
SLIDING METHODS FOR HÖLDER-SMOOTH CONVEX–CONCAVE MINIMAX OPTIMIZATION WITH BILINEAR COUPLING

A PREPRINT

Nhat Trung Nguyen
MIRIAI
nguyen.n@miriai.org

Alexander Gasnikov
Innopolis University, MIRIAI, IITP
gasnikov@yandex.ru

ABSTRACT

We study convex-concave minimax optimization problems with bilinear coupling of the form $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x) + \langle y, \mathbf{B}x \rangle - g(y)$, where the functions f and g have Hölder continuous (sub)gradients. This setting covers a broad range of regimes, from nonsmooth problems with bounded subgradient variation to smooth problems with Lipschitz continuous gradients; for a smooth component used in the coupling-induced regularizer, its Lipschitz-gradient constant is assumed to hold in the ambient space. We propose a sliding method that exploits the composite structure of the problem by querying the oracles associated with f , g , and the bilinear coupling operator at prescribed frequencies determined by their individual properties. The method is based on a recursive sliding scheme for monotone variational inequalities. We establish convergence guarantees under Hölder continuity and show how the resulting complexity bounds depend explicitly on the Hölder exponents, Hölder constants, strong convexity parameters, and spectral properties of the coupling matrix. Our analysis covers nonstrongly convex and partially strongly convex regimes. For stochastic function oracles, we prove componentwise Hölder guarantees with separate variance costs: a uniform expected-gap bound on bounded domains and, when both functions are strongly convex on their feasible sets, a restarted Lyapunov guarantee without an anchor noise floor. Numerical experiments reproduce the predicted Hölder exponents and confirm that the number of gradient evaluations required for each function separates according to its own smoothness level rather than the worse of the two. A tomographic benchmark shows runtime gains when gradient evaluations are more expensive than the additional matrix-vector products.

1 Introduction

Convex-concave minimax optimization plays a central role in many areas of optimization, machine learning, game theory, and signal processing. In this work, we focus on the important class of saddle-point problems in which the interaction between the primal and dual variables is bilinear. Such problems arise naturally in constrained optimization, decentralized optimization, regularized optimal transport, inverse problems, and matrix games. We discuss several applications of this problem class in Appendix B. Specifically, we study minimax problems of the form

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F(x, y) = f(x) + \langle y, \mathbf{B}x \rangle - g(y), \quad (1)$$

where $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ are nonempty closed convex sets, and $\mathbf{B} \in \mathbb{R}^{d_y \times d_x}$ is the coupling matrix.

We consider the setting where $f: \mathcal{X} \rightarrow \mathbb{R}$ and $g: \mathcal{Y} \rightarrow \mathbb{R}$ are convex functions whose (sub)gradients satisfy Hölder continuity conditions. This assumption provides a unified framework that interpolates between nonsmooth problems with bounded subgradient variation and smooth problems with Lipschitz continuous gradients. The formal assumptions on f , g , and the spectral structure of $\mathbf{B}$ are stated in Section 4.

In the smooth case, Problem (1) was studied by Borodich et al. (2025), who obtained separated accelerated complexities for evaluating ∇f , evaluating ∇g , and applying $\mathbf{B}$ and $\mathbf{B}^\top$. This paper investigates how these guarantees extend to the Hölder-continuous setting.

1.1 Main Contributions

The main contributions of this paper are summarized as follows.

Componentwise sliding for variational inequalities. We develop a recursive sliding algorithm for finite-sum monotone variational inequalities with Hölder-continuous function components. The method extends the sliding framework of Borodich et al. (2025) beyond the smooth setting by querying different components at different frequencies, which yields componentwise oracle complexities. In the composite optimization case, this removes the cross-component terms present in existing gradient sliding bounds (Lan 2016; Lan and Ouyang 2022).

Hölder-smooth bilinear minimax optimization. We apply the recursive sliding framework to the Hölder-smooth convex-concave minimax problem (1). By formulating the problem as a structured variational inequality with components corresponding to f , g , and the bilinear coupling, the method separates the oracle calls associated with the different parts of the problem. We prove convergence guarantees with component-wise complexity bounds for (sub)gradient evaluations of f , (sub)gradient evaluations of g , and matrix-vector products involving $\mathbf{B}$ and $\mathbf{B}^\top$. Each count depends on the Hölder exponent and constant of the corresponding function, on the strong convexity parameters, and on the spectrum of $\mathbf{B}$. We also extend the deterministic results to nonstrongly convex and partially strongly convex settings.

Stochastic Hölder guarantees. For conditionally unbiased, bounded-variance function oracles and exact bilinear products, we retain the componentwise Hölder complexities and add separate variance costs. We prove a uniform expected primal–dual gap bound on bounded domains. When $\mu_x, \mu_y > 0$, a decomposition without the anchor regularizer yields convergence in the paper’s Lyapunov metric with variance costs $\sigma_x^2/(\mu_x\varepsilon)$ and $\sigma_y^2/(\mu_y\varepsilon)$, without an anchor noise floor. The Hölder assumptions are imposed on the feasible sets.

Numerical experiments. Synthetic tests recover the predicted componentwise exponents and separated gradient counts. A three-regime Huber–TV tomography benchmark against four primal–dual baselines shows runtime gains when gradient calls dominate the additional matrix-vector cost, including a regime with a genuinely Hölder dual gradient.

1.2 Related Work

Convex-concave saddle-point problems. The study of convex-concave minimax problems has a long history in optimization. Classical first-order methods include the extragradient method of Korpelevich (1976), the mirror-prox algorithm of Nemirovski (2004) achieving an $O(1/t)$ convergence rate for Lipschitz monotone operators, and the primal-dual algorithm of Chambolle and Pock (2011). These general-purpose methods do not exploit any particular structure in the coupling between the primal and dual variables, and hence yield suboptimal rates for problems with bilinear coupling.

Bilinear saddle-point problems. For bilinear coupling $\langle y, \mathbf{B}x \rangle$, spectral structure yields sharper rates. Prior work gives lower bounds (Zhang et al. 2022), bilinear-induced linear convergence (Du and Hu 2019; Ibrahim et al. 2020), and optimal smooth algorithms (Borodich et al. 2025). The latter, closest to this paper, uses condition numbers combining the smoothness of f, g with the spectrum of $\mathbf{B}$. Kovalev et al. (2022) give an optimal Accelerated Primal-Dual Gradient method for the strongly convex–strongly concave regime, but charge ∇f , ∇g , and $\mathbf{B}$ jointly rather than separately.

Universal primal–dual methods already treat Hölder-smooth affine-constrained optimization with bilinear Lagrangians: Yurtsever et al. (2015) adapt to unknown dual Hölder regularity, and Luo (2024) give an accelerated universal method for Lipschitz and Hölder gradients. Their guarantees count joint iterations. In this work, the two function oracles and the coupling oracle are charged separately.

Complexity separation. Per-component budgets are established in the smooth regime: Tominin et al. (2021) and Lan and Ouyang (2021) obtain separated rates for strongly-convex–strongly-concave and two-component composite problems, respectively. Closest to us, Borodich et al. (2025) give an n -level recursive sliding method for composite monotone variational inequalities with separated budgets, plus matching lower bounds. This paper adds the layer above: exact gradients are replaced by inexact oracles, which lets each function carry its own level of smoothness, from Lipschitz gradients to bounded subgradient variation, within a single analysis. For composite minimization, the parameter-free universal gradient-sliding method of Wu et al. (2026) also separates the (sub)gradient counts of a Hölder component and a smooth component without knowing their constants. It does not contain a bilinear coupling oracle. Appendix M gives the detailed comparison.

Variational inequality framework. The reduction of saddle-point problems to monotone variational inequalities (VIs) is a classical technique (Facchinei and Pang 2003). Nesterov (2007) developed the dual extrapolation method for solving

VIs, and Juditsky and Nemirovski (2016) studied VIs on domains given by linear minimization oracles. In this work, we formulate the bilinear saddle-point problem as a structured VI with a finite-sum decomposition into $n = 3$ components, which enables the application of multi-level sliding techniques.

Sliding and multi-level methods. The gradient sliding technique, introduced by Lan (2016) for composite optimization, allows different components of the objective to be updated at different frequencies, thereby reducing the total number of expensive gradient evaluations. This idea was further developed by Lan and Zhou (2018) for incremental gradient methods. Sliding methods have also been applied in distributed optimization (Dvinskikh and Gasnikov 2021; Kovalev et al. 2020). Closest to Section 3 is the Mirror-Prox Sliding method of Lan and Ouyang (2021), which treats a smooth convex function together with a monotone Lipschitz operator and charges each of the two its own number of evaluations, with a stochastic variant in which the operator is sampled; our two-component case with smooth components reproduces those rates.

Hölder continuous gradients and inexact oracles. The framework of Hölder continuous gradients, parameterized by $\nu \in [0, 1]$, provides a unified treatment that interpolates between non-smooth ($\nu = 0$) and smooth ($\nu = 1$) optimization. Nesterov (2015) introduced universal gradient methods that automatically adapt to the unknown smoothness level, and Nesterov (2005) developed smoothing techniques for non-smooth problems. For monotone VIs, the generalized Mirror-Prox method of Stonyakin et al. (2022) is universal for Hölder-continuous operators. Separately, the inexact oracle model of Devolder et al. (2014) provides a principled way to analyze first-order methods when gradient and function value computations are approximate.

Stochastic Hölder optimization. Rodomanov et al. (2024) develop universal accelerated methods for stochastic convex optimization with the rate $O(HR^{1+\nu}/N^{(1+3\nu)/2} + \sigma R/\sqrt{N})$. Our stochastic extension has this dependence for each function component, while separately charging the exact bilinear products. It uses prescribed Hölder and variance bounds to choose the recursive budgets. The analysis combines the Hölder model error with a martingale bound that remains valid for the uniform primal–dual gap.

Paper organization. Section 2 collects definitions and notation. Section 3 develops a recursive sliding algorithm for monotone variational inequalities with Hölder function components and Lipschitz monotone operators. Section 4 specializes it to the bilinear saddle-point problem (1) and states the deterministic and stochastic guarantees. Section 5 presents numerical experiments. All proofs are deferred to the appendix.

2 Definitions and Notations

We denote by $\|\cdot\|$ the standard Euclidean norm on $\mathbb{R}^d$ and by $\langle \cdot, \cdot \rangle$ the standard inner product. Throughout this section, $\mathcal{U} \subseteq \mathbb{R}^d$ denotes a nonempty closed convex set, typically a feasible set. For a positive definite matrix $\mathbf{P} \in \mathbb{S}_{++}^d$, the weighted norm is $\|z\|_{\mathbf{P}} = \langle \mathbf{P}z, z \rangle^{1/2}$. For a symmetric matrix $\mathbf{M}$, we write $\lambda_{\max}(\mathbf{M})$, $\lambda_{\min}(\mathbf{M})$, and $\lambda_{\min}^+(\mathbf{M})$ for its largest eigenvalue, smallest eigenvalue, and smallest positive eigenvalue, respectively. We write $\text{range}(\mathbf{M})$ for the column space of $\mathbf{M}$.

Definition 1. Let $h: \mathbb{R}^d \rightarrow \mathbb{R}$ be convex. A vector $\xi \in \mathbb{R}^d$ is called a *subgradient* of h at the point $u \in \mathbb{R}^d$ if for any $v \in \mathbb{R}^d$ we have

$$h(v) \geq h(u) + \langle \xi, v - u \rangle. \quad (2)$$

The set of all subgradients of h at u , denoted by $\partial h(u)$, is called the *subdifferential* of h at u .

Notation and convention. Throughout the paper, when h is subdifferentiable at u , we write $h'(u) \in \partial h(u)$ to denote an arbitrary subgradient of h at u . Any condition involving $h'(u)$ must hold for every choice from $\partial h(u)$. When such a condition is required for every $u \in \mathcal{U}$, we also assume that a subgradient exists at every point of $\mathcal{U}$, that is,

$$\partial h(u) \neq \emptyset, \quad \forall u \in \mathcal{U}.$$

If h is differentiable at u , then $\partial h(u) = \{\nabla h(u)\}$, and hence $h'(u) = \nabla h(u)$.

Definition 2. A function $h: \mathcal{U} \rightarrow \mathbb{R}$ is called μ -strongly convex on $\mathcal{U} \subseteq \mathbb{R}^d$, for some $\mu \geq 0$, if for any $u, v \in \mathcal{U}$ and $h'(u) \in \partial h(u)$ we have

$$h(v) \geq h(u) + \langle h'(u), v - u \rangle + \frac{\mu}{2} \|v - u\|^2. \quad (3)$$

When $\mu = 0$, the function h is said to be (non-strongly) convex.

Definition 3. We say that $h: \mathcal{U} \rightarrow \mathbb{R}$ is equipped with a first-order (δ, L) -oracle with respect to the norm $\|\cdot\|_{\mathbf{P}}$ if for any $v \in \mathcal{U}$ we can compute a pair $(\tilde{h}(v), \tilde{\nabla}h(v))$ such that for all $u \in \mathcal{U}$,

$$0 \leq h(u) - \tilde{h}(v) - \langle \tilde{\nabla}h(v), u - v \rangle \leq \frac{L}{2} \|u - v\|_{\mathbf{P}}^2 + \delta. \quad (4)$$

Definition 4. Let $0 \leq \nu \leq 1$ and $H > 0$. We say that a function $h: \mathcal{U} \rightarrow \mathbb{R}$ satisfies the (ν, H) -Hölder condition on $\mathcal{U} \subseteq \mathbb{R}^d$ with respect to $\|\cdot\|_{\mathbf{P}}$ if for any $u, v \in \mathcal{U}$ and $h'(u) \in \partial h(u)$, $h'(v) \in \partial h(v)$,

$$\|h'(u) - h'(v)\|_{\mathbf{P}^{-1}} \leq H\|u - v\|_{\mathbf{P}}^{\nu}. \quad (5)$$

For $\nu > 0$ the condition implies Hölder continuity of the gradient, the limit case $\nu = 1$ corresponds to Lipschitz-continuous gradients, while $\nu = 0$ requires only that subgradients have bounded variation.

Definition 5. An operator $G: \mathcal{U} \rightarrow \mathbb{R}^d$ is called μ -strongly monotone with respect to the norm $\|\cdot\|_{\mathbf{P}}$ for $\mu \geq 0$ if the following inequality holds for all $u, v \in \mathcal{U}$:

$$\langle G(u) - G(v), u - v \rangle \geq \mu\|u - v\|_{\mathbf{P}}^2. \quad (6)$$

An operator $G: \mathcal{U} \rightarrow \mathbb{R}^{d_z}$ is called *monotone* if the same inequality holds with $\mu = 0$.

Definition 6. An operator $G: \mathcal{U} \rightarrow \mathbb{R}^d$ is called M -Lipschitz with respect to the norm $\|\cdot\|_{\mathbf{P}}$ for $M_i > 0$ if the following inequality holds for all $u, v \in \mathcal{U}$:

$$\|G(u) - G(v)\|_{\mathbf{P}^{-1}} \leq M\|u - v\|_{\mathbf{P}}. \quad (7)$$

3 General Variational Inequalities

In this section, we consider a general monotone variational inequality problem of finding $z^* \in \mathcal{Z}$ such that

$$p(z^*) - p(z) + \langle Q(z), z^* - z \rangle \leq 0, \quad \forall z \in \mathcal{Z}, \quad (8)$$

where $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ is nonempty, closed, and convex, $p: \mathcal{Z} \rightarrow \mathbb{R}$, and $Q: \mathcal{Z} \rightarrow \mathbb{R}^{d_z}$. We assume that both p and Q have the finite-sum structure:

$$p(z) = \sum_{i=1}^n p_i(z), \quad Q(z) = \sum_{i=1}^n Q_i(z). \quad (9)$$

For each $1 \leq i \leq n$, the function $p_i: \mathcal{Z} \rightarrow \mathbb{R}$ is convex, and the operator $Q_i: \mathcal{Z} \rightarrow \mathbb{R}^{d_z}$ is continuous and monotone.

