhiyuan Huang, Henry Lam, and Ding Zhao. Rare-event simulation for neural network and random forest predictors. *ACM Trans. Model. Comput. Simul.*, 32(3), jul 2022. ISSN 1049-3301. doi: 10.1145/3519385. URL `https://doi.org/10.1145/3519385`.

A Additional Assumptions

We introduce the asddditional assumptions as follows.

Assumption 3 (Relative gradient/Hessian bounds)**.** *Let $p : \mathcal{X} \to [0, \infty)$ be twice continuously differentiable on $\mathcal{X}$. There exist constants $L_{p,1}, L_{p,2} \geq 0$ such that for all $\boldsymbol{x} \in \mathcal{X}$,*

$$\|\nabla p(\boldsymbol{x})\| \;\leq\; L_{p,1}\, p(\boldsymbol{x})$$

and

$$\|\nabla^2 p(\boldsymbol{x})\| \;\leq\; L_{p,2}\, p(\boldsymbol{x}).$$

Next, we assume strong convexity and Lipschitz smoothness, which are common assumptions in first-order optimization [31].

Assumption 4 (Smooth and strong convexity)**.** *Assume that the domain $\mathcal{X} = \mathbb{R}^n$. Let f be twice continuously differentiable and $L-$smooth, i.e., $\|\nabla^2 f(\boldsymbol{x})\| \leq L_f$ for all $\boldsymbol{x} \in \mathcal{X}$. Furthermore, for any rarity level $\gamma > 0$, the combined risk $F(\cdot; \gamma) : \mathcal{X} \to \mathbb{R}$ is μ-strongly convex, i.e., $\nabla^2 F(\boldsymbol{x}; \gamma) \succeq \mu I$ for all $\boldsymbol{x} \in \mathcal{X}$.*

Lastly, we assume that the average-case objective alone attains its minimum at $\tilde{\mathbf{x}}^*$.

Assumption 5 (Lower-boundedness)**.** *The expected loss $f : \mathcal{X} \to \mathbb{R}$ is bounded below and attains its minimum on $\mathcal{X}$, i.e., there exists $\tilde{\mathbf{x}}^* \in \mathcal{X}$ such that $f(\tilde{\mathbf{x}}^*) = \inf_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$. Without loss of generality, we shift f so that $f(\tilde{\mathbf{x}}^*) = 0$.*

B Details of Experiments

We demonstrate that safe start is critical for rare-event optimization, by reporting three representative tasks. Moreover, these examples illustrate how to *identify* and *configure* safe starts. In the first experiment, we consider the estimation of VaR for a portfolio, which can be formulated as a one-dimensional optimization problem. The second experiment complements the first by showing that the notion of safe starts also extend to high-dimensional settings. Lastly, we build a classifier robust to tiny input perturbation and show that safe starts greatly facilitate the training process.

B.1 Value-at-Risk

Similar to [5] and [30], we consider the following VaR estimation problem of a derivatives portfolio.

Example 1 (Value-at-Risk)**.** *Assume a 250-day trading year and a 10-day risk horizon ($t = 0.04$ years). Set the risk-free rate to $r = 5\%$. Consider ten pairwise uncorrelated underlying assets, each starting at price 100 with volatility 0.3. The portfolio holds short positions of 10 at-the-money calls and 5 at-the-money puts on each asset, with all options expiring in half a year. We want to estimate the q-VaR of the portfolio loss where $q = 1 - 10^{-5}$.*

By Rockafellar–Uryasev CVaR representation, we can solve the following optimization for the VaR:

$$\min_{\theta \in \mathbb{R}} \left\{ \theta + \frac{1}{1-q} \mathbb{E}_\xi \left[(L(t, \xi) - \theta)_+ \right] \right\} \tag{4}$$

where $L(t, \xi)$ is the discounted loss of the portfolio at time t. The risk term (the hinge loss) is small for large θ (**safe start**). Importance sampling (IS) is constructed from the quadratic approximation of the loss as in [30]. We solve the above optimization problem using three optimizers: fixed-step SGD without IS, fixed-step SGD with IS and harmonic-step SGD with IS. Each algorithm is configured with many combinations of step sizes and initial solutions. Without IS, no choice of initialization/step size yields an accurate estimate; with IS, both constant-step and harmonic-step SGDs reach the true VaR (≈ 233) within 10^3–2×10^3 iterations from safe starts (e.g., $\theta_0 \geq 500$). The result is summarized in Table 1.

Table 1: SGDs with different configurations for finding the VaR. The true VaR is 233. The table shows θ_{1000} (θ_{2000}, respectively), the quantile estimate after running for $1,000$ (2000, respectively) iterations for fixed-step (harmonic-step, respectively) SGDs. Colored cells: green $\in [233 \pm 10]$ indicating estimates close to the true VaR, red otherwise.

Step size \ θ_0	-200	0	100	400	500	700	900
SGD without IS, fixed step							
10^2	99	99	99	399	499	699	899
10^3	790	990	90	390	490	690	890
10^4	9700	9900	0	300	400	600	800
10^5	59215.6	59216.6	59316.6	59616.6	59716.6	59916.6	60116.6
10^6	50224.6	50234.6	50334.6	50634.6	50734.6	50934.6	51134.6
SGD with IS, fixed step							
10^2	213.9	213.8	214.6	399.0	499.0	699.0	899.0
10^3	774.8	1044.2	251.2	390.0	490.0	690.0	890.0
10^4	1208.6	10256.7	2658.5	300.0	400.0	600.0	800.0
10^5	9219.3	59217.6	10334.0	238.9	233.0	233.3	226.7
10^6	50234.6	50234.6	50224.6	253.8	14607.8	383.4	893.3
SGD with IS, harmonic step, $\eta_t = \eta_0/t$							
10^4	5243.7	10294.0	721.5	399.2	499.2	699.2	899.2
10^5	60207.4	54215.3	15999.1	391.8	491.8	691.8	891.8
10^6	60147.8	60142.8	60142.8	318.2	418.2	618.2	818.2
10^7	59546.7	59580.1	59580.1	17128.5	231.7	232.8	233.0
10^8	-8378.4	53036.2	-8078.4	-7776.4	-7678.4	53536.2	53536.2

Notes: we implement projected SGDs: once the iterates are over $60,000$, they get projected back. Therefore, many incorrect estimates center around $60,000$.

B.2 Extreme Quantile Regression

This numerical session applies our insights to multidimensional settings. In addition to IS, we emphasize that two key requirements for optimization under extreme risk—*safe start* and *accurate gradient estimates*—ensures the convergence of SGD in multidimensional extreme conditional quantile regression. We examine three optimization configurations: **fixed-step SGD**, **harmonic-step SGD**, and **Adam** [32] which is widely used in machine learning. The problem setting is defined as follows:

Example 2 (Multi-dimensional Extreme Quantile Regression)**.** *Assume* $(\boldsymbol{X}, Y)$ *follows a joint distribution* $F_{\boldsymbol{\theta}^*}$ *parameterized by* $\boldsymbol{\theta}^* \in \mathbb{R}^{n-1}$*, where* $\boldsymbol{X} \sim \mathrm{N}\big(0, I_{(n-1)\times(n-1)}\big)$ *and*

$$Y \overset{d}{=} \boldsymbol{\theta}^{*\top} \boldsymbol{X} + \mathcal{N}(0, 1^2).$$

Our goal is to learn the true q*-quantile of* $Y \mid \boldsymbol{X}$*, namely*

$$\boldsymbol{\theta}^{*\top} \boldsymbol{X} \; + \; F^{-1}_{\mathrm{N}(0,1^2)}(q),$$

where $F^{-1}_{\mathrm{N}(0,1^2)}(\cdot)$ *denotes the inverse CDF of the standard normal distribution and* $q = 1 - \frac{1}{\gamma}$ *for* $\gamma > 1$.

Using an auxiliary formulation, the q*-quantile of* $Y \mid \boldsymbol{X}$ *solves*

$$\min_{y} \Big\{ y \; + \; \gamma \, \mathbb{E}_{Y|\boldsymbol{X}} \left[(Y - y)_+ \right] \Big\}. \tag{5}$$

Define

$$\tilde{\boldsymbol{X}} \; = \; [1 \quad \boldsymbol{X}] \qquad \textit{and} \qquad \tilde{\boldsymbol{\theta}} \in \mathbb{R}^n.$$

Using the linear model $Q_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{X}) := \tilde{\boldsymbol{\theta}}^\top \tilde{\boldsymbol{X}}$ *for each* $\boldsymbol{X}$ *as an estimator for the conditional quantile, the population objective is*

$$\min_{\tilde{\boldsymbol{\theta}}} \; \mathbb{E}_{(\boldsymbol{X},Y)} \Big[Q_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{X}) \; + \; \gamma \left(Y - Q_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{X}) \right)_+ \Big]. \tag{6}$$

To solve (6), we draw $(\boldsymbol{X}, Y)$ from $F_{\boldsymbol{\theta}^*}$, and its stochastic gradient with respect to $\tilde{\boldsymbol{\theta}}$ is

$$G_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{X}, Y) \;=\; \tilde{\boldsymbol{X}} \;-\; \gamma\, \mathbb{I}\left\{ Y > Q_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{X}) \right\} \tilde{\boldsymbol{X}}. \tag{7}$$

Hence, the $K-$average stochastic gradient with step size η_t (without IS) is

$$\tilde{\boldsymbol{\theta}}_{t+1} \;=\; \tilde{\boldsymbol{\theta}}_t \;-\; \frac{1}{K} \sum_{k=1}^{K} \eta_t \Big(\tilde{\boldsymbol{X}}_t^{(k)} \;-\; \gamma\, \mathbb{I}\Big\{ Y_t^{(k)} > Q_{\tilde{\boldsymbol{\theta}}_t}\left(\boldsymbol{X}_t^{(k)} \right) \Big\} \, \tilde{\boldsymbol{X}}_t^{(k)} \Big). \tag{8}$$

Variance reduction via importance sampling (IS). To *reduce gradient variance*, for each sample $\boldsymbol{X}_t$ and parameter $\tilde{\boldsymbol{\theta}}_t$, draw

$$Y_t' \sim \mathrm{N}\big(Q_{\tilde{\boldsymbol{\theta}}_t}(\boldsymbol{X}_t),\, 1^2\big),$$

and multiply by the likelihood ratio to obtain an unbiased gradient estimate. That is, we sample Y_t' from a normal distribution centered at the current estimate of the conditional extreme quantile. We update the current solution using $K-$average, IS-embedded gradient descent as follows:

$$\tilde{\boldsymbol{\theta}}_{t+1} \;=\; \tilde{\boldsymbol{\theta}}_t \;-\; \frac{1}{K} \sum_{k=1}^{K} \eta_t \Big(\tilde{\boldsymbol{X}}_t^{(k)} \;-\; \gamma\, L_{\tilde{\boldsymbol{\theta}}_t}\left(Y_t'^{(k)} \mid \boldsymbol{X}_t^{(k)} \right) \mathbb{I}\Big\{ Y_t'^{(k)} > Q_{\tilde{\boldsymbol{\theta}}_t}\left(\boldsymbol{X}_t^{(k)} \right) \Big\} \, \tilde{\boldsymbol{X}}_t^{(k)} \Big), \tag{9}$$

where the likelihood ratio is

$$L_{\tilde{\boldsymbol{\theta}}_t}(Y \mid \boldsymbol{X}) \;:=\; \frac{f_{\mathrm{N}(\boldsymbol{\theta}^{*\top}\boldsymbol{X},\, 1^2)}(Y)}{f_{\mathrm{N}(Q_{\tilde{\boldsymbol{\theta}}_t}(\boldsymbol{X}),\, 1^2)}(Y)}. \tag{10}$$

Experimental setup. Set $n = 101$. Draw the true parameter $\boldsymbol{\theta}^* \in \mathbb{R}^{100}$ from $\mathrm{N}(\mathbf{0}, I_{100\times 100})$ and set $\gamma = 100{,}000$. The corresponding quantile offset equals 4.26. Thus, the true conditional quantile function for any $\boldsymbol{X}$ is

$$\tilde{\boldsymbol{\theta}}^{*\top} \tilde{\boldsymbol{X}} \;=\; [4.26 \quad 1.62 \quad -0.61 \quad -0.53 \quad \ldots \quad 0.70]\, \tilde{\boldsymbol{X}}.$$

We evaluate each optimizer by plotting the progression of the Euclidean distance to the optimum, $\|\tilde{\boldsymbol{\theta}}_t - \tilde{\boldsymbol{\theta}}^*\|$. Consider the 101-dimensional initializations

$$\tilde{\boldsymbol{\theta}}_{\text{safe}} = [40 \quad 0 \quad 0 \quad \ldots \quad 0], \qquad \tilde{\boldsymbol{\theta}}_{\text{unsafe}} = [-30 \quad 0 \quad 0 \quad \ldots \quad 0],$$

as safe and unsafe starts, respectively. This is because the extreme risk is much smaller with $\tilde{\boldsymbol{\theta}}_{\text{safe}}$, i.e.,

$$\mathbb{E}_{(\boldsymbol{X},Y)} \left[\big(Y - Q_{\tilde{\boldsymbol{\theta}}_{\text{safe}}}(\boldsymbol{X})\big)_+ \right] \ll \mathbb{E}_{(\boldsymbol{X},Y)} \left[\big(Y - Q_{\tilde{\boldsymbol{\theta}}_{\text{unsafe}}}(\boldsymbol{X})\big)_+ \right].$$

Note that the Euclidean distance from the unsafe initialization to the optimum is *shorter* than from the safe initialization, yet the subsequent convergence behavior differs markedly. To reduce gradient variance, we *average* $K = 4{,}000$ stochastic gradients per update in the baseline configuration.

