WALL-CLOCK COMPLEXITY FOR ZEROTH-ORDER OPTIMIZATION WITH TUNABLE ORACLE FIDELITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Zeroth-order (black-box) optimization is commonly applied when gradients are unavailable and objective evaluations rely on expensive simulations. In many such applications, the oracle fidelity is tunable: higher-accuracy queries reduce noise but incur higher computational costs. To capture this trade-off, we study an accuracy-aware wall-clock model where each query with fidelity δ has a cost $c(\delta)$, and we minimize the total time $T_{\text{total}} \propto \sum_{k=1}^{N} c(\delta_k)$, subject to a target accuracy constraint. We show how the choice of oracle type, noise model, and optimization scheme induces explicit wall-clock-optimal choices for the algorithmic parameters. For instance, we demonstrate that accelerated methods can be wall-clock inferior to non-accelerated schemes. Furthermore, we characterize the conditions under which a constant fidelity strategy is optimal in the Big-O sense. Our framework provides a unified methodology to translate convergence guarantees into practical fidelity and batching recommendations. We apply this framework to several optimization schemes under various noise settings, including Tsybakov and adversarial noise.

1 INTRODUCTION

We study the joint design of zeroth-order algorithms and their function-value oracles. The oracle's accuracy is controllable: more accurate evaluations require more computation. The objective is to choose the optimization method, gradient estimator, and fidelity schedule that minimize the computational cost of reaching a target accuracy.

These choices are coupled. Finite differences and randomized smoothing transform function-value noise into gradient-estimation bias and variance. The optimization method determines how these errors accumulate. A method with fewer iterations may therefore require more expensive evaluations, and the preferred method can change once evaluation cost is included.

Tunable accuracy arises in Monte Carlo simulations, internal search procedures, and stationary-distribution estimation (Shirobokov et al., 2020; Hoffman et al., 2022; Bogolubsky et al., 2016). Prior work has developed optimal schedules for tunable inexact oracles (Van Dessel & Glineur, 2024) and cost guarantees for certified multi-fidelity global optimization (de Montbrun & Gerchinovitz, 2024). We address the interaction between evaluation accuracy, gradient estimation, and optimization dynamics through a new wall-clock complexity framework for ZO methods. **Our contributions are as follows:**

- **A wall-clock framework for tunable-fidelity ZO.** We introduce a framework for jointly designing the iteration budget N and the oracle fidelity schedule $\{\delta_k\}$. It connects the function-value noise model, gradient estimator, and optimization method to total computational cost. For polynomial oracle costs $c(\delta) \propto \delta^{-\gamma}$, this yields explicit rules for method selection and parameter tuning.
- **A master allocation result turning convergence bounds into wall-clock-optimal designs**. We isolate a structural property—fidelity separability—shared by various methods, and give closed-form optimal schedules in one step (Prop. 1). We also pin down exactly when a uniform schedule is already optimal up to constants.
- **The role of γ depends on the noise model.** Under adversarial noise, γ selects the algorithm (acceleration can hurt once $\gamma \geq 1$) but doesn't tune parameters within it. Under

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Tsybakov noise, the opposite holds: γ tunes the parameters, separating standard tuning from overbatching.

- **An intermediate gradient method (IGM)**. We propose a new first-order method (Alg. 1) interpolating between GM and FGM, and prove its convergence under inexact gradients with time-varying oracle inexactness (Thm. 1). IGM is what enables the application of Proposition 1 to the adversarial setting (Prop. 5).
- **End-to-end recipes.** We illustrate and empirically validate the framework.

Paper organization. Sec. 2 introduces the wall-clock model for tunable-fidelity zeroth-order oracles and formulates the fidelity-allocation problem. Sec. 3 explains how standard ZO convergence guarantees can be converted into the fidelity-separable form. Sec. 4 and 5 illustrate the framework under adversarial and Tsybakov noise, respectively. Sec. 6 discusses the main takeaways and limitations. All missing proofs and additional experimental details are deferred to the appendix.

Notations. For an integer $N \geq 1$, we write $[N] := \{1, \dots, N\}$. The Euclidean norm is denoted by $\|\cdot\|$, and $\langle \cdot, \cdot \rangle$ denotes the inner product. We use $\mathcal{F}_k$ for the filtration generated by the algorithm up to iteration k. Expression $a \asymp_{c_1,\dots,c_m} b$ means that the equivalence constants may depend only on $c_1, \dots, c_m$. The terms $O(\cdot)$ and $\tilde{O}(\cdot)$ suppress constant factors and logarithmic factors, respectively.

2 Wall-clock complexity

We consider the problem of minimizing an objective function $\min_{x\in\Omega} f(x)$, where $f : \Omega \to \mathbb{R}$. For the results analyzed here, we take $\Omega = \mathbb{R}^d$. We assume that the algorithm accesses f through a noisy black-box oracle: upon receiving a query x, the oracle returns an estimate corrupted by an error ξ,

$$\hat{f}_\delta(x) = f(x) + \xi_\delta(x). \tag{1}$$

Throughout the paper, we use the term oracle exclusively for the noisy function-value oracle $\hat{f}_\delta(x)$. Zeroth-order methods construct a gradient estimator g_k from one or more queries to this oracle; the gradient estimator itself is not an oracle.

Assumption 1 (Tunable function-value oracle)**.** For every query point x and fidelity level $\delta > 0$, the oracle returns $\hat{f}_\delta(x)$ as defined in 1. The fidelity level δ is chosen by the algorithm for each query, with smaller δ corresponding to higher oracle accuracy. The precise dependence of $\xi_\delta(x)$ on δ is specified by the noise model below.

To formalize the precise dependence of $\xi(x)$ on δ, let $\mathcal{F}_k$ denote the filtration representing the algorithm's history up to step k, and let the current iterate x_k be measurable with respect to $\mathcal{F}_k$. We consider two widely used models for the oracle inexactness that fit this framework: stochastic and deterministic noise.

Definition 1 (Tsybakov noise)**.** The noise satisfies $\mathbb{E}[\xi_\delta^2 \mid \mathcal{F}_k] \leq \delta^2$. If the algorithm queries x_k multiple times, the respective noise realizations are mutually independent conditionally on $\mathcal{F}_k$.

Definition 2 (Adversarial noise)**.** The noise is determined by a deterministic (but unknown) function of the queried point bounded by the fidelity level, $\xi_\delta = \xi_\delta(x_k)$ and $|\xi_\delta(x_k)| \leq \delta$.

Crucially, we assume a direct trade-off between accuracy and computation: a smaller δ yields less noise but incurs a higher per-query cost $c(\delta)$. Let $\{(x_k, \delta_k)\}_{k\in[N]}$ be a sequence of queries performed by the algorithm. The wall-clock time is

$$T_{\text{total}}(N, \{\delta_k\}_{k\in[N]}) := \sum_{k=1}^{N} c(\delta_k). \tag{2}$$

Remark 1 (Batching). Let $g_k(x_k; w_{k,i})$ denote the i-th realization of a gradient estimator constructed from one or more queries to the noisy function-value oracle equation 1, where $w_{k,i}$ collects the estimator-specific randomness. For a batch size $B \geq 1$, the batched gradient estimate at iteration k is defined as $g_k^{(B)} := \frac{1}{B}\sum_{i=1}^{B} g_k(x_k; w_{k,i})$, where, conditionally on x_k, the random variables $\{w_{k,i}\}_{i=1}^{B}$ are independent and identically distributed.

The averaging reduces variance by $1/B$ but leaves bias unchanged, while wall-clock time scales as $T_{\text{total}} = B\sum_{k=1}^{N} c(\delta_k)$. Our master analysis treats B as fixed, but case studies (Sec. 4 and 5) optimize $(N, B, \{\delta_k\})$, showing phase transitions between $B = 1$ and $B \gg 1$.

Choice of $c(\delta)$. We impose an assumption on how the computational cost of a single query scales.

Assumption 2 (Polynomial fidelity cost)**.** A query to the function-value oracle at fidelity level $\delta > 0$ incurs computational cost $c(\delta) = C\,\delta^{-\gamma}$, for some constants $C > 0$ and $\gamma > 0$. Equivalently, throughout the complexity analysis we write $c(\delta) \asymp \delta^{-\gamma}$. Thus, higher-fidelity queries (smaller δ) are more computationally expensive.

Three regimes motivate the polynomial model. If $c(\delta) = \Theta(\ln(1/\delta))$, the fidelity affects the wall-clock complexity only through logarithmic factors. If $c(\delta) = e^{\Theta(1/\delta)}$), the optimizer remains close to the coarsest fidelity compatible with the target accuracy. If $c(\delta) = \Theta(\delta^{-\gamma})$, there is an explicit interior trade-off, and the phase transitions, e.g., a Monte Carlo estimator with M samples achieves $\delta \sim M^{-1/2}$ via the central limit theorem, yielding $c(\delta) \sim \delta^{-2}$. Sec. 4 and 5 study the effect.

Wall-clock optimization problem. To connect T_{total} with the algorithm's choice, we use a generic convergence metric Gap_N. For instance, Gap_N can be $\mathbb{E}[f(x_N) - f(x^*)]$ for convex objectives, or $\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}[\|\nabla f(x_k)\|^2]$ for non-convex ones. Usually Gap_N is bounded by a deterministic function $\mathcal{E}$ depending on the iteration count N, the fidelities $\{\delta_k\}_k$, and the algorithmic hyperparameters Θ,

$$\text{Gap}_N \le \mathcal{E}\big(N, \{\delta_k\}_{k\in[N]}, \Theta\big). \tag{3}$$

An example is ZO-GD under the adversarial noise model (Def. 2). If $\delta_k = \delta$, the gap $\text{Gap}_N := f(x_N) - f^*$ is bounded by $\mathcal{E}(N, \delta, \Theta) = \frac{c_1}{N} + c_2\delta^2$, with constants c_1, c_2 depending on Θ (e.g., step size and Lipschitz constant).

One achieves accuracy $\mathcal{E}\big(N, \{\delta_k\}_{k\in[N]}, \Theta\big) \le \varepsilon$ through various parameter configurations resulting in a different cumulative wall-clock time T_{total}. So, we formulate the following optimization problem,

$$\min_{N, \{\delta_k\}_{k\in[N]}} T_{\text{total}}(N, \{\delta_k\}_{k\in[N]}) \quad \text{subject to} \quad \mathcal{E}(N, \{\delta_k\}_{k\in[N]}, \Theta) \le \varepsilon.$$

2.1 Optimal Fidelity Allocation

Standard ZO analyses derive error bounds $\mathcal{E}(N, \{\delta_k\}, \Theta)$ assuming uniform fidelity ($\delta_k = \delta$). *Sec. 3 provides a methodology to extend these bounds to time-varying δ_k*, revealing a common structural property we term *fidelity-separability*.

Definition 3 (Fidelity-separable bound)**.** We say that the error bound $\mathcal{E}(N, \{\delta_k\}_{k\in[N]}, \Theta)$ is *fidelity-separable* if, for any N and Θ, it can be written as

$$\mathcal{E}(N, \{\delta_k\}_{k\in[N]}, \Theta) = \mathcal{E}_0(N, \Theta) + \sum_{k=1}^{N} \alpha_k(N, \Theta)\,\delta_k^p,$$

with $\mathcal{E}_0(\cdot)$ and $\alpha_k(\cdot)$ depending on the assumptions on f and the optimization method.

Remark 2. Many algorithms admit the decomposition of Def. 3 either directly or through a finite sum of fidelity-separable terms with different exponents: $\mathcal{E}(\cdot) = \mathcal{E}_0(\cdot) + \sum_{p\in\mathcal{P}}\sum_{k=1}^{N}\alpha_{k,p}\delta_k^p$. We refer to each $\sum_{k=1}^{N}\alpha_{k,p}\delta_k^p$ as a *fidelity channel*. In such cases, the wall-clock analysis reduces to determining which channel governs the complexity in the high-precision regime.

The following result provides the closed-form solution for the optimal schedule and the resulting time complexity. Appx. D contains all proofs for this section.

Proposition 1. *Fix N, $p > 0$, and $\gamma > 0$. After rescaling time, set $C = 1$ in the polynomial fidelity cost, so that $c(\delta) = \delta^{-\gamma}$ for $\delta > 0$. Let $\mathcal{E}$ be as in Def. 3, with fidelities $\delta_i > 0$. Assume that $\alpha_i(N, \Theta) > 0$ for all $i \in [N]$, and define the remaining fidelity budget $\eta_N := \varepsilon - \mathcal{E}_0(N, \Theta)$. If $\eta_N > 0$, then the optimal fidelity allocation $\delta_i^*(N)$ and the total wall-clock time $T_{\text{total}}(N)$ are*

$$A_N := \sum_{j=1}^{N} \alpha_j^{\frac{\gamma}{p+\gamma}}(N, \Theta), \quad \delta_i^*(N) = \left(\frac{\eta_N}{A_N}\right)^{\frac{1}{p}} \alpha_i^{-\frac{1}{p+\gamma}}(N, \Theta), \quad T_{\text{total}}(N) = \eta_N^{-\frac{\gamma}{p}} A_N^{\frac{p+\gamma}{p}}. \tag{4}$$

Remark 3. In practice, one may require $c(\delta) = \max\{1, \delta^{-\gamma}\}$ to prevent the per-query cost from vanishing. However, this does not alter the asymptotic scaling of T_{total} (see Lem. 1).

Prop. 1 assumes that δ_k can be reset at every iteration at no extra cost. In practice, changing the fidelity often incurs overhead (re-seeding a Monte Carlo simulator, rebuilding an inner solver, recalibrating a physical oracle), making piecewise-constant or uniform schedules preferable. Cor. 2 bounds the wall-clock penalty of restricting to a uniform schedule; analysis of equation 5 shows that the penalty is at most constant-order whenever the coefficients $\{\alpha_j\}$ are mildly varying.

High-precision regime. The large-N assumption is natural in the high-precision regime: as $\varepsilon \to 0$, the constraint $\mathcal{E}_0(N, \Theta) < \varepsilon$ with $\mathcal{E}_0(N, \Theta) \asymp N^{-\beta}$ forces $N \gtrsim \varepsilon^{-1/\beta} \to \infty$.

Corollary 1 (Uniform versus nonuniform fidelity)**.** *Under Prop. 1, the optimal uniform fidelity is* $\delta^* = (\eta_N / \sum_{k=1}^N \alpha_k)^{1/p}$*, with cost* $T^{\text{unif}}_{\text{total}}(N) = N(\delta^*)^{-\gamma}$.

Suppose $\mathcal{E}_0(N, \Theta) \asymp N^{-r}$*,* $A_N \asymp N^{\rho}$*, and* $\sum_{k=1}^N \alpha_k \asymp N^{\sigma}$*, where* $r, \rho > 0$ *and* $p + \sigma\gamma > 0$*. If the optimal fidelities remain below* 1*, then, as* $\varepsilon \to 0$*, both optimal iteration counts scale as* $\varepsilon^{-1/r}$*, and*

$$\frac{\min_N T_{\text{total}}(N)}{\min_N T^{\text{unif}}_{\text{total}}(N)} \asymp \varepsilon^{(p+\sigma\gamma-\rho(p+\gamma))/(rp)}, \tag{5}$$

with minima over feasible N*. Thus, uniform fidelity is optimal up to constants exactly when* $\rho(p+\gamma) = p + \sigma\gamma$*; a strict inequality* $\rho(p+\gamma) < p + \sigma\gamma$ *yields a polynomial gain.*

Remark 4. The condition $p + \sigma\gamma > 0$ is automatic for $\sigma \geq 0$, but if $\sigma < 0$, then $\gamma < \frac{p}{-\sigma}$ should hold. If this fails, the optimum is attained at $\delta^* = 1$.

Jensen's inequality gives $A_N^{p+\gamma} \leq N^p (\sum_{k=1}^N \alpha_k)^{\gamma}$, so the exponent in equation 5 is nonnegative. Polynomial gains require sufficiently concentrated coefficients. For mildly decaying sequences $\alpha_k \asymp k^{-a}$ with $0 < a < 1$, this exponent vanishes, and nonuniform fidelity yields at most a constant-factor improvement. See Appx. D.1 for the proof and Sec. A.3 for empirical validation.

3 From ZO error bounds to fidelity-separable form

To apply Prop. 1 to a zeroth-order method, one must show that its convergence bound $\mathcal{E}$ falls into the fidelity-separable class (Def. 3). We establish this reduction in three steps: we first fix the ingredients of a ZO setup; we then formulate a universal error decomposition shared by a broad family of ZO methods; finally, we illustrate this using the kernel two-point estimator.

We characterize a ZO scenario by three interacting components. **(i) Oracle interface.** The way of constructing the gradient approximation g_k, e.g., one-point, symmetric or forward two-point, ℓ-point, and kernel-smoothed finite differences. **(ii) Noise model.** This work focuses on Tsybakov's noise (Def. 1) or adversarial (Def. 2). **(iii) Algorithmic template.** The update rule—e.g. (accelerated) GD, SGD—which governs how estimation errors propagate across iterations.

Different combinations of (i)–(iii) produce structurally different bounds $\mathcal{E}(\cdot)$. We expose their common structure through the conditional bias $b_k(x) \in \mathbb{R}^d$ and a conditional second-moment bound $V_k(x) > 0$ of the gradient estimate. Specifically, we assume that there are such functions $b_k(x) \in \mathbb{R}^d$ and $V_k(x) > 0$, that

$$\mathbb{E}[g_k(x, \cdot) \mid \mathcal{F}_k] := \nabla f(x) + b_k(x), \qquad \mathbb{E}[\|g_k(x, \cdot)\|^2 \mid \mathcal{F}_k] \leq V_k(x).$$

The key intermediate decomposition is

$$\text{Gap}_N \;\leq\; \tilde{\mathcal{E}}_0(N, \Theta) + \sum_{k=1}^{N} \Big[a_k(N, \Theta)\, \mathbb{E}[\|b_k(x_k)\|] + c_k(N, \Theta)\, \mathbb{E}[\|b_k(x_k)\|^2] + e_k(N, \Theta)\, \mathbb{E}[V_k(x_k)] \Big], \tag{6}$$

where $\tilde{\mathcal{E}}_0, a_k, c_k, e_k$ depend on the algorithmic hyperparameters Θ and number of steps N. The result below lists some methods satisfying equation 6.

Proposition 2. *Gradient descent, accelerated gradient descent, and SGD schemes satisfy equation 6.*

Appx. E contains the proof. This structure suggests a broader scope for the framework.

Conjecture 1 (Broader applicability)**.** Decomposition equation 6 extends to a broad class of first-order methods whose convergence analysis admits a telescoping inequality in which the gradient estimator enters through linear inner-product and squared-norm terms. Prop. 2 establishes representative instances of this principle.

Example: the kernel two-point estimator. We illustrate the reduction by expressing b_k and V_k in terms of the oracle fidelity δ_k.

Definition 4 (Kernel two-point estimator)**.** Let $h > 0$ be the smoothing radius, $\zeta \sim \mathrm{Unif}(S_2^{d-1})$ a random direction, $r \sim \mathrm{Unif}([-1, 1])$ an independent scalar, and $K\colon [-1, 1] \to \mathbb{R}$ a smoothing kernel (Def. 5). Given access to the noisy oracle equation 1, the estimator at x is

$$g_k(x; h, r, \zeta) \;:=\; \frac{d}{2h}\left(\hat{f}(x + hr\zeta) - \hat{f}(x - hr\zeta)\right) K(r)\,\zeta. \tag{7}$$

Assumption 3 (Higher-order Smoothness)**.** Fix $\beta \geq 2$ and let $\ell = \lfloor \beta \rfloor$. We assume $f \in \mathcal{F}_\beta(L_\beta)$, meaning f is ℓ-times continuously differentiable and the ℓ-th derivative satisfies the Hölder condition, $\|f^{(\ell)}(x) - f^{(\ell)}(z)\| \leq L_\beta \|x - z\|^{\beta-\ell}, \quad \forall x, z \in \mathbb{R}^d$.

Proposition 3. *Let x_k be $\mathcal{F}_k$-measurable, and let r_k and ζ_k be sampled conditionally independently of $\mathcal{F}_k$ as in Def. 4. Define $x_k^{\pm} := x_k \pm h r_k \zeta_k$, $\Delta_k := \xi_{\delta_k}(x_k^+) - \xi_{\delta_k}(x_k^-)$. Suppose that $f \in \mathcal{F}_\beta(L_\beta)$ and that $\mathbb{E}[\Delta_k^2 \mid \mathcal{F}_k] \lesssim \delta_k^2$, $\mathbb{E}[K(r_k)^2 \Delta_k^2 \mid \mathcal{F}_k] \lesssim \kappa \delta_k^2$. Then, for any fixed $h > 0$,*

$$\|b_k(x_k)\| \lesssim \kappa_\beta L_\beta h^{\beta-1} + \frac{d\sqrt{\kappa}}{h}\delta_k, \tag{8}$$

$$V_k(x_k) \lesssim \frac{d^2}{h^2}\mathbb{E}\Big[\big(f(x_k^+) - f(x_k^-)\big)^2 K(r_k)^2 \,\Big|\, \mathcal{F}_k\Big] + \frac{\kappa d^2}{h^2}\delta_k^2. \tag{9}$$

Appx. F contains the proof. Substituting these bounds into equation 6, taking expectations, and collecting powers of δ_k yields a fidelity-separable bound with two channels (Rem. 2). This completes the reduction: the estimator supplies the fidelity dependence, the optimization method supplies the propagation weights, and the resulting bound enters the allocation framework.

Assumption 4 (Regularity of f)**.** For all $x, y \in \mathbb{R}^d$, the differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies:

(i) L**-Smoothness:** $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

(ii) μ**-Strong Convexity:** $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$ (convex if $\mu = 0$).

4 Adversarial noise

This section studies wall-clock complexity under adversarial noise (Def.2). Our main finding is that the cost exponent γ acts as a template selector: it decides whether acceleration is wall-clock-optimal ($\gamma = 1$ phase transition) but does not change the leading-order choices of N, B, δ.

4.1 Wall-clock complexity of gradient methods

Let f satisfy Asm. 4 and let the oracle's answer $\hat{f}(x)$ be corrupted by the adversarial noise (Def. 2). Let g_δ be a finite-difference gradient estimator constructed from $\hat{f}(x)$ at fidelity δ,

$$g_\delta(x, h) = \frac{1}{h}\sum_{k=1}^{d}\big(\hat{f}_\delta(x + he^k) - \hat{f}_\delta(x)\big)e^k, \tag{10}$$

where $\{e^k\}_{k=1}^d$ is an orthonormal basis. Setting $h = 2\sqrt{\delta/L}$ yields $\|b(x)\|_2 \leq 2\sqrt{d\delta L}$ (Gasnikov et al., 2023). First, consider the case $\delta_k = \delta$. If f is strongly convex (Asm. 4, $\mu > 0$),

$$f(x_N) - f^* \lesssim \mathcal{E}(N, \delta, \Theta) := LR^2 \exp\left(-\left(\tfrac{\mu}{L}\right)^{\frac{1}{p}} N\right) + \left(\tfrac{L}{\mu}\right)^{\frac{2p-1}{p}} d\delta, \tag{11}$$

with $\Theta = (L, p)$. We use the family of gradient methods with intermediate convergence introduced in Devolder et al. (2013). The parameter $p \in [1, 2]$ controls the rate of convergence: $p = 1$ recovers GM, and $p = 2$ recovers fast GM (FGM) (Vasin et al., 2023).