A point $z^* \in \mathcal{Z}$ satisfying (8) is called a *weak solution* of the monotone variational inequality. Since p is convex and Q is continuous and monotone, the weak and strong formulations are equivalent. In particular, every weak solution also satisfies

$$p(z^*) - p(z) + \langle Q(z^*), z^* - z \rangle \leq 0, \quad \forall z \in \mathcal{Z}. \quad (10)$$

3.1 Algorithm for Inexact Oracle

To exploit the finite-sum structure in (9), we allow each component to have its own operator Lipschitz constant and inexact oracle parameters. Throughout this section, Lipschitz continuity of operators and first-order inexact oracles are understood in the sense of Definitions 6 and 3.

Assumption 3.1. For every $1 \leq i \leq n$, the operator Q_i is monotone and M_i -Lipschitz on $\mathcal{Z}$ with respect to the norm $\|\cdot\|_{\mathbf{P}}$.

Assumption 3.2. For every $1 \leq i \leq n$, the function p_i admits a first-order (δ_i, L_i) -oracle on $\mathcal{Z}$ with respect to the norm $\|\cdot\|_{\mathbf{P}}$.

We first describe two special cases to illustrate the construction. When $n = 1$, there is no inner loop, and the method reduces to Algorithm 2. This algorithm recovers several classical methods. If $Q_1 \equiv 0$, it reduces to Nesterov's accelerated gradient method (Nesterov 2013). If $p_1 \equiv 0$, it reduces to the extragradient method of Korpelevich (1976). When both p_1 and Q_1 are present, it coincides with the Accelerated Mirror-Prox method of Chen et al. (2017), specialized to the $\|\cdot\|_{\mathbf{P}}$ -norm.

For $n = 2$, the outer and inner loops handle the first and second components, respectively, allowing query frequencies tailored to L_i and M_i . The n -level method applies this construction recursively. Appendix A gives both algorithms, and Appendix C.1 discusses implementability.

The following theorem gives the main guarantee for the recursive sliding scheme. It bounds the weak variational inequality gap at the output z_{out} uniformly over all comparison points $z \in \mathcal{Z}$. The proof is given in Appendix D. The theorem is used in Section 3.2 to handle Hölder-continuous components.

Theorem 3.3. *Let Assumptions 3.1 and 3.2 hold, where $M_i, L_i \geq 0$ and $M_i + L_i > 0$ for all i . Let $T_1, \dots, T_n$ be positive integers. Suppose that Algorithm 1 uses the sequence*

$$\alpha_0 = 1, \quad \alpha_{t+1} = \frac{2}{1 + \sqrt{1 + 4/\alpha_t^2}}, \quad t \geq 0, \quad (11)$$

and the parameters

$$\eta_{\mathbf{t}}^k = L_k \prod_{\ell=1}^k \alpha_{t_\ell} + M_k \prod_{\ell=1}^k \frac{\alpha_{t_\ell}}{\alpha_{T_\ell-1}}, \quad \mathbf{t} = (t_1, \dots, t_k), \quad (12)$$

for every $1 \leq k \leq n$, where $0 \leq t_\ell \leq T_\ell - 1$ for $\ell = 1, \dots, k$. Then, for every $z \in \mathcal{Z}$, the output z_{out} of Algorithm 1 satisfies

$$p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle \leq \sum_{i=1}^n \left[\frac{2^{2i-1} L_i}{\prod_{j=1}^i T_j^2} \|z_{\text{in}} - z\|_{\mathbf{P}}^2 + \frac{2^{i-1} M_i}{\prod_{j=1}^i T_j} \|z_{\text{in}} - z\|_{\mathbf{P}}^2 + \delta_i \prod_{j=1}^i T_j \right]. \quad (13)$$

3.2 The Hölder Case

We now specialize the inexact-oracle result to the case where the functions p_i satisfy Hölder-type smoothness conditions. This allows us to use exact values and exact (sub)gradients while choosing the oracle accuracy parameters implicitly through the Hölder constants.

Assumption 3.4. *For each $1 \leq i \leq n$, the function p_i is convex and satisfies the (ν_i, H_i) -Hölder condition on $\mathcal{Z}$ with respect to $\|\cdot\|_{\mathbf{P}}$.*

For each $1 \leq i \leq n$ and $\delta > 0$, define

$$L_i(\delta) = \left[\frac{1 - \nu_i}{2(1 + \nu_i)} \cdot \frac{1}{\delta} \right]^{\frac{1-\nu_i}{1+\nu_i}} H_i^{\frac{2}{1+\nu_i}}. \quad (14)$$

For $\nu_i = 1$, we use the convention $L_i(\delta) = H_i$ for all $\delta \geq 0$. As shown in Lemma E.1, taking the exact value p_i and an exact (sub)gradient p'_i as oracle outputs gives a first-order $(\delta, L_i(\delta))$ -oracle for p_i .

The next theorem is obtained by applying Theorem 3.3 with a componentwise choice of the oracle accuracy parameters. Its proof, together with Lemma E.1, is given in Appendix E.

Theorem 3.5. *Let Assumptions 3.1 and 3.4 hold, where $M_i, H_i \geq 0$ and $M_i + H_i > 0$. Let $T_1, \dots, T_n$ be positive integers and define*

$$\Omega_{z_{\text{in}}} := \sup_{z \in \mathcal{Z}} \|z_{\text{in}} - z\|_{\mathbf{P}}^2 < \infty, \quad N_i := \prod_{j=1}^i T_j.$$

Run Algorithm 1 with the sequence $\{\alpha_t\}$ in (11), using exact values and exact (sub)gradients in the induced oracle constructions. Choose

$$\eta_{\mathbf{t}}^k = L_k(\delta_k^\circ) \prod_{\ell=1}^k \alpha_{t_\ell} + M_k \prod_{\ell=1}^k \frac{\alpha_{t_\ell}}{\alpha_{T_\ell-1}}, \quad \mathbf{t} = (t_1, \dots, t_k). \quad (15)$$

where, for $0 \leq \nu_i < 1$,

$$\delta_i^\circ = \left[\frac{(1 - \nu_i) 2^{2i-1} K_i \Omega_{z_{\text{in}}}}{(1 + \nu_i) N_i^3} \right]^{(1+\nu_i)/2}, \quad K_i = \left[\frac{1 - \nu_i}{2(1 + \nu_i)} \right]^{\frac{1-\nu_i}{1+\nu_i}} H_i^{\frac{2}{1+\nu_i}}. \quad (16)$$

For $\nu_i = 1$, set $K_i = H_i$, $\delta_i^\circ = 0$, and $L_i(\delta_i^\circ) = H_i$. Then

$$\sup_{z \in \mathcal{Z}} \{p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle\} \leq \sum_{i=1}^n \left[\frac{2^{i(1+\nu_i)} H_i \Omega_{z_{\text{in}}}^{(1+\nu_i)/2}}{(1 + \nu_i) N_i^{(1+3\nu_i)/2}} + \frac{2^{i-1} M_i \Omega_{z_{\text{in}}}}{N_i} \right]. \quad (17)$$

The following corollary states the resulting componentwise evaluation complexity. The proof is given in Appendix F. A discussion on comparing with existing results is given in Appendix C.2

Corollary 3.6. *Let the assumptions of Theorem 3.5 hold, and set $\Omega := \Omega_{z_{\text{in}}}$. For any $\varepsilon > 0$ and $i = 1, \dots, n$, define*

$$R_i := \max \left\{ \left(\frac{H_i \Omega^{(1+\nu_i)/2}}{\varepsilon} \right)^{\frac{2}{1+3\nu_i}}, \frac{M_i \Omega}{\varepsilon}, 1 \right\}.$$

Relabel the component pairs (p_i, Q_i) so that $R_1 \leq R_2 \leq \dots \leq R_n$. Then there exists a constant $C_n > 0$, depending only on n , such that performing no more than $C_n R_i$ evaluations of the subgradient p'_i and the operator Q_i , for each $1 \leq i \leq n$, is sufficient to ensure

$$\sup_{z \in \mathcal{Z}} \{p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle\} \leq n\varepsilon.$$

4 Bilinear Saddle-Point Problem

We impose the following assumptions on Problem (1).

Assumption 4.1. *The function $f: \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ is convex and finite. Its restriction to $\mathcal{X}$ is μ_x -strongly convex with $\mu_x \geq 0$ and satisfies the (ν_x, H_x) -Hölder condition with respect to the Euclidean norm, with $0 \leq \nu_x \leq 1$ and $H_x > 0$.*

Assumption 4.2. *The function $g: \mathbb{R}^{d_y} \rightarrow \mathbb{R}$ is convex and finite. Its restriction to $\mathcal{Y}$ is μ_y -strongly convex with $\mu_y \geq 0$ and satisfies the (ν_y, H_y) -Hölder condition with respect to the Euclidean norm, with $0 \leq \nu_y \leq 1$ and $H_y > 0$.*

Assumption 4.3. *There exist constants $L_{xy} > 0$ and $\mu_{xy}, \mu_{yx} \geq 0$, with $L_{xy} > \max\{\mu_{xy}, \mu_{yx}\}$. Let*

$$\begin{aligned} R_x &: \partial f(x) \subseteq \text{range}(\mathbf{B}^\top) \quad \forall x \in \mathcal{X}, \\ R_y &: \partial g(y) \subseteq \text{range}(\mathbf{B}) \quad \forall y \in \mathcal{Y}. \end{aligned}$$

Then

$$\begin{aligned} \mu_{xy}^2 &\leq \begin{cases} \lambda_{\min}^+(\mathbf{B}^\top \mathbf{B}), & \text{if } R_x \text{ holds,} \\ \lambda_{\min}(\mathbf{B}^\top \mathbf{B}), & \text{otherwise,} \end{cases} \\ \mu_{yx}^2 &\leq \begin{cases} \lambda_{\min}^+(\mathbf{B} \mathbf{B}^\top), & \text{if } R_y \text{ holds,} \\ \lambda_{\min}(\mathbf{B} \mathbf{B}^\top), & \text{otherwise,} \end{cases} \\ L_{xy}^2 &\geq \lambda_{\max}(\mathbf{B}^\top \mathbf{B}) = \lambda_{\max}(\mathbf{B} \mathbf{B}^\top). \end{aligned}$$

Additionally, if $\mu_{xy} > 0$ and $\mu_{yx} > 0$, then $\mu_{xy} = \mu_{yx}$. Finally, if R_x is used, we require $\mathcal{X} + \ker \mathbf{B} = \mathcal{X}$, and if R_y is used, $\mathcal{Y} + \ker \mathbf{B}^\top = \mathcal{Y}$.

The last requirement is vacuous when the kernel is trivial, e.g. for $\mathcal{X} = \mathbb{R}^{d_x}$; it keeps the iterates in the subspace on which $\lambda_{\min}^+$ lower-bounds $\mathbf{B}$ (Appendix G).

Ambient-space functions and subgradients. In Assumptions 4.1–4.3, $\partial f(x)$ and $\partial g(y)$ denote subdifferentials in the ambient spaces $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, respectively. If they were instead defined relative to $\mathcal{X}$ and $\mathcal{Y}$, their values at boundary points would include contributions from the normal cones $N_{\mathcal{X}}(x)$ and $N_{\mathcal{Y}}(y)$. The range conditions R_x and R_y would then unintentionally restrict these normal cones. The strong-convexity and Hölder conditions are imposed only on $\mathcal{X}$ and $\mathcal{Y}$. The feasible sets enter the algorithm through projections and the optimality conditions (28) through their normal cones.

Ambient smoothness for active anchor terms. Whenever a result below uses the coupling-induced regularizer p_3 , we make the following additional assumption for each of its active anchor terms. If $\nu_x = 1$, then f is differentiable on $\mathbb{R}^{d_x}$ and ∇f is H_x -Lipschitz on $\mathbb{R}^{d_x}$; if $\nu_y = 1$, then g is differentiable on $\mathbb{R}^{d_y}$ and ∇g is H_y -Lipschitz on $\mathbb{R}^{d_y}$. Thus the same constants used in the decomposition are valid in the ambient spaces.

For input $z_{\text{in}} = (x_{\text{in}}, y_{\text{in}})$, choose oracle tolerances δ_f, δ_g , with $\delta_f > 0$ if $\nu_x < 1$ and $\delta_g > 0$ if $\nu_y < 1$. We define

$$L_x := \begin{cases} \left[\frac{1 - \nu_x}{2(1 + \nu_x)\delta_f} \right]^{\frac{1-\nu_x}{1+\nu_x}} H_x^{\frac{2}{1+\nu_x}}, & 0 \leq \nu_x < 1, \\ H_x, & \nu_x = 1, \end{cases} \quad (18)$$

$$L_y := \begin{cases} \left[\frac{1 - \nu_y}{2(1 + \nu_y)\delta_g} \right]^{\frac{1-\nu_y}{1+\nu_y}} H_y^{\frac{2}{1+\nu_y}}, & 0 \leq \nu_y < 1, \\ H_y, & \nu_y = 1. \end{cases} \quad (19)$$

When $\nu_x = 1$, the ambient-smoothness convention makes H_x an ambient-space Lipschitz constant of ∇f whenever its anchor term is active, and we write $L_x := H_x$. Similarly, when $\nu_y = 1$, we write $L_y := H_y$.

For convex h with an ambient L_h -Lipschitz gradient, the standard self-bounding inequality gives, for all u, v ,

$$\|\nabla h(u) - \nabla h(v)\|^2 \leq 2L_h(h(u) - h(v) - \langle \nabla h(v), u - v \rangle). \quad (20)$$

Smoothness restricted to an arbitrary closed convex set is insufficient for (20).

We define

$$\beta_x := \frac{\mathbf{1}_{\{\nu_y=1\}}}{4L_y}, \quad \beta_y := \frac{\mathbf{1}_{\{\nu_x=1\}}}{4L_x}. \quad (21)$$

We use the following effective strong convexity parameters:

$$\delta_x := \mu_x + 4\beta_x \mu_{xy}^2, \quad \delta_y := \mu_y + 4\beta_y \mu_{yx}^2. \quad (22)$$

Assuming that $\min\{\delta_x, \delta_y\} > 0$, we define

$$\kappa_x = \frac{L_x}{\delta_x}, \quad \kappa_y = \frac{L_y}{\delta_y}, \quad \kappa_{xy} = \frac{L_{xy}^2}{\delta_x \delta_y}, \quad (23)$$

and

$$\tilde{H}_x = H_x \delta_x^{-(1+\nu_x)/2}, \quad \tilde{H}_y = H_y \delta_y^{-(1+\nu_y)/2}. \quad (24)$$

In the smooth case, Borodich et al. (2025) obtain separated complexities $\tilde{\mathcal{O}}(\sqrt{\kappa_x})$, $\tilde{\mathcal{O}}(\sqrt{\kappa_y})$, and $\tilde{\mathcal{O}}(\sqrt{\kappa_{xy}})$ for ∇f , ∇g , and products with $\mathbf{B}$, $\mathbf{B}^\top$, respectively, where $\tilde{\mathcal{O}}$ suppresses logarithms in $1/\epsilon$ and the problem parameters. We extend these guarantees to Hölder-continuous f and g .