Experimental Results. In Figure 1, we plot the traces of the three optimizers under safe vs./ unsafe start. These optimizers have one thing in common — the errors go to near zero quickly when they start from a safe solution. Furthermore, each optimizer has its own advantages. **Fixed-step SGD** converges the fastest (error drops below 10^{-2} within $5{,}000$ iterations). The error improves no further due to its constant step size. **Harmonic-step SGD**, on the other hand, achieves a small error (below 10^{-3}) but takes slightly longer iterations. Finally,**Adam** suffers the least from the unsafe start. It takes, however, more than $50,000$ iterations to achieve error below 10^{-1}.

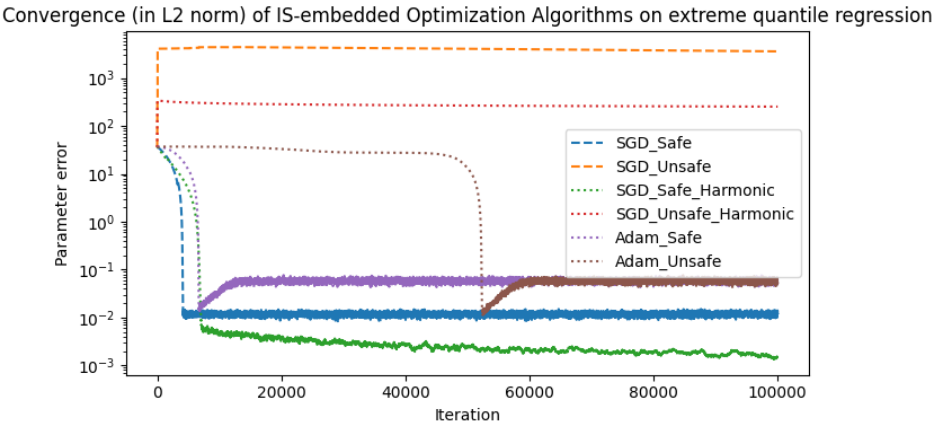

Figure 1: Traces of three optimizers under safe vs. unsafe starts after 100,000 iterations. Safe starts lead all three methods toward the true parameters (small Euclidean distances). In contrast, distances remain large for unsafe starts under both SGD variants.

B.3 Statistically Robust Classification on MAGIC Gamma Telescope

This experiment illustrates that the notion of safe starts extends to model training under extreme risk with neural networks. Here, we demonstrate how to derive safe starts for a robust classification problem. We use these initializations to build a multilayer-perceptron (MLP) classifier that achieves both *high prediction accuracy* and *robustness* to tiny input perturbations. The experiment also reveals that unsafe starts make it harder to find a suitable step size to train a robust classifier. We used the MAGIC Gamma Telescope dataset from the UCI Machine Learning Repository [33] for our experiment.

Robustness is an important consideration for the deployment of machine learning in real-world settings. Standard neural-network classifiers, for example, can be vulnerable to tiny input perturbations that lead to misclassifications [34, 35]. One common robustness metric is the prediction consistency, measuring the probability that input perturbations change the correct prediction [36**?**].

Concretely, we consider a classification model $\boldsymbol{\theta}$, an input $\boldsymbol{z}$, and its true category $y \in \{1, 2, \ldots, C\}$ with C total possible categories. Denote by $\hat{y}(\boldsymbol{z}; \boldsymbol{\theta}) \in [C]$ the predicted category that model $\boldsymbol{\theta}$ makes for a given input $\boldsymbol{z}$, and denote by $\boldsymbol{\epsilon}$ the (random) input perturbation. The *statistical robustness* of $\boldsymbol{\theta}$ at input $\boldsymbol{z}$ with the true category y is denoted by $\mathbb{P}_{\boldsymbol{\epsilon}} \left(\hat{y}(\boldsymbol{z} + \boldsymbol{\epsilon}; \boldsymbol{\theta}) \neq y \right)$. In safety- and security-critical applications, the target robustness is extremely small, and evaluating such rare-event probability requires variance reduction methods [37, 36].

Our goal is to build a statistically robust classifier $\boldsymbol{\theta}$ with high accuracy and robustness measured, respectively, by

1. Prediction Accuracy (PA): $\mathbb{P}_{(\boldsymbol{z},y)} \left[\hat{y}(\boldsymbol{z}; \boldsymbol{\theta}) = y \right]$
2. Conditional Robust Accuracy (CRA): $\mathbb{E}_{(\boldsymbol{z},y)} \left[\mathbb{P}_{\boldsymbol{\epsilon}} \left(\hat{y}(\boldsymbol{z} + \boldsymbol{\epsilon}; \boldsymbol{\theta}) = y \right) \mid \hat{y}(\boldsymbol{z}; \boldsymbol{\theta}) = y \right]$

where $(\boldsymbol{z}, y)$ is drawn from the data-generating distribution. Note that we look at the probability conditional on correct prediction because we do not award consistent incorrectness. To achieve these two criterion spontaneously, we construct a surrogate loss function:

$$\mathcal{L}(\boldsymbol{\theta}) := f(\boldsymbol{\theta}) + \lambda p(\boldsymbol{\theta})$$

where $f(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{z},y)} \left[\ell(\boldsymbol{z}, y; \boldsymbol{\theta}) \right]$ and $p(\boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{z},y)} \left[\mathbf{1}\{ \hat{y}(\boldsymbol{z}; \boldsymbol{\theta}) = y \} \, \mathbb{E}_{\boldsymbol{\epsilon}} \left[m(\boldsymbol{z}, y, \boldsymbol{\epsilon}; \boldsymbol{\theta}) \right]_+ \right]$. Here, ℓ is the classification loss (i.e., cross-entropy loss). $m(\boldsymbol{z}, y, \boldsymbol{\epsilon}; \boldsymbol{\theta}) = \max_{c \neq y} g_c(\boldsymbol{z}+\boldsymbol{\epsilon}; \boldsymbol{\theta}) - g_y(\boldsymbol{z}+\boldsymbol{\epsilon}; \boldsymbol{\theta})$ is the robust marginal at $(\boldsymbol{z}, y)$ where $g_c(\boldsymbol{z}; \boldsymbol{\theta}) \in \mathbb{R}$ is the score function (or logit) for category $c \in [C]$ given by $\boldsymbol{\theta}$ and input $\boldsymbol{z}$. Note that $\hat{y}(\boldsymbol{z}; \boldsymbol{\theta}) = \arg\max_c g_c(\boldsymbol{z}; \boldsymbol{\theta})$. This robust margin term is analogous to the hinge loss in the Value-at-Risk formulation in Equation (4).

The MAGIC Gamma Telescope dataset contains 19020 data points $\left\{ \boldsymbol{z}^{(i)}, y^{(i)} \right\}_{i=1}^{19020}$. $\boldsymbol{z}^{(i)} \in \mathbb{R}^{10}$ represents 10-dimensional features of an atmospheric image i from the telescope, and label $y^{(i)} \in$

Table 2: Baseline models: PA and CRA. Bold = PA $> 80\%$ or CRA $> 98\%$.

Models	$\widehat{\text{PA}}$	$\widehat{\text{CRA}}$	$\widehat{f(\boldsymbol{\theta}_0)}$	$\widehat{p(\boldsymbol{\theta}_0)}$	Distance-Safe	Tail-Risk-Safe
Always Negative (dummy)	36.04%	**100.00%**	–	–	–	–
Always Positive (dummy)	63.96%	**100.00%**	–	–	–	–
RAND (safe)	35.94%	**99.99%**	5.44	0.00	✗	✓
PT (unsafe)	**86.24%**	65.57%	0.28	2.63	✓	✗
PT(+15) (safe)	66.87%	98.24%	3.63	0.076	✓	✓
PT(-15) (safe)	36.18%	99.41%	8.65	0.030	✓	✓

$\{1, 2\}$ corresponding to signal/background. We use these features to distinguish atmospheric images induced from gamma ray (signal) or induced from cosmic ray (background).

Here, we consider a multi-layer perceptron (MLP) classifier with 2 hidden layers and 20 neurons. That is, the parameter $\boldsymbol{\theta}$ is described by $\left\{\mathbf{W}_1^{\boldsymbol{\theta}} \in \mathbb{R}^{10\times 20}, \boldsymbol{b}_1^{\boldsymbol{\theta}} \in \mathbb{R}^{10}, \mathbf{W}_2^{\boldsymbol{\theta}} \in \mathbb{R}^{20\times 1}, b_2^{\boldsymbol{\theta}} \in \mathbb{R}\right\}$. For a binary classification, we can write the score function as

$$g_1(\boldsymbol{z};\boldsymbol{\theta}) := {\mathbf{W}_2^{\boldsymbol{\theta}}}^\top \left(\text{ReLU}\left({\mathbf{W}_1^{\boldsymbol{\theta}}}^\top \boldsymbol{z} + \boldsymbol{b}_1^{\boldsymbol{\theta}}\right)\right) + b_2^{\boldsymbol{\theta}},$$
$$g_2(\boldsymbol{z};\boldsymbol{\theta}) := -{\mathbf{W}_2^{\boldsymbol{\theta}}}^\top \left(\text{ReLU}\left({\mathbf{W}_1^{\boldsymbol{\theta}}}^\top \boldsymbol{z} + \boldsymbol{b}_1^{\boldsymbol{\theta}}\right)\right) - b_2^{\boldsymbol{\theta}}.$$

We sample perturbations $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\, \sigma^2 I_{10})$ with $\sigma = 0.7$ and set $\lambda = 10^4$. The implementation details (optimizer settings, hyperparameters, minibatch sizes) appear in Appendix **??**.

We consider two *unsafe* initializations: **RAND** (randomly initialized) and **PT** (pretrained). The PT parameters are obtained by training for 2000 epochs with Adam on the binary cross-entropy loss only. In Table 2, **PT** is *tail-risk-unsafe* (its tail-risk term exceeds $1/\lambda$), whereas **RAND** attains near-zero tail risk but is *distance-unsafe*—it is a random initialization, qualitatively different from the trained $\boldsymbol{\theta}_{\text{PT}}$ that minimizes expected loss.

We consider two *safe starts*, **PT(+15)** and **PT(-15)**. Starting from the pretrained parameters, we replace only the last-layer bias (originally $b_2^{\textbf{PT}} = 0.89$) by $+15$ and -15, respectively. Thus **PT(+15)** strongly favors the positive ("signal") class, while **PT(-15)** favors the negative ("background") class. As expected, both variants perform poorly in standard prediction accuracy; however, their tail-risk terms are near zero because their outputs are essentially constant under small perturbations. Note that $\boldsymbol{\theta}_{\text{PT}(+15)}$ and $\boldsymbol{\theta}_{\text{PT}(-15)}$ are identical to $\boldsymbol{\theta}_{\text{PT}}$ except for this single bias entry.

We train each of the four initializations (**PT**, **RAND**, **PT(+15)**, **PT(-15)**) with two optimizers—SGD and Adam—using a learning-rate grid $\{10^{-1}, 10^{-2}, \ldots, 10^{-7}\}$. For every (initializer, optimizer, step size) configuration, we run 50 epochs on the surrogate loss and check whether it reaches the targets: prediction accuracy $\geq 80\%$ and conditional robust accuracy (CRA) $\geq 98\%$.

In Table 3, only one configuration (from a safe start) attains both targets. In Table 4, two step sizes achieve the targets for both **PT(+15)** and **PT(-15)**, while only one step size works for **PT**. Thus, safe starts make it easier to find a working step size for optimization under extreme risks.

Table 3: PA and CRA by configuration and step size (95% CIs in brackets). Bold cells: PA $> 80\%$ (*), CRA $> 98\%$ (†).

Initialization	Step Size	PA (95% CI)	RA (95% CI)
RAND	10^{-1}	36.04% [34.72, 37.37]	**100.00%** [100.00, 100.00]†
	10^{-2}	36.04% [34.72, 37.37]	**100.00%** [100.00, 100.00]†
	10^{-3}	63.96% [62.63, 65.28]	**100.00%** [100.00, 100.00]†
	10^{-4}	75.38% [74.17, 76.55]	**99.76%** [99.68, 99.83]†
	10^{-5}	70.72% [69.44, 71.96]	**99.70%** [99.59, 99.78]†
	10^{-6}	35.96% [34.64, 37.29]	**100.00%** [100.00, 100.00]†
	10^{-7}	35.94% [34.62, 37.27]	**100.00%** [100.00, 100.00]†
PT	10^{-1}	63.96% [62.63, 65.28]	**100.00%** [100.00, 100.00]†
	10^{-2}	63.96% [62.63, 65.28]	**100.00%** [100.00, 100.00]†
	10^{-3}	63.96% [62.63, 65.28]	**100.00%** [100.00, 100.00]†
	10^{-4}	63.96% [62.63, 65.28]	**100.00%** [100.00, 100.00]†
	10^{-5}	63.96% [62.63, 65.28]	**99.98%** [99.95, 100.00]†
	10^{-6}	74.06% [72.83, 75.26]	**99.26%** [99.12, 99.42]†
	10^{-7}	72.39% [71.14, 73.61]	**98.44%** [98.25, 98.64]†
PT(+15)	10^{-1}	36.04% [34.72, 37.37]	**100.00%** [100.00, 100.00]†
	10^{-2}	36.04% [34.72, 37.37]	**100.00%** [100.00, 100.00]†
	10^{-3}	63.96% [62.63, 65.28]	**100.00%** [100.00, 100.00]†
	10^{-4}	63.96% [62.63, 65.28]	**100.00%** [100.00, 100.00]†
	10^{-5}	36.04% [34.72, 37.37]	**99.89%** [99.81, 99.96]†
	10^{-6}	**80.42%** [79.30, 81.49]*	**99.18%** [99.04, 99.32]†
	10^{-7}	76.43% [75.24, 77.59]	97.97% [97.76, 98.17]
PT(-15)	10^{-1}	64.02% [62.69, 65.34]	**100.00%** [100.00, 100.00]†
	10^{-2}	63.96% [62.63, 65.28]	**100.00%** [100.00, 100.00]†
	10^{-3}	36.04% [34.72, 37.37]	**100.00%** [100.00, 100.00]†
	10^{-4}	36.04% [34.72, 37.37]	**100.00%** [100.00, 100.00]†
	10^{-5}	73.92% [72.69, 75.12]	**99.55%** [99.43, 99.66]†
	10^{-6}	71.79% [70.53, 73.02]	**99.20%** [99.04, 99.37]†
	10^{-7}	70.70% [69.42, 71.94]	**98.54%** [98.28, 98.75]†

Notes: "PA" and "CRA" are percentages; brackets show 95% confidence intervals as provided. * PA $> 80\%$. † CRA $> 98\%$.

