OPTIMAL STOCHASTIC BILEVEL OPTIMIZATION WITH FIRST-ORDER ORACLES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study nonconvex–strongly-convex bilevel optimization under a stochastic first-order oracle. We introduce MRT-FD, a single-loop first-order method that simultaneously tracks the upper-level variable, the lower-level solution, and the auxiliary response arising from implicit differentiation of the hyperobjective. MRT-FD performs one update of each variable per iteration and approximates the second-order derivative actions using order-p finite differences. For any fixed finite smoothness order $p \geq 1$ in the lower-level variable, MRT-FD finds an ε-stationary point using $\mathcal{O}(\varepsilon^{-4-2/p})$ stochastic gradient queries. We also prove a matching $\Omega(\varepsilon^{-4-2/p})$ oracle lower bound. The lower-bound construction starts from a hard nonconvex minimization chain with a stronger stochastic oracle, and lifts it to a bilevel problem through a sinusoidal coupling with a scalar lower-level variable. Consequently, the dependence on ε is optimal for every fixed finite p, closing the upper–lower complexity gap in this stochastic first-order oracle setting.

1 INTRODUCTION

Bilevel optimization models problems in which the performance of one decision depends on the solution of another optimization problem. We study the stochastic nonconvex–strongly-convex (NC–SC) setting

$$\min_{x\in\mathbb{R}^{d_x}} \Phi(x) := f(x, y^\star(x)), \qquad y^\star(x) := \arg\min_{y\in\mathbb{R}^{d_y}} g(x, y), \tag{1}$$

where $g(x,\cdot)$ is strongly convex and the hyperobjective Φ may be nonconvex. This formulation captures applications including hyperparameter optimization (Franceschi et al., 2018), data reweighting (Pan et al., 2025), and meta-learning (Finn et al., 2017).

Under standard smoothness conditions, Φ is a smooth nonconvex function. Writing $H(x,y) := \nabla^2_{yy} g(x,y)$ and $J(x,y) := \nabla^2_{xy} g(x,y)$, implicit differentiation gives

$$\nabla\Phi(x) = \nabla_x f(x, y^\star(x)) - J(x, y^\star(x))H(x, y^\star(x))^{-1}\nabla_y f(x, y^\star(x)). \tag{2}$$

The hypergradient therefore involves second-order derivatives of g, motivating methods based on Hessian- and Jacobian-vector products (Ghadimi & Wang, 2018; Ji et al., 2021). When such second-order stochastic derivative information is available, an $\mathcal{O}(\epsilon^{-4})$ oracle complexity is achievable for finding an ϵ-stationary point, matching the ϵ dependence inherited from the lower bound for stochastic nonconvex optimization (Arjevani et al., 2023).

For large-scale problems, however, even Hessian-vector products may be expensive or unavailable, motivating methods that use only first-order queries. This restriction substantially changes the known complexity. Under standard smoothness assumptions, the best first-order methods achieve an $\mathcal{O}(\epsilon^{-6})$ stochastic oracle complexity (Chen et al., 2025; Gu et al., 2026b). Recently, Chen et al. (2026b) show that additional pth-order smoothness with respect to the lower variable y can improve the upper bound to $\widetilde{\mathcal{O}}(\epsilon^{-4-2/p})$. As p increases, this rate interpolates between the ϵ^{-6} complexity under base smoothness and the ϵ^{-4} benchmark available with second-order information.

This improvement raises a fundamental question. Existing lower bounds do not characterize how the complexity of a first-order stochastic bilevel oracle depends on the available smoothness order p. It remains unclear whether the exponent $4 + 2/p$ reflects an intrinsic information-theoretic barrier.

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.

What is the optimal complexity of stochastic NC–SC bilevel optimization under a first-order oracle, and how does higher-order smoothness in the lower variable change this complexity?

We answer the question above by sharpening both the upper and lower complexity bounds.

A single-loop first-order algorithm. We develop Moving-Target Tracking with Finite Differences (MRT-FD), a single-loop method that requires order-p smoothness only for g with respect to the lower variable y, for any $p \geq 1$. We introduce the auxiliary vector $z^\star(x)$ as the unique solution of

$$H(x, y^\star(x))z^\star(x) = \nabla_y f(x, y^\star(x)).$$

MRT-FD maintains y_k and z_k to track the moving targets $y^\star(x_k)$ and $z^\star(x_k)$ while simultaneously updating the outer variable x_k. It uses order-p finite differences to reconstruct the required second-order derivative actions from first-order queries, while projection keeps z_k bounded and controls the resulting finite-difference errors. Our analysis couples objective descent with the tracking errors $\|y_k - y^\star(x_k)\|^2$ and $\|z_k - z^\star(x_k)\|^2$ through a moving-target Lyapunov function that jointly controls target motion, finite-difference bias, and stochastic variance. For fixed p, it achieves $\mathcal{O}(\epsilon^{-4-2/p})$ complexity, matching the ϵ dependence of Chen et al. (2026b) while removing the logarithmic factor.

Lower bound for finite smoothness order p. We establish a matching lower bound under the same assumptions and oracle model. The construction embeds a hard nonconvex chain into a bilevel problem with a scalar lower-level variable through a sinusoidal coupling. Setting a coupling scale to $\delta = \epsilon^{1/p}$ preserves the required order-p smoothness while keeping the resulting hyperobjective close to the hard chain. We then construct a stochastic oracle in which informative observations occur only through a rare Bernoulli event with probability $\varpi = \Theta(\delta^2\epsilon^2)$. A random-rotation argument limits progress along the hidden chain for arbitrary adaptive randomized algorithms. Combining a chain of length $T = \Theta(\epsilon^{-2})$ with the reveal probability gives the lower bound of $\Omega(T/\varpi) = \Omega(\epsilon^{-4-2/p})$.

1.1 Related Work

Bilevel optimization with second-order information. Classical bilevel methods compute the hypergradient using second-order information. Iterative differentiation (ITD) differentiates through a finite lower-level trajectory (Domke, 2012; Shaban et al., 2019; Bolte et al., 2021), whereas approximate implicit differentiation (AID) approximates the inverse-Hessian–gradient product through linear-system or Neumann-type solvers (Pedregosa, 2016; Ghadimi & Wang, 2018; Grazzi et al., 2020; Ji et al., 2021).

In stochastic NC–SC bilevel optimization, methods such as stocBiO (Ji et al., 2021) and SUSTAIN (Khanduri et al., 2021) construct hypergradient estimators using Hessian- and Jacobian-vector products. SOBA/SABA (Dagréou et al., 2022) and MA-SOBA (Chen et al., 2024b) instead track the associated linear-system solution jointly with the lower and upper-level variables. These methods are structurally related to MRT-FD through linear-system tracking, but require second-order derivative information.

First-order bilevel optimization. Fully first-order methods often rely on value-function or penalty reformulations, including BOME (Liu et al., 2022), penalty-based bilevel gradient methods (Shen & Chen, 2023), and F^{2}SA (Kwon et al., 2023). Subsequent work improves the first-order oracle complexity through refined penalty and two-timescale schemes (Kwon et al., 2024; Chen et al., 2025). More recently, SGHA (Gu et al., 2026b) develops a single-loop method based on a constrained reformulation of lower-level stationarity and finite-difference approximations, achieving $\mathcal{O}(\epsilon^{-6})$ complexity, but requires additional assumptions in the boundedness.

FdeHBO (Yang et al., 2023) is the closest structural comparison to MRT-FD: it tracks the auxiliary linear-system solution using projection and finite-difference Hessian-vector approximations. Its guarantees, however, rely on stronger stochastic assumptions, including stochastic smoothness and unbiased stochastic Hessian estimates with bounded variance. FdeHBO uses variance reduction and reuses the same sample across multiple query points, whereas MRT-FD uses fresh independent samples and order-p finite differences. Their analyses track different auxiliary errors: FdeHBO uses the current lower-level linear-system solution, while we directly control $\|z_k - z^\star(x_k)\|^2$.

Chen et al. (2026b) show that pth-order smoothness in the lower variable can improve the complexity from $\widetilde{\mathcal{O}}(\epsilon^{-6})$ to $\widetilde{\mathcal{O}}(\epsilon^{-4-2/p})$. Their F^{2}SA-$p$ method uses high-order finite differences within a penalty formulation over perturbed lower-level problems, and assumes higher-order smoothness of both f and g in y. In contrast, our additional order-p smoothness condition is imposed only on g.

Lower bounds. Ji & Liang (2023) established lower complexity bounds for convex–strongly-convex and strongly-convex–strongly-convex bilevel problems, while Dagréou et al. (2024) derived matching lower bounds for finite-sum bilevel empirical-risk minimization. Beyond the strongly-convex lower-level setting, Chen et al. (2024a) showed hardness results for finding stationary points of the hyperobjective when the lower-level problem is only convex. For stochastic NC–SC bilevel optimization, Kwon et al. (2024) establish lower bounds under a $y^\star$-aware oracle, where unbiased bounded-variance stochastic gradients are guaranteed only locally around $y^\star(x)$, making their setting not directly comparable to ours. More recently, Chen et al. (2026a) prove an $\Omega(\kappa^{4.5}\epsilon^{-4})$ lower bound for NC–SC bilevel optimization, where κ denotes the condition number, while concurrent work by Gu et al. (2026a) establishes an $\Omega(\kappa^8\epsilon^{-6})$ lower bound. Our result complements these works by characterizing the effect of higher-order smoothness in the lower variable: for a fixed p, we prove a $\Omega(\epsilon^{-4-2/p})$ lower bound.

2 Problem Setting and Oracle Model

We consider the bilevel optimization problem

$$\min_{x\in\mathbb{R}^{d_x}} \Phi(x) := f(x, y^\star(x)), \qquad y^\star(x) := \arg\min_{y\in\mathbb{R}^{d_y}} g(x, y), \tag{3}$$

where f and g are the upper- and lower-level objectives, respectively. We use the notation

$$H(x,y) := \nabla^2_{yy} g(x,y), \qquad J(x,y) := \nabla^2_{xy} g(x,y), \qquad b(x,y) := \nabla_y f(x,y).$$

Thus $H(x,y) \in \mathbb{R}^{d_y\times d_y}$ and $J(x,y) \in \mathbb{R}^{d_x\times d_y}$, so both matrices act on vectors in $\mathbb{R}^{d_y}$. We define the auxiliary linear-system solutions at a generic point (x,y) and at the lower-level minimizer, respectively, as

$$z(x,y) := H(x,y)^{-1} b(x,y), \qquad z^\star(x) := z(x, y^\star(x)).$$

Under Assumption 2.1 below, $H(x,y) \succeq \mu I$, and implicit differentiation yields the hypergradient

$$\nabla\Phi(x) = \nabla_x f(x, y^\star(x)) - J(x, y^\star(x)) z^\star(x). \tag{4}$$

Our goal is to find an approximately stationary point of Φ. For a possibly random output x_R, we refer to $\mathbb{E}\,\|\nabla\Phi(x_R)\|^2 \le \epsilon^2$ as an ϵ-stationarity guarantee.

2.1 Assumptions

We study smooth nonconvex–strongly-convex (NC–SC) bilevel optimization under the following regularity assumptions.

Assumption 2.1 (Base NC–SC regularity)**.** The following conditions hold for constants $\ell_f, \ell_g, \rho_1, \mu > 0$:

(i) f is differentiable with a jointly ℓ_f-Lipschitz gradient, and g is twice differentiable with a jointly ℓ_g-Lipschitz gradient.

(ii) $g(x,\cdot)$ is μ-strongly convex for every x.

(iii) The Hessian block $\nabla^2_{yy} g$ and the mixed derivative $\nabla^2_{xy} g$ are each ρ_1-Lipschitz in the joint variable (x,y) with respect to the operator norm.

(iv) $\|\nabla_y f(x,y)\| \le B_f$ for all (x,y), and Φ is bounded below.

These assumptions are standard in the literature (Kwon et al., 2024; Chen et al., 2025; 2026a) and are weaker than those used in some prior work (Kwon et al., 2023). Assumption 2.1 implies that, for every (x,y),

$$\|z(x,y)\| \le B_f/\mu, \qquad \|z^\star(x)\| \le B_f/\mu.$$

We define the closed Euclidean ball centered at the origin with radius R_z by

$$Q := \{z \in \mathbb{R}^{d_y} : \|z\| \le R_z\}, \qquad R_z := 4B_f/\mu.$$

The algorithm will project its auxiliary updates onto Q. This choice of radius ensures that both the current-point linear-system solution $z(x, y)$ and the target $z^\star(x)$ lie in Q.

Assumption 2.2 (Lower-variable order-p smoothness)**.** Fix an integer $p \ge 1$ and a constant $L_p > 0$. For each $\ell \in \{x, y\}$, the map $\nabla_\ell g(x, \cdot)$ is p times differentiable for every x and satisfies

$$\left\|\nabla_y^p \nabla_\ell g(x, y) - \nabla_y^p \nabla_\ell g(x, y')\right\|_{\mathrm{op}} \le L_p \|y - y'\|, \text{ for all } x, y, y'. \tag{5}$$

This assumption imposes order-p smoothness on the gradients of the lower-level objective with respect to the lower variable. For $p = 1$, it follows from Assumption 2.1(iii) with $L_1 = \rho_1$, recovering the standard setting. Beyond the base regularity conditions, we impose an explicit Lipschitz condition only at order p, without separately imposing such conditions at each intermediate derivative order. The additional higher-order condition concerns only g; in contrast, prior work also imposes higher-order smoothness on f with respect to y (Chen et al., 2026b).

2.2 Stochastic gradient oracle

We now specify the stochastic gradient oracle used for both the upper and lower complexity bounds. At a query point (x, y), a single oracle call returns four gradient estimates:

$$\mathcal{O}(x, y) = \big(\hat{F}_x(x, y), \hat{F}_y(x, y), \hat{G}_x(x, y), \hat{G}_y(x, y)\big).$$

For $\ell \in \{x, y\}$, $\hat{F}_\ell(x, y)$ estimates $\nabla_\ell f(x, y)$, and $\hat{G}_\ell(x, y)$ estimates $\nabla_\ell g(x, y)$. These estimates are assumed to be unbiased and to have bounded variance, as specified below.

Assumption 2.3 (Unbiased oracle with bounded variance)**.** The query rules and oracle responses are measurable. Let $\mathcal{H}$ denote the history before an oracle call, and let (x, y) be an $\mathcal{H}$-measurable query point. For each $\ell \in \{x, y\}$, the returned estimates satisfy

$$\mathbb{E}[\hat{F}_\ell(x, y) \mid \mathcal{H}] = \nabla_\ell f(x, y), \qquad \mathbb{E}[\hat{G}_\ell(x, y) \mid \mathcal{H}] = \nabla_\ell g(x, y), \tag{6}$$

$$\sum_{\ell \in \{x,y\}} \mathbb{E}\Big[\|\hat{F}_\ell(x, y) - \nabla_\ell f(x, y)\|^2 + \|\hat{G}_\ell(x, y) - \nabla_\ell g(x, y)\|^2 \,\Big|\, \mathcal{H}\Big] \le \sigma^2, \tag{7}$$

almost surely, for a constant $\sigma \ge 0$. Each call uses fresh randomness. Calls at query points fixed by a common conditioning history are conditionally independent given that history. No independence is assumed among the four estimates $\hat{F}_x, \hat{F}_y, \hat{G}_x, \hat{G}_y$ returned by a single call.

For the upper-bound analysis, we use the following shorthand for minibatch estimates obtained by averaging m fresh oracle calls at a fixed query point (x, y):

$$\bar{F}_\ell(x, y) := \frac{1}{m}\sum_{i=1}^{m} \hat{F}_{\ell,i}(x, y), \qquad \bar{G}_\ell(x, y) := \frac{1}{m}\sum_{i=1}^{m} \hat{G}_{\ell,i}(x, y), \qquad \ell \in \{x, y\}.$$

3 A Single-Loop First-Order Algorithm

In this section, we introduce our single-loop first-order algorithm and establish its convergence guarantee. We begin with a finite-difference procedure for estimating the matrix–vector products appearing in (4), together with the bias and variance of the finite-difference approximation.

3.1 Estimating matrix–vector products from first-order queries

We approximate the indirect term in (4) by tracking $z^\star(x)$ and forming $J(x, y)z$. At an available pair (x, y), the tracking update for z is driven by the linear-system residual $H(x, y)z - b(x, y)$. Thus, both $H(x, y)z$ and $J(x, y)z$ must be estimated using only first-order information.

For $z \ne 0$, define $u := z/\|z\|$, and set $u := 0$ when $z = 0$. The two matrix–vector products admit the directional-derivative representations

$$H(x, y)z = \|z\| \lim_{t \to 0} \frac{\nabla_y g(x, y + tu) - \nabla_y g(x, y - tu)}{2t}, \tag{8}$$

$$J(x, y)z = \|z\| \lim_{t \to 0} \frac{\nabla_x g(x, y + tu) - \nabla_x g(x, y - tu)}{2t}. \tag{9}$$

Evaluating these quotients at a finite radius $\tau > 0$ yields a first-order finite-difference approximation. For example, $H(x,y)z \approx \frac{\|z\|}{2\tau}\big[\nabla_y g(x, y+\tau u) - \nabla_y g(x, y-\tau u)\big]$, and replacing $\nabla_y g$ by $\nabla_x g$ gives the corresponding approximation of $J(x,y)z$.

To exploit the higher-order regularity in Assumption 2.2, we use a higher-order finite-difference stencil that combines gradients evaluated at several points $y + j\tau u$ along the same line. Such interpolation-based finite-difference formulas are classical; see, e.g., (Atkinson & Han, 2001; Khan & Ohba, 1999; Fornberg, 1988). Let $\mathcal{I}_1 = \{-1, 1\}$, and, for $p \geq 2$, let $\mathcal{I}_p = \{-\lfloor p/2 \rfloor, \ldots, \lceil p/2 \rceil\}$. Choose the unique interpolation weights $\{a_j\}_{j\in\mathcal{I}_p}$ satisfying

$$\sum_{j\in\mathcal{I}_p} a_j = 0, \qquad \sum_{j\in\mathcal{I}_p} j a_j = 1, \qquad \sum_{j\in\mathcal{I}_p} j^r a_j = 0 \quad (2 \leq r \leq p). \tag{10}$$

These moment conditions preserve the first-order term in the Taylor expansion while canceling all polynomial terms of orders $0, 2, \ldots, p$.

For $\ell \in \{x, y\}$, define

$$\mathrm{FD}_\ell(x, y; z) := \frac{\|z\|}{\tau} \sum_{j\in\mathcal{I}_p} a_j \nabla_\ell g(x, y + j\tau u). \tag{11}$$

We suppress the dependence on p and τ in the notation. For $p = 1$ and $p = 2$, the zero coefficients can be omitted, and (11) reduces to the same two-point central difference. For every p, the stencil contains at most $p + 1$ query locations. A similar high-order finite-difference construction is used by Chen et al. (2026b), but for a different purpose. Their differences are applied to quantities arising from perturbed lower-level problems within a penalty-based construction, whereas our differences act directly on the original lower-level gradients along the direction z to estimate the actions Hz and Jz.