We set $n = 3$ and

$$\begin{aligned} p_1(x, y) &= f(x), \quad p_2(x, y) = g(y), \\ p_3(x, y) &= \frac{\beta_x}{2} \|\mathbf{B}x - g'(y_{\text{in}})\|^2 + \frac{\beta_y}{2} \|\mathbf{B}^\top y + f'(x_{\text{in}})\|^2. \end{aligned} \quad (25)$$

and the operators

$$\begin{aligned} Q_1(x, y) &= 0, \quad Q_2(x, y) = 0, \\ Q_3(x, y) &= \begin{pmatrix} \mathbf{O}_{d_x} & \mathbf{B}^\top \\ -\mathbf{B} & \mathbf{O}_{d_y} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \end{aligned} \quad (26)$$

We define the norm matrix $\mathbf{P} \in \mathbb{S}_{++}^{d_z}$ as

$$\mathbf{P} = \text{diag}(\delta_x \mathbf{I}_{d_x}, \delta_y \mathbf{I}_{d_y}). \quad (27)$$

Algorithm 4 in Appendix A specializes Algorithm 1 to Problem (1). Since level i is evaluated $\prod_{j \leq i} T_j$ times and the additive decomposition is permutation invariant, we place the cheapest component outermost.

Let $\mathcal{S} \neq \emptyset$ be the solution set of Problem (1). By Lemma G.1 we may fix $z^* = (x^*, y^*) \in \mathcal{S}$ whose x -block satisfies $x_{\text{in}} - x^* \in \text{range}(\mathbf{B}^\top)$ whenever R_x is invoked, and whose y -block satisfies $y_{\text{in}} - y^* \in \text{range}(\mathbf{B})$ whenever R_y is invoked; on a block whose range condition is not invoked no restriction is placed on z^* , and Assumption 4.3 supplies the unrestricted bound $\lambda_{\min}$ there. By first-order optimality there are $f'(x^*) \in \partial f(x^*)$, $g'(y^*) \in \partial g(y^*)$ and normal vectors $\xi \in N_{\mathcal{X}}(x^*)$, $\zeta \in N_{\mathcal{Y}}(y^*)$ with

$$f'(x^*) + \mathbf{B}^\top y^* + \xi = 0, \quad g'(y^*) - \mathbf{B}x^* + \zeta = 0; \quad (28)$$

we fix such a choice. Define the Bregman divergences

$$\begin{aligned} D_f(x, x^*) &= f(x) - f(x^*) - \langle f'(x^*), x - x^* \rangle, \\ D_g(y, y^*) &= g(y) - g(y^*) - \langle g'(y^*), y - y^* \rangle. \end{aligned} \quad (29)$$

We denote by N_f , N_g , and N_B the numbers of evaluations of f' , of g' , and of matrix-vector products with $\mathbf{B}$ or $\mathbf{B}^\top$. The next theorem gives a one-step contraction for a single call of Algorithm 1 initialized at z_{in} ; the proof is in Appendix H.

Theorem 4.4. *Let Assumptions 4.1–4.3 hold, together with the ambient-smoothness condition above for every active smooth anchor term. Thus (20) is available whenever it is used below. Assume $\min\{\delta_x, \delta_y\} > 0$ and $\beta_x \delta_y \leq \frac{1}{4}, \beta_y \delta_x \leq \frac{1}{4}$. Assume further that the normal vectors in the optimality conditions (28) satisfy $\zeta = 0$ if $\beta_x > 0$ and $\xi = 0$ if $\beta_y > 0$. This condition is automatic at interior saddle points and vacuous when $\nu_x, \nu_y < 1$ (Remark 2). Let Ω_{in} be any upper bound satisfying $\Psi(z_{\text{in}}) \leq \Omega_{\text{in}}$, where*

$$\Psi(z) := \|z - z^*\|_{\mathbf{P}}^2 + 12 D_f(x, x^*) + 12 D_g(y, y^*). \quad (30)$$

Associate with p_1, p_2, p_3 the constants and componentwise budgets

$$\begin{aligned} L_1 &= \kappa_x, & L_2 &= \kappa_y, & L_3 &= \kappa_{xy}, & M_1 &= M_2 = 0, & M_3 &= \sqrt{\kappa_{xy}}, \\ R_1 &= \max\{\hat{R}_1, 1\}, & R_2 &= \max\{\hat{R}_2, 1\}, & R_3 &= \max\{\sqrt{\kappa_{xy}}, 1\}. \end{aligned} \quad (31)$$

where $\hat{R}_1 := \tilde{H}_x^{\frac{2}{1+3\nu_x}} \Omega_{\text{in}}^{\frac{\nu_x-1}{1+3\nu_x}}$ and $\hat{R}_2 := \tilde{H}_y^{\frac{2}{1+3\nu_y}} \Omega_{\text{in}}^{\frac{\nu_y-1}{1+3\nu_y}}$. Choose δ_f, δ_g as in the proof and apply Algorithm 1 to (25)–(26) with the components sorted along the levels, i.e. with $(p_{\pi(i)}, Q_{\pi(i)})$ at level i for a permutation π of $\{1, 2, 3\}$ satisfying $R_{\pi(1)} \leq R_{\pi(2)} \leq R_{\pi(3)}$. Then the output z_{out} satisfies $\Psi(z_{\text{out}}) \leq \frac{3}{4} \Omega_{\text{in}}$, with $N_f = \mathcal{O}(R_1)$, $N_g = \mathcal{O}(R_2)$, and $N_B = \mathcal{O}(R_3)$.

We apply a standard restarting scheme. This yields the following global complexity bound. The proof is given in Appendix I.

Corollary 4.5. *Let the assumptions of Theorem 4.4 hold. Given $z^0 \in \mathcal{Z}$ and $\varepsilon \in (0, \Psi(z^0)]$, run the restarted Sliding Algorithm with $\Omega_s := (3/4)^S \Psi(z^0)$ for $S := \left\lceil \frac{\log(\Psi(z^0)/\varepsilon)}{\log(4/3)} \right\rceil$ restarts. Then the output z^S satisfies $\Psi(z^S) \leq \varepsilon$, and the total evaluation complexity is given in Table 1.*

Oracle	Complexity
f' evaluations	$\begin{cases} \mathcal{O}\left(\tilde{H}_x^{\frac{2}{1+3\nu_x}} \varepsilon^{\frac{\nu_x-1}{1+3\nu_x}}\right), & 0 \leq \nu_x < 1, \\ \tilde{\mathcal{O}}(\sqrt{\kappa_x}), & \nu_x = 1 \end{cases}$
g' evaluations	$\begin{cases} \mathcal{O}\left(\tilde{H}_y^{\frac{2}{1+3\nu_y}} \varepsilon^{\frac{\nu_y-1}{1+3\nu_y}}\right), & 0 \leq \nu_y < 1, \\ \tilde{\mathcal{O}}(\sqrt{\kappa_y}), & \nu_y = 1 \end{cases}$
$\mathbf{B}, \mathbf{B}^\top$ products	$\tilde{\mathcal{O}}(\sqrt{\kappa_{xy}})$

Table 1: Evaluation complexity for the bilinear saddle-point problem (1). The notation $\tilde{\mathcal{O}}$ hides logarithmic factors in Ψ_0/ε .

Table 1 interpolates componentwise between Hölder and smooth strongly convex rates. For $0 \leq \nu_x < 1$, $N_f = \mathcal{O}\left(\tilde{H}_x^{\frac{2}{1+3\nu_x}} \varepsilon^{-\frac{1-\nu_x}{1+3\nu_x}}\right)$, with the analogous bound for N_g . At $\nu_x = 1$ or $\nu_y = 1$, the corresponding rate becomes $\tilde{\mathcal{O}}(\sqrt{\kappa_x})$ or $\tilde{\mathcal{O}}(\sqrt{\kappa_y})$, with logarithmic accuracy dependence. Coupling requires $\tilde{\mathcal{O}}(\sqrt{\kappa_{xy}})$ products independently of ν_x, ν_y . Thus $(\nu_x, \nu_y) = (1, 1)$ recovers the smooth bilinear complexity.

The case $\delta_x = \delta_y = 0$ We finally discuss the case $\delta_x = \delta_y = 0$. Assuming that $\mathcal{Z}$ is bounded, we apply Theorem 3.5 directly with respect to the Euclidean product norm. Since no strong convexity is available, we set $p_3 \equiv 0$. Consequently, this result requires only the Hölder conditions on $\mathcal{X}, \mathcal{Y}$, not the ambient-smoothness condition for active anchor terms.

Thus only the bilinear coupling remains in Q_3 , which is L_{xy} -Lipschitz. Hence, with $\Omega := \sup_{z \in \mathcal{Z}} \|z^0 - z\|^2$, Corollary 3.6 gives $N_f = \mathcal{O}\left((H_x \Omega^{(1+\nu_x)/2} / \varepsilon)^{2/(1+3\nu_x)}\right)$. Analogously, $N_g = \mathcal{O}\left((H_y \Omega^{(1+\nu_y)/2} / \varepsilon)^{2/(1+3\nu_y)}\right)$. The coupling complexity is $N_B = \mathcal{O}(L_{xy} \Omega / \varepsilon)$. These bounds describe the genuinely non-strongly-convex–non-strongly-concave regime. The H_x - and H_y -terms match the known lower-bound scaling for first-order convex optimization with Hölder-continuous gradients (Nemirovskij and Yudin 1983; Nesterov 2013; 2015), interpolating from $\mathcal{O}(\varepsilon^{-2})$ at $\nu = 0$ to the accelerated smooth rate $\mathcal{O}(\varepsilon^{-1/2})$ at $\nu = 1$. The bilinear term $\mathcal{O}(L_{xy} \Omega / \varepsilon)$ matches the optimal Mirror-Prox rate for monotone Lipschitz variational inequalities and is consistent with lower bounds for bilinear saddle-point problems (Nemirovski 2004; Zhang et al. 2022).

Mixed cases. We next consider the regimes $\delta_x > 0, \delta_y = 0$ and $\delta_x = 0, \delta_y > 0$, where only one effective strong-convexity parameter is positive. In contrast to the fully degenerate case $\delta_x = \delta_y = 0$, we do not discard the coupling-induced regularization p_3 in (25). Instead, we add regularization only in the block whose effective curvature is zero and rebuild p_3 using the regularized component's Hölder/smoothness constant and anchor whenever that part of p_3 is active. The corrected parameters and compatibility conditions are given in Appendix J. After simplification, the rates are summarized in Table 2.

Oracle	Complexity
f' evaluations	$\begin{cases} \mathcal{O}\left(\tilde{H}_x^{\frac{2}{1+3\nu_x}} \varepsilon^{\frac{\nu_x-1}{1+3\nu_x}}\right), & 0 \leq \nu_x < 1, \\ \tilde{\mathcal{O}}(\sqrt{\kappa_x}), & \nu_x = 1 \end{cases}$
g' evaluations	$\begin{cases} \mathcal{O}\left(\left(\frac{H_y \Omega_y^{(1+\nu_y)/2}}{\varepsilon}\right)^{\frac{2}{1+3\nu_y}}\right), & 0 \leq \nu_y < 1, \\ \tilde{\mathcal{O}}\left(\sqrt{\frac{H_y \Omega_y}{\varepsilon}}\right), & \nu_y = 1 \end{cases}$
$\mathbf{B}, \mathbf{B}^\top$ products	$\tilde{\mathcal{O}}\left(L_{xy} \sqrt{\frac{\Omega_y}{\delta_x \varepsilon}}\right)$

Table 2: Mixed-regime complexity for $\delta_x > 0, \delta_y = 0$.

4.1 Stochastic oracle setting

We allow $0 \leq \nu_x, \nu_y \leq 1$ and access f and g through conditionally unbiased stochastic (sub)gradient oracles with Euclidean conditional variances bounded by σ_x^2 and σ_y^2 . The bilinear products are evaluated exactly. At $\nu_q = 0$, the oracle mean is a measurable selected subgradient; the precise filtration and moment conditions are stated in Assumption K.1. The parameter choices use valid Hölder and variance bounds.

The stochastic recursive estimate combines the deterministic Hölder model error with the martingale error. Optimizing their parameters separately gives the componentwise terms

$$\frac{H_i R^{1+\nu_i}}{N_i^{(1+3\nu_i)/2}} + \frac{M_i R^2}{N_i} + \frac{\sigma_i R}{\sqrt{N_i}},$$

up to constants depending only on the number of levels. Here R^2 bounds the expected initial squared distance to a fixed comparator; for the uniform gap we use $R^2 = \Omega$, defined below. The uniform version is proved before taking expectation of the supremum in Theorem K.8; a fixed-comparator estimate alone does not provide a uniform gap guarantee.

Table 3 summarizes two consequences. For the bounded convex–concave regime, let

$$\Omega := \sup_{z \in \mathcal{Z}} \|z^0 - z\|^2 < \infty, \quad \mathcal{C}_q(\varepsilon) := \left(\frac{H_q \Omega^{(1+\nu_q)/2}}{\varepsilon}\right)^{2/(1+3\nu_q)}, \quad q \in \{x, y\}.$$

For the strongly convex–strongly concave regime, assume $\mu_x, \mu_y > 0$, set $\mathbf{P} = \text{diag}(\mu_x I, \mu_y I)$, and let $\Gamma_0 > 0$ be a known upper bound on $\Psi(z^0)$, where Ψ is defined by (30). For $0 < \varepsilon \leq \Gamma_0$, write

$$\ell_\varepsilon := 1 + \log \frac{\Gamma_0}{\varepsilon}, \quad \theta_q := \frac{1 - \nu_q}{1 + 3\nu_q}, \quad \mathcal{K}_q := H_q^{2/(1+3\nu_q)} \mu_q^{-(1+\nu_q)/(1+3\nu_q)},$$

and define the deterministic restart budget

$$\mathcal{D}_q(\varepsilon) := \begin{cases} \ell_\varepsilon + \frac{\mathcal{K}_q \varepsilon^{-\theta_q}}{1 - 2^{-\theta_q}}, & 0 \leq \nu_q < 1, \\ \left(1 + \sqrt{H_q / \mu_q}\right) \ell_\varepsilon, & \nu_q = 1. \end{cases}$$

The finite geometric sum in Theorem K.11 gives a sharper bound near $\nu_q = 1$. For each fixed $\nu_q < 1$, the factor $(1 - 2^{-\theta_q})^{-1}$ is a constant and the Hölder accuracy exponent is $(1 - \nu_q)/(1 + 3\nu_q)$.