Table 4: ADAM: PA and CRA by configuration and step size (95% CIs in brackets). Bold = Test Acc $> 80\%$ or CRA $> 98\%$.

Initialization	Step Size	PA (95% CI)	RA (95% CI)
RAND	10^{-1}	63.96% [62.63, 65.28]	**100.00%** [100.00, 100.00]
	10^{-2}	76.33% [75.14, 77.49]	**99.72%** [99.62, 99.81]
	10^{-3}	74.20% [72.97, 75.39]	**99.66%** [99.56, 99.76]
	10^{-4}	35.90% [34.58, 37.23]	**99.81%** [99.67, 99.93]
	10^{-5}	35.96% [34.64, 37.29]	**100.00%** [99.99, 100.00]
	10^{-6}	35.94% [34.62, 37.27]	**100.00%** [99.99, 100.00]
	10^{-7}	35.94% [34.62, 37.27]	**99.99%** [99.97, 100.00]
PT	10^{-1}	72.85% [71.60, 74.06]	**99.76%** [99.67, 99.84]
	10^{-2}	**82.01%** [80.93, 83.05]	**99.14%** [98.98, 99.28]
	10^{-3}	79.56% [78.42, 80.65]	97.45% [97.16, 97.71]
	10^{-4}	72.03% [70.77, 73.26]	97.51% [97.31, 97.72]
	10^{-5}	73.35% [72.11, 74.55]	87.09% [86.74, 87.50]
	10^{-6}	**84.62%** [83.60, 85.59]	67.93% [67.46, 68.43]
	10^{-7}	**86.33%** [85.36, 87.26]	65.71% [65.10, 66.24]
PT(+15)	10^{-1}	73.41% [72.17, 74.61]	**99.69%** [99.59, 99.78]
	10^{-2}	**82.61%** [81.54, 83.63]	**99.54%** [99.42, 99.66]
	10^{-3}	**80.34%** [79.22, 81.41]	**99.45%** [99.34, 99.56]
	10^{-4}	77.89% [76.72, 79.01]	97.42% [97.17, 97.63]
	10^{-5}	71.93% [70.67, 73.16]	97.42% [97.22, 97.62]
	10^{-6}	68.69% [67.39, 69.95]	97.60% [97.42, 97.77]
	10^{-7}	68.55% [67.25, 69.82]	97.34% [97.16, 97.50]
PT(-15)	10^{-1}	63.96% [62.63, 65.28]	**100.00%** [100.00, 100.00]
	10^{-2}	**81.69%** [80.60, 82.74]	**99.29%** [99.16, 99.44]
	10^{-3}	**80.42%** [79.30, 81.49]	**98.99%** [98.82, 99.14]
	10^{-4}	68.76% [67.47, 70.03]	97.82% [97.55, 98.08]
	10^{-5}	37.27% [35.94, 38.62]	**98.32%** [97.98, 98.66]
	10^{-6}	36.49% [35.17, 37.84]	**99.27%** [99.05, 99.45]
	10^{-7}	36.47% [35.15, 37.82]	**99.27%** [99.06, 99.48]

C Proofs of Negative Results

Proof. Proof of Theorem 1. With $\lambda(\gamma) = e^{\gamma}$, we solve the first-order optimality condition of $\mathcal{P}(\gamma)$ and obtain

$$(x^*(\gamma), y^*(\gamma)) = \left(\frac{1}{2}, \gamma - \frac{1}{2}\right).$$

This gives

$$F(x^*(\gamma), y^*(\gamma); \gamma) = \gamma + \frac{3}{4} \quad \text{and} \quad p(x^*(\gamma), y^*(\gamma)) = e^{-\gamma}.$$

That is, the decay rate I, defined in Assumption 1, is 1. We consider a sequence of initial solutions $\{(x_0(\gamma), y_0(\gamma)) = \left(\frac{\gamma}{4}, 0\right)\}_{\gamma > 0}$. Since

$$\limsup_{\gamma \to \infty} \frac{1}{\gamma} \log p\big((x_0(\gamma), y_0(\gamma))\big) = -\frac{1}{4},$$

this sequence is not a safe solution as defined in Definition 1. Note that these initial solutions are close to the optimal solution in the Euclidean distance, i.e.,

$$\begin{aligned}
&\limsup_{\gamma\to\infty} \frac{1}{\gamma}\|(x_0(\gamma), y_0(\gamma)) - (x^*(\gamma), y^*(\gamma))\| \\
&= \limsup_{\gamma\to\infty} \frac{1}{\gamma}\sqrt{\left(\frac{\gamma}{4} - \frac{1}{2}\right)^2 + \left(\gamma - \frac{1}{2}\right)^2} \\
&= 0.
\end{aligned}$$

We want to show that, with these unsafe starts, the number of iterations to achieve $2-$optimality using gradient descent with a fixed step size grows exponentially in γ.

Fix rarity level $\gamma \geq 40$ and a step size η. Denote by (x_t, y_t) the $t-$iteration that is initialized at $(x_0(\gamma), y_0(\gamma))$. The number of required iterations is defined as the first iterate that reaches the target accuracy, i.e.,

$$\tau := \min_t \{t \geq 0 | F(x_t, y_t; \gamma) \leq 2F^*(\gamma)\}.$$

We consider the following two cases:

Case 1: Large step size $(\eta \geq e^{-\gamma/2})$**.** On the y-coordinate, we have

$$y_1 = y_0 - \eta\nabla_y F(x_0, y_0; \gamma) = \eta e^{3\gamma/4} - \eta.$$

Since $\nabla_y F_(x_t, y_t; \gamma) \leq 1$ for all $x_t, y_t \geq 0$, each iteration reduces y_t by at most η. After t iterations, we have

$$y_t \geq \eta e^{3\gamma/4} - \eta t.$$

Because $2\gamma + 2 > 2F^*(\gamma)$ and $F(x_t, y_t; \gamma) > y_t$ for all $t \geq 0$, the necessary condition for τ is

$$2\gamma + 2 \geq y_\tau \geq \eta e^{3\gamma/4} - \eta\tau.$$

This implies

$$\tau \geq \frac{\eta e^{3\gamma/4} - 2\gamma - 2}{\eta} = \frac{e^{3\gamma/4}}{2} + \frac{\eta e^{3\gamma/4}/2 - 2\gamma - 2}{\eta}.$$

The second term in the RHS is greater than 0 since $\eta \geq e^{-\eta/2}$ and $e^{\gamma/4}/2 \geq 2\gamma + 2$ for all $\gamma \geq 40$. Therefore, τ is at least $\frac{e^{3\gamma/4}}{2}$, growing exponentially.

Case 2: Small step size $(\eta < e^{-\gamma/2})$**.** At iteration τ, we know $2\gamma + 2 > 2F^*(\gamma)$ and $F(x_\tau, y_\tau; \gamma) \geq \inf_{y\in\mathbb{R}} F(x_\tau, y_\tau; \gamma) = F(x_\tau, \gamma - x_\tau; \gamma)$ by the first-order optimality condition. Therefore, we have the necessary condition

$$2\gamma + 2 \geq F(x_\tau, \gamma - x_\tau; \gamma) = x_\tau^2 - x_\tau + \gamma + 1.$$

We can show that this necessary condition, for $\gamma > 40$, implies

$$x_\tau < \frac{\gamma}{5},$$

otherwise we would have $x_\tau^2 - x_\tau + \gamma + 1 \geq x_\tau(x_\tau - 1) + \gamma + 1 \geq \frac{\gamma}{5}\left(\frac{40}{5} - 1\right) + \gamma + 1 > 2\gamma + 2$.

On the $x-$coordinate, we have $\nabla_x F(x_t, y_t; \gamma) \leq 2x_t$ for all $x_t, y_t \geq 0$. For $\gamma > 40$, we know $1 - 2\eta \geq 0$. Therefore,

$$x_{t+1} = x_t - \eta\nabla_x F(x_t, y_t; \gamma) \geq (1 - 2\eta)x_t,$$

and thus

$$x_\tau \geq (1 - 2\eta)^\tau x_0 \geq (1 - 2\tau\eta)x_0.$$

With the necessary condition, we need

$$\frac{\gamma}{5} > x_\tau \geq (1 - 2\tau\eta)\frac{\gamma}{4} \Rightarrow \tau > \frac{1}{10\eta} > \frac{e^{\gamma/2}}{10}.$$

That is, τ grows at an exponential rate of γ.

□

Next, we prove the second negative result.

Proof. Proof of Theorem 2. Fix $\gamma > 0$. Solving the first-order optimality condition for $\mathcal{P}(\gamma)$, we obtain

$$x^*(\gamma) = \gamma \quad \text{and} \quad F(x^*(\gamma); \gamma) = \gamma + 1.$$

Let x_t be the $t-$iteration of a (projected) SGD with a fixed step size η starting at x_0 where

$$x_{t+1} := \max\left\{0, x_t - \eta\left\{G_f(x, \xi) + e^{\gamma} G_p(x_t, \xi)\right\}\right\}.$$

Assume a noiseless estimator $G_f(x, \xi) = 1$ and assume that $G_p(x, \xi)$ is without weak efficiency, that is, $\mathbb{E}\left[\left\|G_p(x, \xi) + e^{-x}\right\|^2\right] \leq e^{-x}$ for all $x \geq 0$. Specifically, we construct the stochastic gradient such that, for x inside the *trap region* $\mathcal{T}_\gamma := (1.5(\gamma+1), 1.6(\gamma+1))$,

$$G_p(x, \xi) = \begin{cases} -e^{-x} + e^{-x/2} & \text{with probability } 0.5 \\ -e^{-x} - e^{-x/2} & \text{with probability } 0.5 \end{cases}.$$

When $x \notin \mathcal{T}_\gamma$, the gradient estimate is noiseless, i.e.,

$$G_p(x, \xi) = -e^{-x} \quad \text{as surely.}$$

Let τ be the stopping time when the SGD reaches $(1 + 0.5)-$optimality, i.e.,

$$\tau := \min_t \left\{t \geq 0 | F(x_t; \gamma) \leq (1.5)(\gamma + 1)\right\}.$$

We have a necessary condition that

$$\gamma - \log\left(1.5(\gamma+1)\right) \leq x_\tau \leq 1.5(\gamma+1). \tag{11}$$

For a sufficiently large γ, we have $\gamma - \log\left(1.5(\gamma+1)\right) > 0$. We consider a starting region S' of size $O(\gamma)$ where

$$S' = [1.6(\gamma+1), 3.3(\gamma+1) + 1].$$

Note that these are considered *safe starts* because $p(x_0(\gamma)) \leq e^{-1.6\gamma}$ and the decay rate of $p(x^*(\gamma))$ is $I = 1$.

We show that for any range of the step size η, we can choose $x_0 \in S'$ such that the sample complexity is exponential in γ. The following is the summary of SGD's behaviors under different step sizes:

1. **Large η:** The iterate is immediately projected to $x_1 = 0$, and subsequentially x_2 would be extremely large. The recovery time takes an exponential number of steps.
2. **Small η:** The iterates progress towards the target region slowly, taking exponential iterations.
3. **Intermediate η:** The iterate lands in the trap region $\mathcal{T}_\gamma$. From there, the stochastic noise either projects the next iterate to 0 or pushes it exponentially far away, ensuring an exponential runtime.

Before we show these results, we prove the following preliminary result when the iterate is at zero and the step size is not small:

Preliminary Result on Recovery Time from Zero with $\eta > 3.4(\gamma+1)e^{-\gamma}$: We show that if an iterate lands at $x_t = 0$ and $\eta > 3.4(\gamma+1)e^{-\gamma}$, then the recovery (to the target accuracy) is exponentially slow. At $x_t = 0$, the gradient is noise-free, so the next iterate is $x_{t+1} = \eta(e^\gamma - 1)$. With sufficiently large γ, $x_{t+1} > 1.6(\gamma+1)$, landing outside the trap. Beyond $x > 1.6(\gamma+1)$, each iteration decreases by at most η as surely. Therefore, in order to be below $1.6(\gamma+1)$ (i.e., the necessary condition for $1.5-$optimality), it needs at least

$$\begin{aligned} \frac{x_{t+1} - 1.6(\gamma+1)}{\eta} &= \frac{\eta(e^\gamma - 1) - 1.6(\gamma+1)}{\eta} \\ &= e^\gamma - 1 - \frac{1.6(\gamma+1)}{3.4(\gamma+1)} e^\gamma = \Omega(e^\gamma) \end{aligned} \tag{12}$$

iterations.

Next, we analyze each case.

Case 1: Large Step Size ($\eta > 1.7(\gamma+1)+1$). We pick $x_0 = 1.6(\gamma+1)+1 \in S'$. For a sufficiently large γ, the update rule gives:

$$x_1 = \max\{0, 1.6(\gamma+1)+1-\eta(1-e^{\gamma-(1.6(\gamma+1)+1)})\} = 0.$$

This requires an exponential number of iterations to reach the target accuracy, by the above preliminary result.