Stochastic estimates. At each node $j \in \mathcal{I}_p$, we average m fresh observations of $\hat{G}_\ell$ into $\bar{G}_{\ell,j}$, using separate batches across nodes. We then define the stochastic finite-difference estimator

$$\widehat{\mathrm{FD}}_\ell(x, y; z) := \frac{\|z\|}{\tau} \sum_{j\in\mathcal{I}_p} a_j \bar{G}_{\ell,j}(x, y + j\tau u). \tag{12}$$

Let $\mathcal{H}$ be a history that fixes (x, y, z) before these batches are drawn. Then

$$\mathbb{E}[\widehat{\mathrm{FD}}_\ell(x, y; z) \mid \mathcal{H}] = \mathrm{FD}_\ell(x, y; z).$$

Define $A_p := \sum_{j\in\mathcal{I}_p} a_j^2$. Under Assumptions 2.2 and 2.3, for each $\ell \in \{x, y\}$,

$$\|\mathrm{FD}_\ell(x, y; z) - \nabla^2_{\ell y} g(x, y) z\| \leq C_p L_p \tau^p \|z\|, \tag{13}$$

$$\mathbb{E}\Big[\|\widehat{\mathrm{FD}}_\ell(x, y; z) - \mathrm{FD}_\ell(x, y; z)\|^2 \,\Big|\, \mathcal{H}\Big] \leq \frac{A_p \sigma^2 \|z\|^2}{m\tau^2}, \tag{14}$$

where C_p depends only on the stencil nodes and weights. Explicit expressions for C_p and A_p are given in Lemma A.3. Related expressions follow from classical finite-difference formulas; see, e.g., Li (2005); Bai et al. (2009). For completeness, we provide a self-contained proof in the appendix.

Equations (13) and (14) reveal the bias–variance tradeoff induced by finite differencing. A smaller τ decreases the deterministic bias at rate τ^p, but amplifies the stochastic variance at rate τ^{-2}. Increasing the batch size m controls this variance amplification. The dependence of C_p and A_p on the stencil therefore enters the constants in the resulting complexity bounds.

3.2 Algorithm

We present our algorithm, *Moving-Target Tracking with Finite Differences* (MRT-FD), in Algorithm 1. The intuition behind the single-loop design is similar to that of SOBA (Dagréou et al., 2022): rather than solving the lower-level problem and the auxiliary linear system to high accuracy at every outer iteration, the algorithm maintains and updates the variables x, y, and z simultaneously. Compared with SOBA, MRT-FD requires only an unbiased, bounded-variance first-order oracle. Finite differences replace Hessian- and Jacobian-vector products without requiring samplewise smoothness or sample reuse across query points. Projection controls the resulting estimation errors.

Algorithm 1 MRT-FD: projected single-loop tracking with order-p differences

Require: Fixed x_0, y_0, initialize $z_0 = 0$. Stepsizes η_x, η_y, η_z. Difference radius τ, batch size m, weights $(\mathcal{I}_p, a)$. Projection radius R_z; $Q = \{z \in \mathbb{R}^{d_y} : \|z\| \le R_z\}$.

1: **for** $k = 0, \dots, K-1$ **do**

2: $\quad z_{k+1} \leftarrow \mathrm{Proj}_Q\Big[z_k - \eta_z\big(\widehat{\mathrm{FD}}_y(x_k, y_k; z_k) - \bar{F}_y(x_k, y_k)\big)\Big]$.

3: $\quad x_{k+1} \leftarrow x_k - \eta_x\big[\bar{F}_x(x_k, y_k) - \widehat{\mathrm{FD}}_x(x_k, y_k; z_{k+1})\big]$.

4: $\quad y_{k+1} \leftarrow y_k - \eta_y \bar{G}_y(x_k, y_k)$.

5: **end for**

6: Draw R uniformly from $\{0, \dots, K-1\}$ and return x_R.

The x update: estimate the outer gradient. The outer update uses a stochastic approximation of $\nabla\Phi(x_k)$. In (4), replacing the unavailable $y^\star(x_k)$ and $z^\star(x_k)$ by the current lower iterate y_k and the newly updated auxiliary iterate z_{k+1} gives

$$\nabla_x f(x_k, y_k) - J(x_k, y_k) z_{k+1}. \tag{15}$$

The two terms in (15) are estimated by $\bar{F}_{x,k} := \bar{F}_x(x_k, y_k)$ and $\widehat{\mathrm{FD}}_x(x_k, y_k; z_{k+1})$, respectively. Accordingly, define $\hat{h}_k := \bar{F}_{x,k} - \widehat{\mathrm{FD}}_x(x_k, y_k; z_{k+1})$ to approximate hypergradient, and then the outer update can be written as $x_{k+1} = x_k - \eta_x \hat{h}_k$. The discrepancy between $\hat{h}_k$ and $\nabla\Phi(x_k)$ comes from four sources: the tracking errors of y_k and z_{k+1}, the finite-difference bias, and the stochastic sampling noise. The y- and z-updates are designed to control the two tracking errors. Since the outer iterate x_k changes over time, the corresponding targets $y^\star(x_k)$ and $z^\star(x_k)$ also evolve; MRT-FD therefore treats them as moving targets.

The z update: track the inverse-Hessian–gradient product. At the available pair (x_k, y_k), the vector $z(x_k, y_k)$ minimizes the strongly convex quadratic

$$z \;\longmapsto\; \tfrac{1}{2} z^\top H(x_k, y_k) z - b(x_k, y_k)^\top z. \tag{16}$$

The gradient of this quadratic is the linear-system residual $H(x_k, y_k)z - b(x_k, y_k)$, whose unique zero is $z(x_k, y_k)$. Moreover, $z(x_k, y_k)$ approaches the target $z^\star(x_k)$ as y_k approaches $y^\star(x_k)$. With $\bar{F}_{y,k} := \bar{F}_y(x_k, y_k)$, Line 2 estimates the residual at z_k using $\widehat{\mathrm{FD}}_y(x_k, y_k; z_k) - \bar{F}_{y,k}$ and performs one projected stochastic gradient step. Since $\|z(x_k, y_k)\| \le B_f/\mu$, choosing $R_z \ge B_f/\mu$ ensures that the current linear-system solution lies in Q. The projection also guarantees $\|z_k\| \le R_z$ for every k, which is crucial for uniformly controlling the finite-difference bias and variance in (13)–(14).

The y update: track the lower solution. With $\bar{G}_{y,k} := \bar{G}_y(x_k, y_k)$, Line 4 performs one stochastic gradient step on $g(x_k, \cdot)$ at y_k, namely $y_{k+1} = y_k - \eta_y \bar{G}_{y,k}$, to track $y^\star(x_k)$. This tracking controls the evaluation-point error in both the outer direction and the coefficients $H(x_k, y_k)$ and $b(x_k, y_k)$ of the auxiliary linear system.

3.3 Convergence guarantee

Let $\Phi_{\inf} := \inf_x \Phi(x)$, and let Δ_0 denote an upper bound on the initial objective gap and tracking errors:

$$\Delta_0 \ge \Phi(x_0) - \Phi_{\inf} + \|y_0 - y^\star(x_0)\|^2 + (B_f/\mu)^2,$$

where the last term bounds the squared auxiliary tracking error at $z_0 = 0$. Let $\alpha > 0$ and $C_\star \ge 1$ be the constants specified in Theorem 3.1. We set the projection radius to $R_z = 4B_f/\mu$, choose the tracking stepsizes $\eta_y = \eta_z = 1/(2\ell_g)$, and set the outer stepsize $\eta_x = \alpha$. For a target accuracy $0 < \epsilon \le 1$, choose the finite-difference radius, minibatch size, and number of iterations as

$$\tau = \left(\frac{\epsilon^2}{3C_\star}\right)^{1/(2p)}, \qquad m = \left\lceil \frac{3C_\star}{\epsilon^2 \tau^2} \right\rceil, \qquad K = \left\lceil \frac{3C_\star(1 + \Delta_0)}{\epsilon^2} \right\rceil. \tag{17}$$

The role of the smoothness order p is apparent from these choices. Since the finite-difference bias scales as τ^{2p}, a larger p allows a larger difference radius τ while maintaining the same target bias. This reduces the variance amplification $1/\tau^2$ and, consequently, the required minibatch size m. The resulting convergence and oracle complexity are summarized below.

Theorem 3.1. *Suppose Assumptions 2.1–2.3 hold. There exist constants $\alpha > 0$ and $C_\star \geq 1$, depending only on p and $\ell_f, \ell_g, \rho_1, \mu, B_f, L_p, \sigma$, such that the following holds. Run Algorithm 1 from fixed x_0, y_0 and $z_0 = 0$, with $R_z = 4B_f/\mu$, $\eta_y = \eta_z = 1/(2\ell_g)$, and $\eta_x = \alpha$. For any $0 < \tau \leq 1$ and positive integers m, K, the output satisfies*

$$\mathbb{E}\|\nabla\Phi(x_R)\|^2 \leq C_\star \left(\frac{\Delta_0}{K} + \tau^{2p} + \frac{1}{m\tau^2} \right). \tag{18}$$

With the parameter choices in (17), we obtain $\mathbb{E}\|\nabla\Phi(x_R)\|^2 \leq \epsilon^2$ using at most

$$(2p+5)mK = \mathcal{O}\left(\epsilon^{-4-2/p}\right)$$

oracle calls, for any fixed finite p.

Appendix A provides the proof and explicit admissible choices of α and $C_\star$. For $p = 1$, Theorem 3.1 gives an $\mathcal{O}(\epsilon^{-6})$ sample complexity, matching existing upper bounds (Chen et al., 2025; Gu et al., 2026a). The single-loop method of Gu et al. (2026b) attains the same dependence under additional boundedness assumptions. For $p \geq 2$, our rate matches the $\epsilon^{-4-2/p}$ dependence of Chen et al. (2026b) while removing the logarithmic factor. Moreover, Chen et al. (2026b) assumes higher-order smoothness of both f and g in y, as their finite-difference estimator is applied to the gradient of a penalty function that combines the upper- and lower-level objectives.

4 Matching Lower Bound under First-Order Oracles

We first recall a hard instance construction for stochastic nonconvex minimization and establish its lower bound under a stronger oracle notion. We then lift this construction to a bilevel instance and show that the bilevel oracle call can be simulated by the oracle in the minimization hard instance.

4.1 A stochastic nonconvex minimization hard instance

Consider a smooth objective $P : \mathbb{R}^d \to \mathbb{R}$ whose gradient Lipschitz constant is bounded by a numerical constant and $P(0) - \inf P \leq 1$. An ordinary stochastic-gradient oracle returns $\widehat{\nabla P}(x)$ such that, for each query x selected from the history $\mathcal{H}$,

$$\mathbb{E}[\widehat{\nabla P}(x) \mid \mathcal{H}] = \nabla P(x), \qquad \mathbb{E}[\|\widehat{\nabla P}(x) - \nabla P(x)\|^2 \mid \mathcal{H}] \leq 1.$$

The construction below recovers the standard minimization lower bound under a strengthened oracle and provides the functions and mechanism used later in our bilevel hard instance.

The objective and its decomposition. The standard smooth chain $C_T(v) = \sum_{i=1}^{T} \psi_i(v)$, $v \in \mathbb{R}^T$, couples neighboring coordinates: each ψ_i depends only on (v_{i-1}, v_i), with $v_0 = 1$ (Carmon et al., 2020). At a point supported on the first k coordinates, its gradient is supported on at most the first $k+1$. Moreover, $|v_T| \leq 1$ implies $\|\nabla C_T(v)\| \geq 1$. Thus a small gradient requires reaching the last coordinate, while gradients expose at most one new coordinate at a time along such queries.

Following (Arjevani et al., 2023), we introduce smooth weights $\Theta_i(v) \in [0,1]$ to isolate the component that can reveal the next coordinate:

$$W_T(v) := -\sum_{i=1}^{T} \Theta_i(v)\psi_i(v), \qquad B_T(v) := C_T(v) + W_T(v). \tag{19}$$

The weight Θ_i equals one if $|v_j| \leq 1/4$ for every $j \geq i$, and zero if some $|v_j| \geq 1/2$ with $j \geq i$. For queries supported on the first k coordinates, ∇B_T stays within those coordinates, whereas ∇W_T can reach coordinate $k+1$. The specific weights also keep W_T, its gradient, and its Hessian bounded independently of T. Appendix B.1 gives the formulas for ψ_i, Θ_i and proves these properties.

Choose $T = \Theta(\epsilon^{-2})$ and numerical scalings $a = \Theta(\epsilon^2)$ and $\lambda_{\mathrm{lb}} = \Theta(\epsilon)$. Let $U \in \mathbb{R}^{d\times T}$ satisfy $U^\top U = I_T$; in the lower-bound argument, U is drawn uniformly at random to hide the chain directions from an adaptive algorithm (Arjevani et al., 2023; Woodworth & Srebro, 2017). Define the soft compression map

$$\rho_R(w) := \frac{w}{\sqrt{1 + \|w\|^2/R^2}}, \qquad R = 230\sqrt{T},$$

and the corresponding latent coordinate $v(x) := U^\top \rho_R(x/\lambda_{\mathrm{lb}})$. The hard objective and the pair (A, u) are

$$\begin{aligned} P(x) &:= aC_T(v(x)) + \frac{a}{10\lambda_{\mathrm{lb}}^2}\|x\|^2, \\ u(x) &:= aW_T(v(x)), \qquad A(x) := P(x) + u(x). \end{aligned} \tag{20}$$

Here $u : \mathbb{R}^d \to \mathbb{R}$ carries the next-coordinate information and $P = A - u$. The soft compression bounds the points seen by the chain, while the quadratic term prevents large-norm queries from having a small gradient. Although A, u, P depend on the chosen U through $v(x)$, we omit this dependence from the notation. Appendix B.2 verifies the scaling constants.

The gradient oracle and a comparison oracle. At each call, draw a fresh $\zeta \sim \mathrm{Bernoulli}(\varpi)$ and define the stochastic gradient estimator

$$\widehat{\nabla P}(x) := \nabla A(x) - \frac{\zeta}{\varpi}\nabla u(x). \tag{21}$$

Since $\mathbb{E}[\zeta/\varpi] = 1$, this estimator is unbiased, while the informative correction $\nabla u(x)$ is suppressed with probability $1 - \varpi$. For the lower-bound argument, we introduce a richer oracle response,

$$\mathcal{R}_\varpi(x) := \left(\nabla A(x), \frac{\zeta}{\varpi}u(x), \frac{\zeta}{\varpi}\nabla u(x)\right), \tag{22}$$

which we call the *comparison oracle*. It is stronger than (21), since subtracting its third component from its first recovers the stochastic gradient estimate in one call. The additional scalar component $\frac{\zeta}{\varpi}u(x)$ is included because it will be used to reproduce the lower-gradient oracle $\hat{G}_y$ in the bilevel construction. The next proposition establishes a lower bound directly for this stronger oracle.

Proposition 4.1 (Hard objective and comparison oracle)**.** *There are numerical constants $c, C, \epsilon_0 > 0$ such that, for every $0 < \epsilon < \epsilon_0$ and $0 < \varpi \le 1/2$, the construction (20) admits $T = \Theta(\epsilon^{-2})$, $d \ge T$, and a distribution of U with the following properties. Both ∇A and ∇u are C-Lipschitz, $P(0) - \inf P \le 1/4$, and*

$$\sup_x |u(x)| \le C\epsilon^2, \qquad \sup_x \|\nabla u(x)\| \le C\epsilon. \tag{23}$$

Every adaptive randomized algorithm using at most cT/ϖ calls to $\mathcal{R}_\varpi$, with arbitrary queries and final output, satisfies

$$\mathbb{P}\big(\|\nabla P(x_{\mathrm{out}})\| \ge 2\epsilon\big) \ge \frac{3}{4}. \tag{24}$$

The rotation U is independent of the algorithm seed and oracle draws; probability includes all three sources of randomness.

The two vector responses in $\mathcal{R}_\varpi$ arise from $U\nabla B_T(v)$ and $U(\zeta/\varpi)\nabla W_T(v)$. Along an exposed coordinate prefix, these vectors can add at most one direction, and only when $\zeta = 1$ (Corollary B.3), giving a mean waiting time $1/\varpi$ between opportunities for progress. For arbitrary queries, the proof conditions on the past latent coordinates $v(x_t)$, oracle draws, and exposed directions, which together determine all past scalar responses. The residual parts of unrevealed directions orthogonal to past queries and exposed directions remain spherically symmetric, allowing high-dimensional concentration to control accidental progress (Lemmas B.11–B.13). The same geometric bound covers the final output as an additional point without an oracle call (Corollary B.14). Appendix B.4 gives the full proof.

Choosing $\varpi = 2(1 + C^2)\epsilon^2$ recovers the classical $\Omega(\epsilon^{-4})$ lower bound for stochastic nonconvex minimization under an ordinary first-order oracle; see Appendix B.4.1. More importantly, the comparison-oracle construction enables a stronger bilevel lower bound: after lifting the hard instance to a bilevel problem, we can choose a even smaller reveal probability ϖ while preserving bounded variance, which increases the number of oracle calls required.

4.2 A BILEVEL HARD INSTANCE

We use the same pair (A, u) to construct a bilevel hard instance and its primitive first-order oracle. For a scale $\epsilon \le \delta \le 1$, let $x \in \mathbb{R}^d$ and let the lower-level variable be scalar, $y \in \mathbb{R}$. Define the upper- and lower-level objectives by

$$f_\delta(x, y) := A(x) + y, \qquad g_\delta(x, y) := \tfrac{1}{2}y^2 + \delta \sin(y/\delta)u(x). \tag{25}$$

At a query point (x, y), draw one response from the comparison oracle, $(r_1, r_2, r_3) = \mathcal{R}_\varpi(x) = \Big(\nabla A(x), \frac{\zeta}{\varpi} u(x), \frac{\zeta}{\varpi} \nabla u(x)\Big)$, and transform it into the primitive bilevel oracle response:

$$\hat{F}_x = r_1, \qquad \hat{F}_y = 1, \qquad \hat{G}_x = \delta \sin(y/\delta) r_3, \qquad \hat{G}_y = y + \cos(y/\delta) r_2. \tag{26}$$

These are exactly the stochastic gradients of f_δ and g_δ, with the same Bernoulli variable ζ appearing in both lower-level gradient blocks. Hence, one comparison-oracle call exactly simulates one primitive bilevel oracle call, including its joint randomness.

The terms carrying the hidden correction u are $\delta \sin(y/\delta)\nabla u(x)$ and $\cos(y/\delta)u(x)$. Their joint norm is $O(\delta\epsilon)$ uniformly over (x, y), since $\|\nabla u\| = O(\epsilon)$, $|u| = O(\epsilon^2)$, and $\delta \geq \epsilon$. We refer to this quantity as the *signal size*: it measures the magnitude of the deterministic terms involving u before multiplication by ζ/ϖ. These terms contain the information needed to recover the correction $-\nabla u$ in $\nabla P = \nabla A - \nabla u$.

The hyperobjective is $\Phi_\delta(x) = A(x) + y^\star_\delta(x)$, where

$$y^\star_\delta + u(x)\cos(y^\star_\delta/\delta) = 0, \qquad |y^\star_\delta| \leq |u(x)|.$$

Since $|u|/\delta = O(\epsilon)$, the lower solution is close to $-u$. Implicit differentiation gives the corresponding gradient estimate. The following proposition verifies that the bilevel construction satisfies Assumption 2.1 and that its hyperobjective remains close to the hard minimization objective P.

Proposition 4.2 (Bilevel gradient and primitive variance)**.** *For each pair in Proposition 4.1, every $\epsilon \leq \delta \leq 1$, and sufficiently small ϵ, the family (25) satisfies Assumption 2.1 with uniformly bounded constants, has $\Phi_\delta(0) - \inf \Phi_\delta \leq 1$ and $|y^\star_\delta(0)| \leq 1$, and obeys*

$$\sup_x \|\nabla \Phi_\delta(x) - \nabla P(x)\| \leq C_{\mathrm{lift}} \epsilon^5/\delta^2 \leq C_{\mathrm{lift}} \epsilon^3 \tag{27}$$

for a numerical $C_{\mathrm{lift}} > 0$. The oracle (26) is conditionally unbiased, has joint variance at most $C_{\mathrm{lift}} \delta^2 \epsilon^2/\varpi$, and is simulated by one call to $\mathcal{R}_\varpi$.