Oracle	Bounded convex-concave	$\mu_x, \mu_y > 0$
f' samples	$\mathcal{O}\left(1 + \mathcal{C}_x(\varepsilon) + \frac{\sigma_x^2 \Omega}{\varepsilon^2}\right)$	$\mathcal{O}\left(\mathcal{D}_x(\varepsilon) + \frac{\sigma_x^2}{\mu_x \varepsilon}\right)$
g' samples	$\mathcal{O}\left(1 + \mathcal{C}_y(\varepsilon) + \frac{\sigma_y^2 \Omega}{\varepsilon^2}\right)$	$\mathcal{O}\left(\mathcal{D}_y(\varepsilon) + \frac{\sigma_y^2}{\mu_y \varepsilon}\right)$
$\mathbf{B}, \mathbf{B}^\top$ products	$\mathcal{O}\left(1 + \frac{L_{xy} \Omega}{\varepsilon}\right)$	$\mathcal{O}\left(\left(1 + \frac{L_{xy}}{\sqrt{\mu_x \mu_y}}\right) \ell_\varepsilon\right)$

Table 3: Stochastic Hölder oracle complexities for $0 \leq \nu_x, \nu_y \leq 1$. The bounded-domain column guarantees $\mathbb{E}[\text{Gap}(z^{\text{out}})] \leq \varepsilon$; the strongly convex column guarantees $\mathbb{E}[\Psi(z^S)] \leq \varepsilon$. The latter uses intrinsic curvature $\mu_x, \mu_y > 0$ and no anchor regularizer. The quantities $\mathcal{C}_q, \mathcal{D}_q$ are defined above.

For bounded $\mathcal{Z}$, we set $p_3 \equiv 0$, use the Euclidean product norm, and obtain the uniform expected primal–dual gap in Theorem K.10. This result includes the degenerate regime $\delta_x = \delta_y = 0$. Under intrinsic strong convexity, we again set $p_3 \equiv 0$, now in the weighted norm above, and restart Algorithm 1. The fixed saddle comparator controls Ψ , giving

$$\mathbb{E}[\Psi(z^S)] \leq \varepsilon. \quad (32)$$

No anchor samples are used, so this guarantee has no anchor noise floor. The variance terms are free of restart logarithms. At every call or restart, the recursive levels are ordered by their complete target budgets, including the variance contributions; this preserves the separate coupling count.

The strongly convex theorem imposes Hölder regularity on the feasible sets, as in Assumptions 4.1–4.2. Global strong convexity and global Hölder regularity with exponent below one cannot hold simultaneously on an unbounded domain. The stochastic strongly convex column uses $\mu_x, \mu_y > 0$; it does not assert the same result for curvature supplied only by the coupling. At $\nu_x = \nu_y = 1$, the bounds recover the separated smooth stochastic rates, and at $\sigma_x = \sigma_y = 0$ they recover the corresponding deterministic Hölder rates.

5 Experiments

We report exact gaps and oracle counts on synthetic and tomography instances. Appendix L gives the setup and tuning.

Synthetic results. Across the tested ν_x , the fitted N_f exponent is within 4% of $2/(1 + 3\nu_x)$, with $R^2 \geq 0.9999$ and per-seed spread below 0.04. All 90 runs satisfy (17), attaining 8%–27% of the bound (Table 5). On $(\nu_x, \nu_y) = (0.75, 0.25)$, recursion reaches gap below 0.1 with 48 f' -calls, versus 707 for the lockstep ablation and 2,477 for Universal Mirror-Prox. Relative to the two-level ablation, it keeps 48 f' -calls and trades 1,584 $\rightarrow$ 1,920 g' -calls for 6,336 $\rightarrow$ 3,840 $\mathbf{B}, \mathbf{B}^\top$ -products (Figure 2 in Appendix L). The appendix also covers other regularities. On a smooth instance, sliding uses 4.7–8.5 $\times$ fewer gradients but 18–40 $\times$ more products, with a gradient-to-product cost crossover of 22–44 (Appendix L.1).

Tomography. At relative gap $5 \cdot 10^{-4}$, we compare against Mirror-Prox, Accelerated Mirror-Prox, and Mirror-Prox Sliding (MPS) on five unseen instances. Standard Huber–TV is the negative control. Figure 1 shows that MPS reaches the target first with diagonal g' , whereas recursion has the lowest weighted cost with nonlocal and nonlocal Hölder g' . Across the five instances, MPS wins 2.43 vs. 3.18 s in the diagonal regime. A nonlocal graph makes g' 48 $\times$ costlier. Recursion uses 480 rather than 3840 g' -calls and takes 3.96 rather than 8.39 s with the same 160 data gradients, reaching 30.81 dB at 128². Adding a $\nu_y = \frac{1}{2}$ power term makes g' genuinely Hölder. The times are 3.98 s for recursion, 8.56 s for MPS, and 10.8 s for Generalized Universal Mirror-Prox at its target-matched 576-iteration budget. Appendix L.2 gives diagnostics and curves. Coupling calls exceed g' -calls by three orders of magnitude, so tomography does not reveal the ν_y -dependent exponent.

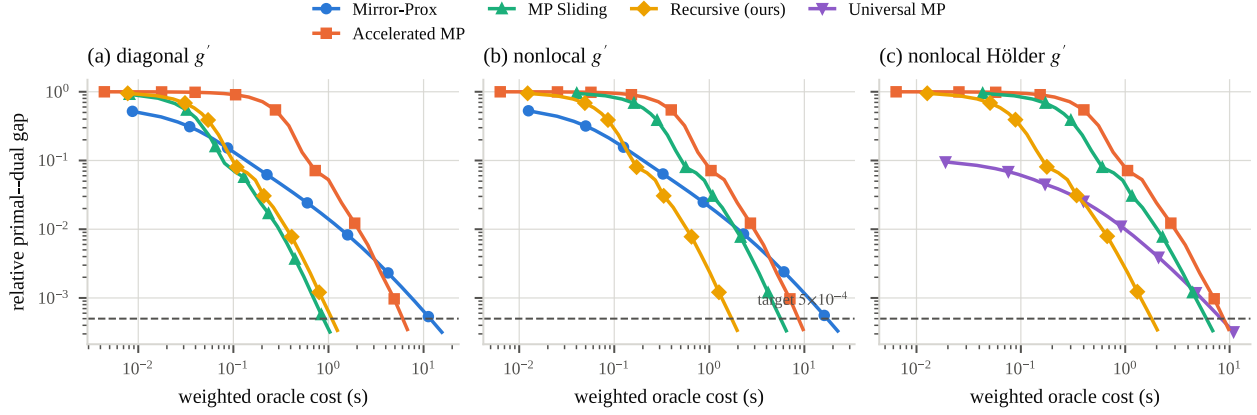


Figure 1: Tomographic convergence on the first unseen instance versus calibrated cost $c_f N_f + c_g N_g + c_B N_B$, where N_f, N_g, N_B count f', g' , and $B/B^\top$ calls and c_f, c_g, c_B are their measured times per call. Panels (a)–(c) use diagonal, nonlocal, and nonlocal Hölder g' , with $(\nu_x, \nu_y) = (1, 1), (1, 1), (1, \frac{1}{2})$, respectively. The dashed line marks the relative-gap target $5 \cdot 10^{-4}$.

6 Conclusions

We extended recursive componentwise sliding to componentwise inexact first-order oracles and applied it to Hölder-smooth bilinear saddle-point problems. The resulting bounds charge f', g' , and $B, B^\top$ separately, with the two gradient counts governed by their respective (H_x, ν_x) and (H_y, ν_y) . At $\nu_x = \nu_y = 1$, they reduce to the known optimal smooth rates. The analysis also covers nonstrongly convex, mixed, and strongly curved deterministic regimes under the stated conditions. For stochastic function oracles, the Hölder model error and the variance contribution separate in the recursive bound. This gives a uniform expected-gap guarantee on bounded domains for all $\nu_x, \nu_y \in [0, 1]$. With intrinsic strong convexity $\mu_x, \mu_y > 0$, restarts of the decomposition with $p_3 = 0$ give an expected Lyapunov guarantee and separate $\sigma_q^2/(\mu_q \varepsilon)$ variance costs without an anchor noise floor. Hölder regularity is required on the feasible sets, and the coupling products remain exact. Synthetic experiments recover the predicted componentwise exponents. The tomography benchmark favors recursive sliding for expensive nonlocal g' , including the Hölder case, and MPS for cheap diagonal g' . Future work includes parameter-free recursive schedules, stochastic coupling oracles, and heterogeneous-exponent lower bounds.

References

- E. Borodich, A. Gasnikov, and D. Kovalev. On linear convergence in smooth convex-concave bilinearly-coupled saddle-point optimization: Lower bounds and optimal algorithms. In *Forty-second International Conference on Machine Learning*, 2025.
- A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.
- Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.
- O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.
- S. S. Du and W. Hu. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 196–205. PMLR, 2019.
- D. Dvinskikh and A. Gasnikov. Decentralized and parallel primal and dual accelerated methods for stochastic convex programming problems. *Journal of Inverse and Ill-posed Problems*, 29(3):385–405, 2021.
- P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn’s algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.
- F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.
- A. Ibrahim, W. Azizian, G. Gidel, and I. Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International conference on machine learning*, pages 4583–4593. PMLR, 2020.
- A. Juditsky and A. Nemirovski. Solving variational inequalities with monotone operators on domains given by linear minimization oracles. *Mathematical Programming*, 156(1):221–256, 2016.
- G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

- D. Kovalev, A. Salim, and P. Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems*, 33:18342–18352, 2020.
- D. Kovalev, A. Gasnikov, and P. Richtárik. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. *Advances in Neural Information Processing Systems*, 35:21725–21737, 2022.
- G. Lan. Gradient sliding for composite optimization. *Mathematical Programming*, 159(1):201–235, 2016.
- G. Lan and Y. Ouyang. Mirror-prox sliding methods for solving a class of monotone variational inequalities. *arXiv preprint arXiv:2111.00996*, 2021.
- G. Lan and Y. Ouyang. Accelerated gradient sliding for structured convex optimization. *Computational Optimization and Applications*, 82(2):361–394, 2022.
- G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, 171(1):167–215, 2018.
- G. Lan, S. Lee, and Y. Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, 180(1):237–284, 2020.
- H. Luo. A universal accelerated primal–dual method for convex optimization problems. *Journal of Optimization Theory and Applications*, 201(1):280–312, 2024.
- A. Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.
- A. S. Nemirovskij and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.
- Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.
- Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.
- Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.
- A. Rodomanov, A. Kavis, Y. Wu, K. Antonakopoulos, and V. Cevher. Universal gradient methods for stochastic convex optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42620–42646. PMLR, 2024. URL <https://proceedings.mlr.press/v235/rodomanov24a.html>.
- K. Scaman, F. Bach, S. Bubeck, L. Massoulié, and Y. T. Lee. Optimal algorithms for non-smooth distributed optimization in networks. *Advances in Neural Information Processing Systems*, 31, 2018.
- F. Stonyakin, A. Gasnikov, P. Dvurechensky, A. Titov, and M. Alkousa. Generalized mirror prox algorithm for monotone variational inequalities: Universality and inexact oracle. *Journal of Optimization Theory and Applications*, 194(3):988–1013, 2022.
- V. Tominin, Y. Tominin, E. Borodich, D. Kovalev, A. Gasnikov, and P. Dvurechensky. On accelerated methods for saddle-point problems with composite structure. *arXiv preprint arXiv:2103.09344*, 2021.
- Y. Wu, Y. Ouyang, Z. Zhang, and Q. Luo. Universal and parameter-free gradient sliding for composite optimization. *arXiv preprint arXiv:2603.23492*, 2026.
- A. Yurtsever, Q. Tran Dinh, and V. Cevher. A universal primal-dual convex optimization framework. In *Advances in Neural Information Processing Systems*, volume 28, 2015.
- J. Zhang, M. Hong, and S. Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1):901–935, 2022.

A The Sliding Algorithm and Explicit Special Cases

The general recursive scheme referred to throughout the paper is Algorithm 1. Its one- and two-level cases and the three-level bilinear specialization follow it.

Algorithm 1 Sliding Algorithm

```

1: Input:  $z_{\text{in}} \in \mathcal{Z}$ .
2: Parameters:  $\{\alpha_t\}, \{T_i\}_{i=1}^n, \{\eta_t^k\}$ 
3: for  $k = 1, \dots, n$  do
4:    $z_{\mathbf{0}_k}^k = z_{\text{in}}$ , where  $\mathbf{0}_k = (0, \dots, 0) \in \mathbb{N}^k$ .
5: end for
6: for  $i = 1, \dots, n$  do
7:    $p_i^{0, \emptyset}(z) \equiv p_i(z)$ .
8: end for
9:  $z_{\text{out}} = \text{RECURSIVEPROCEDURE}(0, \emptyset)$ 
10: return  $z_{\text{out}}$ 
11:
12: function RECURSIVEPROCEDURE( $k, \mathbf{t}$ )  $\triangleright z_{\mathbf{t}, 0, \dots, 0}^\ell$  is initialized for all  $\ell > k$ 
13:   if  $k = n$  then
14:     return  $\arg \min_{z \in \mathcal{Z}} \sum_{i=1}^n p_i^{n, \mathbf{t}}(z)$ 
15:   end if
16:   Set  $\bar{z}_{\mathbf{t}, 0}^{k+1} = z_{\mathbf{t}, 0}^{k+1}$ .
17:   for  $j = 0, \dots, T_{k+1} - 1$  do
18:     Set  $\mathbf{t}' = (\mathbf{t}, j)$ ,  $\mathbf{t}'' = (\mathbf{t}, j + 1)$ 
19:     for  $i = 1, \dots, n$  do
20:        $\hat{p}_i^{k+1, \mathbf{t}'}(z) = \begin{cases} \alpha_j^{-1} p_i^{k, \mathbf{t}}(\alpha_j z + (1 - \alpha_j) \bar{z}_{\mathbf{t}', 0}^{k+1}), & i > k, \\ p_i^{k, \mathbf{t}}(z), & i \leq k, \end{cases}$ 
21:     end for
22:     for  $i = 1, \dots, n$  do
23:        $p_i^{k+1, \mathbf{t}'}(z) = \begin{cases} \frac{\eta_{\mathbf{t}'}^{k+1}}{2} \|z - z_{\mathbf{t}'}^{k+1}\|_{\mathbf{P}}^2 + \langle \tilde{\nabla} \hat{p}_{k+1}^{k+1, \mathbf{t}'}(z_{\mathbf{t}'}^{k+1}) + Q_{k+1}(z_{\mathbf{t}'}^{k+1}), z \rangle, & i = k + 1, \\ \hat{p}_i^{k+1, \mathbf{t}'}(z), & i \neq k + 1, \end{cases}$ 
24:     end for
25:      $\bar{z}_{\mathbf{t}'}^{k+1} = \text{RECURSIVEPROCEDURE}(k + 1, \mathbf{t}')$ 
26:      $\bar{z}_{\mathbf{t}''}^{k+1} = \alpha_j \bar{z}_{\mathbf{t}'}^{k+1} + (1 - \alpha_j) \bar{z}_{\mathbf{t}'}^{k+1}$ 
27:      $z_{\mathbf{t}''}^{k+1} = \arg \min_{z \in \mathcal{Z}} \left\{ \langle Q_{k+1}(\bar{z}_{\mathbf{t}'}^{k+1}) - Q_{k+1}(z_{\mathbf{t}'}^{k+1}), z \rangle + \frac{\eta_{\mathbf{t}''}^{k+1}}{2} \|z - \bar{z}_{\mathbf{t}'}^{k+1}\|_{\mathbf{P}}^2 \right\}$ 
28:   if  $k + 2 \leq n$  then
29:     for  $\ell = k + 2, \dots, n$  do
30:       Set  $z_{\mathbf{t}'', 0, \dots, 0}^\ell = z_{\mathbf{t}', T_{k+2}, 0, \dots, 0}^\ell$ .
31:     end for
32:   end if
33: end for
34: return  $\bar{z}_{\mathbf{t}, T_{k+1}}^{k+1}$ 
35: end function