Proposition 4 (Uniform schedule δ)**.** *Consider the gradient estimate in Eq. 10. Let f satisfy Assm. 4. The wall-clock time complexity required to achieve $\mathcal{E}(N, \delta, \Theta) \leq \varepsilon$ is*

$$T_{\text{total}} = \begin{cases} \tilde{O}\left(\left(\frac{L}{\mu}\right)^{\frac{1}{p}(1-\gamma)+2\gamma}\left(\frac{\varepsilon}{d}\right)^{-\gamma}\right), & \text{if } \mu > 0, \\ O\left(\left(\frac{\varepsilon}{9d}\right)^{-\gamma}\left(\frac{3LR^2}{\varepsilon}\right)^{\gamma+\frac{1}{p}}\right), & \text{if } \mu = 0. \end{cases} \tag{12}$$

Appx. G provides the proof. Per equation 12, the optimal p for strongly convex problems ($\mu > 0$) depends on γ: $p = 2$ (FGM) is best for $\gamma < 1$, and $p = 1$ (GM) for $\gamma > 1$. Thus, accelerated methods may actually be wall-clock slower if $\gamma > 1$; Sec. A.1 provides experiments. In the convex regime ($\mu = 0$, $LR^2/\varepsilon > 1$), however, $p = 2$ is optimal across all γ.

To analyze $\{\delta_k\}_k$ selected using Prop. 1, we introduce an intermediate gradient method (IGM) (Alg. 1) and prove its convergence (Thm. 1),

$$f(x^N) - f^* \leqslant \left(1 - \tfrac{1}{16}\left(\tfrac{\mu}{L}\right)^{\frac{1}{p}}\right)^N LR^2 + 12d\left(\tfrac{\mu}{L}\right)^{2\frac{1-p}{p}} \sum_{k=0}^{N-1}\left(1 - \tfrac{1}{16}\left(\tfrac{\mu}{L}\right)^{\frac{1}{p}}\right)^{N-k-1}\delta_k.$$

Thus, we can obtain the following result.

Proposition 5 (Time-varying fidelity schedule for IGM)**.** *Fix $p \in [1, 2]$ and $\gamma > 0$. Under the assumptions of Prop. 4, there exists a time-varying fidelity schedule such that $T_{\text{total}}^{\text{IGM}} = O_\gamma\left(\left(\frac{d}{\varepsilon}\right)^\gamma\left(\frac{L}{\mu}\right)^{2\gamma+\frac{1-\gamma}{p}}\right)$, with $\mu > 0$. If $\mu = 0$, the regularized IGM construction gives $T_{\text{total}}^{\text{IGM}} = O_\gamma\left(\left(\frac{d}{\varepsilon}\right)^\gamma\left(\frac{LR^2}{\varepsilon}\right)^{2\gamma+\frac{1-\gamma}{p}}\right)$.*

4.2 Accelerated ZO-SGD with adversarial noise

Let $f(x, \zeta)$ be the stochastic sample function, where ζ is a r.v. with an unknown distribution. Assume $f(x, \zeta)$ satisfies Assm. 3 uniformly in ζ with some $\beta \geq 2$ and constant L_β. Let the population objective $F(x) := \mathbb{E}_\zeta[f(x, \zeta)]$. The oracle output is corrupted by deterministic adversarial noise, $\hat{f}(x, \zeta) := f(x, \zeta) + \xi(x)$ and $|\xi(x)| \leq \delta$. Assume there exists $\sigma_* > 0$—a standard-deviation proxy for the stochastic gradient at the optimum—s.t. $\mathbb{E}_\zeta \|\nabla f(x^*, \zeta) - \nabla F(x^*)\|^2 \leq \sigma_*^2$, $x^* := \arg\min_x F(x)$.

We use the accelerated zero-order SGD by Lobanov & Gasnikov (2023) with a two-point scheme (Def. 4) and batching (see Rem. 1). Let B be the batch size and let $h_t \equiv h$. Setting $Q := BN$, we use Thm. 3.1 by Bychkov et al. (2024), and write for $\text{Gap}_N := \mathbb{E}F(x_N) - F^*$,

$$\begin{aligned} \text{Gap}_N \lesssim \mathcal{E}(N, B, \delta, \Theta) := \frac{LR^2}{N^2} + \frac{LR^2}{Q} + \frac{R}{\sqrt{Q}}\Big(\sqrt{d\kappa}\,\sigma_* + \sqrt{d\kappa Lh} + \sqrt{\kappa\,\frac{d\delta}{h}}\Big) \\ + \underbrace{R\Big(\kappa_\beta L_\beta h^{\beta-1} + \frac{d\delta}{h}\Big)}_{(I)} + \underbrace{N\Big(\frac{\kappa_\beta^2 L_\beta^2}{L}\,h^{2(\beta-1)} + \frac{d^2\delta^2}{Lh^2}\Big)}_{(II)}, \end{aligned}$$

with κ, κ_β defined in equation 15, and $\Theta := (d, \sigma_*, L, L_\beta, \beta, R, h)$. Note that for a fixed δ, (I) and (II) are balanced by choosing $h^*(\delta) = \left(\frac{d\delta}{\kappa_\beta L_\beta}\right)^{1/\beta}$ (see Appx. H).

Proposition 6. *Let $Q := BN$, $p := (\beta - 1)/\beta$, and $c(\Delta) = \Delta^{-\gamma}$. Set $a_1 := (\kappa_\beta L_\beta)^{1/\beta} d^p$ and $a_2 := a_1^2/L$. In the large-N, large-Q regime, the optimal design satisfies*

$$N^* \asymp \frac{LR^2}{\varepsilon}, \quad Q^* \asymp \frac{R^2\kappa d\sigma_*^2}{\varepsilon^2}, \quad B^* = \frac{Q^*}{N^*} \asymp \frac{\kappa d\sigma_*^2}{L\varepsilon}, \quad \delta^* \asymp \left(\frac{\varepsilon}{Ra_1}\right)^{1/p}, \quad h^* = h^*(\delta^*).$$

The resulting wall-clock cost is

$$T_{\text{total}}^* \asymp \kappa\sigma_*^2 d^{1+\gamma}(\kappa_\beta L_\beta)^{\gamma/(\beta-1)}\left(\frac{R}{\varepsilon}\right)^{2+\gamma/p}.$$

Take-away message. After balancing h, the adversarial model reduces to two fidelity channels, $Ra_1\delta^p$ and $Na_2\delta^{2p}$. In the large-N regime the second channel is active, which determines δ^*. The total number of oracle calls is instead dictated by σ_*, $Q^* = BN^* \asymp \frac{R^2\kappa d\sigma_*^2}{\varepsilon^2}$, and $B^* \asymp \frac{\kappa d\sigma_*^2}{L\varepsilon}$. Thus the batching regime is controlled by comparing B^* with N^*: if $B^* \le N^*$, then $\kappa d\sigma_*^2 \lesssim L^2R^2$. Hence, small σ_* leads to moderate batching, while large σ_* leads to overbatching.

In this adversarial large-N regime, γ affects the final wall-clock cost through $(\delta^*)^{-\gamma}$, but it does not change the leading-order choices of N^*, B^*, δ^*. *The next section shows that under another noise regime, the oracle-cost exponent γ can also change the optimal tuning parameters.*

5 TSYBAKOV'S NOISE

This section considers the batched setting (see Rem. 1). So, the objective is $T_{\text{total}}(N, B, \delta) = BN\delta^{-\gamma}$. In contrast to Sec. 4, under Tsybakov noise (Def. 1) the cost exponent γ acts as a parameter tuner controlling batching, with threshold $\gamma = 2$.

5.1 STRONGLY CONVEX CASE: THE BATCHED BASELINE

Let f satisfy Asm. 4 with $\mu > 0$, Asm. 3 with $\beta \ge 2$, and assume $\|f^{(2)}(x) - f^{(2)}(z)\| \le L$ for all x, z. We use the batched ℓ_2-randomized estimator of Akhavan et al. (2024). Using Cor. 19 by (Akhavan et al., 2024), we derive the following bound for the batched case (see Appx. J for details). Setting $\text{Gap}_N := \mathbb{E}[f(\hat{x}_N) - f^*]$,

$$\text{Gap}_N \lesssim \underbrace{\frac{dL^2R^2}{\mu N}}_{\mathcal{E}_0(N,\Theta)} + \underbrace{\frac{L_\beta^{2/\beta}}{\mu}\left(\frac{d^2\delta^2}{BN}\right)^{\frac{\beta-1}{\beta}}}_{\text{(I)}} + \underbrace{\frac{d^{1+2/\beta}L^2}{\mu N^{1+1/\beta}}\left(\frac{\delta^2}{BL_\beta^2}\right)^{1/\beta}}_{\text{(II)}}, \tag{13}$$

where $\Theta = (d, L, R, \mu, L_\beta, \beta)$ and L_β come from Asm. 3. Rem. 2 applies to 13 with two channels. In the high-precision regime, (I) is an active channel; see Appx. J.

Proposition 7 (Optimal batched design under Tsybakov noise)**.** *Let the assumptions of Sec. 5.1 hold, with $\mathcal{E}(N, B, \delta)$ given by equation 13. For $0 < \gamma < 2$, batching offers no leading-order gain: $B^* = 1$, and*

$$N^* \asymp \frac{d(LR)^2}{\mu\varepsilon}\,\frac{2\beta - 2 + \gamma}{(\beta-1)(2-\gamma)}, \quad \delta^* \asymp \frac{LR}{\sqrt{d}}\left(\frac{\mu\varepsilon}{L_\beta^2}\right)^{\frac{1}{2(\beta-1)}} C_\beta(\gamma),$$

where $C_\beta(\gamma) := \left(\frac{\gamma\beta}{2\beta-2+\gamma}\right)^{\frac{\beta}{2(\beta-1)}}\left(\frac{2\beta-2+\gamma}{(\beta-1)(2-\gamma)}\right)^{1/2}$. The resulting wall-clock cost is

$$T_{\text{total}}^* \asymp \frac{d^{1+\gamma/2}(LR)^{2-\gamma}L_\beta^{\gamma/(\beta-1)}}{(\mu\varepsilon)^{1+\frac{\gamma}{2(\beta-1)}}}\,\frac{2\beta-2+\gamma}{(\beta-1)(2-\gamma)}C_\beta(\gamma)^{-\gamma}.$$

For $\gamma \ge 2$, the optimal fidelity reaches the boundary:

$$\delta^* = 1, \qquad T_{\text{total}}^* \asymp \frac{d^2L_\beta^{2/(\beta-1)}}{(\mu\varepsilon)^{\beta/(\beta-1)}}.$$

The same order of total work can be achieved with $B = 1$ or with larger batches and fewer sequential iterations.

Take-away message. The Tsybakov model couples the iteration budget, fidelity and batch size through the main noise channel $(BN)^{-(\beta-1)/\beta}\delta^{2(\beta-1)/\beta}$. For $0 < \gamma < 2$, reducing δ is cheaper than batching, hence $B^* \asymp 1$. For $\gamma \ge 2$, reducing δ is too expensive and the optimum moves to $\delta^* = 1$. In this regime, batching does not improve the leading total oracle work, but it can trade parallelism for smaller sequential depth.

5.2 TSYBAKOV NOISE, ACCELERATED CONVEX CASE: BATCHING AS A NEW TUNING PARAMETER

We use the accelerated ZO-SGD scheme (Lobanov et al., 2024) with the batched two-point estimator (Rem. 1). The convergence bound is

$$\mathcal{E}(N,B,\delta,h,\Theta) \lesssim \frac{LR^2}{N^2}\max\left\{1,\left(\frac{\kappa d}{B}\right)^2\right\} + N\min\left\{\frac{B}{\kappa},\frac{\kappa d^2}{B}\right\}\left(\frac{Lh^2}{d}+\frac{\delta^2}{Lh^2}\right) + \widetilde{R}\kappa_\beta L_\beta h^{\beta-1} + \frac{N(\kappa_\beta L_\beta h^{\beta-1})^2}{L}.$$

with κ_β, L_β defined in Def. 4 and $\Theta = (\kappa, R, \tilde{R})$.

Proposition 8 (Optimal design under accelerated Tsybakov noise)**.** *Assume $\widetilde{R} = O(R)$, and fix some suitable $h > 0$. In the small-batch regime $B \le 4\kappa d$, set $Q := BN$. Then we get*

$$Q^* \asymp \kappa d R\sqrt{L}\,\varepsilon^{-1/2}\left(\frac{2+3\gamma}{2+\gamma}\right)^{1/2}, \quad \delta^* \asymp hL^{1/4}(dR)^{-1/2}\varepsilon^{3/4}G(\gamma), \tag{14}$$

with $G(\gamma) := \frac{(2\gamma)^{1/2}(2+\gamma)^{1/4}}{(2+3\gamma)^{3/4}}$. Let C_γ be a constant depending only on γ. We get,

$$T^* \asymp \kappa d^{\frac{2+\gamma}{2}} R^{\frac{2+\gamma}{2}} L^{\frac{2-\gamma}{4}} \varepsilon^{-\frac{2+3\gamma}{4}} h^{-\gamma} C_\gamma.$$

In the large-batch regime $B > 4\kappa d$, the active constraints imply $\delta^2 \asymp \frac{\varepsilon B L h^2}{N\kappa d^2}$, $T^*(B,h) \propto B^{1-\gamma/2}$. Hence $\gamma = 2$ is the batching threshold. For $0 < \gamma \le 2$, overbatching is not beneficial at leading order. For $\gamma > 2$, the optimum increases B until $\delta^* = 1$, yielding

$$N^* \asymp RL^{\frac{1}{2}}\varepsilon^{-\frac{1}{2}}, \qquad B^* \asymp \kappa d^2 R L^{-\frac{1}{2}}\varepsilon^{-\frac{3}{2}}h^{-2}, \quad T^* \asymp \kappa d^2 R^2 \varepsilon^{-2} h^{-2}.$$

Appx. I provides the proof.

Choice of h. Prop. 8 should be read conditionally on an admissible smoothing radius h. This parameter is not arbitrary: in the small-batch regime, the optimal choice is $h^*_{\mathrm{sm}}(B) = \min\{H_1, H_2, H_3(B)\}$, with $H_1 \asymp \varepsilon^{3/4}$, $H_2 \asymp \varepsilon^{1/(\beta-1)}$, and $H_3(B) \asymp (\frac{B\varepsilon^{3/2}}{d})^{1/(2\beta-2)}$; Appx. I provides the details. Therefore the small-batch wall-clock bound should be $T^*_{\mathrm{sm}}(B) \asymp \frac{\kappa d^{\frac{2+\gamma}{2}} R^{\frac{2+\gamma}{2}} L^{\frac{2-\gamma}{4}} C_\gamma}{\varepsilon^{\frac{2+3\gamma}{4}}\left(h^*_{\mathrm{sm}}(B)\right)^\gamma}$.

In the overbatching regime $\gamma > 2$, the optimal smoothing radius is again the largest admissible one, $h^*_{\mathrm{lg}} = \min\{K_1, K_2, K_3\}$, where $K_1 \asymp d^{\frac{1}{4}}$, $K_2 \asymp \varepsilon^{\frac{1}{\beta-1}}$, and $K_3 \asymp \varepsilon^{\frac{3}{2(2\beta-2)}}$; Appx. I provides the detail. The optimized overbatching complexity is $T^*_{\mathrm{overbatch}} \asymp \kappa d^2 R^2 \varepsilon^{-2}\left(h^*_{\mathrm{lg}}\right)^{-2}$.

Take-away message. This case has the same logic as Sec. 5.1,but now batching and smoothing provide additional tuning parameters. In the small-batch regime equation 14, the optimal solution is $Q^* \asymp dR\sqrt{L}\,\varepsilon^{-1/2}$, $\delta^* \asymp h^* G(\gamma) L^{1/4}(dR)^{-1/2}\varepsilon^{3/4}$, and h^* is the largest smoothing radius allowed by the constraints. Thus, as in Sec. 5.1, the fidelity is not chosen from ε alone: its prefactor depends on the oracle-cost exponent γ, and it is also coupled to the smoothing choice h^*. Notably, in this regime, batching does not improve the leading active trade-off between Q and δ, because everything depends on $Q = BN$. However, batching can indirectly help if the smoothing constraint, $h^* = H_3(B)$, is active.

The difference appears when B is optimized. In the large-batch regime, the wall-clock cost scales as $T(B) \propto B^{1-\gamma/2}$. Therefore $\gamma = 2$ governs batching. For $0 < \gamma \le 2$, increasing B does not improve the leading active trade-off, so one may take $B^* = 1$ unless a larger batch is needed to relax the admissible smoothing radius. For $\gamma > 2$, fidelity is too expensive; instead of decreasing δ, the method increases the batch size until $\delta^* = 1$. Thus, in this setting, the expensive-fidelity regime is handled by overbatching rather than by increasing N.

5.3 REGULARIZATION VERSUS ACCELERATION

We compare two approaches to convex optimization under Tsybakov noise. REGULARIZED ZO-SGD applies the strongly convex baseline of Sec. 5.1 to $f_\mu(x) = f(x) + \frac{\mu}{2}\|x - x_0\|^2$, with $\mu \asymp \varepsilon/R^2$.

ACCELERATED ZO-SGD uses the convex method of Sec. 5.2. Denote their wall-clock complexities by T_{reg} and T_{acc}, respectively. Appx. J.1 provides details.

For $0 < \gamma < 2$, Prop. 7 and Sec. 5.2 give

$$T_{\mathrm{reg}} \asymp \frac{d^{1+\gamma/2} L^{2-\gamma} L_\beta^{\gamma/(\beta-1)} R^{4-\gamma+\gamma/(\beta-1)}}{\varepsilon^{2+\gamma/(\beta-1)}},$$

$$T_{\mathrm{acc}} \asymp_{d,\varepsilon} \frac{d^{1+\gamma/2}}{\varepsilon^{1/2+3\gamma/4+\gamma s}}, \qquad s := \max\left\{\frac{3}{4}, \frac{1}{\beta-1}\right\}.$$

Acceleration has the better ε-scaling throughout this regime when $\beta \le 7/3$. For $\beta > 7/3$, the preferred method changes at $\gamma_{\mathrm{crit}}(\beta) := \frac{3(\beta-1)}{3\beta-5}$: acceleration is favored for $\gamma < \gamma_{\mathrm{crit}}(\beta)$, and regularization for $\gamma_{\mathrm{crit}}(\beta) < \gamma < 2$. For $\gamma \ge 2$, both methods have the same leading (d, ε)-scaling:

$$T_{\mathrm{reg}} \asymp \frac{d^2 L_\beta^{2/(\beta-1)} R^{2\beta/(\beta-1)}}{\varepsilon^{2+2/(\beta-1)}} \asymp_{d,\varepsilon} T_{\mathrm{acc}}.$$

Acceleration retains an advantage in sequential depth: $N_{\mathrm{acc}} \asymp \frac{R\sqrt{L}}{\sqrt{\varepsilon}}$, $N_{\mathrm{reg}} \asymp \frac{dL^2R^4}{\varepsilon^2}$.

6 DISCUSSION AND LIMITATIONS

Discussion. Our framework reveals that the oracle-cost exponent γ behaves differently depending on the noise model. Under adversarial noise, it acts as a template selector: a phase transition at $\gamma = 1$ makes non-accelerated GM wall-clock optimal over FGM, while optimal N, B, δ remain γ-independent. Under Tsybakov noise, γ is a parameter tuner, setting a batching threshold ($\gamma = 2$) that induces overbatching. Neither template dominates universally (Sec. 5.3); the choice depends on β and γ via $\gamma_{\mathrm{crit}}(\beta) = \frac{3(\beta-1)}{3\beta-5}$. Finally, the gap $\Delta = p + \sigma\gamma - \rho(p+\gamma)$ in equation 5 precisely diagnoses when adaptive fidelity yields polynomial (rather than constant) wall-clock gains, matching empirical observations.

Limitations. (i) We assume a polynomial oracle cost $c(\delta) \propto \delta^{-\gamma}$. While this drives the non-trivial trade-offs, real oracles may have fixed overheads or regime-dependent exponents. (ii) We assume zero switching costs for fidelity (Prop. 1); explicit switching penalties remain unmodeled. (iii) Our closed-form bounds target the high-precision limit ($\varepsilon \to 0$); moderate ε requires boundary analysis. (iv) Fidelity-separability is proven for first-order templates (GD, FGM, MD, SGD; Prop. 2) under both noise models; extending this to trust-region, surrogate, or second-order ZO methods is left open. (v) Optimal schedules require known problem constants (L, μ, σ_*).

Future research. Practical applications motivate extending our framework from absolute noise δ to composite noise models (supported empirically in Appx. A.4). Future work will explore composite adversarial noise, $|\hat{f}(x) - f(x)| \le \beta(f(x) - f^*)^q + \delta$ (Vasin et al., 2025), and Tsybakov noise under a strong growth condition, $\mathbb{E}(\hat{f}(x) - f(x))^2 \le \rho(f(x) - f^*)^q + \delta^2$ (Vaswani et al., 2019). Developing adaptive estimation for unknown problem constants is another key next step.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing and for retrieval and discovery (e.g., finding related work). We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide complete proofs for theoretical contributions of the paper. Experimental setups and data sources are described in Appendix A. We also provide the code alongside the text submission.

REFERENCES

Arya Akhavan, Evgenii Chzhen, Massimiliano Pontil, and Alexandre B Tsybakov. Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm. *Journal of Machine Learning Research*, 25(370):1–50, 2024.

Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In *Conference on Learning Theory*, pp. 257–283. PMLR, 2016.

Lev Bogolubsky, Pavel Dvurechenskii, Alexander Gasnikov, Gleb Gusev, Yurii Nesterov, Andrei M Raigorodskii, Aleksey Tikhonov, and Maksim Zhukovskii. Learning supervised pagerank with gradient-based and gradient-free optimization methods. *Advances in neural information processing systems*, 29, 2016.

Georgii Bychkov, Darina Dvinskikh, Anastasia Antsiferova, Alexander Gasnikov, and Aleksandr Lobanov. Accelerated zero-order sgd under high-order smoothness and overparameterized regime. *arXiv preprint arXiv:2411.13999*, 2024.

Étienne de Montbrun and Sébastien Gerchinovitz. Certified multifidelity zeroth-order optimization. *SIAM/ASA Journal on Uncertainty Quantification*, 12(4):1135–1164, 2024.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Olivier Devolder, François Glineur, Yurii Nesterov, et al. Intermediate gradient methods for smooth convex problems with inexact oracle. Technical report, Technical report, CORE-2013017, 2013.

Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171(1): 121–145, 2016.

Alexander Gasnikov, Darina Dvinskikh, Pavel Dvurechensky, Eduard Gorbunov, Aleksandr Beznosikov, and Alexander Lobanov. Randomized gradient-free methods in convex optimization. In *Encyclopedia of Optimization*, pp. 1–15. Springer, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Samuel C Hoffman, Vijil Chenthamarakshan, Kahini Wadhawan, Pin-Yu Chen, and Payel Das. Optimizing molecules using efficient queries from property evaluations. *Nature Machine Intelligence*, 4(1):21–31, 2022.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Aleksandr Lobanov and Alexander Gasnikov. Accelerated zero-order sgd method for solving the black box optimization problem under "overparametrization" condition. In *International Conference on Optimization and Applications*, pp. 72–83. Springer, 2023.

Aleksandr Lobanov, Nail Bashirov, and Alexander Gasnikov. The "black-box" optimization problem: Zero-order accelerated stochastic method via kernel approximation. *Journal of Optimization Theory and Applications*, 203(3):2451–2486, 2024.