Appendix B.5 proves the proposition. Choosing $\varpi_\delta = \Theta(\delta^2\epsilon^2)$ keeps the primitive variance at most one. On the event $\|\nabla P\| \geq 2\epsilon$, the gradient estimate gives $\|\nabla \Phi_\delta\| \geq 3\epsilon/2$ for small enough ϵ. The one-call simulation therefore transfers Proposition 4.1 to the bilevel objective, giving

$$\Omega(T/\varpi_\delta) = \Omega(\epsilon^{-4}\delta^{-2}) \tag{28}$$

primitive calls. The intuition is that the sinusoidal lower-level construction attenuates the primitive signal while preserving the relevant hypergradient information, allowing the same variance budget to support a smaller reveal probability.

4.3 Lower bound and minimax complexity

Decreasing δ increases the higher-order smoothness constants of g with respect to the lower variable. For the construction in (25), the order-p smoothness constants in Assumption 2.2 are bounded by $C\epsilon\delta^{-p}$ for $\ell = x$ and $C\epsilon^2\delta^{-(p+1)}$ for $\ell = y$, where C is a numerical constant independent of p. Choosing $\delta = \epsilon^{1/p}$ keeps these quantities, as well as the base regularity constants, uniformly bounded. Substituting this choice into (28) yields the exponent $4 + 2/p$.

For $M \geq 1$, let $\mathcal{P}_p(M)$ be the class of problem–oracle pairs satisfying Assumptions 2.1–2.3, in arbitrary finite dimensions, with

$$\max\{\ell_f, \ell_g, \rho_1, L_p, B_f, \sigma, \mu^{-1}\} \leq M, \tag{29}$$

$$\Phi(0) - \Phi_{\inf} \leq 1, \qquad \|y^\star(0)\| \leq 1. \tag{30}$$

Queries and the final output may be arbitrary measurable functions of past oracle responses and the algorithm's internal randomness.

Theorem 4.3 (Lower bound for bilevel problems)**.** *There exist numerical constants $M_0 \geq 1$, $c > 0$, and $\epsilon_0 > 0$ such that the following holds for every integer $p \geq 1$ and $0 < \epsilon < \epsilon_0$. For any adaptive randomized algorithm using at most $c\epsilon^{-4-2/p}$ primitive oracle calls, there is a pair in $\mathcal{P}_p(M_0)$ on which its output satisfies*

$$\mathbb{E}\|\nabla \Phi(x_{\mathrm{out}})\|^2 > \epsilon^2. \tag{31}$$

Appendix B.6 gives the proof. The hard instances have $d_y = 1$ and can be constructed with $d_x = O(\epsilon^{-6-2/p}\log(1/\epsilon))$. Theorem 4.3 shows that the upper bound in Theorem 3.1 is tight in its dependence on ϵ for every fixed finite p and fixed problem constants.

AI USE STATEMENT

We used OpenAI GPT-5.5 to assist with literature search and reference checking, manuscript drafting and revision, and selected mathematical computations and proof checking. The algorithmic design and the main convergence proof were developed by the authors; GPT-5.5 was used to help verify intermediate calculations, and improve the presentation. The initial lower-bound construction emerged through iterative discussions between the authors and GPT-5.5. The authors subsequently simplified the argument, reorganized the proof, and developed the current comparison-oracle and bilevel-lifting framework. The authors take full responsibility for the final manuscript, including all mathematical claims, proofs, references, and interpretations.

REFERENCES

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2023. URL https://arxiv.org/abs/1912.02365.

Kendall Atkinson and Weimin Han. Finite difference method. In *Theoretical Numerical Analysis: A Functional Analysis Framework*, pp. 171–192. Springer, 2001.

Honghuan Bai, Aimin Xu, and Feng Cui. Representation for the lagrangian numerical differentiation formula involving elementary symmetric functions. *Journal of computational and applied mathematics*, 231(2):907–913, 2009.

Jérôme Bolte, Tam Le, Edouard Pauwels, and Tony Silveti-Falls. Nonsmooth implicit differentiation for machine-learning and optimization. *Advances in neural information processing systems*, 34: 13537–13549, 2021.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184:71–120, 2020. URL https://arxiv.org/abs/1710.11606.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 947–980. PMLR, 2024a.

Lesi Chen, Yaohua Ma, and Jingzhao Zhang. Near-optimal nonconvex–strongly-convex bilevel optimization with fully first-order oracles. *Journal of Machine Learning Research*, 26(109):1–56, 2025. URL https://www.jmlr.org/papers/v26/23-1104.html.

Lesi Chen, Kaiyi Ji, and Jingzhao Zhang. On the condition number dependency in bilevel optimization. arXiv preprint, 2026a. URL https://arxiv.org/abs/2511.22331. Version 4, 11 August 2026.

Lesi Chen, Junru Li, El Mahdi Chayti, and Jingzhao Zhang. Faster gradient methods for highly-smooth stochastic bilevel optimization. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2509.02937.

Xuxing Chen, Tesi Xiao, and Krishnakumar Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *Journal of Machine Learning Research*, 25(151):1–51, 2024b. URL https://www.jmlr.org/papers/v25/23-1323.html.

Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/aa84ec1ac3f5fdcf77bce2c22705ab77-Abstract-Conference.html.

Mathieu Dagréou, Thomas Moreau, Samuel Vaiter, and Pierre Ablin. A lower bound and a near-optimal algorithm for bilevel empirical risk minimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 82–90. PMLR, 2024.

Justin Domke. Generic methods for optimization-based modeling. In *Artificial Intelligence and Statistics*, pp. 318–326. PMLR, 2012.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Bengt Fornberg. Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of computation*, 51(184):699–706, 1988.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International conference on machine learning*, pp. 1568–1577. PMLR, 2018.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. arXiv preprint, 2018. URL https://arxiv.org/abs/1802.02246.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pp. 3748–3758. PMLR, 2020.

Zhihao Gu, Qilong Wu, and Junchi Yang. An $\Omega(\kappa_y^8 \epsilon^{-6})$ lower bound for stochastic NC-SC bilevel optimization with first-order oracles. arXiv preprint, 2026a. URL https://arxiv.org/abs/2609.21905.

Zhihao Gu, Qilong Wu, and Junchi Yang. SGHA: A single-loop fully first-order algorithm for nonconvex–strongly-convex bilevel optimization. arXiv preprint, 2026b. URL https://arxiv.org/abs/2608.23211.

Kaiyi Ji and Yingbin Liang. Lower bounds and accelerated algorithms for bilevel optimization. *Journal of machine learning research*, 24(22):1–56, 2023.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4882–4892, 2021. URL https://proceedings.mlr.press/v139/ji21c.html.

Ishtiaq Rasool Khan and Ryoji Ohba. Closed-form expressions for the finite difference approximations of first and higher derivatives based on taylor series. *Journal of Computational and Applied Mathematics*, 107(2):179–193, 1999.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D. Nowak. A fully first-order method for stochastic bilevel optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18083–18113, 2023. URL https://proceedings.mlr.press/v202/kwon23c.html.

Jeongyeol Kwon, Dohyun Kwon, and Hanbaek Lyu. On the complexity of first-order methods in stochastic bilevel optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25784–25811, 2024. URL https://proceedings.mlr.press/v235/kwon24b.html.

Jianping Li. General explicit difference formulas for numerical differentiation. *Journal of Computational and Applied Mathematics*, 183(1):29–52, 2005.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in neural information processing systems*, 35: 17248–17262, 2022.

Rui Pan, Dylan Zhang, Hanning Zhang, Xingyuan Pan, Minrui Xu, Jipeng Zhang, Renjie Pi, Xiaoyu Wang, and Tong Zhang. Scalebio: Scalable bilevel optimization for llm data reweighting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31959–31982, 2025.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746. PMLR, 2016.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1723–1732. PMLR, 2019.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International conference on machine learning*, pp. 30992–31015. PMLR, 2023.

Blake Woodworth and Nathan Srebro. Lower bound for randomized first order convex optimization. *arXiv preprint arXiv:1709.03594*, 2017.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2312.03807.

A Proof of Theorem 3.1

This appendix proves the guarantee for Algorithm 1 with a common difference radius and batch size. We first collect constants, then establish regularity, control the three updates, and sum a Lyapunov inequality. Throughout, Assumptions 2.1–2.3 hold, and the initial points x_0, y_0 are fixed with $z_0 = 0$.

A.1 Constants and parameter choices

Write $L = \ell_g$ and use $\Phi_{\inf} = \inf_x \Phi(x)$, as in Section 3.3. Define the geometry constants

$$q_0 := \frac{B_f}{\mu}, \quad L_y := \frac{L}{\mu}, \quad B := \ell_f + \rho_1 q_0, \quad L_z := \frac{(1+L_y)B}{\mu}, \quad L_\Phi := (1+L_y)B + LL_z. \tag{32}$$

The projection radius is $R_z = 4q_0$, as in Section 2. The chosen nodes and weights define finite constants

$$A_p := \sum_{j \in \mathcal{I}_p} a_j^2, \qquad C_p := \frac{1}{(p+1)!} \sum_{j \in \mathcal{I}_p} |a_j|\,|j|^{p+1}. \tag{33}$$

For the lower and auxiliary steps, set

$$\eta := \frac{1}{2L}, \qquad r := \mu\eta, \qquad \omega := 1 + \frac{36B^2}{\mu^2}, \qquad \rho := \frac{r}{4}. \tag{34}$$

The remaining constants used in the proof are

$$C_{\mathrm{a}} := 3\left(1 + \frac{3L}{\mu}\right), \qquad C := 3L^2 + \frac{B^2 C_{\mathrm{a}}}{\omega}, \qquad D := \frac{3}{r}\left(L_z^2 + \omega L_y^2\right), \tag{35}$$

$$N_{\mathrm{b}} := \frac{9\eta}{\mu}, \qquad N_{\mathrm{v}} := \frac{3}{2}(1+\omega)\eta^2. \tag{36}$$

All step-size denominators below are positive. In particular, strong convexity and gradient Lipschitzness imply $0 < \mu \le L$, and $0 < r \le 1/2$. Choose

$$\eta_y = \eta_z = \eta, \qquad \eta_x = \alpha := \min\left\{\frac{1}{2L_\Phi}, \sqrt{\frac{\rho}{16CD}}\right\}, \qquad \lambda := \frac{2\alpha C}{\rho}. \tag{37}$$

Define the error and accuracy constants

$$S_{\mathrm{b}} := C_{\mathrm{a}} + \frac{2CN_{\mathrm{b}}}{\rho}, \qquad S_{\mathrm{v}} := 2 + \frac{2CN_{\mathrm{v}}}{\rho}, \tag{38}$$

$$C_{\mathrm{b}} := 4S_{\mathrm{b}}(C_p L_p R_z)^2, \qquad C_{\mathrm{v}} := 4S_{\mathrm{v}}\sigma^2(1 + A_p R_z^2). \tag{39}$$

For the constant in Theorem 3.1, take

$$C_\star := \max\left\{1, \frac{4}{\alpha}\max\{1, \lambda\omega\}, C_{\mathrm{b}}, C_{\mathrm{v}}\right\}.$$

These quantities depend only on the problem parameters and p, not on ϵ, the batch size, or the iteration count.

The errors below use the target at x_k, including after the z update:

$$\delta_k := y_k - y^\star(x_k), \qquad e_k := z_k - z^\star(x_k), \qquad \bar{e}_k := z_{k+1} - z^\star(x_k), \tag{40}$$

$$T_k := \|e_k\|^2 + \omega\|\delta_k\|^2, \qquad \mathcal{V}_k := \Phi(x_k) - \Phi_{\inf} + \lambda T_k. \tag{41}$$

Since $z_0 = 0$, we have $\|e_0\| \le q_0$. Using $\omega \ge 1$ and the initial bound in Section 3.3,

$$\begin{aligned} \mathcal{V}_0 &\le \max\{1, \lambda\omega\}\left(\Phi(x_0) - \Phi_{\inf} + q_0^2 + \|y_0 - y^\star(x_0)\|^2\right) \\ &\le \max\{1, \lambda\omega\}\Delta_0. \end{aligned} \tag{42}$$

An upper bound suffices; the algorithm does not need to query the exact lower solution or the objective value to implement the parameter rule.

We will prove, for $0 < \tau \le 1$ and positive integers m, K, that

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|\nabla\Phi(x_k)\|^2 \le C_\star\left(\frac{\Delta_0}{K} + \tau^{2p} + \frac{1}{m\tau^2}\right). \tag{43}$$

A.2 GEOMETRY AND ERRORS FROM THE LOWER ITERATE

Lemma A.1 (Regularity of the lower solution and auxiliary target)**.** *Under Assumption 2.1, the lower solution exists uniquely for every x. Moreover,*

$$\|y^\star(x) - y^\star(x')\| \le L_y\|x - x'\|, \tag{44}$$

$$\|z(x,y)\|,\ \|z^\star(x)\| \le q_0, \tag{45}$$

$$\|z^\star(x) - z^\star(x')\| \le L_z\|x - x'\|. \tag{46}$$

The hypergradient identity (4) holds, and $\nabla\Phi$ is L_Φ-Lipschitz.

Proof. Strong convexity gives a coercive quadratic lower bound for $g(x,\cdot)$, so continuity gives existence and strong convexity gives uniqueness. The first-order condition is $\nabla_y g(x, y^\star(x)) = 0$. For $a = y^\star(x)$ and $a' = y^\star(x')$, strong monotonicity implies

$$\mu\|a - a'\|^2 \le \langle \nabla_y g(x,a) - \nabla_y g(x,a'), a - a'\rangle.$$

Substitute the two first-order conditions and apply the Lipschitz bound in x. The right-hand side is at most $L\|x - x'\|\|a - a'\|$, proving (44), including the case $a = a'$.

Differentiability of the Lipschitz gradient and lower strong convexity give

$$\mu I \preceq H(x,y) \preceq LI, \qquad \|J(x,y)\|_{\mathrm{op}} \le L, \qquad \|H(x,y)^{-1}\|_{\mathrm{op}} \le \mu^{-1}.$$

The bound on $\nabla_y f$ now proves (45). The derivatives of $(x,y) \mapsto \nabla_y g(x,y)$ are $J(x,y)^\top$ and $H(x,y)$; their continuity follows from the matrix Lipschitz bounds in Assumption 2.1(iii). The implicit-function theorem therefore applies locally at each lower solution. Uniqueness makes these local solution maps agree. It follows that

$$D_x y^\star(x) = -H(x, y^\star(x))^{-1} J(x, y^\star(x))^\top.$$

The chain rule proves (4).

For brevity in the remaining estimates, let $s = (x, y^\star(x))$, $s' = (x', y^\star(x'))$, $q = z^\star(x)$, and $q' = z^\star(x')$. Subtracting the two auxiliary systems gives

$$H(s)(q - q') = b(s) - b(s') - [H(s) - H(s')]q'.$$

By Assumption 2.1(iii), $\|H(s) - H(s')\|_{\mathrm{op}} \le \rho_1\|s - s'\|$. Use this bound, $\|q'\| \le q_0$, $\|H(s)^{-1}\|_{\mathrm{op}} \le \mu^{-1}$, and

$$\|s - s'\| \le \|x - x'\| + \|y^\star(x) - y^\star(x')\| \le (1 + L_y)\|x - x'\|.$$

The resulting bound is $\|q - q'\| \le B(1 + L_y)\|x - x'\|/\mu$, proving (46).

Finally, the difference of the hypergradients equals

$$\nabla_x f(s) - \nabla_x f(s') - [J(s) - J(s')]q' - J(s)(q - q').$$

The first two terms have combined norm at most $B(1 + L_y)\|x - x'\|$, and the last term has norm at most $LL_z\|x - x'\|$. Their sum gives L_Φ. □

Lemma A.2 (Errors caused by the current lower iterate)**.** *For every x, y, z, under Assumption 2.1,*

$$\|H(x,y)z^\star(x) - b(x,y)\| \le B\|y - y^\star(x)\|, \tag{47}$$

$$\|\nabla_x f(x,y) - J(x,y)z - \nabla\Phi(x)\| \le L\|z - z^\star(x)\| + B\|y - y^\star(x)\|. \tag{48}$$

Proof. Write $y^\star = y^\star(x)$ and $q = z^\star(x)$. The identity $H(x,y^\star)q = b(x,y^\star)$ gives

$$H(x,y)q - b(x,y) = [H(x,y) - H(x,y^\star)]q + b(x,y^\star) - b(x,y).$$

The operator-norm Lipschitz bound on H and $\|q\| \le q_0$ give the bound $(\rho_1 q_0 + \ell_f)\|y - y^\star\|$, proving (47). For the second bound, subtract (4) to obtain

$$\begin{aligned}&\nabla_x f(x,y) - J(x,y)z - \nabla\Phi(x)\\ &\quad= \nabla_x f(x,y) - \nabla_x f(x,y^\star) - [J(x,y) - J(x,y^\star)]q - J(x,y)(z - q).\end{aligned}$$

The three terms have norms bounded respectively by $\ell_f\|y - y^\star\|$, $\rho_1 q_0\|y - y^\star\|$, and $L\|z - q\|$. Their sum is (48). □

A.3 FINITE DIFFERENCES AND CONDITIONAL SAMPLING

Lemma A.3 (Normalized finite-difference error)**.** *The weights in (10) exist uniquely for every integer $p \geq 1$. Under Assumptions 2.2 and 2.3, the finite differences and their sampled estimates satisfy (13)–(14) with C_p, A_p from (33). More precisely, if x, y, z are measurable with respect to a history $\mathcal{F}$ and the endpoint batches use fresh calls after that history, then, for $\ell \in \{x, y\}$,*

$$\mathbb{E}[\widehat{\mathrm{FD}}_\ell(x, y; z) \mid \mathcal{F}] = \mathrm{FD}_\ell(x, y; z), \tag{49}$$

$$\|\mathrm{FD}_\ell(x, y; z) - \nabla^2_{\ell y} g(x, y) z\| \leq C_p L_p \tau^p \|z\|, \tag{50}$$

$$\mathbb{E}[\|\widehat{\mathrm{FD}}_\ell(x, y; z) - \mathrm{FD}_\ell(x, y; z)\|^2 \mid \mathcal{F}] \leq \frac{A_p \sigma^2 \|z\|^2}{m\tau^2}. \tag{51}$$

These statements include $z = 0$.

Proof. The set $\mathcal{I}_p$ has $p + 1$ distinct nodes, including for $p = 1$. Let $l_j(t)$ be its Lagrange basis polynomials. Differentiating the interpolation identity for a polynomial of degree at most p shows that $a_j = l_j'(0)$ satisfies (10). The Vandermonde matrix of distinct nodes is invertible, giving uniqueness.