```

Algorithm 2 One-Level Sliding ($n = 1$)

```

1: Choose  $z_0 \in \mathcal{Z}$  and set  $\bar{z}^0 = z_0$ .
2: for  $t = 0, \dots, T - 1$  do
3:    $z^t = \alpha_t z_t + (1 - \alpha_t) \bar{z}^t$ 
4:    $\tilde{z}_t = \arg \min_{z \in \mathcal{Z}} \left\{ \langle \tilde{\nabla} p(z^t) + Q(z_t), z \rangle + \frac{\beta_t}{2} \|z - z_t\|_{\mathbf{P}}^2 \right\}$ 
5:    $z_{t+1} = \arg \min_{z \in \mathcal{Z}} \left\{ \langle Q(\tilde{z}_t) - Q(z_t), z \rangle + \frac{\beta_t}{2} \|z - \tilde{z}_t\|_{\mathbf{P}}^2 \right\}$ 
6:    $\bar{z}^{t+1} = \alpha_t \tilde{z}_t + (1 - \alpha_t) \bar{z}^t$ 
7: end for
8: return  $\bar{z}^T$ 

```

Algorithm 3 Two-Level Sliding (special case $n = 2$)

```

1: Choose  $z_0 \in \mathcal{Z}$  and set  $\bar{z}_0 = z_0^0 = z_0$ .
2: for  $k = 0, \dots, T_1 - 1$  do
3:    $\tilde{z}_k = \alpha_k z_k + (1 - \alpha_k) \bar{z}_k$ 
4:   Set  $\bar{z}_k^0 = z_k^0$ 
5:   for  $t = 0, \dots, T_2 - 1$  do
6:      $\hat{z}_k^t = \alpha_t z_k^t + (1 - \alpha_t) \bar{z}_k^t$ 
7:      $\underline{z}_k^t = \alpha_k \hat{z}_k^t + (1 - \alpha_k) \tilde{z}_k$ 
8:      $\tilde{z}_k^t = \arg \min_{z \in \mathcal{Z}} \left\{ \langle \tilde{\nabla} p_1(\underline{z}_k) + \tilde{\nabla} p_2(\underline{z}_k^t) + Q_1(z_k) + Q_2(z_k^t), z \rangle + \frac{\beta_k}{2} \|z - z_k\|_{\mathbf{P}}^2 + \frac{\eta_k^t}{2} \|z - z_k^t\|_{\mathbf{P}}^2 \right\}$ 
9:      $\bar{z}_k^{t+1} = \alpha_t \tilde{z}_k^t + (1 - \alpha_t) \bar{z}_k^t$ 
10:     $z_k^{t+1} = \arg \min_{z \in \mathcal{Z}} \left\{ \langle Q_2(\tilde{z}_k^t) - Q_2(z_k^t), z \rangle + \frac{\eta_k^t}{2} \|z - \tilde{z}_k^t\|_{\mathbf{P}}^2 \right\}$ 
11:   end for
12:   Set  $\tilde{z}_k = \bar{z}_k^{T_2}$ 
13:    $\bar{z}_{k+1} = \alpha_k \tilde{z}_k + (1 - \alpha_k) \bar{z}_k$ 
14:    $z_{k+1} = \arg \min_{z \in \mathcal{Z}} \left\{ \langle Q_1(\tilde{z}_k) - Q_1(z_k), z \rangle + \frac{\beta_k}{2} \|z - \tilde{z}_k\|_{\mathbf{P}}^2 \right\}$ 
15:   Set  $z_{k+1}^0 = \bar{z}_{k+1}^{T_2}$ 
16: end for
17: return  $\bar{z}_{T_1}$ 

```

Algorithm 4 Three-Level Sliding for the Bilinear Saddle-Point Problem

```

1: Choose  $z_0 = (x_0, y_0) \in \mathcal{Z}$  and set

$$\bar{z}_0 = z_0^0 = z_0^{0,0} = z_0.$$


2: for  $k = 0, \dots, T_1 - 1$  do
3:    $\tilde{z}_k = \alpha_k z_k + (1 - \alpha_k) \bar{z}_k.$ 
4:   Set  $\bar{z}_k^0 = z_k^0.$ 
5:   for  $t = 0, \dots, T_2 - 1$  do
6:      $\hat{z}_k^t = \alpha_t z_k^t + (1 - \alpha_t) \bar{z}_k^t.$ 
7:      $\underline{z}_k^t = \alpha_k \hat{z}_k^t + (1 - \alpha_k) \bar{z}_k.$ 
8:     Set  $\bar{z}_k^{t,0} = z_k^{t,0}.$ 
9:     for  $r = 0, \dots, T_3 - 1$  do
10:       $\hat{z}_k^{t,r} = \alpha_r z_k^{t,r} + (1 - \alpha_r) \bar{z}_k^{t,r}.$ 
11:       $\underline{z}_k^{t,r} = \alpha_k (\alpha_t \hat{z}_k^{t,r} + (1 - \alpha_t) \bar{z}_k^t) + (1 - \alpha_k) \bar{z}_k.$ 
12:      Write

$$\underline{z}_k = (\underline{x}_k, \underline{y}_k), \quad \underline{z}_k^t = (\underline{x}_k^t, \underline{y}_k^t).$$


13:      Define

$$G_k^{t,r} := \begin{pmatrix} f'(\underline{x}_k) + \beta_x \mathbf{B}^\top (\mathbf{B} \underline{x}_k^{t,r} - g'(y_{\text{in}})) + \mathbf{B}^\top y_k^{t,r} \\ g'(\underline{y}_k^t) + \beta_y \mathbf{B} (\mathbf{B}^\top \underline{y}_k^{t,r} + f'(x_{\text{in}})) - \mathbf{B} x_k^{t,r} \end{pmatrix}.$$


14:      
$$\tilde{z}_k^{t,r} = \arg \min_{z \in \mathcal{Z}} \left\{ \langle G_k^{t,r}, z \rangle + \frac{\lambda_k}{2} \|z - z_k\|_{\mathbf{P}}^2 + \frac{\eta_k^t}{2} \|z - z_k^t\|_{\mathbf{P}}^2 + \frac{\gamma_k^{t,r}}{2} \|z - z_k^{t,r}\|_{\mathbf{P}}^2 \right\}.$$


15:      
$$\bar{z}_k^{t,r+1} = \alpha_r \tilde{z}_k^{t,r} + (1 - \alpha_r) \bar{z}_k^{t,r}.$$


16:      
$$z_k^{t,r+1} = \arg \min_{z=(x,y) \in \mathcal{Z}} \left\{ \left\langle \begin{pmatrix} \mathbf{B}^\top (\tilde{y}_k^{t,r} - y_k^{t,r}) \\ -\mathbf{B} (\tilde{x}_k^{t,r} - x_k^{t,r}) \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \right\rangle + \frac{\gamma_k^{t,r}}{2} \|z - \tilde{z}_k^{t,r}\|_{\mathbf{P}}^2 \right\}.$$


17:     end for
18:     Set  $\bar{z}_k^t = \bar{z}_k^{t,T_3}.$ 
19:     
$$\bar{z}_k^{t+1} = \alpha_t \bar{z}_k^t + (1 - \alpha_t) \bar{z}_k^t.$$


20:     Set  $z_k^{t+1} = \bar{z}_k^t.$ 
21:     Initialize the next inner loop by setting

$$z_k^{t+1,0} = z_k^{t,T_3}.$$


22:   end for
23:   Set  $\bar{z}_k = \bar{z}_k^{T_2}.$ 
24:   
$$\bar{z}_{k+1} = \alpha_k \bar{z}_k + (1 - \alpha_k) \bar{z}_k.$$


25:   Set  $z_{k+1} = \bar{z}_k.$ 
26:   Initialize the next middle and inner loops by setting

$$z_{k+1}^0 = z_k^{T_2}, \quad z_{k+1}^{0,0} = z_k^{T_2,0}.$$


27: end for
28: return  $\bar{z}_{T_1}.$ 

```

B Applications of the Bilinear Saddle-Point Problem

The bilinear saddle-point formulation (1) arises naturally in a variety of settings. We describe several important examples below.

Linearly constrained optimization. Consider the problem of minimizing a convex function subject to linear equality constraints:

$$\min_{x \in \mathcal{X}} f(x) \quad \text{subject to} \quad \mathbf{A}x = b, \quad (33)$$

where $\mathbf{A} \in \mathbb{R}^{m \times d_x}$ and $b \in \mathbb{R}^m$. Introducing Lagrange multipliers $y \in \mathbb{R}^m$ yields the bilinear saddle-point problem $\min_{x \in \mathcal{X}} \max_{y \in \mathbb{R}^m} f(x) + \langle y, \mathbf{A}x \rangle - \langle y, b \rangle$, an instance of (1) with $\mathbf{B} = \mathbf{A}$, $g(y) = \langle y, b \rangle$, and $\mathcal{Y} = \mathbb{R}^m$. Adding a strongly convex regularizer $r(y)$ to the dual variable yields the augmented Lagrangian form, which also fits the framework.

Decentralized optimization over networks. In decentralized optimization (Lan et al. 2020; Scaman et al. 2018), m agents connected by a communication network collaboratively solve

$$\min_{x \in \mathcal{X}} \frac{1}{m} \sum_{i=1}^m f_i(x), \quad (34)$$

where each agent i has access only to its local objective f_i . Introducing local copies $\mathbf{x} = \text{col}(x_1, \dots, x_m)$ and a symmetric positive semidefinite *gossip matrix* $\mathbf{W} \in \mathbb{R}^{m \times m}$ with $\ker(\mathbf{W}) = \text{span}\{\mathbf{1}_m\}$, the consensus constraint $x_1 = \dots = x_m$ is equivalent to $(\mathbf{W} \otimes \mathbf{I}_{d_x})\mathbf{x} = 0$. Its Lagrangian dual is a bilinear saddle-point problem of the form (1) with $f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m f_i(x_i)$, $\mathbf{B} = \mathbf{W} \otimes \mathbf{I}_{d_x}$, and $g \equiv 0$.

Regularized optimal transport. Given discrete probability vectors $a \in \mathbb{R}^{n_1}$ and $b \in \mathbb{R}^{n_2}$, the regularized optimal transport problem (Dvurechensky et al. 2018) is

$$\begin{aligned} \min_{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}} \quad & \langle \mathbf{C}, \mathbf{X} \rangle + r(\mathbf{X}) \\ \text{subject to} \quad & \mathbf{X}\mathbf{1}_{n_2} = a, \\ & \mathbf{X}^\top \mathbf{1}_{n_1} = b. \end{aligned} \quad (35)$$

where $\mathbf{C}$ is the cost matrix and $r(\cdot)$ is a strongly convex regularizer (e.g., squared Frobenius norm). Introducing Lagrange multipliers $u \in \mathbb{R}^{n_1}$ and $v \in \mathbb{R}^{n_2}$ for the marginal constraints yields a bilinear saddle-point problem of the form (1) with $f(\mathbf{X}) = \langle \mathbf{C}, \mathbf{X} \rangle + r(\mathbf{X})$, $g(u, v) = \langle u, a \rangle + \langle v, b \rangle$, and the coupling operator $\mathbf{B}: \mathbf{X} \mapsto \text{col}(\mathbf{X}\mathbf{1}_{n_2}, \mathbf{X}^\top \mathbf{1}_{n_1})$.

Inverse problems. Recovering a signal $x \in \mathbb{R}^d$ from noisy measurements $b = \mathbf{A}x + \xi$ with an ℓ_p -penalty on a linear transform $\mathbf{D}$ takes the form

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{A}x - b\|^2 + \frac{\lambda}{p} \|\mathbf{D}x\|_p^p, \quad (36)$$

where $\mathbf{D} \in \mathbb{R}^{m \times d}$, $p \geq 2$, and $\lambda > 0$. Using the Fenchel conjugate of $\frac{1}{p} \|\cdot\|_p^p$, this admits a bilinear saddle-point reformulation of the form (1) with $f(x) = \frac{1}{2} \|\mathbf{A}x - b\|^2$, $\mathbf{B} = \mathbf{D}$, and a dual regularizer g whose gradient is $(q-1)$ -Hölder continuous with $1/p + 1/q = 1$.

Matrix games. A two-player zero-sum game with strongly convex strategy regularizers f and g takes the form

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x) + \langle y, \mathbf{A}x \rangle - g(y), \quad (37)$$

where $\mathbf{A} \in \mathbb{R}^{d_y \times d_x}$ is the payoff matrix and $\mathcal{X}, \mathcal{Y}$ are convex strategy sets. This is an instance of (1) with $\mathbf{B} = \mathbf{A}$.

C Discussion of Algorithm 1

C.1 Algorithmic Implementability

Assumption C.1. *The projection operator*

$$\Pi_{\mathcal{Z}}^{\mathbf{P}}(v) := \arg \min_{z \in \mathcal{Z}} \frac{1}{2} \|z - v\|_{\mathbf{P}}^2$$

can be computed exactly and at negligible cost compared with evaluations of the component oracles p'_i and Q_i . Equivalently, all proximal subproblems generated by Algorithm 1 are assumed to be solved exactly. If this assumption is not satisfied, then the cost of these proximal subproblems must be added separately to the oracle complexity.

Under Assumption C.1, the oracle complexities reported below count only evaluations of the component first-order oracles p'_i and Q_i . Each minimization step in Algorithm 1 is a proximal step with a linear objective and a quadratic

regularizer. In particular, subproblems of the form

$$\arg \min_{z \in \mathcal{Z}} \left\{ \langle c, z \rangle + \sum_{m=1}^q \frac{\theta_m}{2} \|z - a_m\|_{\mathbf{P}}^2 \right\},$$

$$\theta_m \geq 0, \quad \sum_{m=1}^q \theta_m > 0.$$

can be written as

$$\Pi_{\mathcal{Z}}^{\mathbf{P}} \left(\frac{\sum_{m=1}^q \theta_m a_m}{\sum_{m=1}^q \theta_m} - \frac{1}{\sum_{m=1}^q \theta_m} \mathbf{P}^{-1} c \right).$$

Thus the algorithm is directly implementable whenever projection onto $\mathcal{Z}$ in the $\mathbf{P}$ -norm is cheap, for example for common product sets such as boxes, Euclidean balls, or unconstrained domains. For the bilinear specialization, the proximal subproblems reduce to projections onto $\mathcal{X}$ and $\mathcal{Y}$.

C.2 Special Cases and Mirror-Prox Sliding

The general complexity bound recovers familiar guarantees in its one- and two-component special cases. When $n = 1$ and $Q_1 \equiv 0$, Corollary 3.6 gives

$$O \left(\left(\frac{H_1 \Omega_{z_{\text{in}}}^{(1+\nu_1)/2}}{\varepsilon} \right)^{\frac{2}{1+3\nu_1}} \right),$$

which is the usual universal gradient rate, interpolating between nonsmooth subgradient and accelerated smooth complexity (Nesterov 2015).

When $\nu_i = 1$, the function part of the complexity becomes $O \left(\sqrt{\frac{H_i \Omega}{\varepsilon}} \right)$, which recovers the accelerated smooth rate.

When $\nu_i = 0$, the function part becomes $O \left(\frac{H_i^2 \Omega}{\varepsilon^2} \right)$, matching the standard nonsmooth subgradient complexity up to constants. The operator part contributes the Mirror-Prox-type term $O(M_i \Omega / \varepsilon)$.