Case 2: Small Step Size ($\eta < e^{-0.1\gamma}$). We pick $x_0 = 1.6(\gamma+1)+1 \in S'$. When $x_t > 1.6(\gamma+1)$, the iterate decreases by at most η. To get below $1.6(\gamma+1)$ (i.e., the necessary condition for $1.5-$optimality), it needs at least

$$\frac{(1.6(\gamma+1)+1)-1.6(\gamma+1)}{\eta} = \frac{1}{\eta} > e^{0.1\gamma}$$

iterations.

Lastly, we consider the remaining possibility when $\eta \in [e^{-0.1\gamma}, 1.7(\gamma+1)+1]$

Case 3: Intermediate Step Size ($\eta \in [e^{-0.1\gamma}, 1.7(\gamma+1)+1]$). We first show that, for any intermediate step size η, there exists $x_0 \in S'$ such that $x_1 \in \mathcal{T}_\gamma$. We consider the following two cases. First, if $1.6(\gamma+1)-\eta(1-e^{\gamma-1.6(\gamma+1)}) > 1.5(\gamma+1)$, then we can pick $x_0 = 1.6(\gamma+1)$ and thus $x_1 \in (1.5(\gamma+1), 1.6(\gamma+1))$. Second, assume that $1.6(\gamma+1)-\eta(1-e^{\gamma-1.6(\gamma+1)}) \le 1.5(\gamma+1)$. We know that

$$\begin{aligned}
&(3.3(\gamma+1)+1)-\eta(1-e^{\gamma-(3.3(\gamma+1)+1)}) \\
&> (3.3(\gamma+1)+1)-(1.7(\gamma+1)+1) \\
&> 1.6(\gamma+1).
\end{aligned}$$

By the Intermediate Value Theorem, there must exist $x_0 \in [1.6(\gamma+1), 3.3(\gamma+1)]$ such that $x_1 = x_0 - \eta\left(1-e^{\gamma-x_0}\right) \in (1.5(\gamma+1), 1.6(\gamma+1))$. Therefore, we can always select x_0 so that the next step falls into the trap region.

We consider the two possibilities for x_2 :

- **Case 3.1:** $G_p(x_1, \xi) = -e^{-x} + e^{-x/2}$. We thus have

  $$\begin{aligned}
  x_2 &= \max\left\{0, x_1 - \eta\left(1-e^{\gamma-x_1}+e^{\gamma-x_1/2}\right)\right\} \\
  &\le \max\left\{0, x_1 + \eta e^{\gamma-x_1} - \eta e^{\gamma-x_1/2}\right\}.
  \end{aligned}$$

  We can upper bound the second term by $1.6(\gamma+1)+(1.7(\gamma+1)+1)e^{-0.5\gamma-1.5} - e^{-0.1\gamma}e^{0.2\gamma-0.8}$. With sufficiently large γ, the above is dominated by the exponential term of γ and thus $x_2 = 0$. Since $\eta \ge e^{-0.1\gamma} > 3.4(\gamma+1)e^{-\gamma}$ for sufficiently large γ, the preliminary result implies an exponentially slow runtime.

- **Case 3.2:** $G_p(x_1, \xi) = -e^{-x} - e^{-x/2}$**.** For sufficiently large γ, we have

  $$\begin{aligned}
  x_2 &= \max\left\{0, x_1 - \eta\left(1-e^{\gamma-x_1}-e^{\gamma-x_1/2}\right)\right\} \\
  &\ge \max\left\{0, 1.5(\gamma+1) + \eta\left(e^{\gamma-(1.6(\gamma+1))/2}-1\right)\right\} \\
  &= 1.5(\gamma+1) + \eta\left(e^{0.2\gamma-0.8}-1\right).
  \end{aligned}$$

  Since $\eta \ge e^{-\gamma/10}$, the term above is dominated by the positive exponential term. Thus, $x_2 > 1.6(\gamma+1)$ when γ is sufficiently large. Beyond this point, gradient descent reduces x_t by at most η. Therefore, for x_τ to be less than $1.6(\gamma+1)$ again, it needs at least

  $$\frac{1.5(\gamma+1)+\eta\left(e^{0.2\gamma-0.8}-1\right)-1.6(\gamma+1)}{\eta} \ge e^{0.1\gamma-0.8}$$

  more iterations when γ is sufficiently large.

Finally, we conclude that, for any fixed step size η and with sufficiently large γ, we can select an initialization $x_0 \in S'$ such that the number of iterations grow in an exponential rate of γ.

□

D Preliminaries for Positive Results

Recall the negative result of Theorem 1. The reason for SGD's poor performance is the non-global smoothness over the domain space. That said, we find that restricting this to a smaller subspace allows us to analyze a locally smooth landscape where SGD can perform nicely.

Our remaining results analyze the performance of SGD over (properly selected) sublevel sets.

Definition 3 ((Enlarged) Initial sublevel set)**.** *For a given rarity level γ, an initial solution $\boldsymbol{x}_0 \in \mathcal{X}$, and a positive number $c \geq 1$, define an initial sublevel set $S(\boldsymbol{x}_0, \gamma, c)$ as*

$$S(\boldsymbol{x}_0, \gamma, c) := \{\boldsymbol{x} \in \mathcal{X} \mid F(\boldsymbol{x}; \gamma) \leq c \cdot F(\boldsymbol{x}_0; \gamma)\}. \tag{13}$$

Under Assumptions 3-4, $F(\cdot; \gamma)$ is convex, continuous and $\mathcal{X} = \mathbb{R}^n$. Therefore, $S(\boldsymbol{x}_0, \gamma, c)$ is a convex set. In the remainder of this paper, we simply write the initial sublevel by S when $(\boldsymbol{x}_0, \gamma, c)$ are fixed and clear. We show two nice properties of this region. As these two properties hold for all $\boldsymbol{x}_0 \in \mathcal{X}$ and $\gamma > 0$, we omit γ, e.g., writing $F := F(;\gamma)$ and $\lambda := \lambda(\gamma)$. First, we derive the Lipschitz smoothness of S.

Lemma 1 (Lipschitz smoothness over sublevel set)**.** *Under Assumptions 3-5, let L_f and $L_{p,2}$ be as defined there. Consider the initial sublevel set $S(\boldsymbol{x}_0, \gamma, c)$. For all $\boldsymbol{x} \in S$, we have*

$$\|\nabla^2 F(\boldsymbol{x})\| \leq L_S$$

where $L_S := L_f + cL_{p,2}F(\boldsymbol{x}_0)$ is the Lipschitz constant of F over S. That is, F is L_S-smooth over S.

Proof. Proof. We know that

$$\begin{aligned}
\sup_{\boldsymbol{x}\in S}\|\nabla^2 F(\boldsymbol{x})\| &= \sup_{\boldsymbol{x}\in S}\|\nabla^2 f(\boldsymbol{x}) + \lambda(\gamma)\nabla^2 p(\boldsymbol{x})\| \\
&\leq \sup_{\boldsymbol{x}\in S}\left\{\|\nabla^2 f(\boldsymbol{x})\| + \lambda(\gamma)\|\nabla^2 p(\boldsymbol{x})\|\right\} \\
&\leq \sup_{\boldsymbol{x}\in S}\|\nabla^2 f(\boldsymbol{x})\| + \sup_{\boldsymbol{x}\in S}\left\{\lambda(\gamma)\|\nabla^2 p(\boldsymbol{x})\|\right\}
\end{aligned}$$

by the triangular inequality. Under Assumptions 3 and 4, we have

$$\sup_{\boldsymbol{x}\in S}\|\nabla^2 F(\boldsymbol{x})\| \leq L_f + \sup_{\boldsymbol{x}\in S}\{\lambda(\gamma)L_{p,2}p(\boldsymbol{x})\}.$$

Under Assumption 5, we have that

$$\sup_{\boldsymbol{x}\in S}\{\lambda(\gamma)L_{p,2}p(\boldsymbol{x})\} \leq L_{p,2}\sup_{\boldsymbol{x}\in S}F(\boldsymbol{x}) \leq cL_{p,2}F(\boldsymbol{x}_0)$$

since $\boldsymbol{x} \in S$. Therefore,

$$\sup_{\boldsymbol{x}\in S}\|\nabla^2 F(\boldsymbol{x})\| \leq L_f + cL_{p,2}F(\boldsymbol{x}_0).$$

□

In addition to the local smoothness over S, SGD can work well over this region because the second moments of the stochastic gradient and the gradient noise are bounded.

Lemma 2 (Stochastic-gradient and stochastic-noise second-moment bounds over sublevel set)**.** *Under Assumptions 2- 5, let $G_f(\mathbf{x}, \xi)$, $G_p(\mathbf{x}, \xi)$, σ_f, σ_p, and $L_{p,1}$ be as defined there. Consider the initial sublevel set $S(\boldsymbol{x}_0, \gamma, c)$. Denote by $G(\boldsymbol{x}, \xi)$ and $\boldsymbol{w}(\boldsymbol{x}, \xi)$, respectively, stochastic gradient estimator and stochastic noise for $\nabla F(\boldsymbol{x})$, where*

$$\begin{aligned}
G(\boldsymbol{x}, \xi) &:= G_f(\mathbf{x}, \xi) + \lambda(\gamma)G_p(\mathbf{x}, \xi), \\
\boldsymbol{w}(\boldsymbol{x}, \xi) &:= G(\boldsymbol{x}, \xi) - \nabla F(\boldsymbol{x}).
\end{aligned}$$

Then, for all $\boldsymbol{x} \in S$, we have

$$\mathbb{E}\left[\|\boldsymbol{w}(\boldsymbol{x}, \xi)\|^2\right] \leq \sigma^2$$

and

$$\mathbb{E}\left[\|G(\boldsymbol{x}, \xi)\|^2\right] \leq 2\left(\|\nabla F(\boldsymbol{x})\|^2 + \sigma^2\right)$$

where $\sigma = \sigma_f + c\sigma_p L_{p,1}F(\boldsymbol{x}_0)$.

Proof. Proof. By Cauchy inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\boldsymbol{w}(\boldsymbol{x},\xi)\|^2\right] \\
&= \mathbb{E}\left[\|(G_f(\boldsymbol{x},\xi)-\nabla f(\boldsymbol{x}))+\lambda\left(G_p(\boldsymbol{x},\xi)-\nabla p(\boldsymbol{x})\right)\|^2\right] \\
&\leq \left(\sqrt{\mathbb{E}\left[\|G_f(\boldsymbol{x},\xi)-\nabla f(\boldsymbol{x})\|^2\right]}\right. \\
&\left.+\ \lambda\sqrt{\mathbb{E}\left[\|G_p(\boldsymbol{x},\xi)-\nabla p(\boldsymbol{x})\|^2\right]}\right)^2.
\end{aligned}$$

Under Assumption 2, we have

$$\mathbb{E}\left[\|\boldsymbol{w}(\boldsymbol{x},\xi)\|^2\right] \leq \left(\sigma_f+\lambda\sigma_p\|\nabla p(\boldsymbol{x})\|\right)^2.$$

Using Assumptions 3 and 5 and that $\boldsymbol{x}\in S$, we have

$$\lambda\|\nabla p(\boldsymbol{x})\| \leq L_{p,1}\lambda p(\boldsymbol{x}) \leq L_{p,1}F(\boldsymbol{x}) \leq cL_{p,1}F(\boldsymbol{x}_0).$$

Therefore,

$$\mathbb{E}\left[\|\boldsymbol{w}(\boldsymbol{x},\xi)\|^2\right] \leq \left(\sigma_f+c\sigma_pL_{p,1}F(\boldsymbol{x}_0)\right)^2.$$

Furthermore, we have

$$\begin{aligned}
\mathbb{E}\left[\|G(\boldsymbol{x},\xi)\|^2\right] &= \mathbb{E}\left[\|w(\boldsymbol{x},\xi)+\nabla F(\boldsymbol{x})\|^2\right] \\
&\leq 2\mathbb{E}\left[\|w(\boldsymbol{x},\xi)\|^2+\|\nabla F(\boldsymbol{x})\|^2\right] \\
&\leq 2\left(\sigma^2+\|\nabla F(\boldsymbol{x})\|^2\right)
\end{aligned}$$

by Young's inequality.

□

E Proofs of Positive Results for SGD with Constant Step Size

We first prove a descent lemma for SGD with constant step size over a sublevel set S. This descent property is possible under sufficiently small step size and second moment of gradient noise. Throughout Appendix E, we set $c=2$ for the enlarged initial sublevel set.

Lemma 3 (Descent lemma for SGD with constant step size)**.** *Under Assumptions 2-5, let μ be as defined there. Choose the target accuracy level $\epsilon>0$ and the confidence level $\kappa\in(0,1)$. Let $S:=S(\boldsymbol{x}_0,\gamma,2)$ with L_S and σ as given in Lemmas 1-2. Consider SGD with a constant step size $\eta\leq\frac{1}{L_S}$ and denote $\boldsymbol{x}_t$ as the $t-$iterate. If $F(\boldsymbol{x}_0)>(1+\epsilon)F^*$ and $\boldsymbol{x}_t\in int(S)$ and $\sigma^2\leq\frac{\kappa\mu\epsilon F^*}{4}$, then $\boldsymbol{x}_{t+1}\in int(S)$ with probability at least $1-\kappa$; moreover, $F(\boldsymbol{x}_{t+1})\leq(1+\epsilon)F^*$ or $F(\boldsymbol{x}_{t+1})\leq F(\boldsymbol{x}_t)-\frac{3\eta\epsilon\mu F^*}{8}$.*

Proof. Proof. Let $\boldsymbol{w}_t$ be the stochastic noise at $\boldsymbol{x}_t$, defined as in Lemma 2. By Markov's inequality, we have

$$\mathbb{P}\left(\|\boldsymbol{w}_t\|^2\geq\frac{\mu\epsilon F^*}{4}\right)\leq\frac{\sigma^2}{\mu\epsilon F^*/4}\leq\kappa. \tag{14}$$

That is, with probability at least $1-\kappa$, $\|\boldsymbol{w}_t\|^2<\frac{\mu\epsilon F^*}{4}$. In what follows, we condition on this event.