Suppose first that $z \neq 0$ and condition on $\mathcal{F}$. Write $u = z/\|z\|$ and $\varphi(t) = \nabla_\ell g(x, y + tu)$. By Assumption 2.2, $\nabla_y^p \nabla_\ell g(x, \cdot)$ is L_p-Lipschitz. Since $\|u\| = 1$, the pth derivative of φ is also L_p-Lipschitz. Taylor's formula with an integral remainder gives, for every real t,

$$\varphi(t) = \sum_{r=0}^{p} \frac{t^r}{r!} \varphi^{(r)}(0) + \mathcal{R}_p(t),$$

$$\mathcal{R}_p(t) = \frac{t^p}{(p-1)!} \int_0^1 (1 - s)^{p-1} [\varphi^{(p)}(st) - \varphi^{(p)}(0)]\, ds.$$

In particular, including when $t < 0$,

$$\|\mathcal{R}_p(t)\| \leq \frac{L_p |t|^{p+1}}{(p-1)!} \int_0^1 s(1 - s)^{p-1}\, ds = \frac{L_p |t|^{p+1}}{(p+1)!}.$$

Thus this argument does not require a $(p + 1)$st derivative. Insert the expansion at $t = j\tau$ into (11). All polynomial terms except the first derivative vanish by the moment identities. The retained term is $\|z\|\varphi'(0) = \nabla^2_{\ell y} g(x, y) z$. The remainder has norm at most

$$\frac{\|z\|}{\tau} \sum_j |a_j| \frac{L_p |j\tau|^{p+1}}{(p+1)!} = C_p L_p \tau^p \|z\|.$$

For each node j, set $\nu_j := \bar{G}_{\ell,j}(x, y + j\tau u) - \nabla_\ell g(x, y + j\tau u)$. Every summand in (7) is nonnegative, so each block $\hat{G}_\ell(x, y)$ has conditional variance at most σ^2. Averaging m independent calls reduces this bound to σ^2/m. Together with (6), this gives

$$\mathbb{E}[\nu_j \mid \mathcal{F}] = 0, \qquad \mathbb{E}[\|\nu_j\|^2 \mid \mathcal{F}] \leq \sigma^2/m, \qquad \mathbb{E}[\langle \nu_i, \nu_j \rangle \mid \mathcal{F}] = 0 \quad (i \neq j).$$

The cross terms vanish because distinct nodes use distinct calls; no independence between blocks within a single call is required. Consequently,

$$\begin{aligned} \mathbb{E}[\|\widehat{\mathrm{FD}}_\ell - \mathrm{FD}_\ell\|^2 \mid \mathcal{F}] &= \frac{\|z\|^2}{\tau^2} \sum_j a_j^2 \mathbb{E}[\|\nu_j\|^2 \mid \mathcal{F}] \\ &\leq \frac{A_p \sigma^2 \|z\|^2}{m\tau^2}, \end{aligned}$$

and the same centered-noise representation proves (49). If $z = 0$, both differences and both exact matrix–vector products are zero by definition, so every assertion follows without endpoint calls. □

We make the sampling convention for Algorithm 1 explicit. Each occurrence of a center or node average uses a separate batch of m fresh calls, including when query locations coincide. The z update is sampled first. After z_{k+1} is computed, new batches are drawn for the x update; a final fresh batch

is then drawn for the y update. All center batches are evaluated at (x_k, y_k), as in the algorithm. Thus the outer finite difference uses z_{k+1}, whereas the auxiliary finite difference uses z_k. At termination, the output index R is drawn uniformly and independently of all oracle calls.

Let $\mathcal{F}_k$ contain all calls before iteration k, let $\mathcal{F}_k^z$ additionally contain the calls used to compute z_{k+1}, and let $\mathcal{F}_k^x$ additionally contain those used to compute x_{k+1}. The final lower-gradient batch completes $\mathcal{F}_{k+1}$. Thus

$$\mathcal{F}_k \subseteq \mathcal{F}_k^z \subseteq \mathcal{F}_k^x \subseteq \mathcal{F}_{k+1}.$$

In particular, $\mathbb{P}(R = j \mid \mathcal{F}_K) = 1/K$ for $j = 0, \dots, K-1$. Write $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$ and introduce the uniform error bounds

$$\beta := C_p L_p R_z \tau^p, \qquad v := \frac{\sigma^2}{m}\left(1 + \frac{A_p R_z^2}{\tau^2}\right), \qquad v_y := \frac{\sigma^2}{m} \le v. \tag{52}$$

Lemma A.4 (Conditional decomposition of the three updates)**.** *Put $H_k = H(x_k, y_k)$, $J_k = J(x_k, y_k)$, and $b_k = b(x_k, y_k)$. The sampled directions in Algorithm 1 can be written as*

$$\hat{r}_k := \widehat{\mathrm{FD}}_y(x_k, y_k; z_k) - \bar{F}_{y,k} = H_k z_k - b_k + d_k^z + \xi_k^z, \tag{53}$$

$$\hat{h}_k = \nabla_x f(x_k, y_k) - J_k z_{k+1} + d_k^x + \xi_k^x, \tag{54}$$

$$\bar{G}_{y,k} = \nabla_y g(x_k, y_k) + \xi_k^y. \tag{55}$$

Here d_k^z is $\mathcal{F}_k$-measurable, d_k^x is $\mathcal{F}_k^z$-measurable, and $\|d_k^z\|, \|d_k^x\| \le \beta$. The noises satisfy

$$\mathbb{E}[\xi_k^z \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\|\xi_k^z\|^2 \mid \mathcal{F}_k] \le v, \tag{56}$$

$$\mathbb{E}[\xi_k^x \mid \mathcal{F}_k^z] = 0, \qquad \mathbb{E}[\|\xi_k^x\|^2 \mid \mathcal{F}_k^z] \le v, \tag{57}$$

$$\mathbb{E}[\xi_k^y \mid \mathcal{F}_k^x] = 0, \qquad \mathbb{E}[\|\xi_k^y\|^2 \mid \mathcal{F}_k^x] \le v_y. \tag{58}$$

Proof. Every z_k belongs to Q, since $z_0 = 0$ and every auxiliary update is projected. Define

$$d_k^z := \mathrm{FD}_y(x_k, y_k; z_k) - H_k z_k,$$
$$d_k^x := J_k z_{k+1} - \mathrm{FD}_x(x_k, y_k; z_{k+1}).$$

Lemma A.3 gives their measurability and bounds. Notice the opposite signs: the outer direction subtracts its finite difference.

For ξ_k^z, subtract the finite differences and exact center gradients from their sampled counterparts in $\hat{r}_k$. The center term $\bar{F}_{y,k} - b_k$ is conditionally centered with variance at most σ^2/m by the $\hat{F}_y$ block of (6)–(7) and fresh batching. It uses a separate batch from the finite difference, so their conditional variances add, giving at most

$$\frac{A_p \sigma^2 \|z_k\|^2}{m\tau^2} + \frac{\sigma^2}{m} \le v.$$

For the outer direction, condition on $\mathcal{F}_k^z$. Its input z_{k+1} is now fixed and lies in Q. Entirely new center and node batches give the same argument using the $\hat{F}_x$ block and z_{k+1} in place of z_k. Finally, $\bar{G}_{y,k}$ is drawn after $\mathcal{F}_k^x$ at the already selected point (x_k, y_k). The $\hat{G}_y$ block of (6)–(7) gives its conditional mean $\nabla_y g(x_k, y_k)$ and variance bound v_y after averaging. These arguments also cover repeated query locations because their batches are distinct. □

All expectations used below are finite. To see this without assuming bounded iterates, start with the fixed x_0, y_0. The gradients of f and g grow at most linearly by Assumption 2.1(i), while the mean finite differences have norm at most $LR_z + \beta$ by Lemma A.3. The noises above have uniformly bounded second moments. Since z_k is bounded, induction through the three updates shows that x_k, y_k have finite second moments at every finite k. Lemma A.1 then gives finite second moments for the tracking errors and the outer gradient. Finally, smoothness of Φ bounds $|\Phi(x)|$ by a constant times $1 + \|x\|^2$, so the objective and potential are integrable as well.

A.4 One-step descent and tracking estimates

Lemma A.5 (Projected auxiliary correction at the current outer point). *With $\eta_z = \eta = 1/(2L)$, the auxiliary update satisfies*

$$\mathbb{E}_k\|\bar{e}_k\|^2 \le (1-r)\|e_k\|^2 + \frac{6\eta}{\mu}\big(B^2\|\delta_k\|^2 + \beta^2\big) + \eta^2 v. \tag{59}$$

The target in $\bar{e}_k$ is $z^\star(x_k)$, before the outer point moves.

Proof. Set $q = z^\star(x_k)$ and $w_k = H_k q - b_k + d_k^z$. These are $\mathcal{F}_k$-measurable, and Lemmas A.2 and A.4 give

$$\|w_k\|^2 \le 2B^2\|\delta_k\|^2 + 2\beta^2. \tag{60}$$

Because $q \in Q$ and Euclidean projection is nonexpansive,

$$\|\bar{e}_k\|^2 \le \|(I - \eta H_k)e_k - \eta w_k - \eta\xi_k^z\|^2.$$

Take conditional expectation only after this inequality. The noise is centered before projection, so

$$\mathbb{E}_k\|\bar{e}_k\|^2 \le \|(I - \eta H_k)e_k - \eta w_k\|^2 + \eta^2 v. \tag{61}$$

No unbiasedness property of the projected iterate is needed.

The eigenvalues of H_k lie in $[\mu, L]$, and $\eta = 1/(2L)$ implies $\|I - \eta H_k\|_{\mathrm{op}} \le 1 - r$. Apply Young's inequality $\|a+b\|^2 \le (1+s)\|a\|^2 + (1+s^{-1})\|b\|^2$ with $s = r/2$ to the deterministic term in (61). Since $0 < r \le 1/2$,

$$(1 + r/2)(1-r)^2 \le 1 - r, \qquad 1 + 2/r \le 3/r.$$

It follows that

$$\mathbb{E}_k\|\bar{e}_k\|^2 \le (1-r)\|e_k\|^2 + \frac{3\eta^2}{r}\|w_k\|^2 + \eta^2 v.$$

Substitute (60) and use $r = \mu\eta$ to obtain (59). □

Lemma A.6 (Outer-direction error and objective descent). *Let $h_k = \nabla\Phi(x_k)$ and $d_k = \mathbb{E}[\hat{h}_k \mid \mathcal{F}_k^z] - h_k$. With the steps in (37),*

$$\mathbb{E}_k\|d_k\|^2 \le CT_k + C_{\mathrm{a}}\beta^2 + \tfrac{3}{4}v, \tag{62}$$

$$\mathbb{E}_k\|\hat{h}_k\|^2 \le 2\|h_k\|^2 + 2CT_k + 2C_{\mathrm{a}}\beta^2 + \tfrac{5}{2}v, \tag{63}$$

$$\mathbb{E}_k\Phi(x_{k+1}) \le \Phi(x_k) - \tfrac{\alpha}{2}\|h_k\|^2 + \tfrac{\alpha C}{2}T_k + \tfrac{\alpha C_{\mathrm{a}}}{2}\beta^2 + \tfrac{7\alpha}{8}v. \tag{64}$$

In particular, $\mathbb{E}_k\|x_{k+1} - x_k\|^2 = \alpha^2\mathbb{E}_k\|\hat{h}_k\|^2$.

Proof. By Lemmas A.2 and A.4,

$$\|d_k\| \le L\|\bar{e}_k\| + B\|\delta_k\| + \beta.$$

Squaring, taking $\mathbb{E}_k$, and applying Lemma A.5 yields

$$\begin{aligned}
\mathbb{E}_k\|d_k\|^2 &\le 3L^2\mathbb{E}_k\|\bar{e}_k\|^2 + 3B^2\|\delta_k\|^2 + 3\beta^2 \\
&\le 3L^2\|e_k\|^2 + 3B^2\left(1 + \frac{6L^2\eta}{\mu}\right)\|\delta_k\|^2 + 3\left(1 + \frac{6L^2\eta}{\mu}\right)\beta^2 + 3L^2\eta^2 v \\
&= 3L^2\|e_k\|^2 + B^2C_{\mathrm{a}}\|\delta_k\|^2 + C_{\mathrm{a}}\beta^2 + \tfrac{3}{4}v.
\end{aligned}$$

The last equality uses $\eta = 1/(2L)$ and the definition of C_{a}. By definition of C and T_k, $3L^2\|e_k\|^2 + B^2C_{\mathrm{a}}\|\delta_k\|^2 \le CT_k$. This proves (62).

Conditional on $\mathcal{F}_k^z$, the mean of $\hat{h}_k$ is $h_k + d_k$ and its centered second moment is at most v. Therefore

$$\mathbb{E}[\|\hat{h}_k\|^2 \mid \mathcal{F}_k^z] \le \|h_k + d_k\|^2 + v \le 2\|h_k\|^2 + 2\|d_k\|^2 + v.$$

Taking $\mathbb{E}_k$ and using (62) proves (63).

By L_Φ-smoothness and the outer update,

$$\mathbb{E}[\Phi(x_{k+1}) \mid \mathcal{F}_k^z] \le \Phi(x_k) - \alpha\langle h_k, h_k + d_k\rangle + \frac{L_\Phi \alpha^2}{2}\big(\|h_k + d_k\|^2 + v\big).$$

The exact identity

$$-\langle h, h + d\rangle = -\tfrac{1}{2}\|h\|^2 - \tfrac{1}{2}\|h + d\|^2 + \tfrac{1}{2}\|d\|^2$$

and $\alpha L_\Phi \le 1$ give

$$\mathbb{E}[\Phi(x_{k+1}) \mid \mathcal{F}_k^z] \le \Phi(x_k) - \frac{\alpha}{2}\|h_k\|^2 + \frac{\alpha}{2}\|d_k\|^2 + \frac{L_\Phi \alpha^2}{2} v.$$

Take $\mathbb{E}_k$, insert (62), and bound $L_\Phi \alpha^2/2 \le \alpha/2$. The variance coefficient becomes $3\alpha/8 + \alpha/2 = 7\alpha/8$, which proves (64). The movement identity follows directly from $x_{k+1} - x_k = -\alpha \hat{h}_k$. □

Lemma A.7 (Tracking after the outer point moves)**.** *With the steps in (37), the two tracking errors satisfy*

$$\mathbb{E}_k\|e_{k+1}\|^2 \le (1 - r/2)\|e_k\|^2 + \frac{9\eta B^2}{\mu}\|\delta_k\|^2 + \frac{9\eta}{\mu}\beta^2 + \frac{3}{2}\eta^2 v + \frac{3L_z^2}{r}\mathbb{E}_k\|x_{k+1} - x_k\|^2, \quad (65)$$

$$\mathbb{E}_k\|\delta_{k+1}\|^2 \le (1 - r/2)\|\delta_k\|^2 + \frac{3}{2}\eta^2 v_y + \frac{3L_y^2}{r}\mathbb{E}_k\|x_{k+1} - x_k\|^2. \quad (66)$$

Consequently, with T_k from (41),

$$\mathbb{E}_k T_{k+1} \le (1 - \rho)T_k + N_{\mathrm{b}}\beta^2 + N_{\mathrm{v}} v + D\mathbb{E}_k\|x_{k+1} - x_k\|^2. \quad (67)$$

Proof. For the auxiliary target, the exact decomposition is

$$e_{k+1} = \bar{e}_k - [z^\star(x_{k+1}) - z^\star(x_k)].$$

Apply Young's inequality with $s = r/2$ and use the L_z-Lipschitz bound:

$$\mathbb{E}_k\|e_{k+1}\|^2 \le (1 + r/2)\mathbb{E}_k\|\bar{e}_k\|^2 + \frac{3L_z^2}{r}\mathbb{E}_k\|x_{k+1} - x_k\|^2.$$

Insert (59) and use $(1+r/2)(1-r) \le 1 - r/2$ and $1 + r/2 \le 3/2$. This gives (65); no independence between $\bar{e}_k$ and the target movement was used.

For the lower variable, write $y_k^\star = y^\star(x_k)$ and define

$$\widetilde{\delta}_k := \delta_k - \eta\nabla_y g(x_k, y_k).$$

By the fundamental theorem of calculus and $\nabla_y g(x_k, y_k^\star) = 0$,

$$\nabla_y g(x_k, y_k) = \bar{H}_k \delta_k, \qquad \bar{H}_k := \int_0^1 H(x_k, y_k^\star + t\delta_k)\, dt.$$

The symmetric matrix $\bar{H}_k$ has eigenvalues in $[\mu, L]$. Thus $\|\widetilde{\delta}_k\| \le (1-r)\|\delta_k\|$. Put $\Delta_k^y = y^\star(x_{k+1}) - y^\star(x_k)$. The lower update gives $\delta_{k+1} = \widetilde{\delta}_k - \Delta_k^y - \eta\xi_k^y$. Both $\widetilde{\delta}_k$ and Δ_k^y are $\mathcal{F}_k^x$-measurable. Centering the final fresh noise at this history therefore yields

$$\mathbb{E}_k\|\delta_{k+1}\|^2 \le \mathbb{E}_k\|\widetilde{\delta}_k - \Delta_k^y\|^2 + \eta^2 v_y.$$

Young's inequality with $s = r/2$ and the L_y-Lipschitz bound imply

$$\mathbb{E}_k\|\delta_{k+1}\|^2 \le (1 + r/2)(1 - r)^2\|\delta_k\|^2 + \frac{3L_y^2}{r}\mathbb{E}_k\|x_{k+1} - x_k\|^2 + \eta^2 v_y.$$

Relaxing $(1 + r/2)(1 - r)^2 \le 1 - r \le 1 - r/2$ and $\eta^2 v_y \le (3/2)\eta^2 v_y$ proves (66).

Finally, multiply (66) by ω and add (65). The only coupling term is absorbed because

$$\frac{\omega r}{4} = \frac{r}{4} + \frac{9\eta B^2}{\mu} \ge \frac{9\eta B^2}{\mu}.$$

The coefficient of $\|\delta_k\|^2$ is therefore at most $\omega(1 - r/4)$, and that of $\|e_k\|^2$ is at most $1 - r/4$. Use $v_y \le v$ and the definitions of $\rho, N_{\mathrm{b}}, N_{\mathrm{v}}, D$ to obtain (67). □

A.5 LYAPUNOV DESCENT

Proposition A.8 (A uniform one-step inequality)**.** *Under Assumptions 2.1–2.3 and the step choices (37), for every $\tau > 0$ and positive integer m,*

$$\mathbb{E}_k \mathcal{V}_{k+1} \le \mathcal{V}_k - \frac{\alpha}{4}\|\nabla\Phi(x_k)\|^2 - \alpha C T_k + \alpha S_{\mathrm{b}}\beta^2 + \alpha S_{\mathrm{v}} v. \tag{68}$$

All constants are given in Appendix A.1; in particular, the step sizes do not depend on the target accuracy.