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

Sergey Shirobokov, Vladislav Belavin, Michael Kagan, Andrei Ustyuzhanin, and Atilim Gunes Baydin. Black-box optimization with local generative surrogates. *Advances in neural information processing systems*, 33:14650–14662, 2020.

Fedor Stonyakin, Alexander Tyurin, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, Darina Dvinskikh, Mohammad Alkousa, Dmitry Pasechnyuk, Sergei Artamonov, and Victorya Piskunova. Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model. *arXiv preprint arXiv:2001.09013*, 2020.

Fedor Stonyakin, Alexander Tyurin, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, Darina Dvinskikh, Mohammad Alkousa, Dmitry Pasechnyuk, Sergei Artamonov, and Victorya Piskunova. Inexact model: A framework for optimization and variational inequalities. *Optimization Methods and Software*, 36(6):1155–1201, 2021.

Guillaume Van Dessel and François Glineur. Optimal inexactness schedules for tunable oracle-based methods. *Optimization Methods and Software*, 39(3):664–698, 2024.

Artem Vasin, Alexander Gasnikov, Pavel Dvurechensky, and Vladimir Spokoiny. Accelerated gradient methods with absolute and relative noise in the gradient. *Optimization Methods and Software*, 38 (6):1180–1229, 2023.

Artem Vasin, Valery Krivchenko, Dmitry Kovalev, Fedyor Stonyakin, Nazarii Tupitsa, Pavel Dvurechensky, Mohammad Alkousa, Nikita Kornilov, and Alexander Gasnikov. On solving minimization and min-max problems by first-order methods with relative error in gradients. *arXiv preprint arXiv:2503.06628*, 2025.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204. PMLR, 2019.

A Experiments

We conducted two experiments: the comparison of the wall-clock performance of gradient descent (GM) and the fast gradient method (FGM) as presented in Sec. 4.1; the comparison of strategies for choosing the fidelity level δ_k across iterations presented in Prop. 1. Estimation of the γ parameter for DL problems is provided in Appendix A.4. Code is available at `https://anonymous.4open.science/r/wall_clock_nips26-EC16`. The aim of the experiments is to test the predicted wall-clock phase transitions and fixed-vs-adaptive fidelity behavior.

A.1 Comparison of GM vs. FGM

We compare GM and FGM on three strongly convex problems: (i) ridge regression on the UCI Superconductivity dataset with 81 features [1], (ii) a synthetic quadratic objective, (iii) regularized logistic regression on a binarized version of the `digits` dataset [2]. All three experiments exhibit the predicted transition from FGM to GM as γ increases. We present the ridge-regression result here and discuss the other two in Appendix. A.2.

In (i), we considered $f(w) = \frac{1}{2n}\|Xw - Y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$. We set the target accuracy $\varepsilon = 10^{-8}$, and use the search of δ over 300 logarithmically spaced points between 10^{-15} and 10^{-2}. Fig. 1 presents the result; the empirical phase transition occurs around $\gamma \approx 0.9$, supporting the claim of Prop. 4

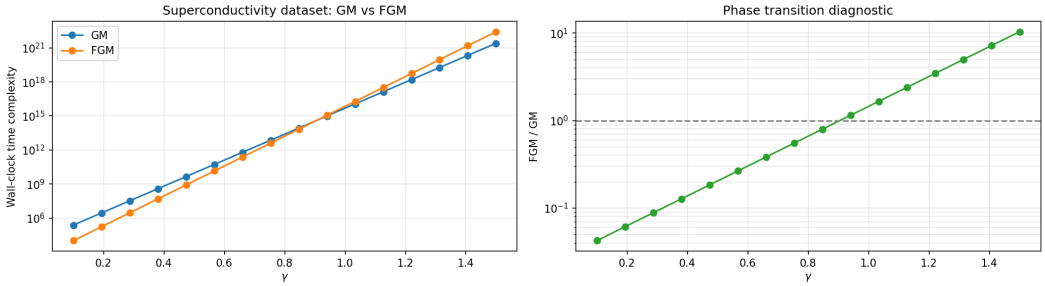

Figure 1: Comparison of the wall-clock time of GM and FGM on ridge regression over the UCI Superconductivity dataset. Left: optimal wall-clock time after tuning δ. Right: the corresponding phase transition as a function of γ.

[1] `https://archive.ics.uci.edu/dataset/464/superconductivty+data`

[2] `https://scikit-learn.org/1.5/auto_examples/datasets/plot_digits_last_image.html`

A.2 Additional details for the GM vs. FGM experiments

We compared standard gradient descent (GM) and the fast gradient method (FGM, implemented as STM in the code) on three problems: a synthetic quadratic objective, ridge regression, and regularized logistic regression. The performance criterion was the wall-clock time required to reach a prescribed target accuracy ε.

In all experiments, the parameter γ was varied over 10 uniformly spaced values between 0.1 and 1.5. For each value of γ, we selected a single fixed noise level δ used throughout the whole run. The value of δ was chosen by grid search over 300 logarithmically spaced points between 10^{-15} and 10^{-2}, with the search truncated at the corresponding theoretical threshold: $\delta \le \mu\varepsilon/(Ld)$ for GM and $\delta \le \mu^{3/2}\varepsilon/(L^{3/2}d)$ for FGM. For every candidate δ, we ran the method and recorded the time needed to reach accuracy ε; the best such wall-clock time was reported.

The synthetic quadratic experiment has the following setup: the objective was $f(x) = \frac{1}{2}x^\top Qx$, where Q is diagonal with eigenvalues ranging from $\mu = 1$ to $L = 10$, and the dimension is $d = 20$. We used coordinate-wise forward finite differences, with gradient estimate $\tilde{\nabla} f(x)_i = \big(\hat{f}(x + he_i) - \hat{f}(x)\big)/h$, where $h = 2\sqrt{\delta/L}$ was chosen according to the theory. The target accuracy was $\varepsilon = 10^{-6}$. The noisy oracle used a deterministic adversarial perturbation with sign given by $\mathrm{sign}(\sin(\sum_i x_i))$. The results are shown in Figure 2.

For ridge regression, we considered $f(w) = \frac{1}{2n}\|Xw - Y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$ on the UCI Superconductivity dataset with 81 features.[3] The target accuracy was $\varepsilon = 10^{-8}$, and the search over δ used the same logarithmic grid as above. The results are shown in Figure 1; the empirical phase transition occurs around $\gamma \approx 0.9$.

For regularized logistic regression, we used $f(x) = \frac{1}{n}\sum_{i=1}^n \log\big(1 + \exp(-y_i a_i^\top x)\big) + \frac{\lambda}{2}\|x\|^2$ on the `digits` dataset from `scikit-learn`, converted into a binary classification problem by grouping digits 0–4 into one class and digits 5–9 into the other.[4] The target accuracy was $\varepsilon = 10^{-4}$, and δ was tuned over the same grid. The results are shown in Figure 3; in this case, the phase transition appears around $\gamma \approx 0.6$.

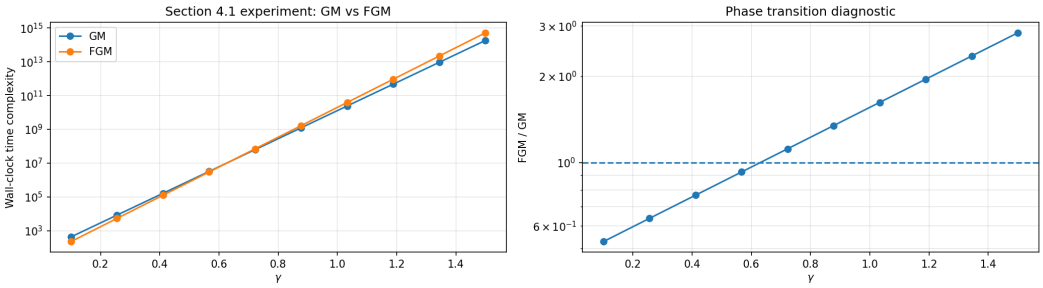

Figure 2: Comparison of GM and FGM on the quadratic objective.

[3] `https://archive.ics.uci.edu/dataset/464/superconductivty+data`

[4] `https://scikit-learn.org/1.5/auto_examples/datasets/plot_digits_last_image.html`

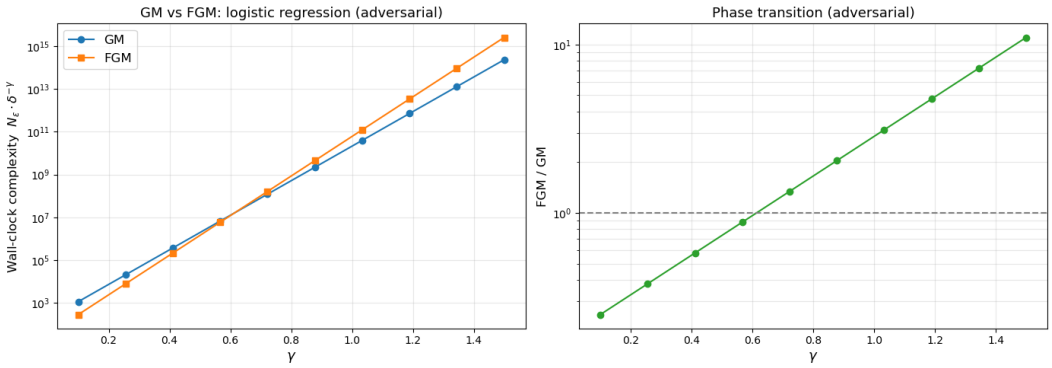

Figure 3: Comparison of the wall-clock time of GM and FGM on regularized logistic regression.

A.3 Comparison of fidelity allocation strategies

This section illustrates the results from Sec. 2.1. We consider FGM on $f(x) = \frac{\mu}{2}x_1^2 + \frac{L}{2}x_2^2$, under assumption of oracle with adversarial noise (Assumption 2). We compare a uniform schedule, $\delta_k = \delta_0$, with the time-varying schedule $\delta_k = \delta_{k-1}\left(1 + \frac{\mu}{4L} + \sqrt{\frac{\mu}{2L}}\right)^{-\frac{1}{1+\gamma}}$ (motivated by Lemma 1). Fig. 4 shows that the adaptive schedule consistently outperforms the fixed-noise baseline. Specifically, the optimal δ_0 is larger for the adaptive strategy, and this strategy achieves lower wall-clock time across the range of target accuracies we tested. The observed improvement is moderate, approximately a factor of 1.7–2.2, which is compatible with Prop. 9: for geometrically varying coefficients, the gain over the best uniform schedule is at most logarithmic rather than polynomial. Full details of the noise model and the tuning procedure are given in Appx. B.

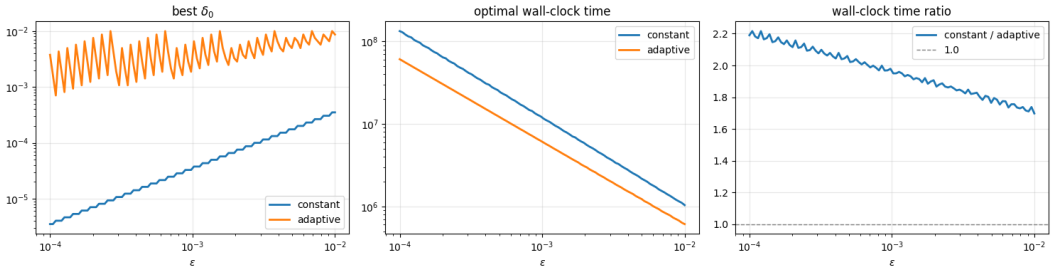

Figure 4: Comparison of fidelity allocation strategies on each iteration

A.4 γ estimation for deep learning problems

Hardware details: Experiments were performed on Intel(R) Core(TM) i9-12900H: 14 cores, 32 GB RAM, NVIDIA RTX 3080 Ti Mobile GPU 16 GiB.

Experiments details: For most experiments below was used Adam (Kingma & Ba, 2014) with learning rate $= 0.001$.

We provide experiments for various deep learning problems. We will estimate the error as the calculation accuracy decreases, thus, at each iteration we will calculate loss using different floating point numbers – $f_{\text{FP16}}(x^k)$ and $f_{\text{FP64}}f(x^k)$. Our goal is to estimate the γ parameter in model $c(\delta) = \delta^{-\gamma}$, assuming that $f_{\text{FP64}}(x^k)$ real function value. Thus

$$\delta_k = f_{\text{FP64}}(x^k) - f_{\text{FP16}}(x^k), \quad \delta^* = \max_{0 \leqslant k \leqslant N-1} \delta_k, \quad \gamma_k = -\log_{\delta^*}(T_k^{\text{FP16}}),$$

T_k^{FP16} – calculation time of loss using FP16.

We will plot training curves: loss on train dataset, accuracy on test part and δ_k, T_k, γ_k. Firstly, we consider a simple custom convolutional neural network and MNIST (Deng, 2012) dataset.

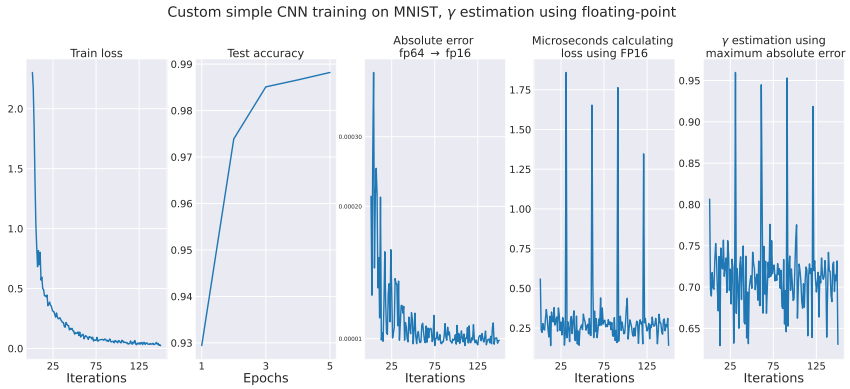

Figure 5: Simple custom CNN, MNIST, batch size – 2048

On the Figure 5 established, that $0.6 \leqslant \gamma \leqslant 1$. Note that the graph shows that the error decreases as the loss decreases; thus, it is motivated to consider relative error $|\hat{f}(x) - f(x)| \leqslant \beta(f(x) - f^*)^q$. The following experiment uses the same idea for ResNet-18 (He et al., 2016) and CIFAR-10 (Krizhevsky et al., 2009).

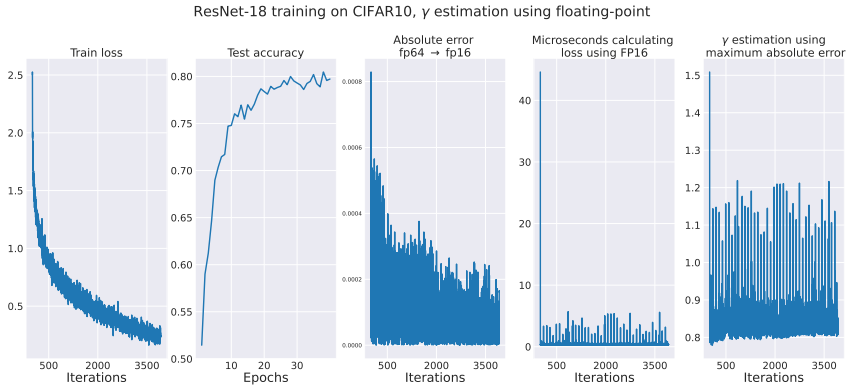

Figure 6: ResNet-18, CIFAR10, batch size – 512

Figure 6 demonstrates, that $0.7 \leqslant \gamma \leqslant 1.6$.

A.5 Sensitivity to fidelity-switching costs

We examine how a cost for changing oracle fidelity affects the allocation selected by Prop. 1. We use the two-dimensional quadratic $f(x) = \frac{\mu}{2}x_1^2 + \frac{L}{2}x_2^2$ from Sec. A.3 with $\mu = 1$, $L = 10$, $x_0 = (1, 1)$, $N = 120$, target accuracy $\varepsilon = 10^{-4}$, and per-query cost $c(\delta) = \delta^{-1}$. The fidelity coefficients follow the geometric profile $\alpha_k = (2L/\mu)q^{N-k}$, where $q = (1 + \mu/(4L) + \sqrt{\mu/(2L)})^{-1}$. We allocate the surrogate fidelity budget $\eta_N = \varepsilon - L\|x_0\|^2 \exp[-N\sqrt{\mu/(2L)}/2]$.

For $M \in \{1, 2, 5, 10, N\}$, we partition the iterations into M contiguous constant-fidelity blocks. We optimize the block boundaries by dynamic programming and the fidelity levels under $\sum_{k=1}^{N} \alpha_k \delta_k \leq \eta_N$. Central finite differences use four function-value queries per iteration, giving evaluation cost

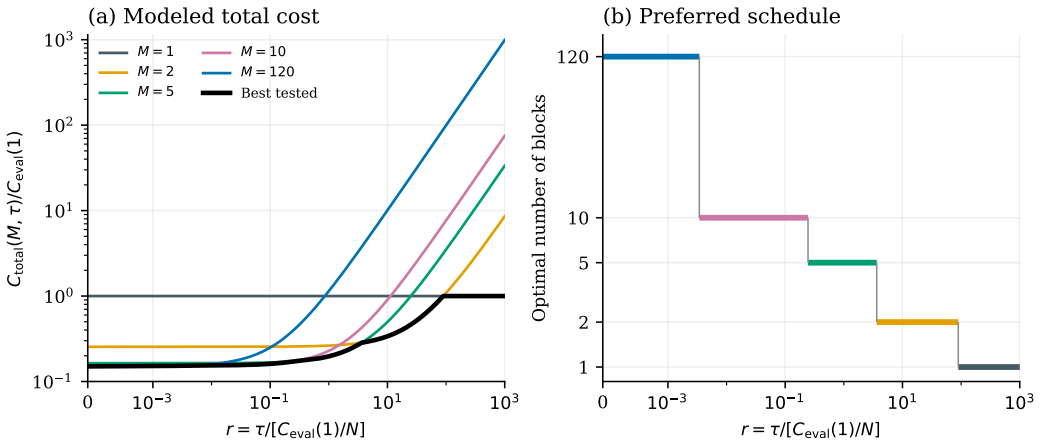

Figure 7: Sensitivity of modeled cost to fidelity switching. (a) Total cost for each tested number M of constant-fidelity blocks; the black curve is the lowest cost among them. (b) The preferred block count as the switching penalty increases. Both horizontal axes normalize τ by the uniform schedule's evaluation cost per iteration. Costs in (a) are normalized by $C_{\text{eval}}(1)$ and are displayed up to 1.05.

$C_{\text{eval}}(M) = 4\sum_{k=1}^{N} c(\delta_k)$. For a switching penalty τ, we define $C_{\text{sw}}(M, \tau) = \tau \sum_{k=2}^{N} \mathbf{1}\{\delta_k \neq \delta_{k-1}\}$ and $C_{\text{total}} = C_{\text{eval}} + C_{\text{sw}}$.

Fig. 7(a) shows $C_{\text{total}}(M, \tau)/C_{\text{eval}}(1)$ as a function of $r = \tau/[C_{\text{eval}}(1)/N]$ over the displayed range $10^{-4} \leq r \leq 10^{3}$; costs above 1.05 are clipped. Panel (b) shows the cost-minimizing block count among the five tested values of M. At $\tau = 0$, which is outside the logarithmic plot, the fully adaptive schedule has 0.150 times the modeled evaluation cost of the uniform schedule. As r increases, the preferred block count changes from 120 to 10, 5, 2, and 1 at approximately 0.00343, 0.245, 3.64, and 89.4.

All five schedules reached the target in an auxiliary rollout with greedy adversarial perturbations. This rollout uses standard strongly convex Nesterov momentum rather than the STM implementation of Fig. 4. Moreover, the convergence profile used to set η_N suppresses constants, and τ is swept rather than measured for a real oracle. Accordingly, this experiment is a sensitivity analysis of the cost model, not a measurement of wall-clock speedup.

B Additional details for the δ_k scheduling experiments

In the second group of experiments, we considered the two-dimensional quadratic objective $f(x) = \frac{\mu}{2}x_1^2 + \frac{L}{2}x_2^2$. The gradient was approximated using central finite differences:

$$\tilde{\nabla} f(x) = \left(\frac{\hat{f}(x + he_1) - \hat{f}(x - he_1)}{2h}, \frac{\hat{f}(x + he_2) - \hat{f}(x - he_2)}{2h} \right).$$

The noise model was greedy and adversarial. At every iteration, the next point x_{k+1} depends on four noisy oracle values, corresponding to the points $x + he_1$, $x - he_1$, $x + he_2$, and $x - he_2$. Since each noisy value can vary independently within the interval $[f(\cdot) - \delta,\ f(\cdot) + \delta]$, the adversary has four degrees of freedom and chooses the perturbation that maximizes $f(x_{k+1})$ at the next iterate.

Because x_{k+1} is obtained from the finite-difference estimate through an affine transformation, the resulting objective is a convex continuous function of these four perturbations. Therefore, its maximum over the box $[-\delta, \delta]^4$ is attained at a vertex, so it is sufficient to enumerate all 16 vertices and choose the worst one.

Figure 8 compares the convergence behavior of GM and FGM under this worst-case noise model, under uniformly random noise from $[-\delta, \delta]$, and in the noiseless setting.

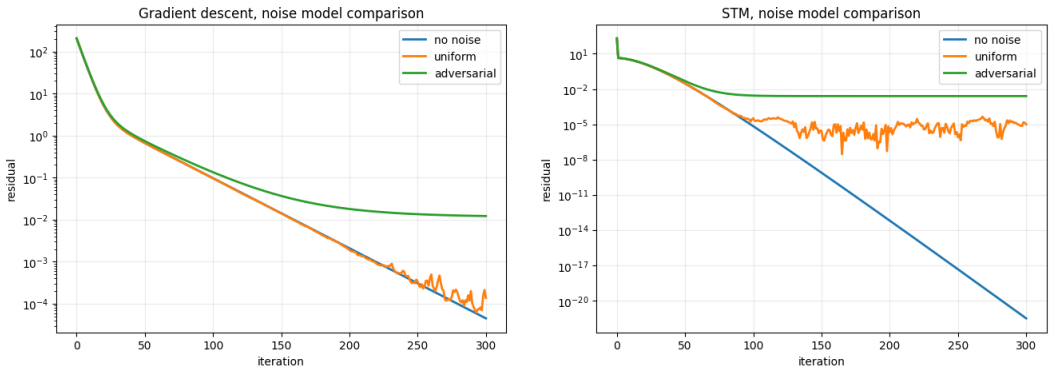

Figure 8: Comparison of convergence in function residual under different noise models.

For this noise model, we considered the following strategies for choosing δ_k at each step of FGM: $\delta_k = \delta_0$, and

$$\delta_k = \delta_{k-1}\left(1 + \frac{\mu}{4L} + \sqrt{\frac{\mu}{2L}}\right)^{-\frac{1}{1+\gamma}},$$

with δ_0 selected over a logarithmic grid from 10^{-8} to 10^{-2}.