The SGD updates as

$$\boldsymbol{x}_{t+1}:=\boldsymbol{x}_t-\eta G(\boldsymbol{x}_t,\xi)=\boldsymbol{x}_t-\eta\left(\nabla F(\boldsymbol{x}_t)+\boldsymbol{w}_t\right).$$

We claim that, if the noise is bounded and $\boldsymbol{x}_t\in int(S)$, then $\boldsymbol{x}_{t+1}\in int(S)$.

We define a line segment between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$:

$$\boldsymbol{x}(\tau):=\boldsymbol{x}_t+\tau\big(\boldsymbol{x}_{t+1}-\boldsymbol{x}_t\big),\qquad\tau\in[0,1].$$

Suppose a contradiction that $\boldsymbol{x}_{t+1} \notin S$. Let $\tau^* := \inf\{\tau \in [0,1] : \boldsymbol{x}(\tau) \in \partial S\}$. Because the line segment between $\boldsymbol{x}_t$ and $\boldsymbol{x}(\tau)$ is contained in S, which is L_S-smooth by Lemma 1, we have a smoothness quadratic upper bound:

$$\begin{aligned}
&F\big(\boldsymbol{x}(\tau^*)\big) \\
&\leq F(\boldsymbol{x}_t) + \langle \nabla F(\boldsymbol{x}_t), \boldsymbol{x}(\tau^*) - \boldsymbol{x}_t \rangle + \frac{L_S}{2}\|\boldsymbol{x}(\tau^*) - \boldsymbol{x}_t\|^2 \\
&= F(\boldsymbol{x}_t) - \eta\tau^* \left(1 - \frac{\eta\tau^* L_S}{2}\right) \|\nabla F(\boldsymbol{x}_t)\|^2 \\
&\quad - \eta\tau^*(1 - \eta\tau^* L_S)\, \nabla F(\boldsymbol{x}_t)^\top \boldsymbol{w}_t + \frac{\eta^2(\tau^*)^2 L_S}{2}\, \|\boldsymbol{w}_t\|^2.
\end{aligned}$$

By Cauchy inequality, we have

$$\begin{aligned}
&F\big(\boldsymbol{x}(\tau^*)\big) \\
&\leq F(\boldsymbol{x}_t) - \eta\tau^* \left(1 - \frac{\eta\tau^* L_S}{2}\right) \|\nabla F(\boldsymbol{x}_t)\|^2 \\
&\quad + \eta\tau^*(1 - \eta\tau^* L_S)\, \|\nabla F(\boldsymbol{x}_t)\| \|\boldsymbol{w}_t\| + \frac{\eta^2(\tau^*)^2 L_S}{2}\, \|\boldsymbol{w}_t\|^2. \qquad (15)
\end{aligned}$$

We then consider the following two possibilities:

Case 1: $\|\nabla F(\boldsymbol{x}_t)\|^2 < \epsilon\mu F^*$**.** Since $\tau^* \leq 1,\ \eta \leq \frac{1}{L_S},\ 1 - \frac{\eta\tau^* L_S}{2} \geq 0,$ and $\sup_{y\in\mathbb{R}} y(1 - yL_S) \leq \frac{1}{4L_S}$, we can refine (15) as

$$F\big(\boldsymbol{x}(\tau^*)\big) \leq F(\boldsymbol{x}_t) + \frac{1}{4L_S}\|\nabla F(\boldsymbol{x}_t)\|\|\boldsymbol{w}_t\| + \frac{1}{2L_S}\|\boldsymbol{w}_t\|^2$$

With the bounded stochastic noise, we have

$$\begin{aligned}
F\big(\boldsymbol{x}(\tau^*)\big) &\leq F(\boldsymbol{x}_t) + \frac{1}{4L_S}\sqrt{\epsilon\mu F^*}\sqrt{\frac{\mu\epsilon F^*}{4}} + \frac{1}{2L_S}\left(\frac{\mu\epsilon F^*}{4}\right) \\
&= F(\boldsymbol{x}_t) + \frac{\mu\epsilon F^*}{4L_S}.
\end{aligned}$$

Since the strong convexity μ is smaller than the Lipschitz constant, i.e., $\mu \leq L_S$, we have

$$F\big(\boldsymbol{x}(\tau^*)\big) \leq F(\boldsymbol{x}_t) + \frac{\epsilon F^*}{4}.$$

Using Assumption 4 that F is $\mu-$strongly convex, we have

$$2\mu\,(F(\boldsymbol{x}_t) - F^*) \leq \|\nabla F(\boldsymbol{x}_t)\|^2 \leq \epsilon\mu F^*.$$

That is, $F(\boldsymbol{x}_t) \leq \left(1 + \frac{\epsilon}{2}\right) F^*$. Therefore, combining the above, we get

$$F\big(\boldsymbol{x}(\tau^*)\big) \leq \left(1 + \frac{3\epsilon}{4}\right) F^* < F(\boldsymbol{x}_0).$$

This leads to a contradiction as $\boldsymbol{x}(\tau^*)$ still resides inside the sublevel set S. Therefore, $\boldsymbol{x}_{t+1} \in int(S)$. With the same derivation as above, we can also conclude that

$$F\big(\boldsymbol{x}_{t+1}\big) \leq \left(1 + \frac{3\epsilon}{4}\right) F^* < (1 + \epsilon)\, F^*.$$

Case 2: $\|\nabla F(\boldsymbol{x}_t)\|^2 \geq \epsilon\mu F^*$**.** We have that $\|\boldsymbol{w}_t\| \leq \|\nabla F(\boldsymbol{x}_t)\|/2$, and thus (15) implies

$$\begin{aligned}
&F\big(\boldsymbol{x}(\tau^*)\big) \\
&\leq F(\boldsymbol{x}_t) - \eta\tau^* \left(1 - \frac{\eta\tau^* L_S}{2}\right) \|\nabla F(\boldsymbol{x}_t)\|^2 \\
&\quad + \frac{\eta\tau^*(1 - \eta\tau^* L_S)}{2}\|\nabla F(\boldsymbol{x}_t)\|^2 + \frac{\eta^2(\tau^*)^2 L_S}{8}\, \|\nabla F(\boldsymbol{x}_t)\|^2 \\
&= F(\boldsymbol{x}_t) - \frac{\eta\tau^*}{2}\|\nabla F(\boldsymbol{x}_t)\|^2 + \frac{\eta^2\,(\tau^*)^2\, L_S}{8}\|\nabla F(\boldsymbol{x}_t)\|^2.
\end{aligned}$$

By the continuity of F and $\boldsymbol{x}_t \in int(S)$, we know $\tau^* > 0$. With $\tau^* \leq 1$ and $\eta \leq \frac{1}{L_S}$, we have

$$F\big(\boldsymbol{x}(\tau^*)\big) \leq F(\boldsymbol{x}_t) - \frac{3\eta\tau^*}{8}\|\nabla F(\boldsymbol{x}_t)\|^2 < 2F(\boldsymbol{x}_0).$$

However, $F(\boldsymbol{x}(\tau^*)) = 2F(\boldsymbol{x}_0)$ as τ^* is the time that the segment cross the sublevel set. This leads to a contradiction, and thus $\boldsymbol{x}_{t+1} \in int(S)$. With the similar analysis, we can conclude that

$$F\big(\boldsymbol{x}_{t+1}\big) \leq F(\boldsymbol{x}_t) - \frac{3\eta}{8}\|\nabla F(\boldsymbol{x}_t)\|^2 = F(\boldsymbol{x}_t) - \frac{3\eta\epsilon\mu F^*}{8}.$$

Therefore, with probability at least $1-\kappa$, $\boldsymbol{x}_{t+1} \in int(S)$; moreover, $F(\boldsymbol{x}_{t+1}) \leq (1+\epsilon)F^*$ or $F(\boldsymbol{x}_{t+1}) \leq F(\boldsymbol{x}_t) - \frac{3\eta\epsilon\mu F^*}{8}$.

□

Next, we use this lemma to show the upper bound on the number of gradient steps.

Lemma 4 (Upper bound on iteration numbers for SGD with constant step size)**.** *Under Assumptions 2-5, let μ be as defined there. Choose the target accuracy level $\epsilon > 0$ and the confidence level $\kappa \in (0,1)$. Let $S := S(\boldsymbol{x}_0, \gamma, 2)$ with L_S and σ as given in Lemmas 1-2. Consider SGD with a constant step size $\eta \in [\frac{1}{10L_S}, \frac{1}{L_S}]$ and denote $\boldsymbol{x}_t$ as the $t-$iterate. If $F(\boldsymbol{x}_0) > (1+\epsilon)F^*$ and we perform SGD for T iterations where*

$$T \geq \left\lceil \frac{80L_S F(\boldsymbol{x}_0)}{3\epsilon\mu F^*} \right\rceil,$$

and $\sigma^2 \leq \frac{\kappa\mu\epsilon F^*}{4T}$, then $F(\boldsymbol{x}_T) \leq (1+\epsilon)F^*$ with probability at least $1-\kappa$.

Proof. Proof. From Lemma 3, if $\boldsymbol{x}_t \in int(S)$, then with probability at least $1 - \frac{\kappa}{T}$, $\boldsymbol{x}_{t+1} \in int(S)$ and its objective either reaches the target accuracy or it diminishes by at least $\frac{3\eta\epsilon\mu F^*}{8} \geq \frac{3\epsilon\mu F^*}{80L_S}$. Therefore, we have that $\mathbb{P}\left(\boldsymbol{x}_0 \in int(S)\right) = 1$ by construction of S and

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \in int(S)\right) \\
&= \mathbb{P}\left(\boldsymbol{x}_0 \in int(S)\right)\mathbb{P}\left(\boldsymbol{x}_1 | \boldsymbol{x}_0 \in int(S)\right)..\mathbb{P}\left(\boldsymbol{x}_T | \boldsymbol{x}_{T-1} \in int(S)\right) \\
&\geq \left(1 - \frac{\kappa}{T}\right)^T \\
&\geq 1 - \kappa.
\end{aligned}$$

Therefore, with probability at least $1-\kappa$, $F(\boldsymbol{x}_T) \leq (1+\epsilon)F^*$. Otherwise, the objective at $\boldsymbol{x}_T$ must decrease from $F(\boldsymbol{x}_0)$ by at least $T \cdot \left(\frac{3\epsilon\mu F^*}{80L_S}\right) \geq F(\boldsymbol{x}_0)$, becoming below zero, which results in a contradiction.

□

Before we go to the main result, we prove that the initial objective value of a safe start grows at a sub-exponential rate.

Lemma 5 (Sub-exponential growth of safe-start objective value)**.** *Suppose Assumptions 1- 5 hold. Consider a sequence of safe starts $\{\boldsymbol{x}_0(\gamma)\}_{\gamma>0}$. We have*

$$\limsup_{\gamma\to\infty} \frac{1}{\gamma}\log F(\boldsymbol{x}_0(\gamma);\gamma) \leq 0.$$

Proof. Proof. Since f is L_f-smooth, the smoothness quadratic upper bound gives

$$f(\boldsymbol{x}_0(\gamma)) \leq f(\tilde{\mathbf{x}}^*) + \frac{L_f}{2}\|\boldsymbol{x}_0(\gamma) - \tilde{\mathbf{x}}^*\|^2 = \frac{L_f}{2}\|\boldsymbol{x}_0(\gamma) - \tilde{\mathbf{x}}^*\|^2$$

where $\tilde{\mathbf{x}}^*$ is defined as in Assumption 5. That is, by the definition of safe start,

$$\begin{aligned}
\limsup_{\gamma\to\infty} \frac{1}{\gamma}\log f(\boldsymbol{x}_0(\gamma)) &\leq \limsup_{\gamma\to\infty} \frac{1}{\gamma}\log\left(\frac{L_f}{2}\|\boldsymbol{x}_0(\gamma) - \tilde{\mathbf{x}}^*\|^2\right) \\
&= 0.
\end{aligned}$$

Moreover, by Assumption 1 and the definition of safe start, we have

$$\begin{aligned}
&\limsup_{\gamma\to\infty} \frac{1}{\gamma} \log(\lambda(\gamma)p(\boldsymbol{x}_0(\gamma))) \\
&= \limsup_{\gamma\to\infty} \frac{1}{\gamma} \log \lambda(\gamma) \\
&+ \limsup_{\gamma\to\infty} \frac{1}{\gamma} \log p(\boldsymbol{x}_0(\gamma)) \leq 0.
\end{aligned}$$

Combining the above gives

$$\begin{aligned}
&\limsup_{\gamma\to\infty} \frac{1}{\gamma} \log F(\boldsymbol{x}_0(\gamma);\gamma) \\
&\leq \limsup_{\gamma\to\infty} \frac{1}{\gamma} \log \left\{2\max\left\{f(\boldsymbol{x}_0(\gamma)), \lambda(\gamma)p(\boldsymbol{x}_0(\gamma))\right\}\right\} \\
&= \max\left\{\limsup_{\gamma\to\infty} \frac{1}{\gamma} \log f(\boldsymbol{x}_0(\gamma)), \right. \\
&\qquad \left. \limsup_{\gamma\to\infty} \frac{1}{\gamma} \log(\lambda(\gamma)p(\boldsymbol{x}_0(\gamma)))\right\} \\
&\leq 0.
\end{aligned}$$

□

Now, we are ready to prove the sub-exponential result of the safe start for SGD with constant step size.