Relation to Mirror-Prox Sliding. The closest two-component comparison is Mirror-Prox Sliding of Lan and Ouyang (2021), for VIs with operator $\nabla G + H$, where G is convex with L -Lipschitz gradient and H is monotone and M -Lipschitz. In our notation, the two component pairs are $(G, 0)$ and $(0, H)$. Mirror-Prox Sliding requires $\mathcal{O}(\sqrt{L\Omega/\varepsilon})$ evaluations of ∇G and $\mathcal{O}(\sqrt{L\Omega/\varepsilon} + M\Omega/\varepsilon)$ evaluations of H . The ascending relabeling of Borodich et al. (2025), used in Corollary 3.6, instead gives $\mathcal{O}(\max\{\sqrt{L\Omega/\varepsilon}, 1\})$ and $\mathcal{O}(\max\{M\Omega/\varepsilon, 1\})$ evaluations, respectively, removing the cross-component square-root term from the H -oracle count.

Stochastic Mirror-Prox Sliding keeps ∇G exact and samples H , adding $\mathcal{O}(\sigma_H^2 \Omega / \varepsilon^2)$ to its H -oracle complexity. Appendix K considers a different model: $\mathbf{B}, \mathbf{B}^\top$ are exact while the function (sub)gradients are stochastic. For $q \in \{x, y\}$, its variance costs are charged separately to N_f, N_g , scaling as $\sigma_q^2 \Omega / \varepsilon^2$ for the bounded-domain uniform expected gap and as $\sigma_q^2 / (\mu_q \varepsilon)$ for the expected Lyapunov guarantee when $\mu_x, \mu_y > 0$. These costs add to the respective deterministic Hölder budgets. The strongly convex result uses $p_3 = 0$, so no anchor residual appears; the Hölder conditions hold on the feasible sets.

D Proof of Theorem 3.3

Lemma D.1. *Let $\{\alpha_t\}_{t \geq 0}$ be the sequence defined in (11). Then $0 < \alpha_t \leq 1$ for all $t \geq 0$, the sequence $\{\alpha_t\}$ is nonincreasing, and the following properties hold:*

$$\frac{1 - \alpha_t}{\alpha_t^2} = \frac{1}{\alpha_{t-1}^2}, \quad t \geq 1, \quad (38)$$

$$\frac{1}{t+1} \leq \alpha_t \leq \frac{2}{t+2}, \quad t \geq 0. \quad (39)$$

Proof. Set $a_t := 1/\alpha_t$. By the recursion in (11),

$$a_{t+1} = \frac{1 + \sqrt{1 + 4a_t^2}}{2},$$

and hence

$$a_{t+1}^2 - a_{t+1} = a_t^2.$$

Replacing $t + 1$ by t gives, for $t \geq 1$,

$$\frac{1 - \alpha_t}{\alpha_t^2} = \frac{1}{\alpha_{t-1}^2}.$$

Moreover, since $a_0 = 1$ and

$$a_{t+1} = \frac{1 + \sqrt{1 + 4a_t^2}}{2} \geq a_t,$$

the sequence $\{a_t\}$ is nondecreasing. Hence $\{\alpha_t\}$ is nonincreasing, and therefore $0 < \alpha_t \leq \alpha_0 = 1$.

It remains to prove (39). Since

$$\sqrt{1 + 4a_t^2} \geq 2a_t,$$

we have

$$a_{t+1} = \frac{1 + \sqrt{1 + 4a_t^2}}{2} \geq a_t + \frac{1}{2}.$$

By induction and $a_0 = 1$, this gives

$$a_t \geq 1 + \frac{t}{2} = \frac{t+2}{2}.$$

Equivalently,

$$\alpha_t = \frac{1}{a_t} \leq \frac{2}{t+2}.$$

On the other hand, since

$$\sqrt{1 + 4a_t^2} \leq 1 + 2a_t,$$

we have

$$a_{t+1} = \frac{1 + \sqrt{1 + 4a_t^2}}{2} \leq a_t + 1.$$

Again by induction and $a_0 = 1$, this gives

$$a_t \leq t + 1.$$

Equivalently,

$$\alpha_t = \frac{1}{a_t} \geq \frac{1}{t+1}.$$

Combining the two inequalities proves (39). $\square$

Lemma D.2. Suppose that $0 < \alpha_t \leq 1$ for all t , and that all averaging points generated by Algorithm 1 belong to $\mathcal{Z}$. For $k \geq 1$, let

$$\mathbf{t} = (t_1, \dots, t_k), \quad 0 \leq t_\ell \leq T_\ell - 1, \quad \ell = 1, \dots, k,$$

and define

$$\Gamma_{\mathbf{t}} := \prod_{\ell=1}^k \alpha_{t_\ell}.$$

Then, for every $1 \leq k \leq i \leq n$, the function $\hat{p}_i^{k,\mathbf{t}}$ is equipped with a first-order $(\delta_i^{k,\mathbf{t}}, L_i^{k,\mathbf{t}})$ -oracle on $\mathcal{Z}$, with respect to $\|\cdot\|_{\mathbf{P}}$, where

$$\delta_i^{k,\mathbf{t}} = \frac{\delta_i}{\Gamma_{\mathbf{t}}}, \quad L_i^{k,\mathbf{t}} = L_i \Gamma_{\mathbf{t}}.$$

Proof. We prove the claim by induction on k . Throughout the proof, the oracle for each transformed function is the oracle induced by the affine transformation in line 20 of Algorithm 1.

Base case $k = 1$. The lemma's index condition $i \geq k$ becomes $i \geq 1$. By lines 7 and 20 of Algorithm 1, with $k = 0$, $\mathbf{t} = \emptyset$, and $\mathbf{t}' = (j)$, for all $1 \leq i \leq n$ and $0 \leq j \leq T_1 - 1$,

$$\begin{aligned}\hat{p}_i^{1,j}(z) &= \alpha_j^{-1} p_i^{0,\emptyset}(\alpha_j z + (1 - \alpha_j) \bar{z}_j^1) \\ &= \alpha_j^{-1} p_i(\alpha_j z + (1 - \alpha_j) \bar{z}_j^1).\end{aligned}$$

The induced oracle is therefore

$$\begin{aligned}\tilde{\hat{p}}_i^{1,j}(z) &= \alpha_j^{-1} \tilde{p}_i(\alpha_j z + (1 - \alpha_j) \bar{z}_j^1), \\ \tilde{\nabla} \hat{p}_i^{1,j}(z) &= \tilde{\nabla} p_i(\alpha_j z + (1 - \alpha_j) \bar{z}_j^1).\end{aligned}$$

Fix $u, v \in \mathcal{Z}$ and define

$$u' = \alpha_j u + (1 - \alpha_j) \bar{z}_j^1, \quad v' = \alpha_j v + (1 - \alpha_j) \bar{z}_j^1.$$

By lines 16 and 26 of Algorithm 1, the averaging point $\bar{z}_j^1$ belongs to $\mathcal{Z}$. Since $\mathcal{Z}$ is convex, we have $u', v' \in \mathcal{Z}$. Hence,

$$\begin{aligned}\hat{p}_i^{1,j}(u) - \left(\tilde{\hat{p}}_i^{1,j}(v) + \left\langle \tilde{\nabla} \hat{p}_i^{1,j}(v), u - v \right\rangle \right) \\ = \alpha_j^{-1} \left[p_i(u') - \left(\tilde{p}_i(v') + \alpha_j \left\langle \tilde{\nabla} p_i(v'), u - v \right\rangle \right) \right] \\ = \alpha_j^{-1} \left[p_i(u') - \left(\tilde{p}_i(v') + \left\langle \tilde{\nabla} p_i(v'), u' - v' \right\rangle \right) \right].\end{aligned}$$

Using the (δ_i, L_i) -oracle bound for p_i , we obtain

$$\begin{aligned}0 &\leq \hat{p}_i^{1,j}(u) - \left(\tilde{\hat{p}}_i^{1,j}(v) + \left\langle \tilde{\nabla} \hat{p}_i^{1,j}(v), u - v \right\rangle \right) \\ &\leq \alpha_j^{-1} \left(\frac{L_i}{2} \|u' - v'\|_{\mathbf{P}}^2 + \delta_i \right) \\ &= \frac{L_i \alpha_j}{2} \|u - v\|_{\mathbf{P}}^2 + \frac{\delta_i}{\alpha_j} \\ &= \frac{L_i^{1,j}}{2} \|u - v\|_{\mathbf{P}}^2 + \delta_i^{1,j}.\end{aligned}$$

Thus $\hat{p}_i^{1,j}$ is equipped with a $(\delta_i^{1,j}, L_i^{1,j})$ -oracle.

Inductive step. Assume the claim holds at level k , namely that $\hat{p}_i^{k,\mathbf{t}}$ is equipped with a $(\delta_i^{k,\mathbf{t}}, L_i^{k,\mathbf{t}})$ -oracle for all $i \geq k$. We prove the claim at level $k + 1$. Let $0 \leq j \leq T_{k+1} - 1$ and set

$$\mathbf{t}' = (\mathbf{t}, j).$$

Fix $u, v \in \mathcal{Z}$ and define

$$u' = \alpha_j u + (1 - \alpha_j) \bar{z}_{\mathbf{t}'}^{k+1}, \quad v' = \alpha_j v + (1 - \alpha_j) \bar{z}_{\mathbf{t}'}^{k+1}.$$

By lines 16 and 26 of Algorithm 1, the averaging point $\bar{z}_{\mathbf{t}'}^{k+1}$ belongs to $\mathcal{Z}$. Since $\mathcal{Z}$ is convex, we have $u', v' \in \mathcal{Z}$.

Now fix $i \geq k + 1$. By line 20 of Algorithm 1,

$$\hat{p}_i^{k+1,\mathbf{t}'}(z) = \alpha_j^{-1} p_i^{k,\mathbf{t}}(\alpha_j z + (1 - \alpha_j) \bar{z}_{\mathbf{t}'}^{k+1}).$$

Since $i \geq k + 1$, we have $i \neq k$. Hence, by line 23 of Algorithm 1 applied at level k ,

$$p_i^{k,\mathbf{t}}(z) = \hat{p}_i^{k,\mathbf{t}}(z).$$

Therefore,

$$\hat{p}_i^{k+1,\mathbf{t}'}(z) = \alpha_j^{-1} \hat{p}_i^{k,\mathbf{t}}(\alpha_j z + (1 - \alpha_j) \bar{z}_{\mathbf{t}'}^{k+1}).$$

The induced oracle is

$$\tilde{\hat{p}}_i^{k+1,\mathbf{t}'}(z) = \alpha_j^{-1} \tilde{\hat{p}}_i^{k,\mathbf{t}}(\alpha_j z + (1 - \alpha_j) \bar{z}_{\mathbf{t}'}^{k+1}),$$

and

$$\tilde{\nabla} \hat{p}_i^{k+1,\mathbf{t}'}(z) = \tilde{\nabla} \hat{p}_i^{k,\mathbf{t}}(\alpha_j z + (1 - \alpha_j) \bar{z}_{\mathbf{t}'}^{k+1}).$$

Consequently,

$$\begin{aligned}
& \hat{p}_i^{k+1, \mathbf{t}'}(u) - \tilde{p}_i^{k+1, \mathbf{t}'}(v) \\
& - \left\langle \tilde{\nabla} \hat{p}_i^{k+1, \mathbf{t}'}(v), u - v \right\rangle \\
& = \alpha_j^{-1} \left[\hat{p}_i^{k, \mathbf{t}}(u') - \tilde{p}_i^{k, \mathbf{t}}(v') \right. \\
& \quad \left. - \alpha_j \left\langle \tilde{\nabla} \hat{p}_i^{k, \mathbf{t}}(v'), u - v \right\rangle \right] \\
& = \alpha_j^{-1} \left[\hat{p}_i^{k, \mathbf{t}}(u') - \tilde{p}_i^{k, \mathbf{t}}(v') \right. \\
& \quad \left. - \left\langle \tilde{\nabla} \hat{p}_i^{k, \mathbf{t}}(v'), u' - v' \right\rangle \right].
\end{aligned}$$

Using the induction hypothesis for $\hat{p}_i^{k, \mathbf{t}}$, we obtain

$$\begin{aligned}
0 & \leq \hat{p}_i^{k+1, \mathbf{t}'}(u) - \left(\tilde{p}_i^{k+1, \mathbf{t}'}(v) + \left\langle \tilde{\nabla} \hat{p}_i^{k+1, \mathbf{t}'}(v), u - v \right\rangle \right) \\
& \leq \alpha_j^{-1} \left(\frac{L_i^{k, \mathbf{t}}}{2} \|u' - v'\|_{\mathbf{P}}^2 + \delta_i^{k, \mathbf{t}} \right) \\
& = \frac{L_i^{k, \mathbf{t}} \alpha_j}{2} \|u - v\|_{\mathbf{P}}^2 + \frac{\delta_i^{k, \mathbf{t}}}{\alpha_j} \\
& = \frac{L_i^{k+1, \mathbf{t}'}}{2} \|u - v\|_{\mathbf{P}}^2 + \delta_i^{k+1, \mathbf{t}'}.
\end{aligned}$$

Thus $\hat{p}_i^{k+1, \mathbf{t}'}$ is equipped with a $(\delta_i^{k+1, \mathbf{t}'}, L_i^{k+1, \mathbf{t}'})$ -oracle. The induction is complete. $\square$

For fixed $k, \mathbf{t} = (t_1, \dots, t_k)$ and $0 \leq t \leq T_k - 1$, we define

$$y_t^{k, \mathbf{t}} = z_{t_1, \dots, t_{k-1}, t}^k \quad (40)$$

Fix $\hat{z} \in \mathcal{Z}$. For $0 \leq k \leq n$ and $\mathbf{t} = (t_1, \dots, t_k)$, define

$$f_{\mathbf{t}}^k(z) := \sum_{i=1}^n p_i^{k, \mathbf{t}}(z), \quad \hat{f}_{\mathbf{t}}^k(z) := \sum_{i=1}^n \hat{p}_i^{k, \mathbf{t}}(z),$$

and

$$g_{\hat{z}}^k(z) := \sum_{i=k+1}^n \langle Q_i(\hat{z}), z \rangle, \quad h_{\mathbf{t}, \hat{z}}^k(z) := f_{\mathbf{t}}^k(z) + g_{\hat{z}}^k(z).$$

For $u, v \in \mathcal{Z}$, write

$$\Delta f_{\mathbf{t}}^k(u, v) := f_{\mathbf{t}}^k(u) - f_{\mathbf{t}}^k(v),$$

$$\Delta \hat{f}_{\mathbf{t}}^k(u, v) := \hat{f}_{\mathbf{t}}^k(u) - \hat{f}_{\mathbf{t}}^k(v),$$

and

$$\Delta g_{\hat{z}}^k(u, v) := g_{\hat{z}}^k(u) - g_{\hat{z}}^k(v),$$

$$\Delta h_{\mathbf{t}, \hat{z}}^k(u, v) := h_{\mathbf{t}, \hat{z}}^k(u) - h_{\mathbf{t}, \hat{z}}^k(v).$$

Finally, define

$$\Phi_{\mathbf{t}, \hat{z}}^k(u, v) := \Delta h_{\mathbf{t}, \hat{z}}^k(u, v) + \frac{H_{\Sigma}^k}{2} \|u - v\|_{\mathbf{P}}^2,$$

where

$$H_{\Sigma}^k := \sum_{\ell=1}^k \eta_{t_1, \dots, t_{\ell}}^{\ell}.$$

We use the convention that empty sums are equal to zero.