We now introduce an adaptive strategy motivated by Proposition 1. Recall that the convergence rate of the accelerated method (FGM) (Vasin et al., 2023) with a noisy gradient, where the gradient is approximated using one of the schemes

$$\tilde{\nabla} f(x)_i = \frac{\hat{f}(x + he_i) - \hat{f}(x)}{h} \qquad \text{or} \qquad \tilde{\nabla} f(x)_i = \frac{\hat{f}(x + he_i) - \hat{f}(x - he_i)}{2h},$$

for N iterations with varying noise levels, has the form

$$f(x) - f(x^*) \lesssim LR^2 \exp\left(-\frac{1}{2}\sqrt{\frac{\mu}{2L}}N\right) + \sum_{i=0}^{N} \frac{dL}{\mu}\left(1 + \frac{\mu}{4L} + \sqrt{\frac{\mu}{2L}}\right)^{i-N} \delta_i.$$

According to Proposition 1,

$$\delta_i^*(N) = \left(\frac{\epsilon}{A_N}\right)\alpha_i^{-\frac{1}{1+\gamma}}, \quad A_N := \sum_{j=1}^{N} \alpha_j^{\frac{\gamma}{1+\gamma}}.$$

Therefore, in the accelerated setting,

$$\frac{\delta_{i+1}^*(N)}{\delta_i^*(N)} = \left(\frac{\alpha_i}{\alpha_{i+1}}\right)^{\frac{1}{1+\gamma}} = \left(\frac{1}{1 + \frac{\mu}{4L} + \sqrt{\frac{\mu}{2L}}}\right)^{\frac{1}{1+\gamma}}.$$

This relation motivates the adaptive strategy

$$\delta_k = \delta_{k-1}\left(1 + \frac{\mu}{4L} + \sqrt{\frac{\mu}{2L}}\right)^{-\frac{1}{1+\gamma}}.$$

As before, we tune δ_0 over the same logarithmic grid. Figure 4 shows three plots. The left panel illustrates how the optimal initial noise depends on the target accuracy. As expected, when δ decreases across iterations, larger values of the initial noise level δ_0 become preferable. The middle panel shows that the adaptive strategy outperforms the strategy with fixed δ at every iteration. Finally, the right panel reports the ratio between the runtime of the strategy $\delta_k = \delta_0$ and that of the adaptive strategy. At the same time, this plot indicates that the gain from adaptivity remains rather modest, which is consistent with the theory:

Proposition 9 (Logarithmic gain for geometrically decaying coefficients)**.** *Fix N and assume that the fidelity-separable channel has coefficients*

$$\alpha_k(N,\Theta) = Cq^{N-k}, \qquad k = 1,\dots,N,$$

where $C > 0$ and

$$q = \frac{1}{1 + \frac{\mu}{4L} + \sqrt{\frac{\mu}{2L}}} \in (0,1).$$

Let

$$T^{\mathrm{var}}_{\mathrm{total}}(N)$$

denote the optimal wall-clock time under the time-varying fidelity schedule of Prop. 1, and let

$$T^{\mathrm{unif}}_{\mathrm{total}}(N)$$

denote the optimal wall-clock time under the uniform restriction $\delta_1 = \cdots = \delta_N$. If

$$N \asymp \sqrt{\frac{L}{\mu}}\,\log\!\left(\frac{LR^2}{\varepsilon}\right),$$

then, in the high-precision regime,

$$\frac{T^{\mathrm{unif}}_{\mathrm{total}}(N)}{T^{\mathrm{var}}_{\mathrm{total}}(N)} = O\!\left(\log\!\left(\frac{LR^2}{\varepsilon}\right)\right).$$

Equivalently, the time-varying fidelity schedule improves the best uniform schedule by at most a logarithmic factor for this geometric coefficient profile.

Proof. For fixed N, Prop. 1 gives

$$T^{\mathrm{var}}_{\mathrm{total}}(N) = \epsilon^{-\gamma/p}\left(\sum_{k=1}^{N} \alpha_k^{\frac{\gamma}{p+\gamma}}\right)^{\frac{p+\gamma}{p}},$$

where $\epsilon = \varepsilon - \mathcal{E}_0(N,\Theta)$ is the accuracy budget allocated to the fidelity channel. By Cor. 2, the best uniform schedule satisfies

$$T^{\mathrm{unif}}_{\mathrm{total}}(N) = N\epsilon^{-\gamma/p}\left(\sum_{k=1}^{N} \alpha_k\right)^{\gamma/p}.$$

Hence the common factor $\epsilon^{-\gamma/p}$ cancels in the ratio. Put

$$s := \frac{\gamma}{p+\gamma} \in (0,1).$$

Since $\alpha_k = Cq^{N-k}$, the constant C also cancels, and after changing indices $r = N - k$ we obtain

$$\frac{T^{\mathrm{unif}}_{\mathrm{total}}(N)}{T^{\mathrm{var}}_{\mathrm{total}}(N)} = N\frac{\left(\sum_{r=0}^{N-1} q^r\right)^{\gamma/p}}{\left(\sum_{r=0}^{N-1} q^{sr}\right)^{(p+\gamma)/p}}.$$

Using the geometric-sum identities,

$$\sum_{r=0}^{N-1} q^r = \frac{1-q^N}{1-q}, \qquad \sum_{r=0}^{N-1} q^{sr} = \frac{1-q^{sN}}{1-q^s},$$

we get

$$\frac{T^{\mathrm{unif}}_{\mathrm{total}}(N)}{T^{\mathrm{var}}_{\mathrm{total}}(N)} = N\frac{(1-q^N)^{\gamma/p}(1-q^s)^{(p+\gamma)/p}}{(1-q)^{\gamma/p}(1-q^{sN})^{(p+\gamma)/p}}.$$

For the chosen value of q,

$$1 - q = \frac{\frac{\mu}{4L} + \sqrt{\frac{\mu}{2L}}}{1 + \frac{\mu}{4L} + \sqrt{\frac{\mu}{2L}}} = \Theta\!\left(\sqrt{\frac{\mu}{L}}\right)$$

as $L/\mu \to \infty$. Moreover, since $s \in (0,1)$ is fixed,

$$1 - q^s = \Theta(1-q) = \Theta\left(\sqrt{\frac{\mu}{L}}\right).$$

Finally,

$$N \asymp \sqrt{\frac{L}{\mu}} \log\left(\frac{LR^2}{\varepsilon}\right)$$

implies that $N(1-q) \asymp \log(LR^2/\varepsilon)$, so in the high-precision regime q^{sN} is bounded away from 1 and $(1-q^N)^{\gamma/p} \leq 1$. Therefore,

$$\frac{T^{\text{unif}}_{\text{total}}(N)}{T^{\text{var}}_{\text{total}}(N)} = O\left(N \frac{(1-q)^{(p+\gamma)/p}}{(1-q)^{\gamma/p}}\right) = O(N(1-q)).$$

Substituting $1 - q = \Theta(\sqrt{\mu/L})$ yields

$$\frac{T^{\text{unif}}_{\text{total}}(N)}{T^{\text{var}}_{\text{total}}(N)} = O\left(\log\left(\frac{LR^2}{\varepsilon}\right)\right),$$

as claimed. □

C Additional definitions

Definition 5 (Smoothing kernel)**.** Let $r \sim \text{Unif}([-1,1])$. We assume that a kernel function $K : [-1,1] \to \mathbb{R}$ is such that for a fixed $\beta \geq 2$ the following occurs Bach & Perchet (2016): $\mathbb{E}[K(r)] = 0$, $\mathbb{E}[rK(r)] = 1$, and $\mathbb{E}[|r|^\beta |K(r)|] < \infty$. Additionally, if $\beta > 2$, we require $\mathbb{E}[r^j K(r)] = 0$ for all $j = 2, \ldots, \lfloor \beta \rfloor$. Further, we set

$$\kappa_\beta = \int_{-1}^{1} |u|^\beta |K(u)|\, du, \quad \kappa = \int_{-1}^{1} |K(u)|^2\, du. \tag{15}$$

D Proofs from Sec. 2.1

Proof of Proposition 1. Since $\mathcal{E}_0(N, \Theta)$ does not depend on $\delta_1, \ldots, \delta_N$, the fidelity-dependent problem is to minimize $\sum_{i=1}^N \delta_i^{-\gamma}$ subject to $\sum_{i=1}^N \alpha_i \delta_i^p \leqslant \eta_N$, where $\eta_N = \varepsilon - \mathcal{E}_0(N, \Theta)$. The objective is strictly decreasing in every δ_i, so the constraint is active at the optimum. Moreover, after the change of variables $y_i = \delta_i^p$, the objective is strictly convex and the constraint is linear; hence the stationary point below is the unique global minimizer. The Lagrangian for the active constraint is

$$\mathcal{L} = \sum_{i=1}^N \delta_i^{-\gamma} + \lambda \left(\sum_{i=1}^N \alpha_i \delta_i^p - \eta_N\right).$$

The stationarity condition $\frac{\partial \mathcal{L}}{\partial \delta_i} = 0$:

$$-\gamma \delta_i^{-\gamma-1} + \lambda p \alpha_i \delta_i^{p-1} = 0 \implies \delta_i = \left(\frac{\gamma}{\lambda \alpha_i p}\right)^{\frac{1}{p+\gamma}}. \tag{16}$$

Substituting the equation for δ_i into the constraint, and then substituting δ_i^* into the objective function, yields the allocation and wall-clock expressions in equation 4, with the available fidelity budget equal to η_N. □

Corollary 2. *Let $\tilde{A}_N := \sum_{i=1}^N \alpha_i(N, \Theta)$. Under the assumptions of Prop. 1, for a uniform fidelity schedule $\delta_k = \delta$, the optimal level δ^* and the resulting total wall-clock time T^{unif}_{total} are $\delta^* = \left(\frac{\eta_N}{\tilde{A}_N}\right)^{\frac{1}{p}}$ and $T^{\text{unif}}_{\text{total}}(N) = N \eta_N^{-\frac{\gamma}{p}} \tilde{A}_N^{\frac{\gamma}{p}}$.*

Lemma 1. *Let*

$$c(\delta) = \max\{\delta^{-\gamma}, 1\}, \qquad \eta_N := \varepsilon - \mathcal{E}_0(N, \Theta) > 0,$$

and assume $\alpha_i > 0$. Without loss of optimality, $0 < \delta_i \le 1$. If $\eta_N \ge \sum_{i=1}^N \alpha_i$, then $\delta_i^* = 1$ for every i. If $0 < \eta_N < \sum_{i=1}^N \alpha_i$, then there exists a unique $K > 0$ such that

$$\delta_i^* = \min\left\{1,\ K\alpha_i^{-1/(p+\gamma)}\right\}.$$

Let

$$I(K) := \left\{i : K\alpha_i^{-1/(p+\gamma)} < 1\right\}, \qquad J(K) := [N] \setminus I(K).$$

The value of K is determined by

$$K^p \sum_{i \in I(K)} \alpha_i^{\gamma/(p+\gamma)} + \sum_{j \in J(K)} \alpha_j = \eta_N.$$

Proof. For any feasible $\delta_i > 1$, replacing it by 1 does not increase the cost and weakly decreases the error. Hence it is sufficient to optimize over $0 < \delta_i \le 1$. If $\eta_N \ge \sum_{i=1}^N \alpha_i$, the choice $\delta_i^* = 1$ is feasible and minimizes every summand of the cost.

Now suppose that $0 < \eta_N < \sum_{i=1}^N \alpha_i$. The error constraint is active because the objective is strictly decreasing in every δ_i. The KKT Lagrangian is

$$\mathcal{L} = \sum_{i=1}^N \delta_i^{-\gamma} + \lambda_0 \left(\sum_{i=1}^N \alpha_i \delta_i^p - \eta_N\right) + \sum_{i=1}^N \lambda_i(\delta_i - 1),$$

where $\lambda_0 > 0$ and $\lambda_i \ge 0$. For an interior coordinate, stationarity gives

$$\delta_i = \left(\frac{\gamma}{\lambda_0 p}\right)^{1/(p+\gamma)} \alpha_i^{-1/(p+\gamma)}.$$

Defining

$$K := \left(\frac{\gamma}{\lambda_0 p}\right)^{1/(p+\gamma)},$$

and applying complementary slackness at the capped coordinates yields

$$\delta_i^* = \min\left\{1, K\alpha_i^{-1/(p+\gamma)}\right\}.$$

Substitution into the active constraint gives

$$K^p \sum_{i \in I(K)} \alpha_i^{\gamma/(p+\gamma)} + \sum_{j \in J(K)} \alpha_j = \eta_N.$$

The left-hand side is continuous and strictly increasing from 0 to $\sum_i \alpha_i$ as K increases, so the required K exists and is unique. □

Corollary 3 (Large N full version)**.** *Under Prop. 1, assume that, as $N \to \infty$,*

$$\mathcal{E}_0(N, \Theta) \sim C_0 N^{-\beta}, \qquad A_N(\gamma) := \sum_{j=1}^N \alpha_j(N, \Theta)^{\gamma/(p+\gamma)} \sim C_A(\gamma) N^{\rho},$$

with $\beta > 0$ and $\rho > 0$. Let

$$D_\gamma := \rho(p + \gamma) + \beta\gamma.$$

Then the continuous interior optimizer is

$$N^* = \left(\frac{C_0 D_\gamma}{\rho(p+\gamma)\varepsilon}\right)^{1/\beta},$$

and

$$\delta_k^* = \left(\frac{\beta\gamma\,\varepsilon}{D_\gamma C_A(\gamma)(N^*)^\rho}\right)^{1/p} \alpha_k(N^*, \Theta)^{-1/(p+\gamma)}.$$

For the uniform schedule $\delta_k \equiv \delta$, assume that

$$\widetilde{A}_N := \sum_{j=1}^{N} \alpha_j(N, \Theta) \sim \widetilde{C}_A N^{\sigma},$$

where $\sigma \in \mathbb{R}$ and $p + \sigma\gamma > 0$. Let

$$\widetilde{D}_\gamma := p + \sigma\gamma + \beta\gamma.$$

Then the continuous interior optimizer under the uniform restriction is

$$\widetilde{N}^* = \left(\frac{C_0 \widetilde{D}_\gamma}{(p + \sigma\gamma)\varepsilon} \right)^{1/\beta},$$

and

$$\delta^* = \left(\frac{\beta\gamma\, \varepsilon}{\widetilde{D}_\gamma \widetilde{C}_A (\widetilde{N}^*)^\sigma} \right)^{1/p}.$$

For integer horizons, the displayed values are continuous relaxations; the integer optimizer is obtained by minimizing the same one-dimensional objective over feasible $N \in \mathbb{N}$, which does not affect the leading ε-scaling.

Proof. We first consider the nonuniform case. For a fixed feasible horizon N, define

$$\eta_N := \varepsilon - \mathcal{E}_0(N, \Theta) = \varepsilon - C_0 N^{-\beta}.$$

By Prop. 1, the fixed-N optimal fidelity allocation is

$$\delta_k^*(N) = \left(\frac{\eta_N}{A_N(\gamma)} \right)^{1/p} \alpha_k(N, \Theta)^{-1/(p+\gamma)},$$

and the corresponding wall-clock cost is

$$T^*(N) \asymp \eta_N^{-\gamma/p} A_N(\gamma)^{(p+\gamma)/p}.$$

Using $A_N(\gamma) \sim C_A(\gamma) N^\rho$, the continuous relaxation is

$$T^*(N) \asymp (\varepsilon - C_0 N^{-\beta})^{-\gamma/p} \big(C_A(\gamma) N^\rho \big)^{(p+\gamma)/p}.$$

Equivalently,

$$\log T^*(N) = -\frac{\gamma}{p} \log(\varepsilon - C_0 N^{-\beta}) + \frac{\rho(p+\gamma)}{p} \log N + \text{const}.$$

The first-order condition gives

$$\frac{\rho(p+\gamma)}{pN} = \frac{\gamma}{p} \frac{\beta C_0 N^{-\beta-1}}{\varepsilon - C_0 N^{-\beta}}.$$

Hence

$$\rho(p+\gamma)(\varepsilon - C_0 N^{-\beta}) = \beta\gamma C_0 N^{-\beta}.$$

Therefore

$$C_0 N^{-\beta} = \frac{\rho(p+\gamma)}{\rho(p+\gamma) + \beta\gamma} \varepsilon = \frac{\rho(p+\gamma)}{D_\gamma} \varepsilon,$$

which yields

$$N^* = \left(\frac{C_0 D_\gamma}{\rho(p+\gamma)\varepsilon} \right)^{1/\beta}.$$

At this horizon,

$$\eta_{N^*} = \varepsilon - C_0 (N^*)^{-\beta} = \frac{\beta\gamma}{D_\gamma} \varepsilon.$$

Substituting this and

$$A_{N^*}(\gamma) \sim C_A(\gamma) (N^*)^\rho$$

into the fixed-N allocation formula gives

$$\delta_k^* = \left(\frac{\beta\gamma\,\varepsilon}{D_\gamma C_A(\gamma)(N^*)^\rho}\right)^{1/p} \alpha_k(N^*,\Theta)^{-1/(p+\gamma)}.$$

In particular,

$$N^* \asymp_\gamma \varepsilon^{-1/\beta}, \qquad \delta_k^* \asymp_\gamma \varepsilon^{(1+\rho/\beta)/p}\alpha_k(N^*,\Theta)^{-1/(p+\gamma)}.$$

We now consider the uniform restriction $\delta_k \equiv \delta$. For fixed N, the error constraint becomes

$$\mathcal{E}_0(N,\Theta) + \widetilde{A}_N \delta^p \le \varepsilon.$$

Thus the largest admissible uniform fidelity level is

$$\delta^*(N) = \left(\frac{\varepsilon - C_0 N^{-\beta}}{\widetilde{A}_N}\right)^{1/p}.$$

The fixed-N wall-clock cost is therefore

$$T^*_{\mathrm{unif}}(N) \asymp N(\delta^*(N))^{-\gamma} = N\left(\frac{\widetilde{A}_N}{\varepsilon - C_0 N^{-\beta}}\right)^{\gamma/p}.$$

Using $\widetilde{A}_N \sim \widetilde{C}_A N^\sigma$, we get

$$T^*_{\mathrm{unif}}(N) \asymp N\left(\frac{\widetilde{C}_A N^\sigma}{\varepsilon - C_0 N^{-\beta}}\right)^{\gamma/p}.$$

Equivalently,

$$\log T^*_{\mathrm{unif}}(N) = \left(1 + \frac{\sigma\gamma}{p}\right)\log N - \frac{\gamma}{p}\log(\varepsilon - C_0 N^{-\beta}) + \mathrm{const}.$$

The condition $p + \sigma\gamma > 0$ ensures that the continuous objective grows for sufficiently large N, so the interior critical point gives the interior minimizer. Differentiating gives

$$1 + \frac{\sigma\gamma}{p} = \frac{\beta\gamma}{p}\frac{C_0 N^{-\beta}}{\varepsilon - C_0 N^{-\beta}}.$$

Hence

$$C_0 N^{-\beta} = \frac{p+\sigma\gamma}{p+\sigma\gamma+\beta\gamma}\varepsilon = \frac{p+\sigma\gamma}{\widetilde{D}_\gamma}\varepsilon.$$

Therefore

$$\widetilde{N}^* = \left(\frac{C_0\widetilde{D}_\gamma}{(p+\sigma\gamma)\varepsilon}\right)^{1/\beta},$$

and

$$\varepsilon - C_0(\widetilde{N}^*)^{-\beta} = \frac{\beta\gamma}{\widetilde{D}_\gamma}\varepsilon.$$

Substituting into the fixed-N formula for $\delta^*(N)$ yields

$$\delta^* = \left(\frac{\beta\gamma\,\varepsilon}{\widetilde{D}_\gamma\widetilde{C}_A(\widetilde{N}^*)^\sigma}\right)^{1/p}.$$

Consequently,

$$\widetilde{N}^* \asymp_\gamma \varepsilon^{-1/\beta}, \qquad \delta^* \asymp_\gamma \varepsilon^{(1+\sigma/\beta)/p}.$$

The argument above covers the interior optimum only. If the resulting fidelity level violates an admissible upper bound, e.g. $\delta^* > 1$, then the optimizer lies on the boundary of the fidelity set and the displayed interior formulas do not apply. □

D.1 Optimized nonuniform-to-uniform cost ratio.

For the nonuniform allocation, Prop. 1 gives, for fixed N,

$$T_{\text{total}}(N) \asymp (\varepsilon - C_0 N^{-\beta})^{-\gamma/p} \big(C_A(\gamma) N^{\rho}\big)^{(p+\gamma)/p}.$$

At the continuous optimum,

$$N^* = \left(\frac{C_0 D_\gamma}{\rho(p+\gamma)\varepsilon}\right)^{1/\beta}, \qquad \varepsilon - C_0 (N^*)^{-\beta} = \frac{\beta\gamma}{D_\gamma}\varepsilon,$$

where $D_\gamma = \rho(p+\gamma) + \beta\gamma$. Hence

$$T_{\text{total}}(N^*) \asymp C_A(\gamma)^{(p+\gamma)/p} \left(\frac{\beta\gamma}{D_\gamma}\right)^{-\gamma/p} \left(\frac{C_0 D_\gamma}{\rho(p+\gamma)}\right)^{\frac{\rho(p+\gamma)}{\beta p}} \varepsilon^{-D_\gamma/(\beta p)}.$$

For the uniform allocation,

$$T^{\text{unif}}_{\text{total}}(N) \asymp N \left(\frac{\widetilde{C}_A N^{\sigma}}{\varepsilon - C_0 N^{-\beta}}\right)^{\gamma/p}.$$

At the continuous optimum,

$$\widetilde{N}^* = \left(\frac{C_0 \widetilde{D}_\gamma}{(p+\sigma\gamma)\varepsilon}\right)^{1/\beta}, \qquad \varepsilon - C_0 (\widetilde{N}^*)^{-\beta} = \frac{\beta\gamma}{\widetilde{D}_\gamma}\varepsilon,$$

where $\widetilde{D}_\gamma = p + \sigma\gamma + \beta\gamma$. Therefore

$$T^{\text{unif}}_{\text{total}}(\widetilde{N}^*) \asymp \widetilde{C}_A^{\gamma/p} \left(\frac{\beta\gamma}{\widetilde{D}_\gamma}\right)^{-\gamma/p} \left(\frac{C_0 \widetilde{D}_\gamma}{p+\sigma\gamma}\right)^{\frac{p+\sigma\gamma}{\beta p}} \varepsilon^{-\widetilde{D}_\gamma/(\beta p)}.$$

Dividing the two estimates yields

$$R_{\text{opt}}(\varepsilon,\gamma) := \frac{T_{\text{total}}(N^*)}{T^{\text{unif}}_{\text{total}}(\widetilde{N}^*)} \asymp K_\gamma \varepsilon^{(\widetilde{D}_\gamma - D_\gamma)/(\beta p)}.$$

Since

$$\widetilde{D}_\gamma - D_\gamma = p + \sigma\gamma - \rho(p+\gamma),$$

we obtain

$$R_{\text{opt}}(\varepsilon,\gamma) \asymp K_\gamma \varepsilon^{\frac{p+\sigma\gamma-\rho(p+\gamma)}{\beta p}}.$$

Thus $R_{\text{opt}}(\varepsilon,\gamma) \to 0$ as $\varepsilon \to 0$ whenever

$$\rho(p+\gamma) < p + \sigma\gamma.$$

E Proof of Prop. 2

The idea of the proof is as follows.

Proof sketch The estimate g_k enters the per-step inequality only through the inner product $\langle g_k, v_k\rangle$ with an $\mathcal{F}_{k-1}$-measurable vector v_k (e.g. $v_k = x_k - x^*$ for GD) and through the squared norm $\|g_k\|^2$. Taking the conditional expectation gives $\mathbb{E}[\langle g_k, v_k\rangle \mid \mathcal{F}_{k-1}] = \langle \nabla f(x_k), v_k\rangle + \langle b_k(x_k), v_k\rangle$; Cauchy–Schwarz on the bias term produces $\|b_k\| \cdot \|v_k\|$, while $\mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_{k-1}] \le V_k$ handles the squared-norm term. A Young-type inequality applied to any cross term involving $\|b_k\| \cdot \|v_k\|$ yields the quadratic contribution $\|b_k\|^2$. Telescoping across iterations gives the three sums in equation 6.