Lemma 6 (Sub-exponential sample complexity for SGD with constant step size and safe start)**.** *Under Assumptions 1- 5, let μ, σ_f, σ_p, and $L_{p,1}$ be as defined there. Choose the target accuracy level $\epsilon > 0$ and the confidence level $\kappa \in (0,1)$. Consider a sequence of safe starts $\{\boldsymbol{x}_0(\gamma)\}_{\gamma>0}$ where $F(\boldsymbol{x}_0(\gamma);\gamma) > (1+\epsilon)F^*(\gamma)$. For each γ, we derive the local Lipschitz constant $L_S(\gamma)$ for the sublevel set $S(\boldsymbol{x}_0(\gamma),\gamma,2)$ as in Lemma 1. We perform SGD with a minibatch size of $B(\gamma)$ i.i.d. samples for $T(\gamma)$ steps, totaling in gradient evaluations of $T(\gamma)\cdot B(\gamma)$. We choose a constant step size $\eta(\gamma) \in [\frac{1}{10L_S(\gamma)}, \frac{1}{L_S(\gamma)}]$ and*

$$\begin{aligned}
T(\gamma) &= \left\lceil \frac{80L_S(\gamma)F(\boldsymbol{x}_0(\gamma);\gamma)}{3\epsilon\mu F^*(\gamma)} \right\rceil \\
B(\gamma) &= \left\lceil \frac{4T(\gamma)(\sigma_f + 2\sigma_p L_{p,1}F(\boldsymbol{x}_0(\gamma);\gamma))^2}{\kappa\mu\epsilon F^*(\gamma)} \right\rceil.
\end{aligned}$$

Then, with probability at least $1-\kappa$, we have $F(\boldsymbol{x}_{T(\gamma)};\gamma) \leq (1+\epsilon)F^*(\gamma)$. Furthermore, the required sample complexity grows sub-exponentially, i.e.,

$$\limsup_{\gamma\to\infty} \frac{1}{\gamma} \log\left(T(\gamma)B(\gamma)\right) \leq 0.$$

Proof. Proof. Fix $\gamma > 0$. We derive the new stochastic-noise second-moment bound $\sigma(\gamma)$ as in Lemma 2 for this $B(\gamma)-$minibatch SGD. Specifically, we take unbiased gradient estimators

$$\begin{aligned}
\bar{G}_f(\boldsymbol{x}) &:= \frac{1}{B(\gamma)} \sum_{i=1}^{B(\gamma)} G_f(\boldsymbol{x},\xi_i) \\
\bar{G}_p(\boldsymbol{x}) &:= \frac{1}{B(\gamma)} \sum_{i=1}^{B(\gamma)} G_p(\boldsymbol{x},\xi_i)
\end{aligned}$$

where ξ_i are i.i.d. We then improve from σ_f and σ_p to $\sigma_f/\sqrt{B(\gamma)}$ and $\sigma_p/\sqrt{B(\gamma)}$, respectively. The new bound becomes

$$\sigma(\gamma) := \frac{\sigma_f + 2\sigma_p L_{p,1} F(\boldsymbol{x}_0(\gamma);\gamma)}{\sqrt{B(\gamma)}}.$$

Taking $B(\gamma) = \left\lceil \frac{4T(\gamma)(\sigma_f + 2\sigma_p L_{p,1} F(\boldsymbol{x}_0(\gamma);\gamma))^2}{\kappa\mu\epsilon F^*(\gamma)} \right\rceil$, we then have

$$\sigma(\gamma) \leq \sqrt{\frac{\kappa\mu\epsilon F^*(\gamma)}{4T(\gamma)}}.$$

Using Lemma 4, we can conclude that $F(\boldsymbol{x}_{T(\gamma)};\gamma) \leq (1+\epsilon)F^*(\gamma)$ with probability at least $1-\kappa$.

So, the required sample complexity is at most $T(\gamma)B(\gamma)$. To show the sub-exponential growth, it is sufficient to show that both $T(\gamma)$ and $B(\gamma)$ also grow at a sub-exponential rate in rarity level.

Omitting all the constants, we have

$$\begin{aligned}
&\limsup_{\gamma\to\infty} \frac{1}{\gamma} \log T(\gamma) \\
&= \limsup_{\gamma\to\infty} \frac{1}{\gamma} \left(\log\left(L_S(\gamma)\right) + \log\left(F(\boldsymbol{x}_0(\gamma);\gamma)\right) - \log\left(F^*(\gamma)\right)\right) \\
&\leq \limsup_{\gamma\to\infty} \frac{1}{\gamma} \left(\log\left(L_S(\gamma)\right) + \log\left(F(\boldsymbol{x}_0(\gamma);\gamma)\right)\right)
\end{aligned}$$

since $F^*(\gamma)$ is non-decreasing as we increase the penalty. Since $L_S(\gamma) := L_f + 2L_{p,2}F(\boldsymbol{x}_0(\gamma);\gamma)$, the exponential growth rate is determined by the term $F(\boldsymbol{x}_0(\gamma);\gamma)$. That is,

$$\begin{aligned}
&\limsup_{\gamma\to\infty} \frac{1}{\gamma} \log T(\gamma) \\
&\leq \limsup_{\gamma\to\infty} \frac{1}{\gamma} \log\left(F(\boldsymbol{x}_0(\gamma);\gamma)\right)
\end{aligned}$$

which grows at a sub-exponential rate by Lemma 5. Similarly, we bound the exponential growth rate of $B(\gamma)$ by

$$\begin{aligned}
&\limsup_{\gamma\to\infty} \frac{1}{\gamma} \log B(\gamma) \\
&\leq \limsup_{\gamma\to\infty} \frac{1}{\gamma} \log T(\gamma) + \limsup_{\gamma\to\infty} \frac{1}{\gamma} \log F(\boldsymbol{x}_0(\gamma);\gamma)^2 \\
&\leq 0
\end{aligned}$$

by Lemma 5 and the previous result on $T(\gamma)$. Therefore, the sample complexity only grows at a sub-exponential rate in the rarity level.

□

F Proofs of Positive Results for SGD with Harmonic Step Size

We write the optimality gap at the t-iteration and rarity level γ by $\Delta_t(\gamma) := F(\boldsymbol{x}_t;\gamma) - F^*(\gamma)$. We omit γ when it is clear that we consider a particular rarity level.

To analyze the behavior of SGD with harmonic step size, we leverage the same preliminary results of the sublevel set as in Lemmas 1 and 2. We introduce the *expected optimality gap over the sublevel set* for a fixed initial solution $\boldsymbol{x}_0$, a rarity level γ, and an enlargement scale c.

Definition 4 (Expected optimality gap over sublevel set)**.** *Consider the sublevel set $S := S(\boldsymbol{x}_0, \gamma, c)$. Let $\boldsymbol{x}_t$ be the t-iterate of SGD. With T iterations, we define the expected optimality gap over S as*

$$\mathrm{EOG}(T) := \mathbb{E}\left[\Delta_T \mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_T \in int(S)\,\}\right]$$

where $int(S)$ is the interior of set S and $\mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\}$ is a (random) indicator which is equal to 1 when all the iterates up to time T remain in the sublevel set interior; otherwise, it is equal to 0. Moreover, this expectation is taken over the entire trajectory given $\boldsymbol{x}_0$.

Next, we derive the upper bound on the expected gap in the next step. This is analogous to the smoothness quadratic upper bound for analyzing first-order methods over a globally smooth region. Throughout Appendix F, we set the enlargement scale to $c = 2$.

Lemma 7 (Upper bound on the next-step expected optimality gap over sublevel set)**.** *Under Assumptions 2-5, let μ be as defined there. Let $S := S(\boldsymbol{x}_0, \gamma, 2)$ with L_S and σ as given in Lemmas 1-2. In each iteration $t \geq 0$, we perform SGD with a harmonic step size η_t. If $\eta_t \leq \frac{1}{2L_S}$, we have*

$$\mathrm{EOG}(t+1) \leq (1 - \eta_t \mu)\,\mathrm{EOG}(t) + \eta_t^2 L_S \sigma^2.$$

Proof. Proof. Our goal is to prove that

$$\begin{aligned}
&\mathbb{E}\left[\Delta_{t+1}\mathbf{1}\{\,\mathbf{x}_0, \ldots, \mathbf{x}_{t+1} \in int(S)\,\}\right] \\
&\quad \leq (1 - \eta_t\mu)\mathbb{E}\left[\Delta_t \mathbf{1}\{\,\mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\}\right] \\
&\qquad + \eta_t^2 L_S \sigma^2.
\end{aligned}$$

It is sufficient to simply show that, for every $(\boldsymbol{x_0}, \ldots \boldsymbol{x}_t) \in \mathcal{X}^{t+1}$, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_{t+1}}\left[\Delta_{t+1}\mathbf{1}\{\,\mathbf{x}_0, \ldots, \mathbf{x}_{t+1} \in int(S)\,\}\,|\boldsymbol{x}_0, ..., \boldsymbol{x}_t\right] \\
&\quad \leq (1 - \eta_t\mu)\Delta_t \mathbf{1}\{\,\mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} \\
&\qquad + \eta_t^2 L_S \sigma^2.
\end{aligned}$$

The above is trivial when $\mathbf{1}\{\,\mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 0$.

In what follows, we condition on $\mathbf{1}\{\,\mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 1$. Denote by $G(\boldsymbol{x}_t, \xi_t)$ the stochastic gradient for $\nabla F(\boldsymbol{x}_t)$. We want to establish that

$$\begin{aligned}
&\Delta_{t+1}\mathbf{1}\{\,\mathbf{x}_0, \ldots, \mathbf{x}_{t+1} \in int(S)\,\} \\
&\leq \underbrace{\left(\Delta_t - \eta_t \nabla F(\boldsymbol{x}_t)^\top G(\boldsymbol{x}_t, \xi_t) + \frac{\eta_t^2 L_S}{2}\|G(\boldsymbol{x}_t, \xi_t)\|^2\right)}_{(\dagger)}
\end{aligned} \tag{16}$$

as surely. We consider the following two cases.

Case 1: $\mathbf{1}\{\,\mathbf{x}_0, \ldots, \mathbf{x}_{t+1} \in int(S)\,\} = 1$. Since $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$ lies inside the locally smooth region S with Lipschitz constant L_S and $\eta \leq \frac{1}{2L_S}$, we then have the smoothness quadratic upper bound

$$\Delta_{t+1} \leq \Delta_t - \eta_t \nabla F(\boldsymbol{x}_t)^\top G(\boldsymbol{x}_t, \xi_t) + \frac{\eta_t^2 L_S}{2}\|G(\boldsymbol{x}_t, \xi_t)\|^2,$$

proving (16).

Case 2: $\mathbf{1}\{\,\mathbf{x}_0, \ldots, \mathbf{x}_{t+1} \in int(S)\,\} = 0$. To establish (16), we just need to show that $(\dagger) \geq 0$. Since F is continuous, we can define $\boldsymbol{x}' := \boldsymbol{x}_t - \tau\eta_t G(\boldsymbol{x}_t, \xi_t)$ as the intersection between the line segment of $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$ and the boundary of the sublevel set. Here $\tau \in (0, 1]$. Using the smoothness quadratic upper bound, as $\boldsymbol{x}'$ and $\boldsymbol{x}_t$ lie in S, we have

$$\begin{aligned}
&F(\boldsymbol{x}') - F(\boldsymbol{x}_t) \\
&\leq \eta_t\tau\left(-\nabla F(\boldsymbol{x}_t)^\top G(\boldsymbol{x}_t, \xi_t) + \frac{\tau\eta_t L_S}{2}\|G(\boldsymbol{x}_t, \xi_t)\|^2\right).
\end{aligned}$$

Since $\boldsymbol{x}'$ is at the boundary of the sublevel set and $\boldsymbol{x}_t$ is in the interior of S, we have that $F(\boldsymbol{x}') > F(\boldsymbol{x}_t)$. Therefore,

$$\eta_t\tau\left(-\nabla F(\boldsymbol{x}_t)^\top G(\boldsymbol{x}_t, \xi_t) + \frac{\tau\eta_t L_S}{2}\|G(\boldsymbol{x}_t, \xi_t)\|^2\right) \geq 0.$$

Since $\eta_t\tau > 0$ and $1 \geq \tau$, we have

$$-\nabla F(\boldsymbol{x}_t)^\top G(\boldsymbol{x}_t, \xi_t) + \frac{\eta_t L_S}{2}\|G(\boldsymbol{x}_t, \xi_t)\|^2 \geq 0.$$

The above and the fact that $\Delta_t \geq 0$ complete the proof that $(\dagger) \geq 0$.