Proof. Add (64) to λ times (67), and subtract $\Phi(x_k) + \lambda T_k$. This gives

$$\begin{aligned}\mathbb{E}_k \mathcal{V}_{k+1} - \mathcal{V}_k \le & -\frac{\alpha}{2}\|h_k\|^2 + \left(\frac{\alpha C}{2} - \lambda\rho\right) T_k + \lambda D \mathbb{E}_k\|x_{k+1} - x_k\|^2 \\ & + \left(\frac{\alpha C_{\mathrm{a}}}{2} + \lambda N_{\mathrm{b}}\right)\beta^2 + \left(\frac{7\alpha}{8} + \lambda N_{\mathrm{v}}\right) v.\end{aligned}$$

Use the movement identity and (63) to bound the third term by

$$\lambda D \alpha^2 \left(2\|h_k\|^2 + 2CT_k + 2C_{\mathrm{a}}\beta^2 + \tfrac{5}{2}v\right).$$

For an explicit check of every coefficient, put $\chi = CD\alpha^2/\rho$. Our choice of α implies $0 < \chi \le 1/16$, and $\lambda = 2\alpha C/\rho$. Substitution yields

$$\begin{aligned}\mathbb{E}_k \mathcal{V}_{k+1} - \mathcal{V}_k \le & \, \alpha\left(-\tfrac{1}{2} + 4\chi\right)\|h_k\|^2 + \alpha C\left(-\tfrac{3}{2} + 4\chi\right) T_k \\ & + \alpha\left[C_{\mathrm{a}}\left(\tfrac{1}{2} + 4\chi\right) + \frac{2CN_{\mathrm{b}}}{\rho}\right]\beta^2 \\ & + \alpha\left[\tfrac{7}{8} + 5\chi + \frac{2CN_{\mathrm{v}}}{\rho}\right] v.\end{aligned}$$

The four numerical coefficients obey

$$-\tfrac{1}{2} + 4\chi \le -\tfrac{1}{4}, \quad -\tfrac{3}{2} + 4\chi \le -\tfrac{5}{4} \le -1, \quad \tfrac{1}{2} + 4\chi \le \tfrac{3}{4} \le 1, \quad \tfrac{7}{8} + 5\chi \le \tfrac{19}{16} \le 2.$$

The definitions of S_{b} and S_{v} now give (68). □

A.6 FINITE-TIME BOUND AND QUERY COMPLEXITY

Proof of Theorem 3.1. Take total expectations in (68), discard the nonpositive $-\alpha C T_k$ term, and sum from $k = 0$ to $K - 1$. Since $\mathcal{V}_K \ge 0$, this gives

$$\frac{\alpha}{4}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(x_k)\|^2 \le \mathcal{V}_0 + K\alpha S_{\mathrm{b}}\beta^2 + K\alpha S_{\mathrm{v}} v.$$

After division by $\alpha K/4$ and substitution of (52),

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla\Phi(x_k)\|^2 &\le \frac{4\mathcal{V}_0}{\alpha K} + 4S_{\mathrm{b}}(C_p L_p R_z)^2\tau^{2p} + \frac{4S_{\mathrm{v}}\sigma^2}{m}\left(1 + \frac{A_p R_z^2}{\tau^2}\right) \\ &\le \frac{4\mathcal{V}_0}{\alpha K} + C_{\mathrm{b}}\tau^{2p} + \frac{C_{\mathrm{v}}}{m\tau^2} \\ &\le C_\star\left(\frac{\Delta_0}{K} + \tau^{2p} + \frac{1}{m\tau^2}\right).\end{aligned}$$

The second inequality uses $\tau \le 1$, and the last follows from (42) and the definition of $C_\star$. The output rule gives

$$\mathbb{E}\big[\|\nabla\Phi(x_R)\|^2 \mid \mathcal{F}_K\big] = \frac{1}{K}\sum_{k=0}^{K-1}\|\nabla\Phi(x_k)\|^2.$$

Taking total expectations proves the finite-time output bound (18) for every $0 < \tau \le 1$ and all positive integers m, K.

Now use the parameters in (17). They satisfy $0 < \tau \le 1$ and

$$\frac{C_\star \Delta_0}{K} \le \frac{\Delta_0}{3(1+\Delta_0)}\epsilon^2 \le \frac{\epsilon^2}{3},$$
$$C_\star \tau^{2p} = \frac{\epsilon^2}{3},$$
$$\frac{C_\star}{m\tau^2} \le \frac{\epsilon^2}{3}.$$

The conditions $C_\star \ge 1$ and $1 + \Delta_0 > 0$ also cover zero variance, zero finite-difference bias, or zero initial gap. Adding the three bounds yields $\mathbb{E}\|\nabla\Phi(x_R)\|^2 \le \epsilon^2$.

It remains to count primitive queries. For fixed p, problem parameters, and Δ_0, the definitions give

$$\tau^{-2} = (3C_\star)^{1/p}\epsilon^{-2/p}, \qquad m = \mathcal{O}_p(\epsilon^{-2-2/p}), \qquad K = \mathcal{O}_p(\epsilon^{-2}).$$

Ceilings preserve these bounds for $0 < \epsilon \le 1$. Each of the two finite differences uses at most $(p+1)m$ calls, and the three center batches use $3m$ more. No samples are shared between these groups. The total number is therefore at most $(2p+5)mK = \mathcal{O}_p(\epsilon^{-4-2/p})$. At $p = 1$ and $p = 2$ the exponents are respectively 6 and 5. □

The constants are not optimized in the condition number or uniformly over p. Section 4 combines this guarantee with a matching lower bound for the same function class and oracle.

B Proof of Theorem 4.3

We construct (A, u) in Sections B.1–B.2 and prove the comparison-oracle bound in Sections B.3–B.4. Lemma B.16 then gives the ordinary first-order benchmark. Section B.5 verifies the bilevel construction, and Section B.6 combines its oracle simulation and gradient estimate with the same comparison bound.

For a scalar function h, $\|h\|_\infty := \sup_w |h(w)|$; for a vector v, $\|v\|_\infty := \max_j |v_j|$. Unsubscripted vector norms are Euclidean. The seed ξ_0 includes all private randomness of the algorithm and is independent of the instance and oracle randomness. The transcript $\mathcal{T}_N$ contains the first N queries and their responses. All conditional laws are taken on standard Borel spaces.

B.1 The chain and its decomposition

B.1.1 The Carmon–Arjevani chain

For $v \in \mathbb{R}^T$ and $\alpha \ge 0$, define

$$\mathrm{prog}_\alpha(v) := \max\big(\{0\} \cup \{j \in [T] : |v_j| > \alpha\}\big).$$

Let $v_0 := 1$, and define

$$\Psi(t) := \begin{cases} 0, & t \le \frac{1}{2}, \\ \exp\left(1 - \dfrac{1}{(2t-1)^2}\right), & t > \frac{1}{2}, \end{cases}$$

and

$$\varphi(t) := \sqrt{e}\int_{-\infty}^{t} e^{-s^2/2}\, ds.$$

For $i \in [T]$, put

$$\psi_i(v) := \Psi(-v_{i-1})\varphi(-v_i) - \Psi(v_{i-1})\varphi(v_i), \tag{69}$$

and

$$C_T(v) := \sum_{i=1}^{T} \psi_i(v). \tag{70}$$

The convention $v_0 = 1$ makes $\psi_1(v) = -\Psi(1)\varphi(v_1)$.

We use the following standard properties of this chain. The constants below are numerical and their precise values are immaterial for our scaling.

Lemma B.1 (Hard-chain properties; Carmon et al. (2020); Arjevani et al. (2023))**.** *There exist numerical constants $\Delta_c, L_0, G_\infty > 0$ such that, for every T:*

(i) $C_T(0) - \inf_v C_T(v) \le \Delta_c T$*;*

(ii) ∇C_T *is* L_0*-Lipschitz;*

(iii) $\|\nabla C_T(v)\|_\infty \le G_\infty$ *for all* v*;*

(iv) $\mathrm{prog}_0(\nabla C_T(v)) \le \mathrm{prog}_{1/2}(v) + 1$*;*

(v) if $k := \mathrm{prog}_1(v) < T$*, then*

$$|\partial_{k+1} C_T(v)| \ge 1, \tag{71}$$

and hence $\|\nabla C_T(v)\| \ge 1$.

One may take $\Delta_c = 12$*,* $L_0 = 152$*, and* $G_\infty = 23$*.*

B.1.2 A SMOOTH DECOMPOSITION FOR THE GRADIENT ORACLE

The decomposition $C_T = B_T - W_T$ separates the part of the gradient that may expose the next coordinate. We define its deterministic weights below; randomness enters through (77).

Fix once and for all a nondecreasing function $\Gamma \in C^\infty(\mathbb{R}; [0,1])$ such that

$$\Gamma(t) = 0 \quad (t \le 1/4), \qquad \Gamma(t) = 1 \quad (t \ge 1/2). \tag{72}$$

Set

$$\gamma(t) := \Gamma(|t|).$$

This is smooth because it vanishes on a neighborhood of the origin. For $i \in [T]$, define

$$r_i(v) := \left(\sum_{j=i}^{T} \gamma(v_j)^2 \right)^{1/2}, \qquad \Theta_i(v) := \Gamma(1 - r_i(v)). \tag{73}$$

The deterministic weight Θ_i equals one when the entire tail $v_{\ge i}$ is below the $1/4$ progress threshold, and it equals zero once that tail contains a coordinate of magnitude at least $1/2$.

Define the quiet base and frontier correction by

$$B_T(v) := \sum_{i=1}^{T} \psi_i(v)(1 - \Theta_i(v)), \tag{74}$$

$$W_T(v) := -\sum_{i=1}^{T} \psi_i(v)\Theta_i(v). \tag{75}$$

Then, identically,

$$C_T = B_T - W_T. \tag{76}$$

Equivalently, for any $\varpi \in (0,1]$, if $\zeta \sim \mathrm{Bernoulli}(\varpi)$, the random scalar function

$$h_T(v;\zeta) := B_T(v) - \frac{\zeta}{\varpi} W_T(v) \tag{77}$$

satisfies $\mathbb{E} h_T(v;\zeta) = C_T(v)$. This is the stochastic function construction of Arjevani et al. (2023), written as a deterministic base and a randomly retained correction.

For $k \in \{0, \dots, T\}$, let P_k denote the orthogonal projection onto $\mathrm{span}\{e_1, \dots, e_k\}$, with $P_0 = 0$ and $P_{T+1} = I_T$. For a C^2 scalar function h, write

$$\mathcal{J}_v^2 h := \big(h(v), \nabla h(v), \nabla^2 h(v)\big).$$

Lemma B.2 (Uniform frontier split)**.** *There exist numerical constants $C_{W,0}, C_{W,1}, C_{W,2}, L_B < \infty$, independent of T, such that:*

(a) Uniform correction bounds.

$$\|W_T\|_\infty \le C_{W,0}, \qquad \sup_v \|\nabla W_T(v)\| \le C_{W,1}, \qquad \sup_v \left\|\nabla^2 W_T(v)\right\|_{\mathrm{op}} \le C_{W,2}, \tag{78}$$

and

$$\sup_v \left\|\nabla^2 B_T(v)\right\|_{\mathrm{op}} \le L_B. \tag{79}$$

(b) Single-frontier jet identity. *If $r = \mathrm{prog}_{1/2}(v)$, then, with $\psi_{T+1} \equiv \Theta_{T+1} \equiv 0$,*

$$\mathcal{J}_v^2 W_T = -\mathcal{J}_v^2\big(\psi_{r+1}\Theta_{r+1}\big). \tag{80}$$

Thus the complete second-order jet of W_T is contributed by at most one chain summand.

(c) Quiet-base truncation. *If $k = \mathrm{prog}_{1/4}(v)$, then*

$$B_T(v) = B_T(P_k v), \tag{81}$$
$$\nabla B_T(v) = P_k \nabla B_T(P_k v), \tag{82}$$
$$\nabla^2 B_T(v) = P_k \nabla^2 B_T(P_k v) P_k. \tag{83}$$

In particular,

$$\operatorname{supp} \nabla B_T(v) \subseteq \{1, \dots, k\}. \tag{84}$$

(d) One-step correction truncation. *For the same k, the following identities hold; when $k = T$, they are understood with the convention $P_{T+1} = I_T$ and are tautological:*

$$W_T(v) = W_T(P_{k+1} v), \tag{85}$$
$$\nabla W_T(v) = P_{k+1} \nabla W_T(P_{k+1} v), \tag{86}$$
$$\nabla^2 W_T(v) = P_{k+1} \nabla^2 W_T(P_{k+1} v) P_{k+1}. \tag{87}$$

Consequently,

$$\operatorname{supp} \nabla W_T(v) \subseteq \{1, \dots, \min\{k+1, T\}\}. \tag{88}$$

(e) Removal of the next-coordinate term from the base. *If $k < T$, then*

$$\Theta_{k+1}(v) = 1, \qquad \nabla\Theta_{k+1}(v) = 0, \qquad \nabla^2\Theta_{k+1}(v) = 0, \tag{89}$$

and hence

$$\mathcal{J}_v^2\big(\psi_{k+1}(1 - \Theta_{k+1})\big) = 0. \tag{90}$$

Proof. We divide the proof into four steps.

Step 1: dimension-free derivative bounds for the weights. Let

$$C_{\Gamma,1} := \|\Gamma'\|_\infty, \qquad C_{\Gamma,2} := \|\Gamma''\|_\infty.$$

Fix i and set $w_j(v) := \gamma(v_j)$ for $j = i, \dots, T$. Write

$$w(v) := (w_i(v), \dots, w_T(v)), \qquad r(v) := \|w(v)\|_2.$$

Then $\Theta_i(v) = \Gamma(1 - r(v))$. The composition is smooth at $r = 0$ because it is constant whenever $r \le 1/2$. Its first and second derivatives vanish for $r \le 1/2$ and for $r \ge 3/4$; at the two endpoints this follows from the smooth flatness of Γ, rather than from a two-sided neighborhood of constancy. Therefore derivatives can be nonzero only on

$$\frac{1}{2} < r < \frac{3}{4}. \tag{91}$$

On this region,

$$\nabla r = Dw(v)^\top \frac{w(v)}{r}, \qquad \|\nabla r\| \le C_{\Gamma,1}.$$

Moreover,

$$\nabla^2 r = \frac{1}{r} Dw(v)^\top \left(I - \frac{w(v)w(v)^\top}{r^2} \right) Dw(v) + \sum_{j=i}^{T} \frac{w_j(v)}{r} \nabla^2 w_j(v).$$

The first term has operator norm at most $2{C_{\Gamma,1}}^2$ by (91). The second term is diagonal, and its operator norm is at most $C_{\Gamma,2}$. Hence

$$\|\nabla r\| \le C_{\Gamma,1}, \qquad \left\|\nabla^2 r\right\|_{\mathrm{op}} \le 2{C_{\Gamma,1}}^2 + C_{\Gamma,2} \tag{92}$$

uniformly in T and i on the only region where the derivatives of Θ_i can be nonzero. The chain rule gives numerical constants $C_{\Theta,1}, C_{\Theta,2}$ such that

$$\sup_{T,i,v} \|\nabla \Theta_i(v)\| \le C_{\Theta,1}, \qquad \sup_{T,i,v} \left\|\nabla^2 \Theta_i(v)\right\|_{\mathrm{op}} \le C_{\Theta,2}. \tag{93}$$

Step 2: only one product has a nonzero second-order jet. Fix v and let $r = \mathrm{prog}_{1/2}(v)$. If $i \le r$, the tail $v_{\ge i}$ contains v_r with $|v_r| > 1/2$. Thus $\gamma(v_r) = 1$, so $r_i(v) \ge 1$ and $1 - r_i(v) \le 0$. Since Γ is identically zero on $(-\infty, 1/4]$,

$$\mathcal{J}_v^2 \Theta_i = 0 \qquad (i \le r). \tag{94}$$

If $i \ge r + 2$, then $|v_{i-1}| \le 1/2$. Both v_{i-1} and $-v_{i-1}$ lie in the flat-zero region of Ψ, and therefore

$$\mathcal{J}_v^2 \psi_i = 0 \qquad (i \ge r + 2). \tag{95}$$

Combining (94) and (95), every summand of W_T has zero second-order jet except possibly the one indexed by $r + 1$. This proves (80).

Step 3: uniform bounds. Each ψ_i is a translate of a fixed two-coordinate smooth template. Hence there are numerical constants $C_{\psi,0}, C_{\psi,1}, C_{\psi,2}$ such that

$$|\psi_i(v)| \le C_{\psi,0}, \qquad \|\nabla \psi_i(v)\| \le C_{\psi,1}, \qquad \left\|\nabla^2 \psi_i(v)\right\|_{\mathrm{op}} \le C_{\psi,2}$$

uniformly in T, i, v. By the single-frontier identity,

$$|W_T(v)| \le C_{\psi,0},$$

$$\|\nabla W_T(v)\| \le C_{\psi,1} + C_{\psi,0} C_{\Theta,1},$$

and

$$\left\|\nabla^2 W_T(v)\right\|_{\mathrm{op}} \le C_{\psi,2} + 2C_{\psi,1} C_{\Theta,1} + C_{\psi,0} C_{\Theta,2}.$$

This proves (78). Since $B_T = C_T + W_T$ and Lemma B.1 gives a dimension-free Hessian bound for C_T, (79) follows.

Step 4: the truncation identities. Let $k = \mathrm{prog}_{1/4}(v)$. For every $j > k$, $|v_j| \le 1/4$, and flatness of Γ gives

$$\gamma(v_j) = \gamma'(v_j) = \gamma''(v_j) = 0. \tag{96}$$

For $i \le k$, the summand ψ_i uses only v_{i-1}, v_i, and all dependence of Θ_i on coordinates after k has zero second-order jet by (96). If $k < T$, then $r_{k+1}(v) = 0$, hence $\Theta_{k+1}(v) = \Gamma(1) = 1$ with zero first and second derivatives. This proves (89)–(90). For $i \ge k + 2$, the predecessor satisfies $|v_{i-1}| \le 1/4 < 1/2$, so $\mathcal{J}_v^2 \psi_i = 0$. If $k = T$, there are no indices after k and the truncation identities are immediate. Thus, in every case, the value, gradient, and Hessian of B_T are unchanged when all coordinates after k are set to zero, proving (81)–(83).

For W_T, if $k < T$, the only potentially active product has index $r + 1$, where $r = \mathrm{prog}_{1/2}(v) \le k$, and its chain summand depends on at most coordinates $r, r + 1$. The projection P_{k+1} deletes only coordinates $j > k + 1$, all of which satisfy (96); in particular it leaves the frontier index $r = \mathrm{prog}_{1/2}(v)$ unchanged. On every deleted coordinate, γ, γ', and γ'' vanish, so evaluating at $P_{k+1} v$ selects the same frontier product and preserves its complete second-order jet. When $k = T$, the claimed identities are tautological under $P_{T+1} = I_T$. This proves (85)–(87). □

Corollary B.3 (Two-vector probability-ϖ zero-chain)**.** *For any $\varpi \in (0,1]$, let $\zeta \sim \mathrm{Bernoulli}(\varpi)$ and define*

$$a^{(1)}(v,\zeta) := \nabla B_T(v), \qquad a^{(2)}(v,\zeta) := \frac{\zeta}{\varpi}\nabla W_T(v).$$

Then, simultaneously for every v,

$$\max_{\ell\in\{1,2\}} \mathrm{prog}_0(a^{(\ell)}(v,\zeta)) \le \mathrm{prog}_{1/4}(v) + 1,$$

and on the event $\{\zeta = 0\}$,

$$\max_{\ell\in\{1,2\}} \mathrm{prog}_0(a^{(\ell)}(v,\zeta)) \le \mathrm{prog}_{1/4}(v).$$

The exceptional event is independent of v and has probability ϖ.

B.2 Construction and scaling of the smooth pair

For $R > 0$, define the soft projection

$$\rho_R(w) := \frac{w}{\sqrt{1 + \|w\|^2/R^2}}, \tag{97}$$

and let $J_R(w) = D\rho_R(w)$. We use the standard choices

$$R := 230\sqrt{T}, \qquad \eta_{\mathrm{r}} := \frac{1}{5}. \tag{98}$$

Let $\mathrm{Ortho}(d,T) := \{U \in \mathbb{R}^{d\times T} : U^\top U = I_T\}$. For $U \in \mathrm{Ortho}(d,T)$, define

$$C_U^{\mathrm{c}}(w) := C_T(U^\top \rho_R(w)) + \frac{\eta_{\mathrm{r}}}{2}\|w\|^2, \tag{99}$$

$$B_U^{\mathrm{c}}(w) := B_T(U^\top \rho_R(w)) + \frac{\eta_{\mathrm{r}}}{2}\|w\|^2, \tag{100}$$

$$W_U^{\mathrm{c}}(w) := W_T(U^\top \rho_R(w)). \tag{101}$$

Then

$$C_U^{\mathrm{c}} = B_U^{\mathrm{c}} - W_U^{\mathrm{c}}. \tag{102}$$

Lemma B.4 (Compressed hard-chain properties)**.** *There is a numerical constant L_{hc} such that, uniformly over T, d, U:*

(i) $C_U^{\mathrm{c}}(0) - \inf_w C_U^{\mathrm{c}}(w) \le \Delta_{\mathrm{c}} T$*;*

(ii) ∇C_U^{c} *is L_{hc}-Lipschitz;*

(iii) if

$$\mathrm{prog}_{1/4}(U^\top \rho_R(w)) < T, \tag{103}$$

then

$$\|\nabla C_U^{\mathrm{c}}(w)\| \ge \frac{1}{2}; \tag{104}$$

(iv) $\|J_R(w)\|_{\mathrm{op}} \le 1$ *and*

$$\|J_R(w) - J_R(w')\|_{\mathrm{op}} \le \frac{3}{R}\|w - w'\|. \tag{105}$$

One may take $L_{\mathrm{hc}} = 155$.