Table 1 provides specific examples of error decomposition equation 6. Namely, a_k multiplies the linear bias term, c_k multiplies the squared bias term, and e_k multiplies the second-moment term of the stochastic gradient estimator. These coefficients depend only on the underlying first-order method and not on the particular zeroth-order oracle.

Table 1: Coefficients a_k, c_k, e_k in the error decomposition equation 6. A dash "—" indicates the term is absent.

Method	$\mathbf{Gap}_N$	a_k	c_k	e_k	**Ref.**
Mirror Descent (cvx)	$\mathbb{E}[f(\bar{x}_N)] - f^*$	$\frac{\alpha_k D_K}{S_N}$	—	$\frac{\alpha_k^2}{2S_N}$	Prop. 10
SGD (nonconvex)	$\mathbb{E}[\frac{1}{S_N}\sum_k \alpha_k \|\nabla f(x_k)\|^2]$	—	$\frac{\alpha_k}{S_N}$	$\frac{L\alpha_k^2}{S_N}$	Prop. 11
SGD (μ-str. cvx)	$f(x_N) - f^*$	—	$\frac{L\rho_N\alpha_k}{\mu\rho_{k+1}}$	$\frac{L\rho_N\alpha_k^2}{\rho_{k+1}}$	Prop. 12
Accelerated SGD (cvx)	$f(x_N^{ag}) - f^*$	$\frac{\gamma_k R}{\beta_N\gamma_N}$	$\frac{\gamma_k^2}{\beta_N\gamma_N}$	$\frac{\gamma_k^2}{\beta_N\gamma_N}$	Prop. 13

Notation. $S_N := \sum_{t=0}^{N-1}\alpha_t$; $\rho_{k+1} := \prod_{s=0}^{k}(1-\alpha_s\mu)$; $D_K := \sup_{x\in K}\|x-x^*\|$; $R := \|x_0-x^*\|$; $\beta_N\gamma_N$ is the normalization of the accelerated scheme; B is the batch size. All equalities hold up to universal constants.

This separation is useful because the oracle-specific estimates enter only through bounds on the conditional bias b_k and second moment V_k. In particular, mirror descent accumulates the bias linearly through the regret inequality, nonconvex SGD converts the bias-gradient cross term into a squared-bias contribution, strongly convex SGD propagates the errors through the contraction weights, and accelerated SGD amplifies them according to the acceleration weights.

Proposition 10 (Mirror descent with biased stochastic gradients)**.** *Let $K \subset \mathbb{R}^d$ be convex and compact, and let $x^* \in K$ be a minimizer of a convex function f. Consider mirror descent*

$$x_{k+1} = \arg\min_{x\in K}\{\alpha_k\langle \hat{g}_k, x\rangle + D_\psi(x, x_k)\},$$

where ψ is 1-strongly convex with respect to a norm $\|\cdot\|$ and D_ψ is the associated Bregman divergence. Suppose

$$\mathbb{E}[\hat{g}_k \mid \mathcal{F}_k] = \nabla f(x_k) + b_k(x_k), \qquad \mathbb{E}[\|\hat{g}_k\|_*^2 \mid \mathcal{F}_k] \le V_k(x_k).$$

Let $D_K := \sup_{x\in K}\|x - x^*\|$ and $S_N := \sum_{k=0}^{N-1}\alpha_k$. Then the averaged point $\bar{x}_N := S_N^{-1}\sum_{k=0}^{N-1}\alpha_k x_k$ satisfies

$$\mathbb{E}[f(\bar{x}_N) - f^*] \le \frac{D_\psi(x^*, x_0)}{S_N} + \sum_{k=0}^{N-1}\left[\frac{\alpha_k D_K}{S_N}\mathbb{E}\|b_k(x_k)\| + \frac{\alpha_k^2}{2S_N}\mathbb{E}V_k(x_k)\right].$$

Thus, in equation 6,

$$a_k = \frac{\alpha_k D_K}{S_N}, \qquad c_k = 0, \qquad e_k = \frac{\alpha_k^2}{2S_N}.$$

Proof. The standard mirror-descent three-point inequality gives, for every $x \in K$,

$$\alpha_k\langle \hat{g}_k, x_k - x\rangle \le D_\psi(x, x_k) - D_\psi(x, x_{k+1}) + \frac{\alpha_k^2}{2}\|\hat{g}_k\|_*^2.$$

Taking $x = x^*$ and conditional expectation with respect to $\mathcal{F}_k$,

$$\alpha_k\langle \nabla f(x_k), x_k - x^*\rangle \le \mathbb{E}[D_\psi(x^*, x_k) - D_\psi(x^*, x_{k+1}) \mid \mathcal{F}_k] + \frac{\alpha_k^2}{2}V_k(x_k) + \alpha_k\langle b_k(x_k), x^* - x_k\rangle.$$

By convexity, $f(x_k) - f^* \le \langle \nabla f(x_k), x_k - x^*\rangle$, and by Cauchy–Schwarz,

$$\langle b_k(x_k), x^* - x_k\rangle \le D_K\|b_k(x_k)\|.$$

Summing over k, taking total expectation, and using convexity once more, $f(\bar{x}_N) \le S_N^{-1}\sum_k \alpha_k f(x_k)$, yields the claim. □

Proposition 11 (Nonconvex SGD with biased stochastic gradients)**.** *Let f be L-smooth and consider the SGD recursion*

$$x_{k+1} = x_k - \alpha_k \hat{g}_k,$$

where

$$\mathbb{E}[\hat{g}_k \mid \mathcal{F}_k] = \nabla f(x_k) + b_k(x_k), \qquad \mathbb{E}[\|\hat{g}_k\|^2 \mid \mathcal{F}_k] \le V_k(x_k).$$

Let $S_N := \sum_{k=0}^{N-1} \alpha_k$. Then

$$\frac{1}{S_N} \sum_{k=0}^{N-1} \alpha_k \mathbb{E}\|\nabla f(x_k)\|^2 \lesssim \frac{f(x_0) - f_{\inf}}{S_N} + \sum_{k=0}^{N-1} \left[\frac{\alpha_k}{S_N} \mathbb{E}\|b_k(x_k)\|^2 + \frac{L\alpha_k^2}{S_N} \mathbb{E} V_k(x_k) \right].$$

Thus, in equation 6,

$$a_k = 0, \qquad c_k = \frac{\alpha_k}{S_N}, \qquad e_k = \frac{L\alpha_k^2}{S_N},$$

up to universal constants.

Proof. By L-smoothness,

$$f(x_{k+1}) \le f(x_k) - \alpha_k \langle \nabla f(x_k), \hat{g}_k \rangle + \frac{L\alpha_k^2}{2} \|\hat{g}_k\|^2.$$

Taking conditional expectation gives

$$\mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] \le f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 - \alpha_k \langle \nabla f(x_k), b_k(x_k) \rangle + \frac{L\alpha_k^2}{2} V_k(x_k).$$

Using Young's inequality,

$$-\langle \nabla f(x_k), b_k(x_k) \rangle \le \frac{1}{2} \|\nabla f(x_k)\|^2 + \frac{1}{2} \|b_k(x_k)\|^2,$$

we obtain

$$\frac{\alpha_k}{2} \|\nabla f(x_k)\|^2 \le f(x_k) - \mathbb{E}[f(x_{k+1}) \mid \mathcal{F}_k] + \frac{\alpha_k}{2} \|b_k(x_k)\|^2 + \frac{L\alpha_k^2}{2} V_k(x_k).$$

Summing and dividing by S_N gives the stated bound, after absorbing absolute constants. □

Proposition 12 (Strongly convex SGD with biased stochastic gradients)**.** *Let f be μ-strongly convex and L-smooth, and let $x^* = \arg\min_x f(x)$. Consider*

$$x_{k+1} = x_k - \alpha_k \widehat{g}_k, \qquad 0 < \alpha_k \le \frac{1}{L},$$

where

$$\mathbb{E}[\widehat{g}_k \mid \mathcal{F}_k] = \nabla f(x_k) + b_k(x_k), \qquad \mathbb{E}[\|\widehat{g}_k\|^2 \mid \mathcal{F}_k] \le V_k(x_k).$$

Define

$$\rho_0 := 1, \qquad \rho_{k+1} := \prod_{s=0}^{k} (1 - \mu\alpha_s).$$

Then

$$\mathbb{E}\|x_N - x^*\|^2 \lesssim \rho_N \|x_0 - x^*\|^2 + \rho_N \sum_{k=0}^{N-1} \left[\frac{\alpha_k}{\mu\rho_{k+1}} \mathbb{E}\|b_k(x_k)\|^2 + \frac{\alpha_k^2}{\rho_{k+1}} \mathbb{E} V_k(x_k) \right].$$

Consequently,

$$\mathbb{E}[f(x_N) - f^*] \lesssim L\rho_N \|x_0 - x^*\|^2 + \sum_{k=0}^{N-1} \left[\frac{L\rho_N \alpha_k}{\mu\rho_{k+1}} \mathbb{E}\|b_k(x_k)\|^2 + \frac{L\rho_N \alpha_k^2}{\rho_{k+1}} \mathbb{E} V_k(x_k) \right].$$

Thus, in the decomposition equation 6, the corresponding coefficients are

$$a_k = 0, \qquad c_k = \frac{L\rho_N \alpha_k}{\mu\rho_{k+1}}, \qquad e_k = \frac{L\rho_N \alpha_k^2}{\rho_{k+1}},$$

up to universal constants.

Proof. Using the update rule,

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha_k \langle \widehat{g}_k, x_k - x^* \rangle + \alpha_k^2 \|\widehat{g}_k\|^2.$$

Taking conditional expectation gives

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \le \|x_k - x^*\|^2 - 2\alpha_k \langle \nabla f(x_k), x_k - x^* \rangle - 2\alpha_k \langle b_k(x_k), x_k - x^* \rangle + \alpha_k^2 V_k(x_k).$$

Since f is μ-strongly convex and x^* is its minimizer,

$$\langle \nabla f(x_k), x_k - x^* \rangle \ge \mu \|x_k - x^*\|^2.$$

For the bias term, Young's inequality gives

$$2\alpha_k \big| \langle b_k(x_k), x_k - x^* \rangle \big| \le \mu \alpha_k \|x_k - x^*\|^2 + \frac{\alpha_k}{\mu} \|b_k(x_k)\|^2.$$

Therefore,

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \le (1 - \mu\alpha_k) \|x_k - x^*\|^2 + \frac{\alpha_k}{\mu} \|b_k(x_k)\|^2 + \alpha_k^2 V_k(x_k).$$

Taking total expectation and unrolling the recursion yields

$$\mathbb{E}\|x_N - x^*\|^2 \le \rho_N \|x_0 - x^*\|^2 + \rho_N \sum_{k=0}^{N-1} \left[\frac{\alpha_k}{\mu \rho_{k+1}} \mathbb{E}\|b_k(x_k)\|^2 + \frac{\alpha_k^2}{\rho_{k+1}} \mathbb{E} V_k(x_k) \right].$$

Finally, by L-smoothness and $\nabla f(x^*) = 0$,

$$f(x_N) - f^* \le \frac{L}{2} \|x_N - x^*\|^2.$$

Multiplying the previous display by $L/2$ and absorbing universal constants gives the claimed bound. □

Proposition 13 (Accelerated SGD with biased stochastic gradients)**.** *Consider the accelerated stochastic scheme used in the ZO-AccSGD analysis, with weights (β_k, γ_k). Let the estimator g_k satisfy*

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla f(y_k) + b_k(y_k), \qquad \mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_k] \le V_k(y_k).$$

Assume that in the accelerated Lyapunov recursion the gradient error enters as

$$\beta_N \gamma_N \big(f(x_N^{ag}) - f^* \big) \le R_0 + \sum_{k=0}^{N-1} \left[\gamma_k \langle g_k - \nabla f(y_k), x^* - z_k \rangle + \gamma_k^2 \|g_k - \nabla f(y_k)\|^2 \right],$$

where R_0 is independent of the oracle fidelity. If $\|z_k - x^*\| \le R$ along the trajectory, then

$$\mathbb{E}[f(x_N^{ag}) - f^*] \le \widetilde{E}_0(N, \Theta) + \sum_{k=0}^{N-1} \left[\frac{\gamma_k R}{\beta_N \gamma_N} \mathbb{E}\|b_k(y_k)\| + \frac{\gamma_k^2}{\beta_N \gamma_N} \mathbb{E}\|b_k(y_k)\|^2 + \frac{\gamma_k^2}{\beta_N \gamma_N} \mathbb{E} V_k(y_k) \right].$$

Thus, in the decomposition (7)*, the corresponding coefficients are*

$$a_k = \frac{\gamma_k R}{\beta_N \gamma_N}, \qquad c_k = \frac{\gamma_k^2}{\beta_N \gamma_N}, \qquad e_k = \frac{\gamma_k^2}{\beta_N \gamma_N}.$$

Proof. Let

$$\varepsilon_k := g_k - \mathbb{E}[g_k \mid \mathcal{F}_k].$$

Then

$$\mathbb{E}[\varepsilon_k \mid \mathcal{F}_k] = 0, \qquad g_k - \nabla f(y_k) = b_k(y_k) + \varepsilon_k.$$

For the linear term in the accelerated recursion, taking conditional expectation gives

$$\mathbb{E}[\langle g_k - \nabla f(y_k), x^* - z_k \rangle \mid \mathcal{F}_k] = \langle b_k(y_k), x^* - z_k \rangle.$$

Hence, by Cauchy–Schwarz and the assumption $\|z_k - x^*\| \le R$,

$$\mathbb{E}[\langle g_k - \nabla f(y_k), x^* - z_k \rangle \mid \mathcal{F}_k] \le R \|b_k(y_k)\|.$$

For the quadratic term, we first note that

$$\begin{aligned}\mathbb{E}\big[\|g_k - \nabla f(y_k)\|^2 \mid \mathcal{F}_k\big] &= \mathbb{E}\big[\|b_k(y_k) + \varepsilon_k\|^2 \mid \mathcal{F}_k\big] \\ &= \|b_k(y_k)\|^2 + \mathbb{E}[\|\varepsilon_k\|^2 \mid \mathcal{F}_k],\end{aligned}$$

because $\mathbb{E}[\varepsilon_k \mid \mathcal{F}_k] = 0$. Moreover,

$$\begin{aligned}\mathbb{E}[\|\varepsilon_k\|^2 \mid \mathcal{F}_k] &= \mathbb{E}\big[\|g_k - \mathbb{E}[g_k \mid \mathcal{F}_k]\|^2 \mid \mathcal{F}_k\big] \\ &= \mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_k] - \|\mathbb{E}[g_k \mid \mathcal{F}_k]\|^2 \\ &\le \mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_k] \le V_k(y_k).\end{aligned}$$

Therefore,

$$\mathbb{E}\big[\|g_k - \nabla f(y_k)\|^2 \mid \mathcal{F}_k\big] \le \|b_k(y_k)\|^2 + V_k(y_k).$$

Taking conditional expectation in the accelerated Lyapunov recursion and using the two bounds above yields

$$\beta_N \gamma_N \mathbb{E}[f(x_N^{ag}) - f^*] \le \widetilde{R}_0 + \sum_{k=0}^{N-1} \left[\gamma_k R \mathbb{E}\|b_k(y_k)\| + \gamma_k^2 \mathbb{E}\|b_k(y_k)\|^2 + \gamma_k^2 \mathbb{E} V_k(y_k)\right],$$

where $\widetilde{R}_0$ is independent of the oracle fidelity. Dividing by $\beta_N \gamma_N$ and absorbing $\widetilde{R}_0/(\beta_N \gamma_N)$ into $\widetilde{E}_0(N, \Theta)$ gives the claim. □

F Proof of Prop. 3

Proof. Introduce the exact-oracle estimator

$$g_k^0 := \frac{d}{2h}\big(f(x_k^+) - f(x_k^-)\big) K(r_k)\zeta_k.$$

The noisy estimator can be decomposed as

$$g_k = g_k^0 + \frac{d}{2h} K(r_k)\zeta_k \Delta_k.$$

For the conditional bias, we have

$$\begin{aligned}\|b_k(x_k)\| &= \|\mathbb{E}[g_k \mid \mathcal{F}_k] - \nabla f(x_k)\| \\ &\le \big\|\mathbb{E}[g_k^0 \mid \mathcal{F}_k] - \nabla f(x_k)\big\| + \frac{d}{2h}\mathbb{E}[|K(r_k)\Delta_k| \,|\, \mathcal{F}_k]\,.\end{aligned}$$

The normalization in Def. 5 gives $\mathbb{E}[r_k K(r_k)] = 1$. Since $\mathbb{E}[\zeta_k \zeta_k^\top] = I_d/d$, the first-order Taylor term is reproduced exactly:

$$\mathbb{E}\left[\frac{d}{2h}\big(2h r_k \langle \nabla f(x_k), \zeta_k\rangle\big) K(r_k)\zeta_k \,\middle|\, \mathcal{F}_k\right] = \nabla f(x_k).$$

The remaining Taylor terms through order $\lfloor \beta \rfloor$ vanish by the kernel moment conditions, while the Hölder remainder yields

$$\big\|\mathbb{E}[g_k^0 \mid \mathcal{F}_k] - \nabla f(x_k)\big\| \lesssim \kappa_\beta L_\beta h^{\beta-1}.$$

By conditional Cauchy–Schwarz and the assumed joint-moment bound,

$$\begin{aligned}\mathbb{E}[|K(r_k)\Delta_k| \,|\, \mathcal{F}_k] &\le \mathbb{E}[K(r_k)^2 \Delta_k^2 \mid \mathcal{F}_k]^{1/2} \\ &\lesssim \sqrt{\kappa}\,\delta_k.\end{aligned}$$

Consequently,

$$\|b_k(x_k)\| \lesssim \kappa_\beta L_\beta h^{\beta-1} + \frac{d\sqrt{\kappa}}{h}\delta_k.$$

For the conditional second moment, using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and $\|\zeta_k\| = 1$, we obtain

$$\begin{aligned} V_k(x_k) &= \mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_k] \\ &\le 2\mathbb{E}[\|g_k^0\|^2 \mid \mathcal{F}_k] + \frac{d^2}{2h^2}\mathbb{E}[K(r_k)^2\Delta_k^2 \mid \mathcal{F}_k] \\ &= \frac{d^2}{2h^2}\mathbb{E}\Big[\big(f(x_k^+) - f(x_k^-)\big)^2 K(r_k)^2 \,\Big|\, \mathcal{F}_k\Big] \\ &\quad + \frac{d^2}{2h^2}\mathbb{E}[K(r_k)^2\Delta_k^2 \mid \mathcal{F}_k]. \end{aligned}$$

Applying the assumed noise bound yields

$$V_k(x_k) \lesssim \frac{d^2}{h^2}\mathbb{E}\Big[\big(f(x_k^+) - f(x_k^-)\big)^2 K(r_k)^2 \,\Big|\, \mathcal{F}_k\Big] + \frac{\kappa d^2}{h^2}\delta_k^2.$$

□

G Proofs and remarks for the adversarial-noise results

Let x_N be an estimate obtained after N steps and let the corresponding approximation error be $\mathcal{E}(N, \delta, \Theta)$. Dvurechensky & Gasnikov (2016) and Devolder et al. (2013) ensure that if f is strongly convex (Asm. 4, $\mu > 0$), the bound is

$$f(x_N) - f^* \lesssim \mathcal{E}(N, \delta, \Theta) := LR^2 \exp\left(-\left(\frac{\mu}{L}\right)^{\frac{1}{p}} N\right) + \left(\frac{L}{\mu}\right)^{\frac{2p-1}{p}} d\delta,$$

where $R := \|x_0 - x^*\|$. Applying Corollary 2 for convergence above (case) one can obtain result provided in Proposition 4 (case $\mu > 0$).

Remark 5. Analysis of equation 12 demonstrates that if $\gamma < 1$, the wall-clock time complexity is optimal (minimal) for $p = 2$. This corresponds to FGM. If $\gamma \ge 1$, the optimal wall-clock time complexity is achieved at $p = 1$. This corresponds to GM. Thus, the upper bound on the wall-clock time complexity is tighter for GM; so, accelerated methods can be wall-clock inferior to non-accelerated schemes.

Furthermore, if f is convex (i.e., Assumption 4 with $\mu = 0$ holds), convergence can be expressed as:

$$f(x_N) - f^* \lesssim \mathcal{E}(N, \delta, \Theta) := \frac{LR^2}{N^p} + \sqrt{d\delta L}\tilde{R}_N + N^{p-1}d\delta, \tag{17}$$

where $\tilde{R}_N = \max\limits_{k \le N-1} \|x^k - x^*\|_2$ is a parameter depending on the convergence trajectory of the chosen algorithm. Using stopping criteria similar to those proposed in Vasin et al. (2023) or setting $\tilde{R}_N = O(R)$ and Corollary 2 the second part of result of the Proposition 4 can be obtained.

Proof of Proposition 4 (case $\mu = 0$)

Proof. Convergence bound equation 17 can be expressed as:

$$\frac{LR^2}{N^p} + 2\max\{\sqrt{d\delta L}R, N^{p-1}d\delta\}.$$

Then we can apply Corollary 2 to both components of error accumulation. Estimating iterations amount:

$$\frac{LR^2}{N^p} \leqslant \frac{\varepsilon}{3} \Rightarrow N = O\left(\left(\frac{3LR^2}{\varepsilon}\right)^{1/p}\right).$$

1. Applying Corollary 2 to component $\sqrt{d\delta L}R$:

$$\begin{aligned} T_{\text{total}}^{(1)} &\leqslant N\left(\frac{2\varepsilon}{3}\right)^{-2\gamma}\left(2\sqrt{dL}R\right)^{2\gamma} \leqslant \left(\frac{3LR^2}{\varepsilon}\right)^{1/p}\left(\frac{9}{4\varepsilon^2}\right)^{\gamma}\left(4dLR^2\right)^{\gamma} \\ &= \left(\frac{3LR^2}{\varepsilon}\right)^{1/p}\left(\frac{9dLR^2}{\varepsilon^2}\right)^{\gamma} = O\left(\left(\frac{\varepsilon}{9d}\right)^{-\gamma}\left(\frac{LR^2}{\varepsilon}\right)^{\gamma+\frac{1}{p}}\right). \end{aligned}$$

2. Applying Corollary 2 to component $N^{p-1}d\delta$:

$$\begin{aligned} T^{(2)}_{\text{total}} &\leqslant N\left(\frac{2\varepsilon}{3}\right)^{-\gamma}\left(N^{p-1}d\right)^{\gamma} \leqslant \left(\frac{3LR^2}{\varepsilon}\right)^{1/p}\left(\frac{3}{2\varepsilon}\right)^{\gamma}\left(\left(\frac{3LR^2}{\varepsilon}\right)^{\gamma(p-1)/p}\right)d^{\gamma} \\ &= O\left(\left(\frac{2\varepsilon}{3d}\right)^{-\gamma}\left(\frac{3LR^2}{\varepsilon}\right)^{\gamma+\frac{1-\gamma}{p}}\right). \end{aligned}$$

Then we can use maximum of $T^{(1)}_{\text{total}}$ and $T^{(2)}_{\text{total}}$ for total estimation of wall clock time for $\mathcal{E}(N,\delta,\Theta) \leqslant \varepsilon$ solution. Note, that we can assume, that $\varepsilon \leqslant LR^2$, since L-smoothness (Assumptions 4) implies $f(x) - f^* \leqslant \frac{1}{2}LR^2$. Thus we can estimate T_{total}:

$$\begin{aligned} T_{\text{total}} &= O\left(\max\left\{T^{(1)}_{\text{total}}, T^{(2)}_{\text{total}}\right\}\right) \\ &= O\left(\max\left\{\left(\frac{\varepsilon}{9d}\right)^{-\gamma}\left(\frac{LR^2}{\varepsilon}\right)^{\gamma+\frac{1}{p}}, \left(\frac{2\varepsilon}{3d}\right)^{-\gamma}\left(\frac{3LR^2}{\varepsilon}\right)^{\gamma+\frac{1-\gamma}{p}}\right\}\right) \\ &= O\left(\left(\frac{\varepsilon}{9d}\right)^{-\gamma}\cdot\max\left\{\left(\frac{LR^2}{\varepsilon}\right)^{\gamma+\frac{1}{p}}, \left(\frac{3LR^2}{\varepsilon}\right)^{\gamma+\frac{1-\gamma}{p}}\right\}\right) \\ &= O\left(\left(\frac{\varepsilon}{9d}\right)^{-\gamma}\left(\frac{3LR^2}{\varepsilon}\right)^{\gamma+\frac{1}{p}}\right) \end{aligned}$$

□

Remark 6. The convex case result of Proposition 4 indicates that for any $\gamma > 0$ the optimal choice of method is FGM (i.e., $p = 2$). Note that full gradient approximation requires d queries to the oracle (Gasnikov et al., 2023). Consequently, under sequential queries, the bounds equation 12 scale by a factor of d. However, the wall-clock time complexity remains invariant under parallel oracle queries.