As (16) holds, we take the expectation over ξ_t and achieve

$$\begin{aligned}
&\mathbb{E}_{\xi_t}\left[\Delta_{t+1}\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_{t+1}\in int(S)\,\}\,|\boldsymbol{x}_0,\ldots,\boldsymbol{x}_t\in int(S)\right]\\
&\leq \mathbb{E}_{\xi_t}\left[\Delta_t-\eta_t\nabla F(\boldsymbol{x}_t)^\top G(\boldsymbol{x}_t,\xi_t)+\frac{\eta_t^2 L_S}{2}\|G(\boldsymbol{x}_t,\xi_t)\|^2\right]\\
&\leq \Delta_t-\eta_t\|\nabla F(\boldsymbol{x}_t)\|^2+\eta_t^2 L_S\left(\sigma^2+\|\nabla F(\boldsymbol{x}_t)\|^2\right).
\end{aligned}$$

since $G(\boldsymbol{x}_t,\xi_t)$ is unbiased and its second moment is bounded as in Lemma 2. As $\eta_t \leq \frac{1}{2L_S}$, we have

$$\begin{aligned}
&\mathbb{E}_{\xi_t}\left[\Delta_{t+1}\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_{t+1}\in int(S)\,\}\,|\boldsymbol{x}_0,\ldots,\boldsymbol{x}_t\in int(S)\right]\\
&\leq \Delta_t-\frac{\eta_t}{2}\|\nabla F(\boldsymbol{x}_t)\|^2+\eta_t^2 L_S\sigma^2
\end{aligned}$$

By strong convexity of F, we have $\|\nabla F(\boldsymbol{x}_t)\|^2 \geq 2\mu\Delta_t$ and thus

$$\begin{aligned}
&\mathbb{E}_{\xi_t}\left[\Delta_{t+1}\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_{t+1}\in int(S)\,\}\,|\boldsymbol{x}_0,\ldots,\boldsymbol{x}_t\in int(S)\right]\\
&\leq \Delta_t\left(1-\eta_t\mu\right)+\eta_t^2 L_S\sigma^2
\end{aligned}$$

completing the proof. □

Next, we show that, with a properly selected harmonic step size, the expected gap over S decays at a sublinear rate.

Lemma 8 (Sublinear decay of the expected optimality gap over sublevel set)**.** *Under Assumptions 2-5, let μ be as defined there. Let $S := S(\boldsymbol{x}_0,\gamma,2)$ with L_S and σ as given in Lemmas 1-2. In each iteration $t \geq 0$, we perform SGD with a harmonic step size $\eta_t = \frac{\alpha}{K+t}$ where $\alpha \geq \frac{2}{\mu}$ and $K \geq 2\alpha L_S$. Then,*

$$\mathrm{EOG}(t) \leq \frac{H}{K+t}$$

where $H = \max\left\{K\Delta_0, \alpha^2 L_S\sigma^2\right\}$.

Proof. Proof. We prove this by induction.

Basic Step: We have that

$$\begin{aligned}
\mathrm{EOG}(0) &= \mathbb{E}\left[\Delta_0\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_0\in int(S)\,\}\right]\\
&= \Delta_0 \leq \frac{H}{K}.
\end{aligned}$$

Inductive Step: Assume there exists $t \geq 0$ such that

$$\mathrm{EOG}(t) \leq \frac{H}{K+t}.$$

Given that $\eta_t = \frac{\alpha}{K+t} \leq \frac{1}{2L_S}$, Lemma 7 implies that

$$\begin{aligned}
\mathrm{EOG}(t+1) &\leq (1-\eta_t\mu)\,\mathrm{EOG}(t)+\eta_t^2 L_S\sigma^2\\
&= \left(1-\frac{\alpha\mu}{K+t}\right)\mathrm{EOG}(t)+\frac{\alpha^2 L_S\sigma^2}{(K+t)^2}.
\end{aligned}$$

Because $\mu\alpha \geq 2$ and $H \geq \alpha^2 L_S\sigma^2$, we have

$$\mathrm{EOG}(t+1) \leq \mathrm{EOG}(t)\left(1-\frac{2}{K+t}\right)+\frac{H}{(K+t)^2}.$$

By the inductive hypothesis, we have

$$\begin{aligned}\mathrm{EOG}(t+1) &\leq \left(\frac{H}{K+t}\right)\left(1-\frac{2}{K+t}\right)+\frac{H}{(K+t)^2}\\ &= \frac{(K+t-2)H}{(K+t)^2}+\frac{H}{(K+t)^2}\\ &= \frac{(K+t-1)H}{(K+t)^2}\\ &\leq \frac{H}{K+t+1},\end{aligned}$$

completing this inductive proof. Therefore, the expected optimality gap decays at a sublinear rate. □

The decay result of the expected optimality gap alone is not meaningful enough. This term, for instance, might decay because of the probability of iterates escaping the sublevel set grows quickly. Therefore, we need an additional result that can upper bound the probability of escaping the sublevel set.

Lemma 9 (Upper bound on first-time escape probability from sublevel set)**.** *Under Assumptions 2-5, let μ be as defined there. Let $S := S(\boldsymbol{x}_0, \gamma, 2)$ with L_S and σ as given in Lemmas 1-2. In each iteration $t \geq 0$, we perform SGD with a harmonic step size $\eta_t = \frac{\alpha}{K+t}$ where $\alpha \geq \frac{2}{\mu}$ and $K \geq 2\alpha L_S$. Then, the escape probability for the first time at iteration $t+1$ is upper bounded, i.e., $\mathbb{P}\left(\mathbf{1}\{\mathbf{x}_0, \ldots, \mathbf{x}_{t+1} \in int(S)\} < \mathbf{1}\{\mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\}\right) \leq \frac{\sigma\mathcal{U}}{2+t}$ where $\mathcal{U} := \frac{\alpha\left(2\sqrt{L_S\gamma_0}+\sigma\right)}{\Delta_0} + \frac{H\left(2\sqrt{2\mu\Delta_0}+\sigma\right)}{\mu\Delta_0^2}$ and $H = \max\left\{K\Delta_0, \alpha^2 L_S \sigma^2\right\}$.*

Proof. Proof. We can write the escape probability from the interior of S for the first time at iteration $t+1$ as

$$\mathbb{P}\left(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0) \text{ and } \mathbf{1}\{\mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\} = 1\right).$$

Denote by $G(\boldsymbol{x}_t, \xi_t)$ and $\boldsymbol{w}_t$ the stochastic gradient and stochastic noise at iteration t for $\nabla F(\boldsymbol{x}_t)$ where $G(\boldsymbol{x}_t, \xi_t) = \nabla F(\boldsymbol{x}_t) + \boldsymbol{w}_t$. Suppose $F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0)$ and $\mathbf{1}\{\mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\} = 1$. Using the similar technique to the previous result, we write $\boldsymbol{x'}$ as the first point where the line segment between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$ cut the boundary of S. We write $\boldsymbol{x}' := \boldsymbol{x} - \tau\eta_t G(\boldsymbol{x}_t, \xi_t)$ where $\tau \in (0, 1]$. Then, we apply the smoothness quadratic upper bound on the sublevel, i.e.,

$$\begin{aligned}2F(\boldsymbol{x}_0) &= F(\boldsymbol{x}')\\ &\leq F(\boldsymbol{x}_t) - \eta_t\tau\nabla F(\boldsymbol{x}_t)^\top G(\boldsymbol{x}_t, \xi_t)\\ &\quad + \frac{\eta_t^2\tau^2 L_S}{2}\|G(\boldsymbol{x}_t, \xi_t)\|^2\end{aligned}$$

Since $F(\boldsymbol{x}_t) < 2F(\boldsymbol{x}_0)$ and $\tau \in (0, 1]$, we have

$$\begin{aligned}&2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t)\\ &\leq \eta_t\tau\left(-\nabla F(\boldsymbol{x}_t)^\top G(\boldsymbol{x}_t, \xi_t) + \frac{\eta_t\tau L_S}{2}\|G(\boldsymbol{x}_t, \xi_t)\|^2\right)\\ &\leq \eta_t\left(-\nabla F(\boldsymbol{x}_t)^\top G(\boldsymbol{x}_t, \xi_t) + \frac{\eta_t L_S}{2}\|G(\boldsymbol{x}_t, \xi_t)\|^2\right).\end{aligned}$$

Therefore, we can upper bound the (conditional on $\mathbf{1}\{\mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\} = 1$) escape probability by

$$\begin{aligned}&\mathbb{P}\left(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0)\right)\\ &\leq \mathbb{P}\Bigg(2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t)\\ &\qquad \leq \eta_t\left(-\nabla F(\boldsymbol{x}_t)^\top G(\boldsymbol{x}_t, \xi_t) + \frac{\eta_t L_S}{2}\|G(\boldsymbol{x}_t, \xi_t)\|^2\right)\Bigg).\end{aligned}$$

Note that we omit the probability condition for readability. Using inequality $\|G(\boldsymbol{x}_t, \xi_t)\|^2 \leq 2\left(\|\nabla F(\boldsymbol{x}_t)\|^2 + \|\boldsymbol{w}_t\|^2\right)$, we have

$$\begin{aligned}
&\mathbb{P}\left(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0)\right) \\
&\leq \mathbb{P}\Bigg(2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) \\
&\qquad \leq \eta_t\Big(-\|\nabla F(\boldsymbol{x}_t)\|^2 - \nabla F(\boldsymbol{x}_t)^\top \boldsymbol{w}_t \\
&\qquad\quad + \eta_t L_S\left(\|\nabla F(\boldsymbol{x}_t)\|^2 + \|\boldsymbol{w}_t\|^2\right)\Big)\Bigg).
\end{aligned}$$

By $\eta_t L_S \leq \frac{1}{2}$ and Cauchy inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0)\right) \\
&\leq \mathbb{P}\Bigg(2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) \\
&\qquad \leq \eta_t\left(-\|\nabla F(\boldsymbol{x}_t)\|^2/2 + \|\nabla F(\boldsymbol{x}_t)\|\|\boldsymbol{w}_t\| + \|\boldsymbol{w}_t\|^2/2\right)\Bigg) \\
&\leq \mathbb{P}\Bigg(2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) + \eta_t\|\nabla F(\boldsymbol{x}_t)\|^2/2 \\
&\qquad \leq \eta_t\left(\|\nabla F(\boldsymbol{x}_t)\|\|\boldsymbol{w}_t\| + \|\boldsymbol{w}_t\|^2/2\right)\Bigg).
\end{aligned}$$

By Markov's inequality and the bounds on the stochastic noise in Lemma 2, we have

$$\begin{aligned}
&\mathbb{P}\left(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0) \;\middle|\; \mathbf{1}\{\,\mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 1\right) \\
&\leq \frac{\eta_t \mathbb{E}\left[\|\nabla F(\boldsymbol{x}_t)\|\|\boldsymbol{w}_t\| + \|\boldsymbol{w}_t\|^2/2\right]}{2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) + \eta_t\|\nabla F(\boldsymbol{x}_t)\|^2/2} \\
&\leq \frac{\eta_t\sigma\left(\|\nabla F(\boldsymbol{x}_t)\| + \sigma/2\right)}{2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) + \eta_t\|\nabla F(\boldsymbol{x}_t)\|^2/2}
\end{aligned}$$

We note that this upper bound captures the essence of the escape probability. Basically, we $\boldsymbol{x}_t$ is deep inside the interior set, the term $2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t)$ in the denominator will be large. On the other hand, if $\boldsymbol{x}_t$ is near the boundary, the gradient will be large by strong convexity. We separate these two cases by whether the $2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) \geq F(\boldsymbol{x}_0)$. We use these behaviors to bound the escape probability further and consider the two cases.

Case 1: $2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) \geq F(\boldsymbol{x}_0)$**.** This is equivalent to $\Delta_t \leq \Delta_0$. We thus bound the escape probability by

$$\begin{aligned}
&\mathbb{P}\Bigg(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0) \\
&\qquad , \mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 1, 2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) \geq F(\boldsymbol{x}_0) \Bigg) \\
&\leq \mathbb{P}\Bigg(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0) \\
&\qquad \Big| \, \mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 1, 2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) \geq F(\boldsymbol{x}_0) \Bigg) \\
&\leq \frac{\eta_t \sigma \left(\|\nabla F(\boldsymbol{x}_t)\| + \sigma/2 \right)}{2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) + \eta_t \|\nabla F(\boldsymbol{x}_t)\|^2/2} \\
&\leq \frac{\eta_t \sigma \left(2\|\nabla F(\boldsymbol{x}_t)\| + \sigma \right)}{2\Delta_0} \\
&\leq \frac{\eta_t \sigma \left(2\sqrt{2L_S \gamma_0} + \sigma \right)}{2\Delta_0}.
\end{aligned}$$

Note that we use $2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) \geq F(\boldsymbol{x}_0) \geq \Delta_0$. The last inequality is true by the fact that$\boldsymbol{x}_t$ is in the L_S-smooth, so $\|\nabla F(\boldsymbol{x}_t)\|^2 \leq 2L_S \Delta_t \leq 2L_S \gamma_0$.