Proof. The compression estimates in Lemma 15 of Arjevani et al. (2023) give (iv) and show that ρ_R is 1-Lipschitz. The nonnegative quadratic term vanishes at zero, so (i) follows from Lemma B.1. For (ii), the chain rule gives the gradient Lipschitz bound

$$L_0 + \frac{3G_\infty\sqrt{T}}{R} + \eta_{\mathrm{r}} = 152 + \frac{3}{10} + \frac{1}{5} < 155.$$

It remains to prove (iii). Put

$$v := \rho_R(w), \qquad h := U^\top v, \qquad v_g := U\nabla C_T(h), \qquad c := \sqrt{1 + \|w\|^2 / R^2}.$$

Then $w = cv$, $\|v_g\| \le G_\infty \sqrt{T} = 23\sqrt{T}$, and

$$\nabla C_U^{\mathrm{c}}(w) = \frac{I - vv^\top/R^2}{c}\, v_g + \eta_{\mathrm{r}} w. \tag{106}$$

First suppose $\|w\| \le R/2$. Condition (103) implies $k := \mathrm{prog}_1(h) < T$. Set $j := k + 1$. Then $|h_j| \le 1$, while Lemma B.1 gives $|\langle u_j, v_g\rangle| = |\partial_j C_T(h)| \ge 1$. Since $c \le \sqrt{5}/2$ and $\|v\| \le R/2$, equation (106) yields

$$\begin{aligned}
\|\nabla C_U^{\mathrm{c}}(w)\| &\ge |\langle u_j, \nabla C_U^{\mathrm{c}}(w)\rangle| \\
&\ge \frac{1}{c} - \frac{|h_j|\,|\langle v, v_g\rangle|}{R^2 c} - \eta_{\mathrm{r}} c |h_j| \\
&\ge \frac{2}{\sqrt{5}} - \frac{23\sqrt{T}}{2R} - \frac{\eta_{\mathrm{r}}\sqrt{5}}{2} \\
&= \frac{2}{\sqrt{5}} - \frac{1}{20} - \frac{\sqrt{5}}{10} > \frac{1}{2},
\end{aligned}$$

where the equality uses $R = 230\sqrt{T}$ and $\eta_{\mathrm{r}} = 1/5$.

Now suppose $\|w\| > R/2$. Lemma 15 of Arjevani et al. (2023) gives $\|J_R(w)\|_{\mathrm{op}} \le c^{-1} \le 2/\sqrt{5}$, so

$$\begin{aligned}
\|\nabla C_U^{\mathrm{c}}(w)\| &\ge \eta_{\mathrm{r}} \|w\| - \|J_R(w)\|_{\mathrm{op}} \|v_g\| \\
&> \frac{R}{10} - \frac{46}{\sqrt{5}}\sqrt{T} \\
&= \left(23 - \frac{46}{\sqrt{5}}\right)\sqrt{T} > \frac{1}{2}.
\end{aligned}$$

This proves (104). The numerical smoothness constant $L_{\mathrm{hc}} = 155$ is the corresponding bound obtained from the same compression estimates. □

Lemma B.5 (Compressed correction bounds)**.** *There are numerical constants $C_{W,0}^{\mathrm{c}}, C_{W,1}^{\mathrm{c}}, C_{W,2}^{\mathrm{c}}$ such that, uniformly in T, d, U,*

$$\|W_U^{\mathrm{c}}\|_\infty \le C_{W,0}^{\mathrm{c}}, \qquad \sup_w \|\nabla W_U^{\mathrm{c}}(w)\| \le C_{W,1}^{\mathrm{c}}, \tag{107}$$

$$\sup_w \left\|\nabla^2 W_U^{\mathrm{c}}(w)\right\|_{\mathrm{op}} \le C_{W,2}^{\mathrm{c}}. \tag{108}$$

Consequently, ∇B_U^{c} is $(L_{\mathrm{hc}} + C_{W,2}^{\mathrm{c}})$-Lipschitz.

Proof. The value bound follows from Lemma B.2. By the chain rule,

$$\nabla W_U^{\mathrm{c}}(w) = J_R(w)^\top U \nabla W_T(U^\top \rho_R(w)),$$

so the gradient bound follows from $\|J_R(w)\|_{\mathrm{op}} \le 1$. For w, w', use Lemma B.2, the 1-Lipschitzness of ρ_R, and (105) to obtain

$$\begin{aligned}
&\|\nabla W_U^{\mathrm{c}}(w) - \nabla W_U^{\mathrm{c}}(w')\| \\
&\quad \le \frac{3C_{W,1}}{R}\|w - w'\| + C_{W,2}\|w - w'\|.
\end{aligned}$$

Since $R \ge 1$ for $T \ge 1$, this gives a numerical Hessian bound. Finally, $B_U^{\mathrm{c}} = C_U^{\mathrm{c}} + W_U^{\mathrm{c}}$. □

Set

$$\ell_\star := L_{\mathrm{hc}} + C_{W,2}^{\mathrm{c}} + 1. \tag{109}$$

For sufficiently small ϵ, choose

$$T := \lfloor c_T \epsilon^{-2} \rfloor, \qquad c_T := \frac{1}{64 \Delta_{\mathrm{c}} \ell_\star}, \tag{110}$$

so $T \geq 4$. Define

$$\lambda_{\mathrm{lb}} := 4\ell_\star \epsilon, \qquad a := \frac{{\lambda_{\mathrm{lb}}}^2}{\ell_\star} = 16 \ell_\star \epsilon^2. \tag{111}$$

Define the base, correction, and nonconvex objective by

$$A_U(x) := a B^{\mathrm{c}}_U(x/\lambda_{\mathrm{lb}}), \tag{112}$$
$$u_U(x) := a W^{\mathrm{c}}_U(x/\lambda_{\mathrm{lb}}), \tag{113}$$
$$P_U(x) := A_U(x) - u_U(x) = a C^{\mathrm{c}}_U(x/\lambda_{\mathrm{lb}}). \tag{114}$$

Proposition B.6 (Properties of the scaled pair)**.** *There are numerical constants $C_{u,0}, C_{u,1}, C_{u,2}$ such that, uniformly in U, T, ϵ,*

$$\|u_U\|_\infty \leq C_{u,0} \epsilon^2, \qquad \sup_x \|\nabla u_U(x)\| \leq C_{u,1} \epsilon, \qquad \sup_x \left\|\nabla^2 u_U(x)\right\|_{\mathrm{op}} \leq C_{u,2}. \tag{115}$$

Moreover,

$$\sup_x \left\|\nabla^2 A_U(x)\right\|_{\mathrm{op}} \leq 1, \tag{116}$$

$$P_U(0) - \inf_x P_U(x) \leq \frac{1}{4}, \tag{117}$$

and

$$\mathrm{prog}_{1/4}\big(U^\top \rho_R(x/\lambda_{\mathrm{lb}})\big) < T \quad \Longrightarrow \quad \|\nabla P_U(x)\| \geq 2\epsilon. \tag{118}$$

Proof. The three correction scales follow from Lemma B.5 and

$$a = 16 \ell_\star \epsilon^2, \qquad \frac{a}{\lambda_{\mathrm{lb}}} = 4\epsilon, \qquad \frac{a}{{\lambda_{\mathrm{lb}}}^2} = \frac{1}{\ell_\star}.$$

The same identities and the definition of $\ell_\star$ give (116). By Lemma B.4 and (110),

$$P_U(0) - \inf P_U \leq a \Delta_{\mathrm{c}} T \leq 16 \ell_\star \epsilon^2 \Delta_{\mathrm{c}} c_T \epsilon^{-2} = \frac{1}{4}.$$

Finally,

$$\nabla P_U(x) = \frac{a}{\lambda_{\mathrm{lb}}} \nabla C^{\mathrm{c}}_U(x/\lambda_{\mathrm{lb}}) = 4\epsilon \nabla C^{\mathrm{c}}_U(x/\lambda_{\mathrm{lb}}),$$

and (118) follows from Lemma B.4. □

B.3 Finite-channel random rotation

The comparison oracle returns two ambient vectors and a scalar. We prove a random-rotation bound that permits these separate channels, including the scalar information, and applies to arbitrary adaptive queries and final outputs.

Definition B.7 (Uniform finite-channel probability-ϖ zero-chain)**.** Fix $\varpi \in [0,1]$. Let $\mathcal{Y}, \Xi, \mathcal{S}$ be standard Borel spaces, let $\xi \sim \nu$, and let

$$s : \mathbb{R}^T \times \mathcal{Y} \times \Xi \to \mathcal{S},$$
$$a^{(\ell)} : \mathbb{R}^T \times \mathcal{Y} \times \Xi \to \mathbb{R}^T, \qquad \ell \in [m],$$

be Borel maps. We call $(s, a^{(1)}, \ldots, a^{(m)})$ a uniform finite-channel probability-ϖ zero-chain if there is a measurable event $E \subseteq \Xi$ with $\nu(E) \leq \varpi$ such that, for every (v, y, ξ),

$$\max_{\ell \in [m]} \mathrm{prog}_0(a^{(\ell)}(v, y, \xi)) \leq \mathrm{prog}_{1/4}(v) + 1, \tag{119}$$

and, whenever $\xi \notin E$,

$$\max_{\ell \in [m]} \mathrm{prog}_0(a^{(\ell)}(v, y, \xi)) \leq \mathrm{prog}_{1/4}(v). \tag{120}$$

The map s collects all non-ambient channels. It may depend on the hidden rotation only through the latent query v.

For $U \in \mathrm{Ortho}(d, T)$, define the canonical rotated response

$$\widetilde{\mathcal{O}}_U(x, y; \xi) := \left(s(U^\top x, y, \xi), Ua^{(1)}(U^\top x, y, \xi), \ldots, Ua^{(m)}(U^\top x, y, \xi)\right). \tag{121}$$

Lemma B.8 (Projection accumulation; cf. Lemma 12 of Arjevani et al. (2023))**.** *Fix integers $n, T \geq 1$, orthonormal vectors $u_1, \ldots, u_T \in \mathbb{R}^d$, and vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ with $\|x_i\| \leq R$. For $i \in \{0, \ldots, n\}$ and $j \in \{0, \ldots, T\}$, let*

$$\mathsf{S}_j^{(i)} := \mathrm{span}\{u_1, \ldots, u_j, x_1, \ldots, x_i\},$$

let $P_j^{(i)}$ be the orthogonal projection onto $(\mathsf{S}_j^{(i)})^\perp$, and put $\Pi_j^{(i)} := I - P_j^{(i)}$. For $i \in [n]$ and $j \in [T]$, define

$$\mathcal{G}_j^{(i)} := \left\{ \left\| \Pi_{j-1}^{(i)} P_{j-1}^{(i-1)} u_j \right\| < \frac{1}{4R\sqrt{n}} \right\}. \tag{122}$$

Then, for every $i \in [n]$ and $j \in [T]$,

$$\bigcap_{q=1}^{i} \mathcal{G}_j^{(q)} \quad \Longrightarrow \quad |\langle u_j, x_i \rangle| \leq \frac{\|x_i\|}{4R} \sqrt{\frac{i}{n}} \leq \frac{1}{4}. \tag{123}$$

Proof. For fixed j, the spaces $\mathsf{S}_{j-1}^{(q)}$ are nested in q. Consequently,

$$\Pi_{j-1}^{(q-1)} + P_{j-1}^{(q-1)} \Pi_{j-1}^{(q)} P_{j-1}^{(q-1)} = \Pi_{j-1}^{(q)}.$$

Iterating this identity gives

$$\Pi_{j-1}^{(i)} = \sum_{r=1}^{j-1} u_r u_r^\top + \sum_{q=1}^{i} P_{j-1}^{(q-1)} \Pi_{j-1}^{(q)} P_{j-1}^{(q-1)}. \tag{124}$$

Since $x_i \in \mathsf{S}_{j-1}^{(i)}$ and $u_j \perp u_r$ for $r < j$, (124) and Cauchy–Schwarz yield

$$\begin{aligned} |\langle u_j, x_i \rangle| \leq & \left(\sum_{q=1}^{i} \left\| \Pi_{j-1}^{(q)} P_{j-1}^{(q-1)} u_j \right\|^2 \right)^{1/2} \\ & \times \left(\sum_{q=1}^{i} \left\| \Pi_{j-1}^{(q)} P_{j-1}^{(q-1)} x_i \right\|^2 \right)^{1/2}. \end{aligned} \tag{125}$$

On $\cap_{q=1}^{i} \mathcal{G}_j^{(q)}$, the first squared factor is less than $i/(16R^2 n)$. Moreover, subtracting the first term on the right of (124) shows that

$$\sum_{q=1}^{i} P_{j-1}^{(q-1)} \Pi_{j-1}^{(q)} P_{j-1}^{(q-1)} \preceq I.$$

The second squared factor in (125) is therefore at most $\|x_i\|^2$. This proves (123). □

For the remaining geometric lemmas, fix an adaptive algorithm and a horizon n. Assume that U has the uniform invariant Stiefel distribution on $\mathrm{Ortho}(d, T)$, independently of the algorithm seed and the i.i.d. oracle seeds, and write $U = [u_1, \ldots, u_T]$ and

$$v_i := U^\top x_i, \qquad a_i^{(\ell)} := a^{(\ell)}(v_i, y_i, \xi_i),$$

$$\gamma_i := \max_{\substack{1 \leq q \leq i \\ 1 \leq \ell \leq m}} \mathrm{prog}_0(a_q^{(\ell)}), \qquad \gamma_0 := 0.$$

Thus γ_i is the largest hidden coordinate index appearing in a canonical vector response by call i. It is a proof bookkeeping quantity; the algorithm need not identify the individual hidden basis vectors or observe γ_i. For $i \in \{0, \ldots, n\}$ and $j \in \{0, \ldots, T\}$, define

$$\mathcal{U}_j^{(i)} := \sigma(\xi_0, u_1, \ldots, u_j, v_1, \xi_1, \ldots, v_i, \xi_i), \tag{126}$$

where the list after u_j is empty for $i = 0$, and define

$$\mathcal{T}_j^{(i)} := \{j > \gamma_i\}. \tag{127}$$

Crucially, $\mathcal{U}_{j-1}^{(i-1)}$ contains only the *past* latent queries $v_1, \dots, v_{i-1}$; it does not contain the current latent query v_i. All conditional laws below are regular conditional laws, whose existence follows from the standard-Borel hypotheses.

Lemma B.9 (Invariant posterior under an invariant observation)**.** *Let $\mathsf{C}, \mathsf{X}, \mathsf{Z}$ be standard Borel spaces, let G be a compact second-countable group with normalized Haar probability λ_G, and let*

$$\tau : \mathsf{C} \times G \times \mathsf{X} \to \mathsf{X}$$

be a Borel map such that $g \mapsto \tau_{c,g}$ is a group action for every c. Let $K(c, dx)$ be a Borel probability kernel satisfying

$$K(c, \tau_{c,g}^{-1} A) = K(c, A) \qquad (c \in \mathsf{C},\ g \in G) \tag{128}$$

for every Borel $A \subseteq \mathsf{X}$. Suppose that the Borel observation $H : \mathsf{C} \times \mathsf{X} \to \mathsf{Z}$ is invariant:

$$H(c, \tau_{c,g} x) = H(c, x). \tag{129}$$

Then there is a Borel regular conditional kernel $L(c, v, dx)$ for x given $v = H(c, x)$ under $K(c, dx)$ such that

$$L(c, v, \tau_{c,g}^{-1} A) = L(c, v, A) \tag{130}$$

for every $g \in G$ and for $H(c, \cdot)_{\#} K(c)$-almost every v.

Proof. Let $L_0(c, v, dx)$ be any Borel disintegration of $K(c, dx)$ over the map $H(c, \cdot)$, and define

$$L(c, v, A) := \int_G L_0(c, v, \tau_{c,g}^{-1} A)\, \lambda_G(dg). \tag{131}$$

This is a Borel kernel. If $M(c, dv) := H(c, \cdot)_{\#} K(c)$, then for Borel $A \subseteq \mathsf{X}$ and $B \subseteq \mathsf{Z}$,

$$\begin{aligned} \int_B L(c, v, A) M(c, dv) &= \int_G K\big(c, \tau_{c,g}^{-1} A \cap H(c, \cdot)^{-1} B\big)\, \lambda_G(dg) \\ &= \int_G K\big(c, \tau_{c,g}^{-1} (A \cap H(c, \cdot)^{-1} B)\big)\, \lambda_G(dg) \\ &= K\big(c, A \cap H(c, \cdot)^{-1} B\big), \end{aligned}$$

where the second line uses (129) and the third uses (128). Hence L is a regular conditional kernel. Equation (130) follows directly from (131) and left/right invariance of Haar probability on the compact group G. □

Lemma B.10 (Sequential posterior-stabilizer invariance)**.** *Let $U = [u_1, \dots, u_T]$ have the uniform invariant Stiefel distribution on $\mathrm{Ortho}(d, T)$, and let ϑ be an independent random element of a standard Borel space. Fix $j \in [T]$ and an integer $r \geq 0$. Suppose that Borel maps χ_q define recursively*

$$x_q = \chi_q(\vartheta, u_1, \dots, u_{j-1}, v_1, \dots, v_{q-1}), \qquad v_q = U^\top x_q, \qquad q \in [r]. \tag{132}$$

Let

$$C_q := (\vartheta, u_1, \dots, u_{j-1}, v_1, \dots, v_q),$$

with $C_0 = (\vartheta, u_1, \dots, u_{j-1})$, and set

$$\mathsf{S}_q := \mathrm{span}\{u_1, \dots, u_{j-1}, x_1, \dots, x_q\}.$$

Regular conditional kernels $K_q(C_q, dU)$ of U given C_q may be chosen so that, for almost every $C_q = c_q$, the kernel is invariant under every orthogonal map fixing $\mathsf{S}_q(c_q)$ pointwise:

$$K_q(c_q, W^{-1} A) = K_q(c_q, A) \quad \textit{whenever } Ws = s \textit{ for all } s \in \mathsf{S}_q(c_q). \tag{133}$$

Proof. We construct the kernels inductively. Conditional on $u_1, \dots, u_{j-1}$, the remaining columns of a uniform Stiefel frame have the standard invariant-completion law. Explicitly, after choosing a Borel orthonormal frame $Q_\perp(u_{<j})$ for $\mathrm{span}(u_1, \dots, u_{j-1})^\perp$, this law is the image of the uniform invariant probability measure on the Stiefel manifold $\mathrm{Ortho}(d-j+1, T-j+1)$ under

$$V \longmapsto [u_1, \dots, u_{j-1}, Q_\perp(u_{<j})V].$$

It is invariant under every orthogonal map fixing $\mathsf{S}_0 = \mathrm{span}\{u_1, \dots, u_{j-1}\}$ pointwise. Independence of ϑ gives the required kernel K_0.