The constants $L_i^{k, \mathbf{t}}$ and $\delta_i^{k, \mathbf{t}}$ are those from Lemma D.2. We define the transformed Lipschitz constants for the operators by

$$M_i^{k, \mathbf{t}} := M_i \prod_{\ell=1}^k \frac{\alpha_{t_{\ell}}}{\alpha_{T_{\ell}-1}}.$$

Lemma D.3. Fix $\hat{z} \in \mathcal{Z}$. For every $0 \leq k \leq n$ and $\mathbf{t} = (t_1, \dots, t_k)$, where for $k \geq 1$

$$0 \leq t_\ell \leq T_\ell - 1, \quad \ell = 1, \dots, k,$$

and where $\mathbf{t} = \emptyset$ when $k = 0$, we have

$$\Phi_{\mathbf{t}, \hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) \leq S_{\mathbf{t}}^k + R_{\mathbf{t}}^k.$$

Here

$$A_i^{k, \mathbf{t}} := L_i^{k, \mathbf{t}} \prod_{j=k+1}^i \alpha_{T_j-1}^2 + M_i^{k, \mathbf{t}} \prod_{j=k+1}^i \alpha_{T_j-1},$$

and

$$S_{\mathbf{t}}^k := \sum_{i=k+1}^n \frac{A_i^{k, \mathbf{t}}}{2} \left(\|z_{\mathbf{t}^i, 0}^i - \hat{z}\|_{\mathbf{P}}^2 - \|z_{\mathbf{t}_+^i, 0}^i - \hat{z}\|_{\mathbf{P}}^2 \right),$$

and

$$R_{\mathbf{t}}^k := \sum_{i=k+1}^n \frac{\delta_i^{k, \mathbf{t}}}{\prod_{j=k+1}^i \alpha_{T_j-1}}.$$

We use the conventions

$$L_i^{0, \emptyset} = L_i, \quad M_i^{0, \emptyset} = M_i, \quad \delta_i^{0, \emptyset} = \delta_i,$$

and empty products are equal to 1.

For $i \geq k+1$, the multi-indices in $S_{\mathbf{t}}^k$ are defined by

$$\mathbf{0}_r := (0, \dots, 0) \in \mathbb{N}^r,$$

$$\mathbf{t}^{i, 0} := (\mathbf{t}, \mathbf{0}_{i-k}) \in \mathbb{N}^i.$$

For $k \geq 1$, set

$$\mathbf{t}_+ := (t_1, \dots, t_{k-1}, t_k + 1).$$

Then

$$\mathbf{t}_+^{i, 0} \in \mathbb{N}^i, \quad \mathbf{t}_+^{i, 0} := \begin{cases} (T_1, \mathbf{0}_{i-1}), & k = 0, \\ (\mathbf{t}_+, \mathbf{0}_{i-k}), & 1 \leq k \leq n-1. \end{cases}$$

When $k = n$, the sums defining $S_{\mathbf{t}}^n$ and $R_{\mathbf{t}}^n$ are empty, so $S_{\mathbf{t}}^n = R_{\mathbf{t}}^n = 0$.

Proof. We prove the claim by backward induction on k .

Base case $k = n$. Let $\mathbf{t} = (t_1, \dots, t_n)$, where $0 \leq t_\ell \leq T_\ell - 1$ for $\ell = 1, \dots, n$. Since the sum defining $g_{\hat{z}}^n$ is empty, we have

$$g_{\hat{z}}^n(z) \equiv 0, \quad h_{\mathbf{t}, \hat{z}}^n(z) = f_{\mathbf{t}}^n(z).$$

Moreover, by line 14 of Algorithm 1,

$$\tilde{z}_{\mathbf{t}}^n = \arg \min_{z \in \mathcal{Z}} f_{\mathbf{t}}^n(z).$$

Since $f_{\mathbf{t}}^n$ is H_{Σ}^n -strongly convex with respect to $\|\cdot\|_{\mathbf{P}}$, it follows that, for every $\hat{z} \in \mathcal{Z}$,

$$f_{\mathbf{t}}^n(\hat{z}) \geq f_{\mathbf{t}}^n(\tilde{z}_{\mathbf{t}}^n) + \frac{H_{\Sigma}^n}{2} \|\tilde{z}_{\mathbf{t}}^n - \hat{z}\|_{\mathbf{P}}^2.$$

Therefore,

$$\Phi_{\mathbf{t}, \hat{z}}^n(\tilde{z}_{\mathbf{t}}^n, \hat{z}) = f_{\mathbf{t}}^n(\tilde{z}_{\mathbf{t}}^n) - f_{\mathbf{t}}^n(\hat{z}) + \frac{H_{\Sigma}^n}{2} \|\tilde{z}_{\mathbf{t}}^n - \hat{z}\|_{\mathbf{P}}^2 \leq 0.$$

On the other hand, when $k = n$, both sums defining $S_{\mathbf{t}}^n$ and $R_{\mathbf{t}}^n$ are empty, so

$$S_{\mathbf{t}}^n = 0, \quad R_{\mathbf{t}}^n = 0.$$

Hence

$$\Phi_{\mathbf{t}, \hat{z}}^n(\tilde{z}_{\mathbf{t}}^n, \hat{z}) \leq S_{\mathbf{t}}^n + R_{\mathbf{t}}^n.$$

Inductive step. Fix $1 \leq k \leq n$. Assume that the claim holds at level k , i.e., for every $\mathbf{t} = (t_1, \dots, t_k)$, where $0 \leq t_\ell \leq T_\ell - 1$ for $\ell = 1, \dots, k$,

$$\Phi_{\mathbf{t}, \hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) \leq S_{\mathbf{t}}^k + R_{\mathbf{t}}^k.$$

Let

$$\mathbf{t}' = (t_1, \dots, t_{k-1}),$$

with the convention that $\mathbf{t}' = \emptyset$ when $k = 1$. We prove

$$\Phi_{\mathbf{t}', \hat{z}}^{k-1}(\tilde{z}_{\mathbf{t}'}^{k-1}, \hat{z}) \leq S_{\mathbf{t}'}^{k-1} + R_{\mathbf{t}'}^{k-1}.$$

For any $\mathbf{t} = (\mathbf{t}', t_k)$, $0 \leq t_k \leq T_k - 1$, we have

$$\Phi_{\mathbf{t}, \hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) = \Delta h_{\mathbf{t}, \hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \frac{H_{\Sigma}^k}{2} \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2.$$

For any $z \in \mathcal{Z}$, by line 23 of Algorithm 1, the only component in which $f_{\mathbf{t}}^k$ and $\hat{f}_{\mathbf{t}}^k$ differ is the k -th component. Hence

$$f_{\mathbf{t}}^k(z) - \hat{f}_{\mathbf{t}}^k(z) = p_k^{k, \mathbf{t}}(z) - \hat{p}_k^{k, \mathbf{t}}(z),$$

and consequently

$$\begin{aligned} \Delta f_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) &= \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \Delta p_k^{k, \mathbf{t}}(\tilde{z}_{\mathbf{t}}^k, \hat{z}) - \Delta \hat{p}_k^{k, \mathbf{t}}(\tilde{z}_{\mathbf{t}}^k, \hat{z}) \\ &= \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \left[p_k^{k, \mathbf{t}}(\tilde{z}_{\mathbf{t}}^k) - \hat{p}_k^{k, \mathbf{t}}(\tilde{z}_{\mathbf{t}}^k) \right] - \left[p_k^{k, \mathbf{t}}(\hat{z}) - \hat{p}_k^{k, \mathbf{t}}(\hat{z}) \right]. \end{aligned}$$

For any function f equipped with a first-order (δ, L) -oracle, we have, for all $x, y, z \in \mathcal{Z}$,

$$f(z) - f(y) \geq \left\langle \tilde{\nabla} f(x), z - y \right\rangle - \frac{L}{2} \|y - x\|_{\mathbf{P}}^2 - \delta. \quad (41)$$

By Lemma D.2, $\hat{p}_k^{k, \mathbf{t}}$ is equipped with a first-order $(\delta_k^{k, \mathbf{t}}, L_k^{k, \mathbf{t}})$ -oracle. Therefore, using line 23 of Algorithm 1, we obtain

$$\begin{aligned} \Delta f_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) &= \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \left[p_k^{k, \mathbf{t}}(\tilde{z}_{\mathbf{t}}^k) - p_k^{k, \mathbf{t}}(\hat{z}) \right] - \Delta \hat{p}_k^{k, \mathbf{t}}(\tilde{z}_{\mathbf{t}}^k, \hat{z}) \\ &= \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \frac{\eta_{\mathbf{t}}^k}{2} (\|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|\hat{z} - \tilde{z}_{\mathbf{t}}^k\|_{\mathbf{P}}^2) \\ &\quad + \left\langle \tilde{\nabla} p_k^{k, \mathbf{t}}(\tilde{z}_{\mathbf{t}}^k) + Q_k(\tilde{z}_{\mathbf{t}}^k), \tilde{z}_{\mathbf{t}}^k - \hat{z} \right\rangle - \Delta \hat{p}_k^{k, \mathbf{t}}(\tilde{z}_{\mathbf{t}}^k, \hat{z}) \\ &\geq \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \frac{\eta_{\mathbf{t}}^k - L_k^{k, \mathbf{t}}}{2} \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 \\ &\quad - \frac{\eta_{\mathbf{t}}^k}{2} \|\hat{z} - \tilde{z}_{\mathbf{t}}^k\|_{\mathbf{P}}^2 + \left\langle Q_k(\tilde{z}_{\mathbf{t}}^k), \tilde{z}_{\mathbf{t}}^k - \hat{z} \right\rangle - \delta_k^{k, \mathbf{t}}. \end{aligned}$$

Define

$$\Delta_{Q_k}^{k, \mathbf{t}} := Q_k(\tilde{z}_{\mathbf{t}}^k) - Q_k(\hat{z}_{\mathbf{t}}^k).$$

By monotonicity of Q_k ,

$$\begin{aligned} \left\langle Q_k(\tilde{z}_{\mathbf{t}}^k), \tilde{z}_{\mathbf{t}}^k - \hat{z} \right\rangle &= \left\langle \Delta_{Q_k}^{k, \mathbf{t}} + Q_k(\tilde{z}_{\mathbf{t}}^k), \tilde{z}_{\mathbf{t}}^k - \hat{z} \right\rangle \\ &\geq \left\langle \Delta_{Q_k}^{k, \mathbf{t}} + Q_k(\hat{z}), \tilde{z}_{\mathbf{t}}^k - \hat{z} \right\rangle. \end{aligned}$$

Moreover, since

$$g_{\hat{z}}^{k-1}(z) = g_{\hat{z}}^k(z) + \langle Q_k(\hat{z}), z \rangle,$$

we have

$$\Delta g_{\hat{z}}^{k-1}(\tilde{z}_{\mathbf{t}}^k, \hat{z}) = \Delta g_{\hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \langle Q_k(\hat{z}), \tilde{z}_{\mathbf{t}}^k - \hat{z} \rangle.$$

Therefore, using $h_{\mathbf{t},\hat{z}}^k = f_{\mathbf{t}}^k + g_{\hat{z}}^k$ and the previous lower bound on $\Delta f_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z})$, we obtain

$$\begin{aligned}\Delta h_{\mathbf{t},\hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) &\geq \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \Delta g_{\hat{z}}^{k-1}(\tilde{z}_{\mathbf{t}}^k, \hat{z}) \\ &\quad + \frac{\eta_{\mathbf{t}}^k - L_k^{k,\mathbf{t}}}{2} \|\tilde{z}_{\mathbf{t}}^k - z_{\mathbf{t}}^k\|_{\mathbf{P}}^2 - \frac{\eta_{\mathbf{t}}^k}{2} \|\hat{z} - z_{\mathbf{t}}^k\|_{\mathbf{P}}^2 \\ &\quad + \langle \Delta_{Q_k}^{k,\mathbf{t}}, \tilde{z}_{\mathbf{t}}^k - \hat{z} \rangle - \delta_k^{k,\mathbf{t}}.\end{aligned}$$

Let

$$\hat{\mathbf{t}} = (t_1, \dots, t_{k-1}, t_k + 1).$$

From line 27 of Algorithm 1, we have

$$z_{\hat{\mathbf{t}}}^k = \arg \min_{z \in \mathcal{Z}} \left\{ \langle Q_k(z_{\hat{\mathbf{t}}}^k) - Q_k(z_{\mathbf{t}}^k), z \rangle + \frac{\eta_{\hat{\mathbf{t}}}^k}{2} \|z - \tilde{z}_{\hat{\mathbf{t}}}^k\|_{\mathbf{P}}^2 \right\}.$$

Since

$$\Delta_{Q_k}^{k,\mathbf{t}} := Q_k(z_{\mathbf{t}}^k) - Q_k(\tilde{z}_{\mathbf{t}}^k),$$

the optimality of $z_{\hat{\mathbf{t}}}^k$ gives, for every $y \in \mathcal{Z}$,

$$\langle -\Delta_{Q_k}^{k,\mathbf{t}}, z_{\hat{\mathbf{t}}}^k - y \rangle \leq \frac{\eta_{\hat{\mathbf{t}}}^k}{2} (\|\tilde{z}_{\hat{\mathbf{t}}}^k - y\|_{\mathbf{P}}^2 - \|z_{\hat{\mathbf{t}}}^k - y\|_{\mathbf{P}}^2 - \|z_{\hat{\mathbf{t}}}^k - \tilde{z}_{\hat{\mathbf{t}}}^k\|_{\mathbf{P}}^2).$$

Equivalently,

$$\langle \Delta_{Q_k}^{k,\mathbf{t}}, z_{\hat{\mathbf{t}}}^k - y \rangle \geq \frac{\eta_{\hat{\mathbf{t}}}^k}{2} (\|z_{\hat{\mathbf{t}}}^k - y\|_{\mathbf{P}}^2 + \|z_{\hat{\mathbf{t}}}^k - \tilde{z}_{\hat{\mathbf{t}}}^k\|_{\mathbf{P}}^2 - \|\tilde{z}_{\hat{\mathbf{t}}}^k - y\|_{\mathbf{P}}^2).$$

Taking $y = \hat{z}$, we have

$$\begin{aligned}\langle \Delta_{Q_k}^{k,\mathbf{t}}, \tilde{z}_{\mathbf{t}}^k - \hat{z} \rangle &= \langle \Delta_{Q_k}^{k,\mathbf{t}}, z_{\hat{\mathbf{t}}}^k - \hat{z} \rangle + \langle \Delta_{Q_k}^{k,\mathbf{t}}, \tilde{z}_{\mathbf{t}}^k - z_{\hat{\mathbf{t}}}^k \rangle \\ &\geq \frac{\eta_{\hat{\mathbf{t}}}^k}{2} (\|z_{\hat{\mathbf{t}}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) \\ &\quad + \frac{\eta_{\hat{\mathbf{t}}}^k}{2} \|z_{\hat{\mathbf{t}}}^k - \tilde{z}_{\mathbf{t}}^k\|_{\mathbf{P}}^2 + \langle \Delta_{Q_k}^{k,\mathbf{t}}, \tilde{z}_{\mathbf{t}}^k - z_{\hat{\mathbf{t}}}^k \rangle.\end{aligned}$$

By Young's inequality,

$$\langle \Delta_{Q_k}^{k,\mathbf{t}}, \tilde{z}_{\mathbf{t}}^k - z_{\hat{\mathbf{t}}}^k \rangle \geq -\frac{1}{2\eta_{\hat{\mathbf{t}}}^k} \|\Delta_{Q_k}^{k,\mathbf{t}}\|_{\mathbf{P}^{-1}}^2 - \frac{\eta_{\hat{\mathbf{t}}}^k}{2} \|\tilde{z}_{\mathbf{t}}^k - z_{\hat{\mathbf{t}}}^k\|_{\mathbf{P}}^2.$$