Stonyakin et al. (2020; 2021) obtained convergence results for fast gradient method and $(\{\delta_k\}_{k=0}^{N-1}, L, \mu)$ inexact models described at Devolder et al. (2013). To study wall-clock complexity for time-varying δ_k we introduce the following intermediate gradient method.

Algorithm 1 IGM (Intermediate Gradient Method).

Input: Starting point x^0, number of steps N, L - smoothness parameter, μ - strong-convexity parameter, $\nu \in [0, 1/2]$ - intermediate parameter.
Set $u^0 = x^0$.
Set $h = \dfrac{1}{4L}$,
Set $s = \left(1 + \dfrac{1}{4}\left(\dfrac{\mu}{2L}\right)^{\nu}\right)$.
Set $m = \left(1 - \dfrac{1}{4}\left(\dfrac{\mu}{2L}\right)^{\nu}\right)$.
Set $q = \frac{\mu}{16L}$.
Set $\omega = \dfrac{(m-s) + \sqrt{(s-m)^2 + 4mq}}{2m}$.
for $k = 0 \ldots N-1$ **do**
  $y^k = \dfrac{\omega u^k + x^k}{1+\omega}$,
  $u^{k+1} = (1-\omega)u^k + \omega y^k - \dfrac{2\omega}{\mu}\widetilde{\nabla} f(y^k)$,
  $x^{k+1} = y^k - h\widetilde{\nabla} f(y^k)$.
Output: x^N.

We will establish convergence of Algorithm 1 with inexact gradient $\widetilde{\nabla} f$, satisfying:

$$(\forall x \in \mathbb{R}^d) \quad \widetilde{\nabla} f(x) = \nabla f(x) + \zeta_a(x), \quad \|\zeta_a(x)\|_2 \leqslant \Delta. \tag{18}$$

Following (Nesterov, 2018, p. 83), let us introduce a parameterized set of functions Ψ, its element defined for $c \in \mathbb{R}$, $\kappa \in \mathbb{R}^{++}$, and $u \in \mathbb{R}^n$, as follows

$$\psi(x|c, \kappa, u) = c + \frac{\kappa}{2}\|x - u\|_2^2, \quad \forall x \in \mathbb{R}^n. \tag{19}$$

According to Nesterov (2018), we use the following properties of the class Ψ.

Lemma 2. *Let $\psi_1, \psi_2 \in \Psi$. Then $\forall \eta_1, \eta_2, c_1, c_2 \in \mathbb{R}, \forall \kappa_1, \kappa_2 \in \mathbb{R}^{++}$, and $\forall u, v \in \mathbb{R}^n$, we have*

$$\eta_1\psi_1(x|c_1, \kappa_1, u) + \eta_2\psi_2(x|c_2, \kappa_2, v) = \psi_3(x|c_3, \kappa_3, w),$$

where

$$c_3 = \eta_1 c_1 + \eta_2 c_2 + \frac{\eta_1\eta_2\kappa_1\kappa_2}{2(\eta_1\kappa_1 + \eta_2\kappa_2)}\|u - v\|_2^2,$$
$$\kappa_3 = \eta_1\kappa_1 + \eta_2\kappa_2, \quad w = \frac{\eta_1\kappa_1 u + \eta_2\kappa_2 v}{\eta_1\kappa_1 + \eta_2\kappa_2}.$$

Lemma 3. *Let $0 < \lambda < 1$, $A > 0$, $\psi_0, \psi \in \Psi$, $z \in \mathbb{R}^n$ such that*

$$f(z) \leqslant \min_{x \in \mathbb{R}^n} \psi(x) + A,$$

and

$$\psi(x) \leqslant \lambda\psi_0(x) + (1 - \lambda)f(x), \quad \forall x \in \mathbb{R}^n.$$

Then

$$f(z) - f^* \leqslant \lambda(\psi_0(x^*) - f^*) + A.$$

Proof.

$$\begin{aligned} f(z) - f^* &\leqslant \min_{x \in \mathbb{R}^n} \psi(x) + A - f^* \\ &\leqslant \psi(x^*) - f^* + A \leqslant \lambda\psi_0(x^*) + (1 - \lambda)f(x^*) - f^* + A \\ &= \lambda(\psi_0(x^*) - f^*) + A. \end{aligned}$$

□

Lemma 4. *Let $\widetilde{\nabla} f$ satisfy error condition equation 18, $0 < \mu \leqslant L$, then:*

$$\forall \nu \geqslant 0 \quad \|\widetilde{\nabla} f(x)\|_2^2 \geqslant \left(1 - \frac{1}{4}\left(\frac{\mu}{2L}\right)^\nu\right)\|\nabla f(x)\|_2^2 - \left(4\left(\frac{2L}{\mu}\right)^\nu - 1\right)\Delta^2,$$
$$\|\widetilde{\nabla} f(x)\|_2^2 \leqslant \left(1 + \frac{1}{4}\left(\frac{\mu}{2L}\right)^\nu\right)\|\nabla f(x)\|_2^2 + \left(1 + 4\left(\frac{2L}{\mu}\right)^\nu\right)\Delta^2.$$

Proof.

$$\begin{aligned} \|\widetilde{\nabla} f(x)\|_2^2 &= \|\nabla f(x) + \zeta_a(x)\|_2^2 = \|\nabla f(x)\|_2^2 + 2\langle\nabla f(x), \zeta_a(x)\rangle + \|\zeta_a(x)\|_2^2 \\ &\overset{\text{Fenchel ineq.}}{\leqslant} \left(1 + \frac{1}{\lambda}\right)\|\nabla f(x)\|_2^2 + (1 + \lambda)\|\zeta_a(x)\|_2^2 \\ &\overset{\lambda = 4(2L/\mu)^\nu}{\leqslant} \left(1 + \frac{1}{4}\left(\frac{\mu}{2L}\right)^\nu\right)\|\nabla f(x)\|_2^2 + \left(1 + 4\left(\frac{2L}{\mu}\right)^\nu\right)\Delta^2. \end{aligned}$$

Similarly for the lower bound. □

Lemma 5. *Let $\widetilde{\nabla} f$ satisfy error condition equation 18, $0 < \mu \leqslant L$, and $0 \leqslant \nu \leqslant 1/2$. Then:*

$$\begin{aligned}
\frac{\mu}{4}\left\|x - z + \frac{2}{\mu}\widetilde{\nabla} f(z)\right\|_2^2 &\leqslant \langle \nabla f(z), x - z\rangle + \frac{\mu}{2}\|x - z\|_2^2 \\
&+ \frac{1}{\mu}\left(1 + \frac{1}{4}\left(\frac{\mu}{2L}\right)^{\nu}\right)\|\nabla f(z)\|_2^2 \\
&+ \frac{1}{\mu}\left(4\left(\frac{2L}{\mu}\right)^{\nu} + 2\right)\Delta^2, \\
\frac{\mu}{4}\left\|x - z + \frac{2}{\mu}\widetilde{\nabla} f(z)\right\|_2^2 &\geqslant \langle \nabla f(z), x - z\rangle \\
&+ \frac{1}{\mu}\left(1 - \frac{1}{4}\left(\frac{\mu}{2L}\right)^{\nu}\right)\|\nabla f(z)\|_2^2 \\
&- \frac{4}{\mu}\left(\frac{2L}{\mu}\right)^{\nu}\Delta^2.
\end{aligned}$$

Proof. We will find bounds for each term of sum:

$$\frac{\mu}{4}\left\|x - z + \frac{2}{\mu}\widetilde{\nabla} f(z)\right\|_2^2 = \frac{\mu}{4}\|x - z\|_2^2 + \langle \widetilde{\nabla} f(z), x - z\rangle + \frac{1}{\mu}\|\widetilde{\nabla} f(z)\|_2^2.$$

Linear form:

$$\begin{aligned}
\langle \widetilde{\nabla} f(z), x - z\rangle &\leqslant \langle \nabla f(z), x - z\rangle + \frac{1}{\mu}\Delta^2 + \frac{\mu}{4}\|x - z\|_2^2. \\
\langle \widetilde{\nabla} f(z), x - z\rangle &\geqslant \langle \nabla f(z), x - z\rangle - \frac{1}{\mu}\Delta^2 - \frac{\mu}{4}\|x - z\|_2^2.
\end{aligned}$$

Quadratic term:

$$\begin{aligned}
\|\widetilde{\nabla} f(z)\|_2^2 &\overset{\text{Lemma 4}}{\leqslant} \left(1 + \frac{1}{4}\left(\frac{\mu}{2L}\right)^{\nu}\right)\|\nabla f(z)\|_2^2 + \left(4\left(\frac{2L}{\mu}\right)^{\nu} + 1\right)\Delta^2. \\
\|\widetilde{\nabla} f(z)\|_2^2 &\overset{\text{Lemma 4}}{\geqslant} \left(1 - \frac{1}{4}\left(\frac{\mu}{2L}\right)^{\nu}\right)\|\nabla f(z)\|_2^2 - \left(4\left(\frac{2L}{\mu}\right)^{\nu} - 1\right)\Delta^2.
\end{aligned}$$

Then we can sum up bounds above:

$$\begin{aligned}
\frac{\mu}{4}\left\|x - z + \frac{2}{\mu}\widetilde{\nabla} f(z)\right\|_2^2 &\leqslant \langle \nabla f(z), x - z\rangle + \frac{\mu}{2}\|x - z\|_2^2 \\
&+ \frac{1}{\mu}\left(1 + \frac{1}{4}\left(\frac{\mu}{2L}\right)^{\nu}\right)\|\nabla f(z)\|_2^2 \\
&+ \frac{1}{\mu}\left(4\left(\frac{2L}{\mu}\right)^{\nu} + 1\right)\Delta^2 + \frac{\Delta^2}{\mu} \\
&\leqslant \langle \nabla f(z), x - z\rangle + \frac{\mu}{2}\|x - z\|_2^2 \\
&+ \frac{1}{\mu}\left(1 + \frac{1}{4}\left(\frac{\mu}{2L}\right)^{\nu}\right)\|\nabla f(z)\|_2^2 \\
&+ \frac{1}{\mu}\left(4\left(\frac{2L}{\mu}\right)^{\nu} + 2\right)\Delta^2. \\
\frac{\mu}{4}\left\|x - z + \frac{2}{\mu}\widetilde{\nabla} f(z)\right\|_2^2 &\geqslant \langle \nabla f(z), x - z\rangle \\
&+ \frac{1}{\mu}\left(1 - \frac{1}{4}\left(\frac{\mu}{2L}\right)^{\nu}\right)\|\nabla f(z)\|_2^2 \\
&- \frac{4}{\mu}\left(\frac{2L}{\mu}\right)^{\nu}\Delta^2.
\end{aligned}$$

□

Lemma 6. *Let f satisfy Assumption 4, let $\widetilde{\nabla} f$ satisfy equation 18, and let $0 \leqslant \nu \leqslant 1/2$. Then for any $z \in \mathbb{R}^n$, there is $\phi_{z,\nu} \in \Psi$ such that:*

$$\phi_{z,\nu}(x) = c_\phi + \frac{\kappa_\phi}{2}\|x - u_\phi\|_2^2 \leqslant f(x), \quad \forall x \in \mathbb{R}^n.$$
$$c_\phi := f(z) - \frac{1}{\mu}\left(1 + \frac{1}{4}\left(\frac{\mu}{2L}\right)^\nu\right)\|\nabla f(z)\|_2^2 - \frac{1}{\mu}\left(4\left(\frac{2L}{\mu}\right)^\nu + 2\right)\Delta^2,$$
$$\kappa_\phi := \mu/2,$$
$$u_\phi := z - \frac{2}{\mu}\widetilde{\nabla} f(z).$$

Proof. Proof follows directly from Lemma 5 and strong convexity Def. equation 4. □

Now, for Algorithm 1, we have $u^0 = x^0, x^0 \in \mathbb{R}^d$. Let u^k be defined in Algorithm 1 and ω defined in Algorithm 1. Let us define the following sequences:

$$\lambda_0 = 1, \quad c_0 = f(x^0), \quad \lambda_{k+1} = (1-\omega)\lambda_k, \quad \forall k \geqslant 0, \tag{20}$$
$$\psi_k(x|c_k, \mu/2, u^k) = c_k + \frac{\mu}{4}\|x - u^k\|_2^2, \quad \forall k \geqslant 0 \tag{21}$$

Also $\phi_{y^k,\nu}$ is the function that is obtained from Lemma 6. We define, recursively

$$\psi_{k+1}(x|c_{k+1}, \mu/2, u^{k+1}) := (1-\omega)\psi_k(x|c_k, \mu/2, u^k) + \omega\phi_{y^k,\nu}(x), \quad \forall k \geqslant 0. \tag{22}$$

Error accumulation sequence also required to be defined:

$$E_0 = 0, \quad E_{k+1} = (1-\omega)E_k + 3\left(\frac{\mu}{L}\right)^{1-2\nu}\frac{\Delta_k^2}{\mu}, \quad \|\widetilde{\nabla} f(y^k) - \nabla f(y^k)\|_2 \leqslant \Delta_k. \tag{23}$$

The following Lemma motivates the definition of function $\phi_{y^k,\nu}$.

Lemma 7. *Let $\{\lambda_k\}_{k\geqslant 0}$ be a sequence defined in equation 20 and $\{\psi_k\}_{k\geqslant 0}$ be the corresponding functions equation 22. Then*

$$\psi_k(x) \leqslant \lambda_k\psi_0(x) + (1-\lambda_k)f(x), \quad \forall k \geqslant 0 \ \text{ and } \ \forall x \in \mathbb{R}^n.$$

Proof. By the principle of induction. For $k = 0$, the inequality is obvious. Assume that inequality takes place for k. From equation 22, and Lemma 6, we get

$$\begin{aligned}\psi_{k+1}(x) &= (1-\omega)\psi_k(x) + \omega\phi_{y^k,\nu}(x)\\ &\leqslant (1-\omega)\left((1-\lambda_k)f(x) + \lambda_k\psi_0(x)\right) + \omega f(x)\\ &= \lambda_{k+1}\psi_0(x) + (1-\lambda_{k+1})f(x).\end{aligned}$$

□

Lemma 8. *Let variables s, m defined at Algorithm 1, ω is the largest root of the quadratic equation*

$$m\omega^2 + (s-m)\omega - q = 0, \tag{24}$$

where $q = \frac{\mu}{16L}, 0 \leqslant \nu \leqslant 1/2$. Then:

$$\frac{1}{8}\left(\frac{\mu}{2L}\right)^{1-\nu} \leqslant \omega \leqslant \frac{1}{4}\left(\frac{\mu}{2L}\right)^{1-\nu}.$$

Proof. By solving the equation equation 24 and choosing the largest root we get

$$\omega = \frac{(m-s) + \sqrt{(s-m)^2 + 4mq}}{2m}. \tag{25}$$

Let us assume that $\omega > \frac{1}{4}\left(\frac{\mu}{2L}\right)^{1-\nu}$. Then

$$
\begin{aligned}
m\omega^2 + (s-m)\omega - q &> \left(1 - \frac{1}{4}\left(\frac{\mu}{2L}\right)^{\nu}\right)\frac{1}{16}\left(\frac{\mu}{2L}\right)^{2(1-\nu)} \\
&+ \frac{1}{2}\left(\frac{\mu}{2L}\right)^{\nu}\cdot\frac{1}{4}\left(\frac{\mu}{2L}\right)^{1-\nu} - \frac{1}{16}\frac{\mu}{L} > 0.
\end{aligned}
$$

which is a contradiction with equation 24. Hence $\omega \leqslant \frac{1}{4}\left(\frac{\mu}{2L}\right)^{1-\nu}$.

One can note, that $m < 1$ (from definition). So we can prove lower bound for ω. Let us prove by contradiction. To that end, we assume $\omega < \frac{1}{8}\left(\frac{\mu}{2L}\right)^{1-\nu}$. Then:

$$
\begin{aligned}
m\omega^2 + (s-m)\omega - q &< \frac{1}{64}\left(\frac{\mu}{2L}\right)^{2(1-\nu)} + \frac{1}{2}\left(\frac{\mu}{2L}\right)^{\nu}\cdot\frac{1}{8}\left(\frac{\mu}{2L}\right)^{1-\nu} - \frac{\mu}{16L} \\
&\overset{0\leqslant\nu\leqslant 1/2}{\leqslant} \frac{1}{128}\frac{\mu}{L} + \frac{1}{32}\frac{\mu}{L} - \frac{1}{16}\frac{\mu}{L} < 0.
\end{aligned}
$$

Thus, we come to a contradiction because ω is a root of the equation $m\omega^2 + (s-m)\omega - q = 0$.

□

Lemma 9. *Let f satisfies equation 4, $\widetilde{\nabla} f$ satisfies noise condition equation 18,*

$$
f(x^{k+1}) \leqslant f(y^k) - \frac{1}{16L}\|\nabla f(y^k)\|_2^2 + \frac{3}{16L}\Delta_k^2.
$$

Where:

$$
h = \frac{1}{4L}. \tag{26}
$$

Proof. Using smoothness of function f we get:

$$
f(x^{k+1}) \leqslant f(y^k) + \langle \nabla f(y^k), x^{k+1} - y^k\rangle + \frac{L}{2}\|x^{k+1} - y^k\|_2^2.
$$

Let us estimate terms. Linear form:

$$
\begin{aligned}
\langle \nabla f(y^k), x^{k+1} - y^k\rangle &= -h\langle \nabla f(y^k), \widetilde{\nabla} f(y^k)\rangle \\
&= -h\left(\|\nabla f(y^k)\|_2^2 + \langle \nabla f(y^k), \zeta_a(y^k)\rangle\right) \\
&\overset{\text{Fenchel ineq.}}{\leqslant} -h\left(\|\nabla f(y^k)\|_2^2 - \frac{\lambda}{2}\|\nabla f(y^k)\|_2^2 - \frac{1}{2\lambda}\Delta_k^2\right) \\
&\overset{\lambda=1}{\leqslant} -h\left(\frac{1}{2}\|\nabla f(y^k)\|_2^2 - \frac{1}{2}\Delta_k^2\right).
\end{aligned}
$$

Quadratic term:

$$
\|x^{k+1} - y^k\|_2^2 = h^2\|\widetilde{\nabla} f(y^k)\|_2^2 \leqslant 2h^2\left(\|\nabla f(y^k)\|_2^2 + \Delta_k^2\right).
$$

Combining this inequalities:

$$
\begin{aligned}
f(x^{k+1}) &\leqslant f(y^k) - h\left(\frac{1}{2}\|\nabla f(y^k)\|_2^2 - \frac{1}{2}\Delta_k^2\right) \\
&+ Lh^2\left(\|\nabla f(y^k)\|_2^2 + \Delta_k^2\right) \\
&= f(y^k) + h\|\nabla f(y^k)\|_2^2\left(Lh - \frac{1}{2}\right) \\
&+ h\Delta_k^2\left(\frac{1}{2} + hL\right) \\
&\overset{h \text{ definition}}{=} f(y^k) - \frac{1}{16L}\|\nabla f(y^k)\|_2^2 + \frac{3}{16L}\Delta_k^2.
\end{aligned}
$$

□

Lemma 10. *Let f satisfies equation 4, $\widetilde{\nabla} f$ satisfies equation 18, $\{x^k\}_{k\geqslant 0}$ be a sequence of points generated by Algorithm 1, and $c_k\,\forall k \geqslant 0$ be a minimal value of the function ψ_k corresponding to Algorithm 1, $0 \leqslant \nu \leqslant 1/2$. Then $c_k \geqslant f(x^k) - E_k\,\forall k \geqslant 0$, where E_k is defined in equation 23.*

Proof. We will use the principle of mathematical induction to prove the statement of this lemma. For $k = 0$, it is obvious that $c_0 \geqslant f(x^0)$. Now, let us assume that $c_k \geqslant f(x^k) - E_k$, and prove the statement for $k + 1$. For this, from Lemma 2, Def. equation 22 and definition of ϕ_{y^k} from Lemma 6 we have

$$
\begin{aligned}
c_{k+1} &= (1-\omega)c_k && (27)\\
&+ \omega\left(f(y^k) - \frac{s}{\mu}\|\nabla f(y^k)\|_2^2 - \frac{1}{\mu}\left(4\left(\frac{2L}{\mu}\right)^{\nu} + 2\right)\Delta_k^2\right) && (28)\\
&+ \frac{\omega(1-\omega)\mu}{4}\left\|y^k - u^k - \frac{2}{\mu}\widetilde{\nabla} f(y^k)\right\|_2^2. && (29)
\end{aligned}
$$

From Lemma 5 we can estimate last term:

$$
\begin{aligned}
\frac{\omega(1-\omega)\mu}{4}\left\|y^k - u^k - \frac{2}{\mu}\widetilde{\nabla} f(y^k)\right\|_2^2 &\geqslant \omega(1-\omega)\langle \nabla f(y^k), u^k - y^k\rangle\\
&+ \frac{m\omega(1-\omega)}{\mu}\|\nabla f(y^k)\|_2^2\\
&- \frac{4\omega(1-\omega)}{\mu}\left(\frac{2L}{\mu}\right)^{\nu}\Delta_k^2.
\end{aligned}
$$

Combining the inequalities above and grouping the corresponding terms:

$$
\begin{aligned}
c_{k+1} &\geqslant (1-\omega)c_k + \omega f(y^k) + \omega(1-\omega)\langle \nabla f(y^k), u^k - y^k\rangle\\
&- \frac{1}{\mu}\left(\omega s - \omega(1-\omega)m\right)\|\nabla f(y^k)\|_2^2\\
&\overset{(i)}{-} \frac{\omega}{\mu}\left(8\left(\frac{2L}{\mu}\right)^{\nu} + 2\right)\Delta_k^2,
\end{aligned}
$$