Case 2: $2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) < F(\boldsymbol{x}_0)$**.** This is equivalent to $\Delta_t > \Delta_0$. We thus bound the escape probability by

$$\begin{aligned}
&\mathbb{P}\Bigg(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0) \\
&\qquad , \mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 1, 2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) < F(\boldsymbol{x}_0) \Bigg) \\
&\leq \mathbb{P}\Bigg(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0) \\
&\qquad \Big| \, \mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 1, 2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) < F(\boldsymbol{x}_0) \Bigg) \\
&\qquad \cdot \underbrace{\mathbb{P}\left(\mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 1, 2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) < F(\boldsymbol{x}_0) \right)}_{(\dagger)}
\end{aligned}$$

The first probability term is upper bounded by

$$\begin{aligned}
&\mathbb{P}\Bigg(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0) \\
&\qquad \Big| \, \mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 1, 2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) < F(\boldsymbol{x}_0) \Bigg) \\
&\leq \frac{\eta_t \sigma \left(\|\nabla F(\boldsymbol{x}_t)\| + \sigma/2 \right)}{\eta_t \|\nabla F(\boldsymbol{x}_t)\|^2/2} \\
&= \frac{2\sigma}{\|\nabla F(\boldsymbol{x}_t)\|} + \frac{\sigma^2}{\|\nabla F(\boldsymbol{x}_t)\|^2} \\
&\leq \frac{\sigma\sqrt{2}}{\sqrt{\mu\Delta_0}} + \frac{\sigma^2}{2\mu\Delta_0}
\end{aligned}$$

since $\boldsymbol{x}_t \in int(S)$ and $2F(\boldsymbol{x}_0) - F(\boldsymbol{x}_t) > 0$. The last inequality is true by strong convexity that $\|\nabla F(\boldsymbol{x}_t)\|^2 \geq 2\Delta_t \geq 2\mu\Delta_0$. Lastly, we use Markov's inequality to bound $(\dagger)$

$$\begin{aligned}(\dagger) &= \mathbb{P}\left(\mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 1 \text{ and } \Delta_t > \Delta_0\right) \\ &= \mathbb{P}\left(\Delta_t \mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} > \Delta_0\right) \\ &\leq \mathbb{E}\left[\Delta_t \mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\}\right] / \Delta_0 \\ &\leq \frac{2H}{(K+t)\Delta_0}\end{aligned}$$

by Lemma 8. Therefore, we bound the escape probability at time $t+1$ by

$$\begin{aligned}&\mathbb{P}\left(F(\boldsymbol{x}_{t+1}) \geq 2F(\boldsymbol{x}_0) \text{ and } \mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_t \in int(S)\,\} = 1\right) \\ &\leq \frac{\eta_t \sigma\left(2\sqrt{L_S\gamma_0} + \sigma\right)}{\Delta_0} + \left(\frac{2H}{(K+t)\Delta_0}\right)\left(\frac{\sigma\sqrt{2}}{\sqrt{\mu\Delta_0}} + \frac{\sigma^2}{2\mu\Delta_0}\right) \\ &= \frac{\alpha\sigma\left(2\sqrt{L_S\gamma_0} + \sigma\right)}{(K+t)\Delta_0} + \left(\frac{2H}{(K+t)\Delta_0}\right)\left(\frac{\sigma\sqrt{2}}{\sqrt{\mu\Delta_0}} + \frac{\sigma^2}{2\mu\Delta_0}\right) \\ &= \frac{\sigma}{(K+t)}\left(\frac{\alpha\left(2\sqrt{L_S\gamma_0} + \sigma\right)}{\Delta_0} + \frac{H\left(2\sqrt{2\mu\Delta_0} + \sigma\right)}{\mu\Delta_0^2}\right) \\ &= \frac{\sigma\mathcal{U}}{K+t}.\end{aligned}$$

We know that $K \geq 2\alpha L_S \geq \frac{4L_S}{\mu} > 4$ since the smoothness L_S must be at least the strong convexity μ. Therefore, the escape probability is at most $\frac{\sigma\mathcal{U}}{2+t}$. □

Now, we are ready to prove that, if the second moment of the stochastic noise σ over the sublevel set is sufficiently small, the iterate of SGD will reach the target accuracy with high probability. We prove this result for a fixed rarity level first.

Lemma 10 (Upper bound on iteration numbers of SGD with harmonic step size)**.** *Under Assumptions 2-5, let μ be as defined there. Choose the target accuracy level $\epsilon > 0$ and the confidence level $\kappa \in (0,1)$. Let $S := S(\boldsymbol{x}_0, \gamma, 2)$ with L_S and σ as given in Lemmas 1-2. In each iteration $t \geq 0$, we perform SGD with a harmonic step size $\eta_t = \frac{\alpha}{K+t}$ where $\alpha \geq \frac{2}{\mu}$ and $K \geq 2\alpha L_S$. We then perform SGD for T iterations where $T \geq \left\lceil \frac{2K\Delta_0}{\kappa\epsilon F^*} \right\rceil$ and the gradient estimates are independent across iterations. If $\sigma^2 \leq \min\left\{ \frac{K\Delta_0}{\alpha^2 L_S}, \Delta_0, \frac{\kappa^2\Delta_0}{4\log^2(1+T)M^2} \right\}$ where $M = 2\alpha\sqrt{L_S} + \alpha + \frac{K\left(2\sqrt{2\mu}+1\right)}{\mu}$ then $F(\boldsymbol{x}_T) \leq (1+\epsilon)F^*$ with probability at least $1-\kappa$.*

Proof. Proof.

By Lemma 8 and that $\sigma^2 \leq \frac{K\Delta_0}{\alpha^2 L_S}$, we upper bound the optimality gap over the sublevel set at $\boldsymbol{x}_T$ by

$$\begin{aligned}\mathrm{EOG}(T) &\leq \frac{\max\left\{K\Delta_0, \alpha^2 L_S \sigma^2\right\}}{K+T} \\ &\leq \frac{K\Delta_0}{T} \\ &\leq \frac{\kappa\epsilon F^*}{2}.\end{aligned}$$

By Markov's inequality, we have

$$\mathbb{P}\left(\Delta_T \mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_T \in int(S)\,\} \geq \epsilon F^*\right) \leq \frac{\mathrm{EOG}(T)}{\epsilon F^*}.$$

That is, the above inequalities give

$$\mathbb{P}\left(\mathbf{1}\{\, \mathbf{x}_0, \ldots, \mathbf{x}_T \in int(S)\,\} = 1 \text{ and } \Delta_T \geq \epsilon F^*\right) \leq \frac{\kappa}{2}.$$

By Lemma 9, we can upper bound the escape probability before iteration $T+1$ by

$$\begin{aligned}
&\mathbb{P}\left(\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_T \in int(S)\,\}=0\right)\\
&=\sum_{t=0}^{T-1}\mathbb{P}\left(\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_{t+1} \in int(S)\,\}<\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_t \in int(S)\,\}\right)\\
&\leq \sigma\mathcal{U}\sum_{t=0}^{T-1}\frac{1}{2+t}\\
&\leq \sigma\mathcal{U}\sum_{t=2}^{T+1}\frac{1}{t}\\
&\leq \sigma\mathcal{U}\log\left(T+1\right)
\end{aligned}$$

where $\mathcal{U}=\frac{\alpha\left(2\sqrt{L_S\gamma_0}+\sigma\right)}{\Delta_0}+\frac{H\left(2\sqrt{2\mu\Delta_0}+\sigma\right)}{\mu\Delta_0^2}$ and $H=\max\left\{K\Delta_0,\alpha^2L_S\sigma^2\right\}$. We upper bound $\mathcal{U}$ as $\sigma^2\leq\Delta_0$ and $H=K\Delta_0$ by

$$\begin{aligned}
\mathcal{U}&\leq\frac{\alpha\left(2\sqrt{L_S\gamma_0}+\sqrt{\Delta_0}\right)}{\Delta_0}+\frac{K\left(2\sqrt{2\mu\Delta_0}+\sqrt{\Delta_0}\right)}{\mu\Delta_0}\\
&=\frac{\left(2\alpha\sqrt{L_S}+\alpha+\frac{K\left(2\sqrt{2\mu}+1\right)}{\mu}\right)}{\sqrt{\Delta_0}}.
\end{aligned}$$

Applying this to the bound on the escape probability using the bound assumption on σ^2 gives

$$\mathbb{P}\left(\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_T \in int(S)\,\}=0\right)\leq\frac{\kappa}{2}.$$

Combing all the results above, we have

$$\begin{aligned}
&\mathbb{P}\left(\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_T \in int(S)\,\}=1 \text{ and } \Delta_T<\epsilon F^*\right)\\
&\quad=\mathbb{P}\left(\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_T \in int(S)\,\}=1\right)\\
&\qquad-\mathbb{P}\left(\mathbf{1}\{\,\mathbf{x}_0,\ldots,\mathbf{x}_T \in int(S)\,\}=1 \text{ and } \Delta_T\geq\epsilon F^*\right)\\
&\quad\geq\left(1-\frac{\kappa}{2}\right)-\frac{\kappa}{2}=1-\kappa.
\end{aligned}$$

That is, the probability of reaching $\epsilon-$accuracy at time $\boldsymbol{x}_T$ is at least $1-\kappa$. □

We now prove the sub-exponential result for SGD with harmonic step size. The outline of this is similar to the proof in Lemma 6.

Lemma 11 (Sub-exponential sample complexity for SGD with harmonic step size and safe start)**.** *Under Assumptions 1- 5, let μ,σ_f,σ_p, and $L_{p,1}$ be as defined there. Choose the target accuracy level $\epsilon>0$ and the confidence level $\kappa\in(0,1)$. Consider a sequence of safe starts $\{\boldsymbol{x}_0(\gamma)\}_{\gamma>0}$ where $F(\boldsymbol{x}_0(\gamma);\gamma)>(1+\epsilon)F^*(\gamma)$. For each γ, we derive the local Lipschitz constant $L_S(\gamma)$ for the sublevel set $S(\boldsymbol{x}_0(\gamma),\gamma,2)$ as in Lemma 1. We perform SGD with a minibatch size of $B(\gamma)$ i.i.d. samples for $T(\gamma)$ steps, totaling in gradient evaluations of $T(\gamma)\cdot B(\gamma)$. We choose a harmonic step size a harmonic step size $\eta_t(\gamma)=\frac{\alpha(\gamma)}{K(\gamma)+t}$ where $\alpha(\gamma)\in\left[\frac{2}{\mu},\frac{20}{\mu}\right]$, $K(\gamma)\in\left[\frac{40L_S(\gamma)}{\mu},\frac{400L_S(\gamma)}{\mu}\right]$, and*

$$\begin{aligned}
T(\gamma)&=\left\lceil\frac{2K(\gamma)\Delta_0(\gamma)}{\kappa\epsilon F^*(\gamma)}\right\rceil\\
B(\gamma)&=\left\lceil\frac{(\sigma_0(\gamma))^2\left(1+\frac{\mu\alpha^2(\gamma)}{40}+\frac{4\log^2(T(\gamma)+1)M(\gamma)^2}{\kappa^2}\right)}{\Delta_0(\gamma)}\right\rceil
\end{aligned}$$

where

$$M(\gamma):=2\alpha(\gamma)\sqrt{L_S(\gamma)}+\alpha(\gamma)+\frac{K(\gamma)\left(2\sqrt{2\mu}+1\right)}{\mu}.$$

and

$$\sigma_0(\gamma) := \sigma_f + 2\sigma_p L_{p,1} F(\boldsymbol{x}_0(\gamma);\gamma).$$

Then, with probability at least $1-\kappa$, we have $F(\boldsymbol{x}_{T(\gamma)};\gamma) \le (1+\epsilon)F^*(\gamma)$. Furthermore, the required sample complexity grows sub-exponentially, i.e.,

$$\limsup_{\gamma\to\infty} \frac{1}{\gamma} \log\left(T(\gamma)B(\gamma)\right) \le 0.$$

Proof. Proof.

Similar to the proof in Lemma 6, the bound of stochastic noise $\sigma(\gamma)$ over the sublevel set $S(\boldsymbol{x}_0(\gamma),\gamma,2)$ with $B(\gamma)$-minibatch SGD is

$$\sigma(\gamma) = \frac{\sigma_0(\gamma)}{\sqrt{B(\gamma)}}.$$

Taking the selected minibatch size $B(\gamma)$ and using the fact $\frac{L_S(\gamma)}{K(\gamma)} \ge \frac{\mu}{40}$, we have that

$$\begin{aligned}
&\sigma^2(\gamma)\\
&\le \frac{\Delta_0(\gamma)}{1+\frac{\alpha^2(\gamma)L_S(\gamma)}{K(\gamma)}+\frac{2\log^2(T(\gamma)+1)(M(\gamma))^2}{\kappa^2}}\\
&\le \min\left\{\Delta_0(\gamma), \frac{K(\gamma)\Delta_0(\gamma)}{\alpha^2(\gamma)L_S(\gamma)}, \frac{\kappa^2\Delta_0(\gamma)}{4\log^2\left(T_\gamma+1\right)M(\gamma)^2}\right\}.
\end{aligned}$$

We also know that $\alpha(\gamma) \ge \frac{2}{\mu}$ and $K(\gamma) \ge 2\alpha(\gamma)L_S(\gamma)$. Therefore, Lemma 10 suggests that $F(\boldsymbol{x}_{T(\gamma)};\gamma) - F^*(\gamma) \le \epsilon F^*(\gamma)$ with probability at least $1-\kappa$.

Next, we show the sub-exponential rate of the sample complexity by showing that both $T(\gamma)$ and $B(\gamma)$ grow at a sub-exponential rate.

We know that

$$\begin{aligned}
&\limsup_{\gamma\to\infty} \frac{1}{\gamma}\log T(\gamma)\\
&= \limsup_{\gamma\to\infty} \frac{1}{\gamma}\log\left(\frac{2K(\gamma)\Delta_0(\gamma)}{\kappa\epsilon F^*(\gamma)}\right)\\
&\le \limsup_{\gamma\to\infty} \frac{1}{\gamma}\log\left(\frac{2\left(\frac{400L_S(\gamma)}{\mu}\right)F(\boldsymbol{x}_0(\gamma);\gamma)}{\kappa\epsilon F^*(\gamma)}\right)
\end{aligned}$$

We also know that $F^*(\gamma)$ is non-decreasing as we increase the penalty term. Moreover, $L_S(\gamma) = L_f + 2L_{p,2}F(\boldsymbol{x}_0(\gamma);\gamma)$ and $F(\boldsymbol{x}_0(\gamma);\gamma)$ both grow at a sub-exponential rate by Lemma 5. We conclude that the number of iterations $T(\gamma)$ also grows at a sub-exponential rate.

Next, we want the minibatch size $B(\gamma)$ to grows at a sub-exponential rate. Since $\Delta_0(\gamma) > \epsilon F^*(\gamma)$, which is non-decreasing, it is sufficient to show that $F(\boldsymbol{x}_0(\gamma);\gamma)$, $K(\gamma)$, $L_S(\gamma)$ and $\log(T(\gamma)+1)$ grow at a sub-exponential rate. Since $K(\gamma) \le \frac{400L_S(\gamma)}{\mu}$, the first three terms grow at a sub-exponential rate. Moreover, since $T(\gamma)$ grows at a sub-exponential rate by the previous result, so does it logarithmic. Therefore, we conclude that $B(\gamma)$ also grows at a sub-exponential rate. □