Assume that K_{q-1} has been chosen. Let $\mathsf{C}_{q-1}^{\mathrm{good}}$ be a Borel full-measure subset of the conditioning space on which the induction hypothesis holds. We first restrict the conditioning parameter and the kernel to this good set; only after constructing the next posterior there will we extend it arbitrarily over the null complement. Once $C_{q-1} = c_{q-1} \in \mathsf{C}_{q-1}^{\mathrm{good}}$ is fixed, the vector x_q in (132) is fixed. Put

$$\mathsf{S}_q = \mathrm{span}(\mathsf{S}_{q-1}, x_q).$$

On each stratum on which $\dim \mathsf{S}_q = s$, choose a Borel orthogonal completion $Q_\perp(c_{q-1})$ of S_q. For $R \in O(d-s)$ define

$$W_{c_{q-1},R} := P_{\mathsf{S}_q} + Q_\perp(c_{q-1}) R Q_\perp(c_{q-1})^\top.$$

This is a Borel action of the fixed compact group $O(d-s)$ on the conditional Stiefel fiber. Because it fixes S_q pointwise, it belongs to the stabilizer from the induction hypothesis. Moreover, the new observation

$$H_q(c_{q-1}, U) := U^\top x_q$$

is invariant under this action:

$$(W_{c_{q-1},R} U)^\top x_q = U^\top W_{c_{q-1},R}^\top x_q = U^\top x_q.$$

Apply Lemma B.9, separately on each Borel rank stratum of $\mathsf{C}_{q-1}^{\mathrm{good}}$, to disintegrate K_{q-1} over H_q. The finitely many resulting kernels paste into a Borel kernel on the good set; extend that kernel arbitrarily over the null complement. The resulting K_q is a regular conditional law given C_q and, for almost every C_q, is invariant under the full pointwise stabilizer of S_q. This closes the induction. Every orthogonal transformation used in this argument is fixed only after C_{q-1} has been conditioned upon; no state-dependent global change of variables is invoked. □

Lemma B.11 (Conditional residual spherical law; cf. Lemma 14 of Arjevani et al. (2023))**.** *Fix $i \in [n]$ and $j \in [T]$. On the event $\mathcal{T}_j^{(i-1)}$, the past transcript and the current query (x_i, y_i) are measurable from $\mathcal{U}_{j-1}^{(i-1)}$. In particular, $\mathcal{T}_j^{(i-1)}$, the space*

$$\mathsf{S} := \mathrm{span}\{u_1, \dots, u_{j-1}, x_1, \dots, x_{i-1}\},$$

and its orthogonal projector are measurable under this conditioning. There is an $\mathcal{U}_{j-1}^{(i-1)}$-measurable radius $r_{i,j} \in [0,1]$ such that, outside a null event and on $\mathcal{T}_j^{(i-1)}$,

$$\mathcal{L}\Big(P_{\mathsf{S}^\perp} u_j \,\Big|\, \mathcal{U}_{j-1}^{(i-1)}\Big) = \begin{cases} \delta_0, & r_{i,j} = 0, \\ \mathrm{Unif}\big(\{v \in \mathsf{S}^\perp : \|v\| = r_{i,j}\}\big), & r_{i,j} > 0. \end{cases} \tag{134}$$

Moreover, $r_{i,j} = \|P_{\mathsf{S}^\perp} u_j\|$ almost surely. Thus normalization is used only on $\{r_{i,j} > 0\}$.

Proof. We first construct globally Borel quiet-simulated queries. For fixed $u_{<j} := (u_1, \dots, u_{j-1})$, define the truncated canonical response

$$\overline{\mathcal{O}}_{<j}(v, y; \xi, u_{<j}) := \Big(s(v, y, \xi), \sum_{r<j} [a^{(1)}(v, y, \xi)]_r u_r, \dots, \sum_{r<j} [a^{(m)}(v, y, \xi)]_r u_r \Big). \tag{135}$$

Feed these responses recursively to the algorithm. Since all algorithm and oracle maps are Borel, there are Borel maps χ_q such that, for $\vartheta := (\xi_0, \xi_1, \dots, \xi_{i-1})$,

$$(\bar{x}_q, \bar{y}_q) = \chi_q(\vartheta, u_{<j}, \bar{v}_1, \dots, \bar{v}_{q-1}), \qquad \bar{v}_q := U^\top \bar{x}_q, \qquad q \le i. \tag{136}$$

For $q < i$, let $\bar{\gamma}_q$ be the progress computed from the corresponding latent coefficients $a^{(\ell)}(\bar{v}_q, \bar{y}_q, \xi_q)$.

A first-disagreement induction gives

$$\{j > \gamma_{i-1}\} = \{j > \bar{\gamma}_{i-1}\}, \qquad (x_q, y_q, v_q) = (\bar{x}_q, \bar{y}_q, \bar{v}_q) \quad (q \le i) \tag{137}$$

on this common event. For the first implication, suppose $j > \gamma_{i-1}$. Then every actual latent coefficient before time i is supported in $\{1, \dots, j-1\}$. Starting from the common first query, the actual ambient response therefore equals (135); induction gives equality of the two trajectories and hence $\bar{\gamma}_{i-1} = \gamma_{i-1} < j$. The reverse implication is identical with the roles of the actual and quiet trajectories interchanged.

For completeness, measurability from the actual conditioning tuple can be checked without referring to the random quiet trajectory. Given $(\xi_0, u_{<j}, v_1, \xi_1, \dots, v_{i-1}, \xi_{i-1})$, recursively compute the query, the non-ambient channel, and the latent coefficients at each past round. If a coefficient first has support reaching j, declare $\mathcal{T}_j^{(i-1)}$ false; while all coefficients remain supported below j, every ambient response is computable from $u_{<j}$ and the recursion continues. The declaration is correct by induction and, on the true event, also reconstructs (x_i, y_i). Hence $\mathcal{T}_j^{(i-1)}$ and the current query are $\mathcal{U}_{j-1}^{(i-1)}$-measurable.

Apply Lemma B.10 to the quiet recursion (136), with $r = i-1$. It yields a regular conditional kernel of U given

$$(\vartheta, u_1, \dots, u_{j-1}, \bar{v}_1, \dots, \bar{v}_{i-1}).$$

Transport this kernel through the fixed Borel coordinate reordering and write the conditioning tuple, in exactly the order used in (126), as

$$\bar{C} := (\xi_0, u_1, \dots, u_{j-1}, \bar{v}_1, \xi_1, \dots, \bar{v}_{i-1}, \xi_{i-1}).$$

Denote the transported kernel by $\bar{K}(\bar{C}, dU)$. It is invariant under every orthogonal map fixing

$$\bar{\mathsf{S}} := \operatorname{span}\{u_1, \dots, u_{j-1}, \bar{x}_1, \dots, \bar{x}_{i-1}\}$$

pointwise. Let

$$C := (\xi_0, u_1, \dots, u_{j-1}, v_1, \xi_1, \dots, v_{i-1}, \xi_{i-1}),$$

the tuple generating $\mathcal{U}_{j-1}^{(i-1)}$. By (137), on $\mathcal{T}_j^{(i-1)}$ we now have the literal equality $\bar{C} = C$, and $\bar{\mathsf{S}} = \mathsf{S}$. Write $\bar{K}\phi(c) := \int \phi(U)\bar{K}(c, dU)$. Because the events and the conditioning tuples agree in (137), for every bounded Borel h and ϕ,

$$\begin{aligned}
\mathbb{E}\Big[\mathbf{1}_{\mathcal{T}_j^{(i-1)}} h(C)\phi(U)\Big] &= \mathbb{E}\big[\mathbf{1}_{\{j > \bar{\gamma}_{i-1}\}} h(\bar{C})\phi(U)\big] \\
&= \mathbb{E}\big[\mathbf{1}_{\{j > \bar{\gamma}_{i-1}\}} h(\bar{C})\bar{K}\phi(\bar{C})\big] \\
&= \mathbb{E}\Big[\mathbf{1}_{\mathcal{T}_j^{(i-1)}} h(C)\bar{K}\phi(C)\Big].
\end{aligned}$$

Thus $\bar{K}(C, \cdot)$ is a version of the actual conditional law on $\mathcal{T}_j^{(i-1)}$. We have therefore obtained, by a sequential posterior-kernel induction, the conditional invariance of U under the pointwise stabilizer of $\bar{\mathsf{S}}$. In particular, the conditional law of u_j is invariant under that stabilizer.

It remains to identify its orbit. The conditioning fixes all inner products of u_j with the spanning vectors of $\bar{\mathsf{S}}$: they are zero against $u_1, \dots, u_{j-1}$, and

$$\langle u_j, x_q \rangle = [v_q]_j, \qquad q < i.$$

These values determine $b := P_{\bar{\mathsf{S}}} u_j$ uniquely, even when the spanning vectors are linearly dependent. Since $\|u_j\| = 1$, they also determine the Borel radius

$$r_{i,j} := \sqrt{1 - \|b\|^2} = \|P_{\bar{\mathsf{S}}^\perp} u_j\|.$$

The residual conditional law is supported on the radius-$r_{i,j}$ sphere in $\bar{\mathsf{S}}^\perp$ and is invariant under its full orthogonal group. For $r_{i,j} > 0$, uniqueness of rotation-invariant spherical probability gives the second case of (134); for $r_{i,j} = 0$, the sphere is the singleton $\{0\}$ and the law is δ_0. □

Lemma B.12 (One-pair guessing estimate; cf. Lemma 13 of Arjevani et al. (2023))**.** *Use the projections and events $\mathcal{G}_j^{(i)}$ from Lemma B.8, with the random query sequence $x_1, \dots, x_n$. If $d > n + T$, then, for every $i \in [n]$ and $j \in [T]$,*

$$\mathbb{P}\left((\mathcal{G}_j^{(i)})^c, \mathcal{T}_j^{(i-1)}\right) \le 2\exp\left(-\frac{d-i-j}{32R^2n}\right) \le 2\exp\left(-\frac{d-n-T}{32R^2n}\right). \tag{138}$$

Proof. Work on $\mathcal{T}_j^{(i-1)}$ and condition on $\mathcal{U}_{j-1}^{(i-1)}$. Put

$$\mathsf{S} = \operatorname{span}\{u_1, \dots, u_{j-1}, x_1, \dots, x_{i-1}\}, \qquad P = P_{\mathsf{S}^\perp} = P_{j-1}^{(i-1)},$$

and let $r = r_{i,j}$ be the conditional radius from Lemma B.11. If $r = 0$, then $Pu_j = 0$ and hence

$$\Pi_{j-1}^{(i)} P_{j-1}^{(i-1)} u_j = 0;$$

the conditional probability of the bad event is therefore zero.

Suppose now that $r > 0$ and define $\widehat{u}_j := Pu_j/r$. By Lemma B.11, $\widehat{u}_j$ is uniform on the unit sphere of $\mathsf{S}^\perp$. The current query x_i is measurable under the conditioning, and the restriction of $\Pi_{j-1}^{(i)}$ to $\mathsf{S}^\perp$ has rank at most one: it is the projection onto the residual direction of x_i modulo S. Since $r \le 1$,

$$\begin{aligned}\mathbb{P}\left(\left\|\Pi_{j-1}^{(i)} P_{j-1}^{(i-1)} u_j\right\| \ge \frac{1}{4R\sqrt{n}} \,\middle|\, \mathcal{U}_{j-1}^{(i-1)}\right) \\ \le \mathbb{P}\left(\left\|\Pi_{j-1}^{(i)} \widehat{u}_j\right\| \ge \frac{1}{4R\sqrt{n}} \,\middle|\, \mathcal{U}_{j-1}^{(i-1)}\right).\end{aligned}$$

Let $d' := \dim(\mathsf{S}^\perp)$. Since $\dim \mathsf{S} \le (i-1) + (j-1)$, we may use the conservative lower bound $d' \ge d - i - j$. For a uniform vector v on $\mathbb{S}^{d'-1}$, the standard spherical-coordinate estimate

$$\mathbb{P}(v_1^2 \ge \alpha) \le 2e^{-\alpha d'/2}$$

gives the first inequality in (138) with $\alpha = (16R^2n)^{-1}$. Averaging over the conditioning, including the $r = 0$ fibers, proves the claim. □

Lemma B.13 (Finite-channel random rotation)**.** *Let U have the uniform invariant Stiefel distribution on $\mathrm{Ortho}(d, T)$, independently of i.i.d. seeds $\xi_1, \xi_2, \dots$ and the private randomness of an adaptive randomized algorithm. Suppose the algorithm makes queries (x_t, y_t) satisfying $\|x_t\| \le R$ almost surely and observes the canonical responses (121). Put*

$$\pi_N := \max_{1 \le t \le N} \mathrm{prog}_{1/4}(U^\top x_t).$$

If the latent kernel is a uniform finite-channel probability-ϖ zero-chain and $d > N + T$, then

$$\mathbb{P}(\pi_N \ge T) \le e^{2\varpi N - T} + 2NT\exp\left(-\frac{d-N-T}{32R^2N}\right). \tag{139}$$

The bound does not depend on the finite number m of vector channels.

Proof. We separate rare progress from geometric guessing.

Rare progress. For $t \in [N]$, let

$$\mathcal{V}_t := \{\mathrm{prog}_{1/4}(v_t) \le \gamma_{t-1}\}.$$

Use the filtration

$$\mathcal{H}_t := \sigma(U, \xi_0, \xi_1, \dots, \xi_t), \qquad \mathcal{H}_0 := \sigma(U, \xi_0). \tag{140}$$

Because the algorithm is deterministic after fixing ξ_0, U, and the past seeds, (x_t, y_t), v_t, γ_{t-1}, and $\mathcal{V}_t$ are $\mathcal{H}_{t-1}$-measurable. On $\mathcal{V}_t$, (119) gives $\gamma_t - \gamma_{t-1} \in \{0, 1\}$. If additionally $\xi_t \notin E$, then (120) gives $\gamma_t = \gamma_{t-1}$. Equivalently,

$$0 \le \gamma_t - \gamma_{t-1} \le \mathbf{1}_{\{\xi_t \in E\}} \qquad \text{on } \mathcal{V}_t. \tag{141}$$

For the Bernoulli constructions, the right side is ζ_t. Thus an open gate is necessary for progress on $\mathcal{V}_t$, but need not be sufficient. Freshness of ξ_t implies

$$\mathbb{P}\left(\gamma_t - \gamma_{t-1} = 1, \mathcal{V}_t \mid \mathcal{H}_{t-1}\right) \le \varpi. \tag{142}$$

Set $\iota_t := \gamma_t - \gamma_{t-1}$. Iterated conditioning gives

$$\mathbb{E}\left[e^{\sum_{t=1}^{N} \iota_t} \mathbf{1}_{\cap_{t=1}^{N} \mathcal{V}_t}\right] \le (1 - \varpi + \varpi e)^N \le e^{2\varpi N}.$$

Hence

$$\mathbb{P}\left(\gamma_N \ge T, \bigcap_{t=1}^{N} \mathcal{V}_t\right) \le e^{2\varpi N - T}. \tag{143}$$

On $\cap_t \mathcal{V}_t$, the event $\{\pi_N \ge T\}$ implies $\{\gamma_N \ge T\}$.

Geometric guessing. Apply Lemmas B.8 and B.12 with horizon $n = N$. Since γ_t is nondecreasing,

$$\bigcap_{i=1}^{N} \bigcap_{j=1}^{T} \left(\mathcal{G}_j^{(i)} \cup (\mathcal{T}_j^{(i-1)})^c\right) \subseteq \bigcap_{i=1}^{N} \mathcal{V}_i. \tag{144}$$

Indeed, fix i and $j > \gamma_{i-1}$. Then $j > \gamma_{q-1}$ for every $q \le i$, so the event on the left of (144) supplies $\cap_{q=1}^{i} \mathcal{G}_j^{(q)}$. Lemma B.8 then gives $|\langle u_j, x_i \rangle| \le 1/4$. This holds for every $j > \gamma_{i-1}$, which is exactly $\mathcal{V}_i$.

A union bound and Lemma B.12 now give

$$\mathbb{P}\left(\left[\bigcap_{i=1}^{N} \mathcal{V}_i\right]^c\right) \le 2NT \exp\left(-\frac{d - N - T}{32R^2 N}\right). \tag{145}$$

Combining (143) and (145) proves (139). The argument never counts the number of vector channels: on $\mathcal{T}_j^{(i-1)}$, every one of them lies in $\mathrm{span}\{u_1, \dots, u_{j-1}\}$, and all non-ambient channels are already determined by the conditioned latent queries and seeds. □

Corollary B.14 (Terminal output)**.** *Under the assumptions of Lemma B.13, suppose $d > N + 1 + T$ and, after N calls, the algorithm outputs an $\sigma(\xi_0, \mathcal{T}_N)$-measurable point x_{out} with $\|x_{\mathrm{out}}\| \le R$. Then*

$$\begin{aligned}\mathbb{P}\left(\max\left\{\max_{t \le N} \mathrm{prog}_{1/4}(U^\top x_t), \mathrm{prog}_{1/4}(U^\top x_{\mathrm{out}})\right\} \ge T\right) \\ \le e^{2\varpi N - T} + 2(N+1)T \exp\left(-\frac{d - (N+1) - T}{32R^2(N+1)}\right).\end{aligned} \tag{146}$$

Proof. Set $x_{N+1} := x_{\mathrm{out}}$ and regard it only as an additional point in the geometric argument. No oracle call is made at this point: no new seed is drawn, no response is returned, and the revealed progress is fixed at $\gamma_{N+1} := \gamma_N$. Thus the rare-progress estimate remains exactly $e^{2\varpi N - T}$. Repeat the projection-accumulation, conditional-residual-law, and union bound arguments with the $N + 1$ vectors $x_1, \dots, x_N, x_{\mathrm{out}}$. For each j, on $\mathcal{T}_j^{(N)}$, the quiet-simulation argument in Lemma B.11 reconstructs the transcript from $\mathcal{U}_{j-1}^{(N)}$. The output is therefore measurable from this field on that event. The same conditional-law argument applies to the final point without a new seed. Taking $n = N + 1$ in the geometric term gives (146). □

Corollary B.15 (Soft projection and known postprocessing)**.** *Let $\rho : \mathbb{R}^d \to \mathbb{R}^d$ be Borel with $\|\rho(x)\| \le R$ for all x. Let $(x, y, o) \mapsto \mathsf{P}_{x,y}(o)$ be a jointly Borel map independent of U, and suppose the actual response is*

$$\mathsf{P}_{x,y}\left(\widetilde{\mathcal{O}}_U(\rho(x), y; \xi)\right).$$

Then Lemma B.13 and Corollary B.14 hold with $U^\top x$ replaced by $U^\top \rho(x)$.

Proof. An algorithm interacting with the actual oracle can be simulated by an algorithm interacting with the richer canonical oracle: the simulator queries $\rho(x)$, receives the canonical response, applies the known postprocessing, and passes the result to the original algorithm. At termination it returns $\rho(x_{\mathrm{out}})$, with no additional call. Joint measurability makes this a valid adaptive simulation. □

B.4 THE COMPARISON ORACLE AND THE FIRST-ORDER BENCHMARK

Proof of Proposition 4.1. Take the scaled pair (A_U, u_U) from Proposition B.6, with T as in (110). For $0 < \varpi \le 1/2$, put $N_{\mathrm{s}} := \lfloor T/(8\varpi) \rfloor$ and choose

$$d := \left\lceil N_{\mathrm{s}} + 1 + T + 32R^2(N_{\mathrm{s}}+1)\log\big(64(N_{\mathrm{s}}+1)T\big) \right\rceil . \tag{147}$$

Draw U uniformly from $\mathrm{Ortho}(d, T)$, independently of the algorithm seed and oracle draws. The functions are smooth; their Lipschitz, value, and gap bounds follow from (115)–(117).