(i) follows from $0 \leqslant \omega < 1$ and Eq. equation 25:

$$
\omega\left(4\left(\frac{2L}{\mu}\right)^{\nu} + 2\right) + 4\omega(1-\omega)\left(\frac{2L}{\mu}\right)^{\nu} \leqslant \omega\left(8\left(\frac{2L}{\mu}\right)^{\nu} + 2\right).
$$

From convexity of function f and induction hypothesis:

$$
\begin{aligned}
(1-\omega)c_k &\geqslant (1-\omega)\left(f(x^k) - E_k\right)\\
&\geqslant (1-\omega)\left(f(y^k) + \langle \nabla f(y^k), x^k - y^k\rangle\right) - (1-\omega)E_k,
\end{aligned}
$$

then, using definition y^k we can eliminate the linear form:

$$
\langle \nabla f(y^k), \omega(1-\omega)(u^k - y^k) + (1-\omega)(x^k - y^k)\rangle = 0
$$

and continue c_{k+1} estimation:

$$
\begin{aligned}
c_{k+1} &\geqslant f(y^k) - (1-\omega)E_k\\
&- \frac{1}{\mu}\left(\omega s - \omega(1-\omega)m\right)\|\nabla f(y^k)\|_2^2\\
&- \frac{2\omega}{\mu}\left(4\left(\frac{2L}{\mu}\right)^{\nu} + 1\right)\Delta_k^2,
\end{aligned}
$$

From ω definition one can simplify the coefficient of the gradient norm:

$$\begin{aligned} c_{k+1} &\geqslant f(y^k) - (1-\omega)E_k - \frac{q}{\mu}\|\nabla f(y^k)\|_2^2 - \frac{2\omega}{\mu}\left(4\left(\frac{2L}{\mu}\right)^{\nu} + 1\right)\Delta_k^2 \\ &\geqslant f(y^k) - (1-\omega)E_k - \frac{1}{16L}\|\nabla f(y^k)\|_2^2 - \frac{2\omega}{\mu}\left(4\left(\frac{2L}{\mu}\right)^{\nu} + 1\right)\Delta_k^2. \\ &\overset{\text{Lemma 8}}{\geqslant} f(y^k) - (1-\omega)E_k - \frac{1}{16L}\|\nabla f(y^k)\|_2^2 - \left(\frac{\mu}{2L}\right)^{1-2\nu}\frac{5\Delta_k^2}{2\mu}. \end{aligned}$$

Since $x^{k+1} = y^k - h\widetilde{\nabla} f(y^k)$, where h selected consistently with the Lemma 9 we can obtain:

$$\begin{aligned} c_{k+1} &\geqslant f(x^{k+1}) - (1-\omega)E_k - \frac{3}{16L}\Delta_k^2 - \left(\frac{\mu}{2L}\right)^{1-2\nu}\frac{5\Delta_k^2}{2\mu} \\ &= f(x^{k+1}) - (1-\omega)E_k - \frac{\mu}{L}\cdot\frac{3\Delta_k^2}{16\mu} - \left(\frac{\mu}{2L}\right)^{1-2\nu}\frac{5\Delta_k^2}{2\mu} \\ &\geqslant f(x^{k+1}) - (1-\omega)E_k - 3\left(\frac{\mu}{L}\right)^{1-2\nu}\frac{\Delta_k^2}{\mu} \\ &\overset{E_k \text{ definition}}{=} f(x^{k+1}) - E_{k+1}. \end{aligned}$$

□

We can now prove the convergence of Algorithm 1.

Theorem 1. *Let f be an L-smooth and μ-strongly convex function (Assumption 4), and let $0 \leqslant \nu \leqslant 1/2$. Suppose that at every iteration k,*

$$\|\widetilde{\nabla} f(y^k) - \nabla f(y^k)\|_2 \leqslant \Delta_k.$$

Then Algorithm 1 generates x^N such that

$$\begin{aligned} f(x^N) - f^* &\leqslant \left(1 - \frac{1}{8}\left(\frac{\mu}{2L}\right)^{1-\nu}\right)^N \left(f(x^0) - f^* + \frac{\mu}{4}R^2\right) \\ &+ 3\left(\frac{\mu}{L}\right)^{1-2\nu}\sum_{k=0}^{N-1}\left(1 - \frac{1}{8}\left(\frac{\mu}{2L}\right)^{1-\nu}\right)^{N-k-1}\frac{\Delta_k^2}{\mu} \\ &\leqslant \left(1 - \frac{1}{16}\left(\frac{\mu}{L}\right)^{1-\nu}\right)^N LR^2 \\ &+ 3\left(\frac{\mu}{L}\right)^{1-2\nu}\sum_{k=0}^{N-1}\left(1 - \frac{1}{16}\left(\frac{\mu}{L}\right)^{1-\nu}\right)^{N-k-1}\frac{\Delta_k^2}{\mu} \end{aligned}$$

where $R := \|x^0 - x^*\|_2$.

Proof. From Lemma 10 we obtain:

$$f(x^N) \leqslant c_N + E_N \quad \text{see denotation equation 23 .}$$

Applying Lemma 7 and then Lemma 3:

$$\begin{aligned}
f(x^N) - f^* &\leqslant \lambda_N(\psi_0(x^*) - f^*) + E_N \\
&\overset{equation\ 23}{\leqslant} \lambda_N \left(f(x^0) - f^* + \frac{\mu}{4} R^2 \right) \\
&+ 3 \left(\frac{\mu}{L} \right)^{1-2\nu} \sum_{k=0}^{N-1} (1-\omega)^{N-k-1} \frac{\Delta_k^2}{\mu} \\
&\overset{\text{Lemma 8}}{\leqslant} \left(1 - \frac{1}{8} \left(\frac{\mu}{2L} \right)^{1-\nu} \right)^N \left(f(x^0) - f^* + \frac{\mu}{4} R^2 \right) \\
&+ 3 \left(\frac{\mu}{L} \right)^{1-2\nu} \sum_{k=0}^{N-1} \left(1 - \frac{1}{8} \left(\frac{\mu}{2L} \right)^{1-\nu} \right)^{N-k-1} \frac{\Delta_k^2}{\mu} \\
&\leqslant \left(1 - \frac{1}{16} \left(\frac{\mu}{L} \right)^{1-\nu} \right)^N LR^2 \\
&+ 3 \left(\frac{\mu}{L} \right)^{1-2\nu} \sum_{k=0}^{N-1} \left(1 - \frac{1}{16} \left(\frac{\mu}{L} \right)^{1-\nu} \right)^{N-k-1} \frac{\Delta_k^2}{\mu}.
\end{aligned}$$

□

So we can move to Prop. 5.

***Proof of Proposition** 5 (case $\mu > 0$)*

Proof. Let $\theta_p := \frac{1}{16} \left(\frac{\mu}{L} \right)^{1/p}$ and $\rho_p := 1 - \theta_p$. Substituting

$$\nu = 1 - \frac{1}{p}, \qquad \Delta_k = 2\sqrt{dL\delta_k}$$

into Theorem 1 gives

$$\begin{aligned}
f(x^N) - f^* &\leq \rho_p^N LR^2 + 3 \left(\frac{\mu}{L} \right)^{1-2\nu} \sum_{k=0}^{N-1} \rho_p^{N-k-1} \frac{\Delta_k^2}{\mu} \\
&= \rho_p^N LR^2 + 12d \left(\frac{\mu}{L} \right)^{2(1-p)/p} \sum_{k=0}^{N-1} \rho_p^{N-k-1} \delta_k.
\end{aligned}$$

Choose

$$N := \left\lceil 16 \left(\frac{L}{\mu} \right)^{1/p} \log \left(\frac{2LR^2}{\varepsilon} \right) \right\rceil.$$

Since $\rho_p^N \leqslant \exp(-\theta_p N)$, this guarantees $\rho_p^N LR^2 \leqslant \varepsilon/2$. Moreover, assuming $\log(2LR^2/\varepsilon) \geqslant 1$,

$$N \leq 17 \left(\frac{L}{\mu} \right)^{1/p} \log \left(\frac{2LR^2}{\varepsilon} \right).$$

It therefore suffices to allocate the remaining accuracy budget

$$\eta := \frac{\varepsilon}{2}$$

to the fidelity-dependent term. This term is linear in δ_k, so the fidelity exponent in Proposition 1 is equal to one. Its coefficients are

$$\alpha_k := 12d \left(\frac{\mu}{L} \right)^{2(1-p)/p} \rho_p^{N-k-1}, \qquad k = 0, \ldots, N-1.$$

Let $s := \frac{\gamma}{1+\gamma}$. Then

$$A_N := \sum_{k=0}^{N-1} \alpha_k^s = (12d)^s \left(\frac{\mu}{L}\right)^{\frac{2\gamma(1-p)}{p(1+\gamma)}} \sum_{j=0}^{N-1} \rho_p^{sj}.$$

The optimal interior schedule is

$$\delta_k^* = \frac{\eta}{A_N} \alpha_k^{-1/(1+\gamma)},$$

and Proposition 1 yields

$$T_{\text{total}}^{\text{IGM}} = \eta^{-\gamma} A_N^{1+\gamma} = \left(\frac{\varepsilon}{24d}\right)^{-\gamma} \left(\frac{\mu}{L}\right)^{\frac{2\gamma(1-p)}{p}} S_N^{1+\gamma},$$

where $S_N := \sum_{j=0}^{N-1} \rho_p^{sj}$.

We now use the two bounds $S_N \le N$ and $S_N \le \frac{1}{1-\rho_p^s}$.

Since $s \in (0,1)$, concavity of $x \mapsto x^s$ gives

$$\rho_p^s = (1-\theta_p)^s \le 1 - s\theta_p.$$

Therefore,

$$\frac{1}{1-\rho_p^s} \le \frac{1}{s\theta_p} = 16\frac{1+\gamma}{\gamma}\left(\frac{L}{\mu}\right)^{1/p}.$$

Consequently,

$$S_N \le \left(\frac{L}{\mu}\right)^{1/p} \min\left\{17\log\left(\frac{2LR^2}{\varepsilon}\right), 16\frac{1+\gamma}{\gamma}\right\}.$$

Substitution into the wall-clock bound gives

$$\boxed{\begin{aligned} T_{\text{total}}^{\text{IGM}} \le\ & \left(\frac{\varepsilon}{24d}\right)^{-\gamma} \left(\frac{L}{\mu}\right)^{2\gamma+\frac{1-\gamma}{p}} \\ & \times \left[\min\left\{17\log\left(\frac{2LR^2}{\varepsilon}\right), 16\frac{1+\gamma}{\gamma}\right\}\right]^{1+\gamma}. \end{aligned}}$$

In particular, for every fixed $\gamma > 0$,

$$T_{\text{total}}^{\text{IGM}} = O_\gamma\left(\left(\frac{\varepsilon}{d}\right)^{-\gamma} \left(\frac{L}{\mu}\right)^{2\gamma+\frac{1-\gamma}{p}}\right).$$

□

Remark 7. Same for Remark 5 one can note, that for $\gamma < 1$, the wall-clock time complexity is minimal for $p = 2$, and optimal for $p = 1$, if $\gamma \ge 1$.

***Proof of Proposition** 5 (convex case)*

Proof. Assume the nontrivial accuracy regime $0 < \varepsilon \leqslant LR^2$, and set

$$\mu_{\text{reg}} := \frac{\varepsilon}{R^2}, \qquad f_{\text{reg}}(x) := f(x) + \frac{\mu_{\text{reg}}}{2}\|x - x^0\|_2^2.$$

The function f_{reg} is μ_{reg}-strongly convex and $\bar{L}$-smooth with

$$\bar{L} := L + \mu_{\text{reg}} \leqslant 2L.$$

Let x_{reg}^* minimize f_{reg} and set $R_{\text{reg}} := \|x_{\text{reg}}^* - x^0\|_2$. Since $f_{\text{reg}}(x_{\text{reg}}^*) \leqslant f_{\text{reg}}(x^*)$ and $f(x_{\text{reg}}^*) \geqslant f(x^*)$, we have $R_{\text{reg}} \leqslant R$. Moreover, the regularized inexact gradient

$$\widetilde{\nabla} f_{\text{reg}}(x) := \widetilde{\nabla} f(x) + \mu_{\text{reg}}(x - x^0)$$

has the same approximation error as $\widetilde{\nabla} f(x)$.

Applying the strongly convex result with target accuracy $\varepsilon/2$ therefore gives

$$T_{\text{total}}^{\text{IGM}} = O_\gamma\left(\left(\frac{d}{\varepsilon}\right)^\gamma \left(\frac{\bar{L}}{\mu_{\text{reg}}}\right)^{2\gamma+\frac{1-\gamma}{p}}\right) = O_\gamma\left(\left(\frac{d}{\varepsilon}\right)^\gamma \left(\frac{LR^2}{\varepsilon}\right)^{2\gamma+\frac{1-\gamma}{p}}\right).$$

Finally, if $f_{\text{reg}}(x^N) - f_{\text{reg}}(x^*_{\text{reg}}) \leqslant \varepsilon/2$, then

$$f(x^N) - f^* \leqslant f_{\text{reg}}(x^N) - f_{\text{reg}}(x^*_{\text{reg}}) + \frac{\mu_{\text{reg}}}{2} R^2 \leqslant \varepsilon.$$

□

H Proof of Proposition 6

We derive the optimal wall-clock design for the adversarial-noise model directly from the bound in Sec. 4.2. Set

$$T := BN, \qquad p := \frac{\beta - 1}{\beta}.$$

The wall-clock objective is $T_{\text{total}} = T\delta^{-\gamma}$.

Step 1: optimizing out the smoothing radius.

The adversarial bound contains two bias terms depending on h and δ, $\kappa_\beta L_\beta h^{\beta-1}$ and $\frac{d\delta}{h}$. For fixed δ, we balance them:

$$\kappa_\beta L_\beta h^{\beta-1} \asymp \frac{d\delta}{h}.$$

Thus

$$h^\beta = \frac{d\delta}{\kappa_\beta L_\beta}, \qquad h^*(\delta) = \left(\frac{d\delta}{\kappa_\beta L_\beta}\right)^{1/\beta}.$$

With this choice,

$$\kappa_\beta L_\beta (h^*)^{\beta-1} \asymp \frac{d\delta}{h^*} \asymp a_1 \delta^p, \quad a_1 := (\kappa_\beta L_\beta)^{1/\beta} d^{(\beta-1)/\beta}.$$

Similarly,

$$\frac{\kappa_\beta^2 L_\beta^2}{L}(h^*)^{2(\beta-1)} \asymp \frac{d^2\delta^2}{L(h^*)^2} \asymp a_2 \delta^{2p}, \quad a_2 := \kappa_\beta^{2/\beta} L_\beta^{\frac{2}{\beta}} L^{-1} d^{\frac{2(\beta-1)}{\beta}}.$$

Therefore, after substituting $h = h^*(\delta)$, the error bound becomes

$$\mathcal{E}(N, T, \delta) \lesssim \mathcal{E}_0(N, T) + \frac{R}{\sqrt{T}}\left(c_1 \delta^{1/(2\beta)} + c_2 \delta^{p/2}\right) + 2Ra_1\delta^p + 2Na_2\delta^{2p},$$

where

$$\mathcal{E}_0(N, T) := \frac{LR^2}{N^2} + \frac{LR^2}{T} + \frac{Rc_0}{\sqrt{T}}, \qquad c_0 := \sqrt{\kappa d}\,\sigma_*, \qquad c_0^2 = \kappa d \sigma_*^2.$$

Step 2: the δ-independent part.

The δ-independent part contains two horizons, N and $T = BN$. With respect to N,

$$\mathcal{E}_{0,N}(N) = \frac{LR^2}{N^2}, \qquad \mathcal{E}_{0,N}(N) \asymp N^{-2}.$$

Thus the N-exponent is

$$\boxed{\beta_0^{(N)} = 2.}$$

With respect to T, the large-T part is

$$\frac{LR^2}{T} + \frac{Rc_0}{\sqrt{T}}.$$

In the high-precision regime and for $c_0 > 0$, the term $Rc_0/\sqrt{T}$ dominates LR^2/T. Hence

$$\mathcal{E}_{0,T}(T) \asymp T^{-1/2}, \qquad \boxed{\beta_0^{(T)} = \frac{1}{2}.}$$

The constraint $\frac{Rc_0}{\sqrt{T}} \lesssim \varepsilon$ gives

$$T \gtrsim T_{\min} := \frac{R^2 c_0^2}{\varepsilon^2} = \frac{R^2 \kappa d \sigma_*^2}{\varepsilon^2}.$$

Step 3: fidelity channels.

After balancing h, the active fidelity channels are $2Ra_1\delta^p$ and $2Na_2\delta^{2p}$. For fixed N, the accuracy constraint imposes

$$\delta \lesssim \delta_D := \left(\frac{\varepsilon}{2Ra_1}\right)^{1/p}, \quad \delta \lesssim \delta_E(N) := \left(\frac{\varepsilon}{2Na_2}\right)^{1/(2p)}.$$

Since $T_{\text{total}} = T\delta^{-\gamma}$, is decreasing in δ for every $\gamma > 0$, the optimal fidelity is the largest admissible one:

$$\delta^*(N) = \min\{\delta_D, \delta_E(N)\}.$$

The transition between the two fidelity channels occurs when $\delta_D = \delta_E(N)$. Equivalently,

$$\left(\frac{\varepsilon}{2Ra_1}\right)^{1/p} = \left(\frac{\varepsilon}{2Na_2}\right)^{1/(2p)}.$$

Solving for N gives

$$N_{\text{cr}} = \frac{2R^2 a_1^2}{a_2 \varepsilon}.$$

Using $\frac{a_1^2}{a_2} = L$, we obtain

$$\boxed{N_{\text{cr}} \asymp \frac{LR^2}{\varepsilon}.}$$

Thus, in the large-N regime $N \gtrsim N_{\text{cr}}$, the accumulated quadratic channel is active:

$$\delta^*(N) = \left(\frac{\varepsilon}{2Na_2}\right)^{1/(2p)}.$$

Step 4: optimizing N.

In the large-N regime,

$$T_{\text{total}}(N, T) = T\left(\frac{2Na_2}{\varepsilon}\right)^{\gamma/(2p)}.$$

For fixed T, this is increasing in N. Therefore the optimal choice is the smallest admissible N,

$$\boxed{N^* \asymp N_{\text{cr}} \asymp \frac{LR^2}{\varepsilon}.}$$

Substituting $N^* = N_{\text{cr}}$ into the active fidelity constraint gives

$$\delta^* = \left(\frac{\varepsilon}{2a_2 N_{\text{cr}}}\right)^{1/(2p)} \asymp \left(\frac{\varepsilon^2}{a_2 LR^2}\right)^{1/(2p)}.$$

Expanding a_2, we obtain

$$\boxed{\delta^* \asymp d^{-1}\kappa_\beta^{-1/(\beta-1)} L_\beta^{-1/(\beta-1)} R^{-\beta/(\beta-1)} \varepsilon^{\beta/(\beta-1)}.}$$

The corresponding smoothing radius is

$$\boxed{h^* = h^*(\delta^*) = \left(\frac{d\delta^*}{\kappa_\beta L_\beta}\right)^{1/\beta} \asymp \kappa_\beta^{-1/(\beta-1)} L_\beta^{-1/(\beta-1)} R^{-1/(\beta-1)} \varepsilon^{1/(\beta-1)}.}$$

Step 5: optimizing $T = BN$.

The δ-independent term $Rc_0/\sqrt{T}$ gives

$$T \gtrsim T_{\min} = \frac{R^2 c_0^2}{\varepsilon^2}.$$

Since $T_{\text{total}} = T\delta^{-\gamma}$ is increasing in T, the optimal choice is

$$\boxed{T^* \asymp T_{\min} = \frac{R^2 \kappa d \sigma_*^2}{\varepsilon^2}.}$$

Therefore

$$\boxed{B^* = \frac{T^*}{N^*} \asymp \frac{\kappa d \sigma_*^2}{L\varepsilon}.}$$

Step 6: optimized wall-clock cost.

The optimized wall-clock cost is $T^*_{\text{total}} = T^*(\delta^*)^{-\gamma}$. Using the expressions above,

$$T^* \asymp \frac{R^2 \kappa d \sigma_*^2}{\varepsilon^2},$$

and

$$(\delta^*)^{-\gamma} \asymp d^\gamma \kappa_\beta^{\gamma/(\beta-1)} L_\beta^{\gamma/(\beta-1)} R^{\gamma\beta/(\beta-1)} \varepsilon^{-\gamma\beta/(\beta-1)}.$$

Hence

$$\boxed{T^*_{\text{total}} \asymp \kappa \sigma_*^2 d^{1+\gamma} \kappa_\beta^{\gamma/(\beta-1)} L_\beta^{\gamma/(\beta-1)} R^{2+\gamma\beta/(\beta-1)} \varepsilon^{-(2+\gamma\beta/(\beta-1))}.}$$

Equivalently, hiding problem-dependent constants,

$$T^*_{\text{total}} = O\left(d^{1+\gamma} \varepsilon^{-\left(2+\frac{\gamma\beta}{\beta-1}\right)}\right).$$

Step 7: batching regimes.

Since

$$B^* = \frac{T^*}{N^*},$$

the batching regime is determined by comparing B^* and N^*, or equivalently comparing T^* with $(N^*)^2$. We have

$$T^* \asymp \frac{R^2 \kappa d \sigma_*^2}{\varepsilon^2}, \quad (N^*)^2 \asymp \frac{L^2 R^4}{\varepsilon^2}.$$

Therefore

$$B^* \le N^* \iff T^* \le (N^*)^2 \iff \kappa d \sigma_*^2 \lesssim L^2 R^2.$$

Similarly,

$$B^* > N^* \iff \kappa d \sigma_*^2 \gtrsim L^2 R^2.$$

Thus the batching transition is

$$\boxed{\kappa d \sigma_*^2 \asymp L^2 R^2.}$$

I PROOF OF PROPOSITION 8

Throughout this proof we use the bounded-trajectory assumption

$$\widetilde{R} = \max_t \{\|z_t - x^*\|, \|z_t - y_t\|\} = O(R).$$

Thus the bias term $\widetilde{R}\kappa_\beta L_\beta h^{\beta-1}$ is of order $R\kappa_\beta L_\beta h^{\beta-1}$.

Small-batch regime: $B \le 4\kappa d$**.**

Set $Q := BN$. Then we get

$$E(Q, B, \delta, h) \lesssim \underbrace{\frac{\kappa^2 d^2 L R^2}{Q^2}}_{\mathcal{E}_0(Q)} + \underbrace{\frac{Q}{\kappa L h^2}\delta^2}_{\text{Tsybakov channel}} + S_{\text{sm}}(Q, B, h),$$

where the smoothing remainder is

$$S_{\text{sm}}(Q, B, h) = \frac{QLh^2}{\kappa d} + R\kappa_\beta L_\beta h^{\beta-1} + \frac{Q}{B}\frac{(\kappa_\beta L_\beta h^{\beta-1})^2}{L}.$$

Hence $\mathcal{E}_0(Q) = \kappa^2 d^2 L R^2\, Q^{-2}, \qquad \mathcal{E}_0(Q) \sim Q^{-2}$. For a fixed admissible h, the active wall-clock subproblem is

$$\min_{Q,\delta} Q\delta^{-\gamma} \qquad \text{s.t.} \qquad \frac{A}{Q^2} + CQ\delta^2 \lesssim \varepsilon,$$

with $A := \kappa^2 d^2 L R^2, C := \frac{1}{\kappa L h^2}$. For fixed Q, the largest feasible δ is optimal:

$$\delta^2(Q) = \frac{\varepsilon - A/Q^2}{CQ}.$$

Substituting this into $T = Q\delta^{-\gamma}$ gives

$$T(Q) = C^{\gamma/2} Q^{1+\gamma/2} \left(\varepsilon - \frac{A}{Q^2}\right)^{-\gamma/2}.$$

Let $s := \frac{A/Q^2}{\varepsilon} \in (0, 1)$. Then $Q = \left(\frac{A}{\varepsilon s}\right)^{1/2}$, and, up to s-independent factors,

$$T(s) \propto s^{-(2+\gamma)/4}(1-s)^{-\gamma/2}.$$

The first-order condition yields

$$s^* = \frac{2+\gamma}{2+3\gamma}.$$

Therefore

$$\frac{A}{(Q^*)^2} = \frac{2+\gamma}{2+3\gamma}\varepsilon, \qquad CQ^*(\delta^*)^2 = \frac{2\gamma}{2+3\gamma}\varepsilon.$$

Thus

$$Q^* = \kappa d R\sqrt{L}\,\varepsilon^{-1/2}\left(\frac{2+3\gamma}{2+\gamma}\right)^{1/2},$$

and

$$\delta^* = hL^{1/4}(dR)^{-1/2}\varepsilon^{3/4}G(\gamma),$$

where

$$G(\gamma) := \frac{(2\gamma)^{1/2}(2+\gamma)^{1/4}}{(2+3\gamma)^{3/4}}.$$

It remains to choose h. Since

$$T^* = Q^*(\delta^*)^{-\gamma} \propto h^{-\gamma},$$

the optimal smoothing radius is the largest one allowed by $S_{\text{sm}}(Q^*, B, h) \lesssim \varepsilon$. Therefore

$$h^*_{\text{sm}}(B) = \min\{H_1, H_2, H_3(B)\},$$

where

$$H_1 \asymp \left(\frac{\varepsilon\kappa d}{Q^* L}\right)^{1/2} \asymp \varepsilon^{3/4} R^{-1/2} L^{-3/4}, \quad H_2 \asymp \left(\frac{\varepsilon}{R\kappa_\beta L_\beta}\right)^{1/(\beta-1)},$$

and

$$H_3(B) \asymp \left(\frac{\varepsilon LB}{Q^* \kappa_\beta^2 L_\beta^2}\right)^{1/(2\beta-2)} \asymp \left(\frac{B\varepsilon^{3/2}\sqrt{L}}{\kappa d R \kappa_\beta^2 L_\beta^2}\right)^{1/(2\beta-2)}.$$

Consequently,

$$T^*_{\mathrm{sm}}(B) \asymp \kappa d^{1+\gamma/2} R^{1+\gamma/2} L^{1/2-\gamma/4} \varepsilon^{-1/2-3\gamma/4} \big(h^*_{\mathrm{sm}}(B)\big)^{-\gamma} C_\gamma,$$

where C_γ depends only on γ.

The active terms depend on N and B only through $Q = BN$. Thus batching does not improve the leading active trade-off itself. It can only affect the smoothing cap through $H_3(B)$. In the regime where H_1 or H_2 is active, one may take $B^* = 1$ and $N^* = Q^*$. If $H_3(B)$ is active, increasing B can enlarge the admissible smoothing radius, but only up to the boundary $B \le 4\kappa d$.

Large-batch regime: $B > 4\kappa d$**.**

In this case, we get

$$E(N, B, \delta, h) \lesssim \underbrace{\frac{LR^2}{N^2}}_{\mathcal{E}_0(N)} + \underbrace{\frac{N\kappa d^2}{BLh^2}\delta^2}_{\text{Tsybakov channel}} + S_{\mathrm{lg}}(N, B, h),$$

where

$$S_{\mathrm{lg}}(N, B, h) = \frac{N\kappa d L h^2}{B} + R\kappa_\beta L_\beta h^{\beta-1} + \frac{N(\kappa_\beta L_\beta h^{\beta-1})^2}{L}.$$

Hence

$$\mathcal{E}_0(N) = LR^2\, N^{-2}, \qquad \mathcal{E}_0(N) \sim N^{-2}$$

For fixed B and admissible h, the active subproblem is

$$\min_{N,\delta} NB\delta^{-\gamma} \qquad \text{s.t.} \qquad \frac{A}{N^2} + C_B N\delta^2 \le \varepsilon,$$

where $A := LR^2$, $C_B := \frac{\kappa d^2}{BLh^2}$. Repeating the one-dimensional optimization above gives

$$N^* = R\sqrt{L}\,\varepsilon^{-1/2}\left(\frac{2+3\gamma}{2+\gamma}\right)^{1/2},$$

and

$$\delta^*(B, h) = B^{1/2} h L^{1/4} (\kappa d^2 R)^{-1/2} \varepsilon^{3/4} G(\gamma).$$

Substituting this into $T = NB\delta^{-\gamma}$ yields

$$T^*_{\mathrm{lg}}(B, h) \asymp B^{1-\gamma/2} \kappa^{\gamma/2} d^\gamma R^{1+\gamma/2} L^{1/2-\gamma/4} \varepsilon^{-1/2-3\gamma/4} h^{-\gamma} C_\gamma.$$

Thus the dependence on B is

$$T^*_{\mathrm{lg}}(B, h) \propto B^{1-\gamma/2}.$$

If $0 < \gamma < 2$, increasing B is not beneficial at the active trade-off level, so the optimum is attained at the smallest admissible large batch and is covered, up to constants, by the small-batch regime. If $\gamma = 2$, the leading dependence on B is flat. If $\gamma > 2$, then $T^*_{\mathrm{lg}}(B, h)$ decreases with B as long as the interior solution has $\delta^*(B, h) < 1$. Therefore the optimal large-batch strategy increases B until the boundary $\delta^* = 1$ is reached.

Boundary solution for $\gamma > 2$**.**

At the boundary $c(\delta^*) = 1$, so the active problem becomes

$$\min_{N,B} NB \qquad \text{s.t.} \qquad \frac{LR^2}{N^2} + \frac{N\kappa d^2}{BLh^2} \le \varepsilon.$$

For fixed N, the smallest feasible B is

$$B(N) = \frac{N\kappa d^2}{Lh^2(\varepsilon - LR^2/N^2)}.$$

Therefore

$$T(N) = NB(N) = \frac{\kappa d^2 N^2}{Lh^2(\varepsilon - LR^2/N^2)}.$$

Minimizing in N gives $\frac{LR^2}{(N^*)^2} \asymp \varepsilon$, and hence

$$N^* \asymp R\sqrt{L}\,\varepsilon^{-1/2}.$$

The corresponding batch size is

$$B^* \asymp \frac{\kappa d^2 N^*}{\varepsilon L h^2} \asymp \kappa d^2 R L^{-1/2} \varepsilon^{-3/2} h^{-2}.$$

Thus

$$T^* = N^* B^* \asymp \kappa d^2 R^2 \varepsilon^{-2} h^{-2}.$$

It remains to choose h. Substituting N^* and $B^*(h)$ into $S_{\mathrm{lg}}(N, B, h) \lesssim \varepsilon$ gives

$$h^*_{\mathrm{lg}} = \min\{K_1, K_2, K_3\},$$

where

$$K_1 \asymp d^{1/4} L^{-1/2}, \quad K_2 \asymp \left(\frac{\varepsilon}{R\kappa_\beta L_\beta}\right)^{1/(\beta-1)}, \quad K_3 \asymp \left(\frac{\varepsilon^{3/2}\sqrt{L}}{R\kappa_\beta^2 L_\beta^2}\right)^{1/(2\beta-2)}.$$

Since $T^* \propto h^{-2}$, the optimal choice is $h = h^*_{\mathrm{lg}}$, yielding

$$T^*_{\mathrm{overbatch}} \asymp \kappa d^2 R^2 \varepsilon^{-2} \big(h^*_{\mathrm{lg}}\big)^{-2}.$$

This proves the large-batch overbatching regime for $\gamma > 2$.

J Proof of Proposition 7

We first justify the batched extension used in equation 13. Let

$$\bar{g}_t = \frac{1}{B}\sum_{i=1}^{B} g_t^{(i)}$$

be the average of B conditionally independent copies of the ℓ_2-randomized estimator. Starting from the proof of Corollary 19 in Akhavan et al. (2024), batching only changes the oracle-noise contributions: the deterministic and smoothing-bias terms remain unchanged, while the Tsybakov variance level is divided by B. Indeed, writing

$$g_t = s_t + n_t, \qquad n_t = \frac{d}{2h_t}(\xi_t - \xi_t')\zeta_t K(r_t),$$

the independent averaging gives

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} n_t^{(i)}\right\|^2 \,\middle|\, x_t\right] = \frac{1}{B}\mathbb{E}[\|n_t\|^2 \mid x_t].$$

Thus every occurrence of the Tsybakov variance level δ^2 in the noise-dependent terms is replaced by δ^2/B. Applying the same strongly convex argument as in Akhavan et al. (2024) with this replacement gives equation 13.

We now solve the leading-order wall-clock problem. Put

$$a := \frac{\beta - 1}{\beta}, \qquad C_0 := \frac{dL^2R^2}{\mu}, \qquad C_A := \frac{L_\beta^{2/\beta}}{\mu} d^{2a}.$$

Ignoring the lower-order residual channel, the active constraint is

$$\frac{C_0}{N} + C_A \delta^{2a} (BN)^{-a} \le \varepsilon.$$

Let

$$Q := BN, \qquad \eta_N := \varepsilon - \frac{C_0}{N}, \qquad S_N := \left(\frac{C_A}{\eta_N}\right)^{1/a}.$$

For fixed N and δ, feasibility requires $\eta_N > 0$, and the smallest feasible total number of oracle calls is

$$Q^*(N, \delta) = \max\{N, S_N \delta^2\}.$$

Thus, after optimizing over Q, the fixed-N objective is

$$T_N(\delta) = \max\{N, S_N \delta^2\} \delta^{-\gamma}, \qquad 0 < \delta \le 1.$$

If $0 < \gamma < 2$, then $T_N(\delta)$ is minimized at the balancing point

$$S_N (\delta^*)^2 = N.$$

Hence

$$(\delta^*)^2 = N \left(\frac{\eta_N}{C_A}\right)^{1/a}, \qquad Q^*(N, \delta^*) = N, \qquad B^* = \frac{Q^*}{N} = 1.$$

Substituting this δ^* into the cost gives

$$T(N) = N^{1-\gamma/2} C_A^{\gamma/(2a)} \left(\varepsilon - \frac{C_0}{N}\right)^{-\gamma/(2a)}.$$

Minimizing the last display over $N > C_0/\varepsilon$ yields

$$N^* = \frac{C_0}{\varepsilon} \frac{2\beta - 2 + \gamma}{(\beta - 1)(2 - \gamma)}.$$

Consequently,

$$\delta^* \asymp d^{-1/2} L R \, \mu^{\frac{1}{2(\beta-1)}} L_\beta^{-\frac{1}{\beta-1}} \varepsilon^{\frac{1}{2(\beta-1)}} C_\beta(\gamma)^{-\gamma},$$

where

$$C_\beta(\gamma) = \left(\frac{\gamma\beta}{2\beta - 2 + \gamma}\right)^{\frac{\beta}{2(\beta-1)}} \left(\frac{2\beta - 2 + \gamma}{(\beta - 1)(2 - \gamma)}\right)^{1/2}.$$

Therefore

$$T^*_{\text{total}} \asymp d^{1+\gamma/2} L^{2-\gamma} R^{2-\gamma} L_\beta^{\frac{\gamma}{\beta-1}} \mu^{-1-\frac{\gamma}{2(\beta-1)}} \varepsilon^{-1-\frac{\gamma}{2(\beta-1)}} \frac{2\beta - 2 + \gamma}{(\beta - 1)(2 - \gamma)} C_\beta(\gamma)^{-\gamma}.$$

If $\gamma \ge 2$, then the fixed-N objective is minimized, up to leading order, at the largest admissible fidelity level $\delta^* = 1$. Indeed, on the region $S_N \delta^2 \le N$, one has

$$T_N(\delta) = N \delta^{-\gamma},$$

which decreases with δ. On the region $S_N \delta^2 \ge N$, one has

$$T_N(\delta) = S_N \delta^{2-\gamma},$$

which is decreasing for $\gamma > 2$ and constant for $\gamma = 2$. Hence $\delta^* = 1$ is leading-order optimal for all $\gamma \ge 2$.

Substituting $\delta = 1$ into $Q^*(N, \delta)$, we obtain

$$Q^*(N, 1) = \max\left\{N, \left(\frac{C_A}{\varepsilon - C_0/N}\right)^{1/a}\right\}.$$

Thus the remaining one-dimensional problem is

$$\min_{N > C_0/\varepsilon} \max\left\{N, \left(\frac{C_A}{\varepsilon - C_0/N}\right)^{1/a}\right\}.$$

The first term is increasing in N, while the second term is decreasing in N. Therefore the continuous leading-order optimum is attained at the balancing point

$$N = \left(\frac{C_A}{\varepsilon - C_0/N}\right)^{1/a}.$$

Equivalently,

$$\varepsilon N - C_0 = C_A N^{1-a}.$$

Since $1 - a = 1/\beta$, the high-precision solution satisfies

$$N^* \asymp C_A^{1/a} \varepsilon^{-1/a}.$$

At the balancing point, $Q^* = N^*$, and hence the sequential leading-order optimizer can be chosen with

$$B^* = \frac{Q^*}{N^*} = 1.$$

Therefore

$$T^*_{\text{total}} = Q^* \asymp C_A^{1/a} \varepsilon^{-1/a}.$$

Substituting

$$C_A = \frac{L_\beta^{2/\beta}}{\mu} d^{2a}, \qquad \frac{1}{a} = \frac{\beta}{\beta - 1},$$

gives

$$N^* \asymp T^*_{\text{total}} \asymp d^2 L_\beta^{2/(\beta-1)} \mu^{-\beta/(\beta-1)} \varepsilon^{-\beta/(\beta-1)}, \qquad B^* = 1.$$

It remains to verify that the residual channel is lower-order. Let A_{main} and A_{res} denote the main noise channel and the residual channel in equation 13. Then

$$\frac{A_{\text{res}}}{A_{\text{main}}} = L^2 L_\beta^{-4/\beta} d^{4/\beta - 1} \delta^{-2(\beta-2)/\beta} B^{(\beta-2)/\beta} N^{-2/\beta}.$$

For $0 < \gamma < 2$, we have

$$B^* = 1, \qquad N^* \asymp \varepsilon^{-1}, \qquad \delta^* \asymp \varepsilon^{1/(2(\beta-1))},$$

and hence

$$\frac{A_{\text{res}}}{A_{\text{main}}} \asymp \varepsilon^{1/(\beta-1)} \to 0.$$

For $\gamma \geq 2$, the sequential leading-order optimizer has

$$\delta^* = 1, \qquad B^* = 1, \qquad N^* \asymp \varepsilon^{-\beta/(\beta-1)}.$$

Therefore

$$\frac{A_{\text{res}}}{A_{\text{main}}} \asymp \varepsilon^{2/(\beta-1)} \to 0.$$

Thus the residual channel is lower-order in the high-precision regime.

Finally, although the sequential leading-order optimizer can be chosen with $B^* = 1$, for $\gamma \geq 2$ the same total-work order can be redistributed across batches. Taking $\delta = 1$, any choice satisfying

$$N \gtrsim \frac{C_0}{\varepsilon}, \qquad Q = BN \gtrsim C_A^{1/a} \varepsilon^{-1/a}$$

achieves the same leading accuracy. Hence one may reduce the sequential depth to $N \asymp C_0/\varepsilon$ by increasing

$$B \asymp \frac{C_A^{1/a} \varepsilon^{-1/a}}{C_0/\varepsilon},$$

without changing the leading total oracle work.

J.1 Details for regularization and comparison

Throughout this section, $A \asymp_{d,\varepsilon} B$ means that A/B is bounded above and below by factors independent of d and ε.

Let f be convex and assume that an optimal solution x^* exists with

$$\|x_0 - x^*\| \le R.$$

For $\mu > 0$, define

$$f_\mu(x) := f(x) + \frac{\mu}{2}\|x - x_0\|^2.$$

Then f_μ is μ-strongly convex. Let x_μ^* be its minimizer. Since

$$f_\mu(x_\mu^*) \le f_\mu(x^*),$$

we have

$$f_\mu(x_\mu^*) - f^* \le \frac{\mu}{2}\|x^* - x_0\|^2 \le \frac{\mu R^2}{2}.$$

Therefore, if

$$f_\mu(\hat{x}) - f_\mu(x_\mu^*) \le \frac{\varepsilon}{2},$$

then

$$f(\hat{x}) - f^* \le f_\mu(\hat{x}) - f^* \le \frac{\varepsilon}{2} + \frac{\mu R^2}{2}.$$

Choosing

$$\mu = \frac{\varepsilon}{R^2}$$

gives $f(\hat{x}) - f^* \le \varepsilon$. Thus it is enough to solve the regularized problem to accuracy $O(\varepsilon)$ with

$$\mu \asymp \frac{\varepsilon}{R^2}.$$

We now substitute this value of μ into the strongly convex Tsybakov baseline used by Regularized ZO-SGD. For $0 < \gamma < 2$,

$$T_{\mathrm{sc}} \asymp d^{1+\gamma/2} L^{2-\gamma} R^{2-\gamma} L_\beta^{\gamma/(\beta-1)} \mu^{-1-\frac{\gamma}{2(\beta-1)}} \varepsilon^{-1-\frac{\gamma}{2(\beta-1)}}.$$

Since

$$\mu^{-1-\frac{\gamma}{2(\beta-1)}} \asymp R^{2+\gamma/(\beta-1)} \varepsilon^{-1-\frac{\gamma}{2(\beta-1)}},$$

we get

$$T_{\mathrm{reg}} \asymp d^{1+\gamma/2} L^{2-\gamma} L_\beta^{\gamma/(\beta-1)} R^{4-\gamma+\gamma/(\beta-1)} \varepsilon^{-2-\gamma/(\beta-1)}.$$

Hence the ε-exponent is

$$p_{\mathrm{reg}} = 2 + \frac{\gamma}{\beta - 1}.$$

For $\gamma \ge 2$,

$$T_{\mathrm{sc}} \asymp d^2 L_\beta^{2/(\beta-1)} \mu^{-\beta/(\beta-1)} \varepsilon^{-\beta/(\beta-1)}.$$

Substituting $\mu \asymp \varepsilon/R^2$ yields

$$T_{\mathrm{reg}} \asymp d^2 L_\beta^{2/(\beta-1)} R^{2\beta/(\beta-1)} \varepsilon^{-2\beta/(\beta-1)} = d^2 L_\beta^{2/(\beta-1)} R^{2\beta/(\beta-1)} \varepsilon^{-2-\frac{2}{\beta-1}}.$$

We now compare this with Accelerated ZO-SGD. In the small-batch regime,

$$T_{\mathrm{acc}} \asymp \kappa d^{1+\gamma/2} R^{1+\gamma/2} L^{1/2-\gamma/4} \varepsilon^{-1/2-3\gamma/4} \left(h_{\mathrm{sm}}^*(B)\right)^{-\gamma}.$$

Here

$$h_{\mathrm{sm}}^*(B) = \min\{H_1, H_2, H_3(B)\},$$

where

$$H_1 \asymp \varepsilon^{3/4}, \qquad H_2 \asymp \varepsilon^{1/(\beta-1)},$$

and

$$H_3(B) \asymp \left(\frac{B\varepsilon^{3/2}}{d}\right)^{1/(2\beta-2)}.$$

Since the bound is decreasing in $h^*_{\mathrm{sm}}(B)$, we take the largest admissible batch size in the small-batch regime:

$$B \asymp 4\kappa d.$$

Then

$$H_3(B) \asymp \varepsilon^{3/(4(\beta-1))}$$

up to κ-dependent constants. Since

$$\frac{3}{4(\beta-1)} < \frac{1}{\beta-1},$$

we have $H_2 \lesssim H_3(B)$ for sufficiently small ε. Therefore $H_3(B)$ is not active for the leading ε-exponent, and

$$h^*_{\mathrm{sm}}(B) \asymp_{d,\varepsilon} \varepsilon^s, \qquad s := \max\left\{\frac{3}{4}, \frac{1}{\beta-1}\right\}.$$

Consequently,

$$T_{\mathrm{acc}} \asymp_{d,\varepsilon} d^{1+\gamma/2}\varepsilon^{-\left(\frac{1}{2}+\frac{3\gamma}{4}+\gamma s\right)}.$$

Thus

$$p_{\mathrm{acc}} = \frac{1}{2} + \frac{3\gamma}{4} + \gamma s.$$

If $\beta \le 7/3$, then $s = 1/(\beta-1)$. Hence

$$p_{\mathrm{acc}} = \frac{1}{2} + \frac{3\gamma}{4} + \frac{\gamma}{\beta-1},$$

and

$$p_{\mathrm{reg}} - p_{\mathrm{acc}} = \frac{3}{4}(2-\gamma) > 0.$$

Thus $p_{\mathrm{acc}} < p_{\mathrm{reg}}$ for all $0 < \gamma < 2$.

If $\beta > 7/3$, then $s = 3/4$, so

$$p_{\mathrm{acc}} = \frac{1}{2} + \frac{3\gamma}{2}.$$

The accelerated method is better when

$$\frac{1}{2} + \frac{3\gamma}{2} < 2 + \frac{\gamma}{\beta-1}.$$

Solving this inequality gives

$$\gamma < \gamma_{\mathrm{crit}}(\beta) := \frac{3(\beta-1)}{3\beta-5}.$$

For $\beta > 7/3$, one has

$$1 < \gamma_{\mathrm{crit}}(\beta) < 2.$$

Therefore, for $\beta > 7/3$, ACCELERATED ZO-SGD has the smaller ε-exponent when $\gamma < \gamma_{\mathrm{crit}}(\beta)$, while REGULARIZED ZO-SGD has the smaller ε-exponent when $\gamma_{\mathrm{crit}}(\beta) < \gamma < 2$.

Finally, in the regime $\gamma \ge 2$, ACCELERATED ZO-SGD gives

$$T_{\mathrm{acc}} \asymp_{d,\varepsilon} d^2\varepsilon^{-2}\left(h^*_{\mathrm{lg}}\right)^{-2}.$$

Using

$$h^*_{\mathrm{lg}} \asymp_{d,\varepsilon} \varepsilon^{1/(\beta-1)}$$

for the leading (d,ε)-scaling, we get

$$T_{\mathrm{acc}} \asymp_{d,\varepsilon} d^2\varepsilon^{-2-\frac{2}{\beta-1}}.$$

This matches the leading (d,ε)-scaling of REGULARIZED ZO-SGD:

$$T_{\mathrm{reg}} \asymp_{d,\varepsilon} T_{\mathrm{acc}} \asymp_{d,\varepsilon} d^2\varepsilon^{-2-\frac{2}{\beta-1}}.$$

The difference is in sequential depth: ACCELERATED ZO-SGD uses

$$N_{\mathrm{acc}} \asymp R\sqrt{L}\,\varepsilon^{-1/2},$$

whereas REGULARIZED ZO-SGD has deterministic sequential scale

$$N_{\mathrm{reg}} \asymp dL^2R^4\varepsilon^{-2}.$$