At a bounded query $\bar{r} \in \mathbb{R}^d$, write $v = U^\top \bar{r}$ and consider

$$(w, v_B, v_W) := \left(\frac{\zeta}{\varpi} W_T(v),\, U\nabla B_T(v),\, U\frac{\zeta}{\varpi}\nabla W_T(v) \right). \tag{148}$$

By Corollary B.3, its two vector channels satisfy Definition B.7 with exceptional event $\{\zeta = 1\}$. The scalar channel is a Borel function of the latent query v and the seed, as allowed by that definition. Hence the full response, including the scalar value, is covered by Corollary B.14.

One comparison call at an arbitrary x is obtained by querying (148) at $\bar{r} = \rho_R(x/\lambda_{\mathrm{lb}})$ and returning

$$\begin{aligned} r_A &:= \frac{a}{\lambda_{\mathrm{lb}}} J_R(x/\lambda_{\mathrm{lb}})^\top v_B + \frac{a\eta_{\mathrm{r}}}{\lambda_{\mathrm{lb}}^2} x, \\ r_u &:= aw, \qquad r_{\nabla u} := \frac{a}{\lambda_{\mathrm{lb}}} J_R(x/\lambda_{\mathrm{lb}})^\top v_W . \end{aligned} \tag{149}$$

Differentiating (112)–(113) shows that $(r_A, r_u, r_{\nabla u})$ is exactly $\mathcal{R}_\varpi(x)$ in (22). These transformations are jointly continuous and independent of U. The compressed final output $\bar{r}_{\mathrm{out}} = \rho_R(x_{\mathrm{out}}/\lambda_{\mathrm{lb}})$ has norm at most R and costs no oracle call.

For any algorithm using at most N_{s} comparison calls, pad a shorter run with ignored calls. Corollaries B.15 and B.14 and the dimension choice give

$$\mathbb{P}\big(\mathrm{prog}_{1/4}(U^\top \bar{r}_{\mathrm{out}}) \ge T\big) \le e^{2\varpi N_{\mathrm{s}} - T} + \frac{1}{32} \le e^{-3T/4} + \frac{1}{32} < \frac{1}{4}, \tag{150}$$

since $T \ge 4$. Equation (118) now gives (24). Thus $c = 1/8$ is admissible, and a common numerical C bounds the remaining quantities. □

B.4.1 VARIANCE AND THE ORDINARY FIRST-ORDER LOWER BOUND

Lemma B.16 (Variance and the ordinary first-order lower bound)**.** *For each pair in Proposition 4.1, (21) is conditionally unbiased and*

$$\mathbb{E}[\|\widehat{\nabla P}(x) - \nabla P(x)\|^2 \mid \mathcal{H}] = \frac{1-\varpi}{\varpi}\|\nabla u(x)\|^2 \le \frac{C^2\epsilon^2}{\varpi}. \tag{151}$$

Choosing $\varpi = 2(1+C^2)\epsilon^2 \le 1/2$ gives a family with variance at most one for which guaranteeing $\mathbb{E}\|\nabla P(x_{\mathrm{out}})\|^2 \le \epsilon^2$ requires $\Omega(\epsilon^{-4})$ queries.

Proof. Fix U and condition on the history $\mathcal{H}$, which fixes the query x. The fresh Bernoulli seed satisfies $\mathbb{E}[\zeta/\varpi] = 1$, so (21) is conditionally unbiased. Moreover,

$$\mathbb{E}[\|\widehat{\nabla P_U}(x) - \nabla P_U(x)\|^2 \mid \mathcal{H}] = \frac{1-\varpi}{\varpi}\|\nabla u_U(x)\|^2 \le \frac{C^2\epsilon^2}{\varpi},$$

where C is the constant in Proposition 4.1. If $\mathcal{R}_\varpi(x) = (r_A, r_u, r_{\nabla u})$, then $r_A - r_{\nabla u}$ has exactly this stochastic-gradient law, with one comparison call per gradient call.

Set $\varpi = 2(1+C^2)\epsilon^2$. For sufficiently small ϵ this lies in $(0, 1/2]$, and the conditional variance is at most $1/2$. The objective has gradient Lipschitz constant at most $2C$ and initial gap at most $1/4$, by Proposition 4.1. Simulating any stochastic-gradient algorithm with at most $T/(8\varpi)$ calls therefore gives a comparison-oracle algorithm with the same budget and output law. Proposition 4.1 yields

$$\mathbb{P}\big(\|\nabla P_U(x_{\mathrm{out}})\| \ge 2\epsilon\big) \ge \frac{3}{4}.$$

Averaging this probability over U gives a fixed rotation on which the probability, now over algorithm and oracle randomness, is at least $3/4$. On that instance, $\mathbb{E}\|\nabla P_U(x_{\mathrm{out}})\|^2 \ge 3\epsilon^2 > \epsilon^2$. Finally, $T/\varpi = \Theta(\epsilon^{-4})$. □

B.5 THE BILEVEL CONSTRUCTION

Fix $0 < \epsilon \le \delta \le 1$ and a scaled pair (A_U, u_U) from Proposition B.6. With $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$, set

$$f_U(x, y) := A_U(x) + y, \tag{152}$$

$$g_U(x, y) := \frac{1}{2}y^2 + \delta \sin(y/\delta) u_U(x). \tag{153}$$

The dependence on δ is suppressed in this appendix.

B.5.1 PRIMITIVE ORACLE: SIMULATION AND VARIANCE

For $0 < \varpi \le 1/2$, draw a fresh $\zeta \sim \mathrm{Bernoulli}(\varpi)$ at each call and return

$$\hat{F}_x := \nabla A_U(x), \qquad \hat{F}_y := 1, \tag{154}$$

$$\hat{G}_x := \delta \sin(y/\delta) \frac{\zeta}{\varpi} \nabla u_U(x), \qquad \hat{G}_y := y + \cos(y/\delta) \frac{\zeta}{\varpi} u_U(x). \tag{155}$$

The seeds are independent across calls and independent of U and the algorithm seed. Define the numerical constant

$$C_{\mathrm{var}} := C_{u,1}^2 + C_{u,0}^2. \tag{156}$$

Lemma B.17 (Primitive simulation and conditional variance)**.** *For every fixed U and selected query (x, y), the four estimates are conditionally unbiased and their joint conditional variance is at most $C_{\mathrm{var}}\delta^2\epsilon^2/\varpi$. One primitive call is a known measurable transformation of one comparison call $\mathcal{R}_\varpi(x)$.*

Proof. Condition on the history, fixing (x, y). Unbiasedness follows from $\mathbb{E}[\zeta/\varpi] = 1$ and direct differentiation of (152)–(153). The upper estimates are exact. The sum of the conditional squared errors in the lower estimates is

$$\begin{aligned} &\frac{1-\varpi}{\varpi}\left(\delta^2 \sin^2(y/\delta)\|\nabla u_U(x)\|^2 + \cos^2(y/\delta)|u_U(x)|^2\right) \\ &\quad \le \frac{C_{u,1}^2\delta^2\epsilon^2 + C_{u,0}^2\epsilon^4}{\varpi} \le \frac{C_{\mathrm{var}}\delta^2\epsilon^2}{\varpi}. \end{aligned} \tag{157}$$

Here Proposition B.6 supplies the uniform bounds, and $\epsilon \le \delta$ controls the scalar contribution.

If $\mathcal{R}_\varpi(x) = (r_A, r_u, r_{\nabla u})$, return

$$(r_A,\ 1,\ \delta \sin(y/\delta) r_{\nabla u},\ y + \cos(y/\delta) r_u)\,. \tag{158}$$

This reproduces the full joint law of (154)–(155), including the common Bernoulli draw. The transformation is continuous and independent of U. Applying it after every comparison call simulates an adaptive primitive-oracle algorithm, with the same query choices and output law. □

B.5.2 LOWER SOLUTION AND OUTER GRADIENT

For $|s|/\delta \le 1/4$, let $Y_\delta(s)$ denote the unique solution of

$$Y + s\cos(Y/\delta) = 0. \tag{159}$$

Lemma B.18 (Lower response)**.** *For $|s|/\delta \le 1/4$, the response $Y_\delta(s)$ is continuously differentiable and obeys*

$$|Y_\delta(s)| \le |s|, \tag{160}$$

$$|Y_\delta(s) + s| \le \frac{|s|^3}{2\delta^2}, \tag{161}$$

$$|Y_\delta'(s) + 1| \le \frac{2s^2}{\delta^2}. \tag{162}$$

For all sufficiently small ϵ, uniformly over $\epsilon \le \delta \le 1$, the lower solution of (153) is

$$y_U^\star(x) = Y_\delta(u_U(x)), \tag{163}$$

and $\nabla_{yy}^2 g_U(x, y) \ge 3/4$ globally.

Proof. The derivative in Y of the left side of (159) is

$$1 - (s/\delta)\sin(Y/\delta) \geq \frac{3}{4}.$$

Hence that function is strictly increasing and crosses zero exactly once; the implicit-function theorem gives differentiability. Equation (159) gives $|Y_\delta(s)| \leq |s|$, and

$$|Y_\delta(s) + s| = |s|\,|1 - \cos(Y_\delta(s)/\delta)| \leq \frac{|s|^3}{2\delta^2}.$$

Implicit differentiation yields

$$Y'_\delta(s) = -\frac{\cos(Y_\delta(s)/\delta)}{1 - (s/\delta)\sin(Y_\delta(s)/\delta)}.$$

With $t = Y_\delta(s)/\delta$ and $\beta = s/\delta$,

$$|Y'_\delta(s) + 1| = \left|\frac{1 - \cos t - \beta \sin t}{1 - \beta \sin t}\right| \leq \frac{\frac{1}{2}\beta^2 + \beta^2}{3/4} = 2\beta^2,$$

where we used $|t| \leq |\beta|$, $|1 - \cos t| \leq t^2/2$, and $|\sin t| \leq |t|$. This proves (162). Finally, Proposition B.6 and $\delta \geq \epsilon$ imply

$$\frac{\|u_U\|_\infty}{\delta} \leq C_{u,0}\epsilon \leq \frac{1}{4}$$

for a universal choice of ϵ_0. Since

$$\nabla_y g_U = y + \cos(y/\delta)u_U, \qquad \nabla^2_{yy} g_U = 1 - \delta^{-1}\sin(y/\delta)u_U,$$

the claims follow. □

Define the actual hyperobjective

$$\Phi_U(x) := f_U(x, y^\star_U(x)) = A_U(x) + Y_\delta(u_U(x)). \tag{164}$$

Corollary B.19 (Controlled perturbation of the hard objective)**.** *There are universal constants $C_F, C_\nabla > 0$ such that*

$$\sup_x |\Phi_U(x) - P_U(x)| \leq C_F \frac{\epsilon^6}{\delta^2} \leq C_F \epsilon^4, \tag{165}$$

$$\sup_x \|\nabla\Phi_U(x) - \nabla P_U(x)\| \leq C_\nabla \frac{\epsilon^5}{\delta^2} \leq C_\nabla \epsilon^3. \tag{166}$$

After decreasing ϵ_0 universally if necessary,

$$\Phi_U(0) - \inf_x \Phi_U(x) \leq 1, \qquad |y^\star_U(0)| \leq 1, \tag{167}$$

and

$$\|\nabla P_U(x)\| \geq 2\epsilon \quad \Longrightarrow \quad \|\nabla\Phi_U(x)\| \geq \frac{3}{2}\epsilon. \tag{168}$$

Proof. The first estimate follows from (161), $P_U = A_U - u_U$, and (115). The chain rule and (162) give

$$\nabla\Phi_U = \nabla A_U + Y'_\delta(u_U)\nabla u_U,$$

and hence the second estimate. Here $\delta \geq \epsilon$ makes both displayed upper bounds uniform over $\epsilon \leq \delta \leq 1$. Combining the value estimate with (117) gives the objective-gap bound. The initial lower solution satisfies $|y^\star_U(0)| \leq C_{u,0}\epsilon^2 \leq 1$ for small enough ϵ, proving (167). If $\|\nabla P_U(x)\| \geq 2\epsilon$, the gradient estimate gives $\|\nabla\Phi_U(x)\| \geq 2\epsilon - C_\nabla\epsilon^3 \geq 3\epsilon/2$ for a numerical sufficiently small ϵ_0. □

The gradient estimate can also be read directly from the implicit hypergradient. In the notation of Section 2,

$$b = 1, \qquad H_U(x,y) = 1 - \delta^{-1}\sin(y/\delta)u_U(x), \qquad J_U(x,y) = \cos(y/\delta)\nabla u_U(x).$$

At $y = y^\star_U(x)$, implicit differentiation gives

$$\begin{aligned} J_U(x, y^\star_U(x))z^\star_U(x) &= -Y'_\delta(u_U(x))\nabla u_U(x), \\ \|J_U(x, y^\star_U(x))z^\star_U(x) - \nabla u_U(x)\| &\leq 2C^2_{u,0}C_{u,1}\frac{\epsilon^5}{\delta^2}. \end{aligned} \tag{169}$$

Thus the mixed-derivative term in the hypergradient retains the correction ∇u_U even though the primitive signal is smaller.

B.5.3 BASE AND HIGHER-ORDER REGULARITY

Proof of Proposition 4.2. The Hessian blocks of g_U are

$$\nabla^2_{xx} g_U = \delta \sin(y/\delta)\nabla^2 u_U(x), \qquad \nabla^2_{xy} g_U = \cos(y/\delta)\nabla u_U(x),$$

$$\nabla^2_{yy} g_U = 1 - \delta^{-1}\sin(y/\delta)u_U(x).$$

They are bounded by numerical constants for all (x, y), by Proposition B.6 and $\epsilon \le \delta \le 1$. Hence the full gradient of g_U is jointly Lipschitz. The x- and y-Lipschitz moduli of J_U are bounded by $C_{u,2}$ and $C_{u,1}\epsilon/\delta$, respectively; those of H_U are bounded by $C_{u,1}\epsilon/\delta$ and $C_{u,0}\epsilon^2/\delta^2$. Combining the two moduli in each case gives a numerical joint Lipschitz bound, uniformly over $\epsilon \le \delta \le 1$. Lemma B.18 gives strong convexity with $\mu = 3/4$. Also $\nabla_y f_U = 1$, and (116) bounds the full Hessian of f_U by one. The objective is bounded below and obeys (30) by Corollary B.19. This proves Assumption 2.1.

The same corollary proves (27), while Lemma B.17 supplies conditional unbiasedness, variance, and one-call simulation. A numerical constant $C_{\mathrm{lift}} \ge \max\{C_\nabla, C_{\mathrm{var}}\}$ bounds both quantities in the proposition. □

For an integer $p \ge 1$, choose

$$\delta := \epsilon^{1/p} \in [\epsilon, 1]. \tag{170}$$

Lemma B.20 (Uniform order-p regularity)**.** *With (170), the family satisfies Assumptions 2.1 and 2.2 with a numerical bound on every constant, independent of T, d, U, ϵ, and p.*

Proof. Proposition 4.2 proves the base regularity. For $p \ge 2$, write $t = y/\delta$ and denote derivatives of sine by $\sin^{(j)}$. Direct differentiation gives

$$\begin{aligned} \nabla^p_y \nabla_x g_U(x, y) &= \delta^{1-p}\sin^{(p)}(t)\nabla u_U(x), \\ \nabla^p_y \nabla_y g_U(x, y) &= \delta^{-p}\sin^{(p+1)}(t)u_U(x). \end{aligned} \tag{171}$$

Every derivative of sine has absolute value at most one. Since the lower variable is scalar, the operator norms in Assumption 2.2 are respectively the Euclidean norm and absolute value. The Lipschitz moduli in y are therefore at most

$$C_{u,1}\epsilon\delta^{-p} = C_{u,1}, \qquad C_{u,0}\epsilon^2\delta^{-(p+1)} = C_{u,0}\epsilon^{1-1/p} \le C_{u,0}.$$

For $p = 1$, the base Lipschitz bounds for J_U and H_U imply the same assumption. All constants are independent of p. □

Finally, set

$$c_{\mathrm{rr}} := 2(1 + C_{\mathrm{var}}), \tag{172}$$

and take

$$\varpi := c_{\mathrm{rr}}\delta^2\epsilon^2 = c_{\mathrm{rr}}\epsilon^{2+2/p} \le \frac{1}{2}. \tag{173}$$

The last inequality holds for a numerical sufficiently small ϵ_0, uniformly in p, since $\delta \le 1$.

Lemma B.21 (Admissible primitive oracle)**.** *With (173), the oracle (154)–(155) satisfies Assumption 2.3 with $\sigma^2 \le 1/2$.*

Proof. Lemma B.17 gives conditional unbiasedness and joint variance at most $C_{\mathrm{var}}\delta^2\epsilon^2/\varpi = C_{\mathrm{var}}/c_{\mathrm{rr}} \le 1/2$. The response is a Borel function of the query and fresh Bernoulli seed; its pushforward law is a Borel kernel. Independent seeds give the required conditional independence between fresh calls. □

B.6 Completion of the lower bound

Proof of Theorem 4.3. Fix an integer $p \geq 1$ and $0 < \epsilon < \epsilon_0$. Take T from (110), $\delta = \epsilon^{1/p}$, and $\varpi = c_{\mathrm{rr}}\delta^2\epsilon^2$ as in (173). Use the distribution of (A_U, u_U) in Proposition 4.1, and form the bilevel pair and primitive oracle in Section B.5. Define

$$N_0 := \left\lfloor \frac{T}{8\varpi} \right\rfloor, \tag{174}$$

and the numerical constant

$$c_{\mathrm{lb}} := \frac{c_T}{64 c_{\mathrm{rr}}}. \tag{175}$$

For a numerical sufficiently small ϵ_0, we have $T \geq (c_T/2)\epsilon^{-2}$ and $T/(8\varpi) \geq 2$, uniformly in p. Consequently,

$$N_0 \geq \frac{T}{16\varpi} \geq \frac{c_T}{32 c_{\mathrm{rr}}}\epsilon^{-4-2/p} \geq c_{\mathrm{lb}}\epsilon^{-4-2/p}. \tag{176}$$

Let an arbitrary adaptive randomized algorithm use at most N_0 primitive calls. Lemma B.17 simulates it with the same number of calls to $\mathcal{R}_\varpi$ and preserves its output law. Proposition 4.1, whose proof permits $c = 1/8$, therefore gives

$$\mathbb{P}\big(\|\nabla P_U(x_{\mathrm{out}})\| \geq 2\epsilon\big) \geq \frac{3}{4}. \tag{177}$$

By (168), on this event, $\|\nabla \Phi_U(x_{\mathrm{out}})\| \geq 3\epsilon/2$. Averaging the event probability over U gives a fixed rotation for which this event has probability at least $3/4$ over the algorithm and oracle randomness. For that fixed instance,

$$\mathbb{E}\|\nabla \Phi_U(x_{\mathrm{out}})\|^2 \geq \frac{3}{4}\left(\frac{3}{2}\epsilon\right)^2 = \frac{27}{16}\epsilon^2 > \epsilon^2.$$

Equation (176) covers every integer budget in Theorem 4.3 with $c = c_{\mathrm{lb}}$.

It remains to check that the fixed instance lies in the stated class. Proposition 4.2 gives the numerical base regularity and initial bounds. Lemma B.20 bounds the order-p constant independently of p, and Lemma B.21 gives $\sigma^2 \leq 1/2$. Taking M_0 to be the maximum of one and these numerical bounds places the pair in $\mathcal{P}_p(M_0)$. Every smallness condition on ϵ used above is numerical and independent of p, so a common $\epsilon_0 > 0$ suffices.

The construction is finite dimensional. In (147), $N_{\mathrm{s}} = N_0$ and $R^2 = \Theta(T)$, hence

$$d = O\left(\frac{T^2}{\varpi}\log\frac{T}{\varpi}\right) = O\Big(\epsilon^{-6-2/p}\log(1/\epsilon)\Big).$$

□