Therefore,

$$\begin{aligned}\langle \Delta_{Q_k}^{k,\mathbf{t}}, \tilde{z}_{\mathbf{t}}^k - \hat{z} \rangle &\geq \frac{\eta_{\hat{\mathbf{t}}}^k}{2} (\|z_{\hat{\mathbf{t}}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) - \frac{1}{2\eta_{\hat{\mathbf{t}}}^k} \|\Delta_{Q_k}^{k,\mathbf{t}}\|_{\mathbf{P}^{-1}}^2 \\ &\geq \frac{\eta_{\hat{\mathbf{t}}}^k}{2} (\|z_{\hat{\mathbf{t}}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) - \frac{(M_k^{k,\mathbf{t}})^2}{2\eta_{\hat{\mathbf{t}}}^k} \|\tilde{z}_{\mathbf{t}}^k - z_{\hat{\mathbf{t}}}^k\|_{\mathbf{P}}^2.\end{aligned}$$

where the last inequality follows from Assumption 3.1 and the definition

$$M_k^{k,\mathbf{t}} = M_k \prod_{\ell=1}^k \frac{\alpha_{t_\ell}}{\alpha_{T_\ell-1}}.$$

Indeed, since $0 \leq t_\ell \leq T_\ell - 1$ and $\{\alpha_t\}$ is positive and nonincreasing, $M_k^{k,\mathbf{t}} \geq M_k$, and hence

$$\|\Delta_{Q_k}^{k,\mathbf{t}}\|_{\mathbf{P}^{-1}} \leq M_k \|z_{\hat{\mathbf{t}}}^k - z_{\mathbf{t}}^k\|_{\mathbf{P}} \leq M_k^{k,\mathbf{t}} \|z_{\hat{\mathbf{t}}}^k - z_{\mathbf{t}}^k\|_{\mathbf{P}}.$$

Therefore,

$$\begin{aligned}\Delta h_{\mathbf{t},\hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) &\geq \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \Delta g_{\hat{z}}^{k-1}(\tilde{z}_{\mathbf{t}}^k, \hat{z}) \\ &\quad + \frac{1}{2} \left(\eta_{\mathbf{t}}^k - L_k^{k,\mathbf{t}} - \frac{(M_k^{k,\mathbf{t}})^2}{\eta_{\mathbf{t}}^k} \right) \|\tilde{z}_{\mathbf{t}}^k - z_{\mathbf{t}}^k\|_{\mathbf{P}}^2 \\ &\quad + \frac{\eta_{\mathbf{t}}^k}{2} (\|z_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|z_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) - \delta_k^{k,\mathbf{t}},\end{aligned}\tag{42}$$

Since $\eta_{\mathbf{t}}^k = L_k^{k,\mathbf{t}} + M_k^{k,\mathbf{t}}$, we have

$$\eta_{\mathbf{t}}^k - L_k^{k,\mathbf{t}} - \frac{(M_k^{k,\mathbf{t}})^2}{\eta_{\mathbf{t}}^k} = M_k^{k,\mathbf{t}} - \frac{(M_k^{k,\mathbf{t}})^2}{L_k^{k,\mathbf{t}} + M_k^{k,\mathbf{t}}} \geq 0.$$

Dropping the corresponding nonnegative term in (42), we obtain

$$\begin{aligned} \Delta h_{\mathbf{t},\hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) &\geq \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \Delta g_{\hat{z}}^{k-1}(\tilde{z}_{\mathbf{t}}^k, \hat{z}) \\ &\quad + \frac{\eta_{\mathbf{t}}^k}{2} (\|z_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|z_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) - \delta_k^{k,\mathbf{t}}. \end{aligned} \quad (43)$$

Using (43) and $H_{\Sigma}^k = H_{\Sigma}^{k-1} + \eta_{\mathbf{t}}^k$, we obtain

$$\begin{aligned} \Phi_{\mathbf{t},\hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) &= \Delta h_{\mathbf{t},\hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \frac{H_{\Sigma}^k}{2} \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 \\ &\geq \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \Delta g_{\hat{z}}^{k-1}(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \frac{H_{\Sigma}^{k-1}}{2} \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 \\ &\quad + \frac{\eta_{\mathbf{t}}^k}{2} (\|z_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) - \delta_k^{k,\mathbf{t}}. \end{aligned} \quad (44)$$

Since

$$\Phi_{\mathbf{t},\hat{z}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) \leq S_{\mathbf{t}}^k + R_{\mathbf{t}}^k,$$

it follows from (44) that

$$\begin{aligned} S_{\mathbf{t}}^k + R_{\mathbf{t}}^k &\geq \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \Delta g_{\hat{z}}^{k-1}(\tilde{z}_{\mathbf{t}}^k, \hat{z}) + \frac{H_{\Sigma}^{k-1}}{2} \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 \\ &\quad + \frac{\eta_{\mathbf{t}}^k}{2} (\|z_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|\tilde{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) - \delta_k^{k,\mathbf{t}}. \end{aligned} \quad (45)$$

Let $\alpha = \alpha_{t_k}$. By line 20 of Algorithm 1, we have

$$\begin{aligned} \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) &= \sum_{i=1}^n \left(p_i^{k,\mathbf{t}}(\tilde{z}_{\mathbf{t}}^k) - \hat{p}_i^{k,\mathbf{t}}(\hat{z}) \right) \\ &= \sum_{i=1}^{k-1} \left(p_i^{k-1,\mathbf{t}'}(\tilde{z}_{\mathbf{t}}^k) - p_i^{k-1,\mathbf{t}'}(\hat{z}) \right) \\ &\quad + \frac{1}{\alpha} \sum_{i=k}^n p_i^{k-1,\mathbf{t}'}(\alpha \tilde{z}_{\mathbf{t}}^k + (1-\alpha)\tilde{z}_{\mathbf{t}}^k) \\ &\quad - \frac{1}{\alpha} \sum_{i=k}^n p_i^{k-1,\mathbf{t}'}(\alpha \hat{z} + (1-\alpha)\tilde{z}_{\mathbf{t}}^k). \end{aligned}$$

By line 26 of Algorithm 1,

$$\tilde{z}_{\mathbf{t}}^k = \alpha \tilde{z}_{\mathbf{t}}^k + (1-\alpha)\tilde{z}_{\mathbf{t}}^k.$$

Therefore,

$$\begin{aligned} \Delta \hat{f}_{\mathbf{t}}^k(\tilde{z}_{\mathbf{t}}^k, \hat{z}) &= \sum_{i=1}^{k-1} \left(p_i^{k-1,\mathbf{t}'}(\tilde{z}_{\mathbf{t}}^k) - p_i^{k-1,\mathbf{t}'}(\hat{z}) \right) \\ &\quad + \frac{1}{\alpha} \sum_{i=k}^n p_i^{k-1,\mathbf{t}'}(\tilde{z}_{\mathbf{t}}^k) \\ &\quad - \frac{1}{\alpha} \sum_{i=k}^n p_i^{k-1,\mathbf{t}'}(\alpha \hat{z} + (1-\alpha)\tilde{z}_{\mathbf{t}}^k). \end{aligned}$$

By convexity of $p_i^{k-1,\mathbf{t}'}$, for $i \geq k$,

$$p_i^{k-1,\mathbf{t}'}(\alpha \hat{z} + (1-\alpha)\tilde{z}_{\mathbf{t}}^k) \leq \alpha p_i^{k-1,\mathbf{t}'}(\hat{z}) + (1-\alpha)p_i^{k-1,\mathbf{t}'}(\tilde{z}_{\mathbf{t}}^k),$$

and, for $i \leq k-1$,

$$p_i^{k-1, \mathbf{t}'}(\bar{z}_{\mathbf{t}}^k) \geq \frac{1}{\alpha} p_i^{k-1, \mathbf{t}'}(\bar{z}_{\mathbf{t}}^k) - \frac{1-\alpha}{\alpha} p_i^{k-1, \mathbf{t}'}(\bar{z}_{\mathbf{t}}^k).$$

Consequently,

$$\begin{aligned} \Delta \hat{f}_{\mathbf{t}}^k(\bar{z}_{\mathbf{t}}^k, \hat{z}) &\geq \frac{1}{\alpha} \sum_{i=1}^n p_i^{k-1, \mathbf{t}'}(\bar{z}_{\mathbf{t}}^k) - \frac{1-\alpha}{\alpha} \sum_{i=1}^n p_i^{k-1, \mathbf{t}'}(\bar{z}_{\mathbf{t}}^k) \\ &\quad - \sum_{i=1}^n p_i^{k-1, \mathbf{t}'}(\hat{z}) \\ &= \frac{1}{\alpha} \Delta f_{\mathbf{t}'}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) - \frac{1-\alpha}{\alpha} \Delta f_{\mathbf{t}'}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}). \end{aligned} \quad (46)$$

Moreover, by line 26 of Algorithm 1, with $\alpha = \alpha_{t_k}$, we have

$$\bar{z}_{\mathbf{t}}^k = \alpha \bar{z}_{\mathbf{t}}^k + (1-\alpha) \bar{z}_{\mathbf{t}}^k.$$

Equivalently,

$$\bar{z}_{\mathbf{t}}^k - \hat{z} = \frac{1}{\alpha} (\bar{z}_{\mathbf{t}}^k - \hat{z}) - \frac{1-\alpha}{\alpha} (\bar{z}_{\mathbf{t}}^k - \hat{z}).$$

Therefore,

$$\begin{aligned} \Delta g_{\hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) &= \sum_{i=k}^n \langle Q_i(\hat{z}), \bar{z}_{\mathbf{t}}^k - \hat{z} \rangle \\ &= \frac{1}{\alpha} \sum_{i=k}^n \langle Q_i(\hat{z}), \bar{z}_{\mathbf{t}}^k - \hat{z} \rangle - \frac{1-\alpha}{\alpha} \sum_{i=k}^n \langle Q_i(\hat{z}), \bar{z}_{\mathbf{t}}^k - \hat{z} \rangle \\ &= \frac{1}{\alpha} \Delta g_{\hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) - \frac{1-\alpha}{\alpha} \Delta g_{\hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}). \end{aligned}$$

By convexity of the squared $\mathbf{P}$ -norm,

$$\|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 \leq \alpha \|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 + (1-\alpha) \|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2.$$

Rearranging gives

$$\frac{H_{\Sigma}^{k-1}}{2} \|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 \geq \frac{H_{\Sigma}^{k-1}}{2\alpha} \|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \frac{(1-\alpha)H_{\Sigma}^{k-1}}{2\alpha} \|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2.$$

Consequently,

$$\begin{aligned} \Delta g_{\hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) &+ \frac{H_{\Sigma}^{k-1}}{2} \|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 \\ &\geq \frac{1}{\alpha} \left[\Delta g_{\hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) + \frac{H_{\Sigma}^{k-1}}{2} \|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 \right] \\ &\quad - \frac{1-\alpha}{\alpha} \left[\Delta g_{\hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) + \frac{H_{\Sigma}^{k-1}}{2} \|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 \right]. \end{aligned}$$

It follows from (45), (46), and the preceding bound on $\Delta g_{\hat{z}}^{k-1}$ that

$$\begin{aligned} S_{\mathbf{t}}^k + R_{\mathbf{t}}^k &\geq \frac{1}{\alpha_{t_k}} \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) - \frac{1-\alpha_{t_k}}{\alpha_{t_k}} \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) \\ &\quad + \frac{\eta_{\mathbf{t}}^k}{2} (\|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) - \delta_k^{k, \mathbf{t}}. \end{aligned} \quad (47)$$

Dividing both sides by α_{t_k} , we obtain

$$\begin{aligned} \frac{S_{\mathbf{t}}^k + R_{\mathbf{t}}^k}{\alpha_{t_k}} &\geq \frac{1}{\alpha_{t_k}^2} \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) - \frac{1-\alpha_{t_k}}{\alpha_{t_k}^2} \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) \\ &\quad + \frac{\eta_{\mathbf{t}}^k}{2\alpha_{t_k}} (\|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|\bar{z}_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) - \frac{\delta_k^{k, \mathbf{t}}}{\alpha_{t_k}}. \end{aligned} \quad (48)$$

Using

$$\frac{1 - \alpha_{t_k}}{\alpha_{t_k}^2} = \frac{1}{\alpha_{t_k-1}^2}, \quad t_k \geq 1,$$

with the convention $1/\alpha_{-1}^2 := 0$, and using $\delta_k^{k,\mathbf{t}} = \delta_k^{k-1,\mathbf{t}'} / \alpha_{t_k}$, we get

$$\begin{aligned} \frac{S_{\mathbf{t}}^k + R_{\mathbf{t}}^k}{\alpha_{t_k}} &\geq \frac{1}{\alpha_{t_k}^2} \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) - \frac{1}{\alpha_{t_k-1}^2} \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\bar{z}_{\mathbf{t}}^k, \hat{z}) \\ &\quad + \frac{\eta_{\mathbf{t}}^k}{2\alpha_{t_k}} (\|z_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|z_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) - \frac{\delta_k^{k-1,\mathbf{t}'}}{\alpha_{t_k}^2}. \end{aligned} \quad (49)$$

Summing (49) over $t_k = 0, \dots, T_k - 1$, we obtain

$$\begin{aligned} \sum_{t_k=0}^{T_k-1} \frac{S_{\mathbf{t}}^k + R_{\mathbf{t}}^k}{\alpha_{t_k}} &\geq \frac{1}{\alpha_{T_k-1}^2} \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\bar{z}_{\mathbf{t}', T_k}^k, \hat{z}) \\ &\quad + \sum_{t_k=0}^{T_k-1} \frac{\eta_{\mathbf{t}}^k}{2\alpha_{t_k}} (\|z_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2 - \|z_{\mathbf{t}}^k - \hat{z}\|_{\mathbf{P}}^2) \\ &\quad - \delta_k^{k-1,\mathbf{t}'} \sum_{t_k=0}^{T_k-1} \frac{1}{\alpha_{t_k}^2}. \end{aligned}$$

Moreover, by the definitions of $L_k^{k,\mathbf{t}}$, $M_k^{k,\mathbf{t}}$, and $\eta_{\mathbf{t}}^k$,

$$\frac{\eta_{\mathbf{t}}^k}{\alpha_{t_k}} = L_k^{k-1,\mathbf{t}'} + \frac{M_k^{k-1,\mathbf{t}'}}{\alpha_{T_k-1}},$$

which is independent of t_k . For brevity, set

$$\begin{aligned} D_k^{\mathbf{t}'} &:= L_k^{k-1,\mathbf{t}'} + \frac{M_k^{k-1,\mathbf{t}'}}{\alpha_{T_k-1}}, \\ C_k^{\mathbf{t}'} &:= \alpha_{T_k-1}^2 D_k^{\mathbf{t}'} \\ &= L_k^{k-1,\mathbf{t}'} \alpha_{T_k-1}^2 + M_k^{k-1,\mathbf{t}'} \alpha_{T_k-1}. \end{aligned}$$

Hence the second sum telescopes:

$$\begin{aligned} \sum_{t_k=0}^{T_k-1} \frac{S_{\mathbf{t}}^k + R_{\mathbf{t}}^k}{\alpha_{t_k}} &\geq \frac{1}{\alpha_{T_k-1}^2} \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\bar{z}_{\mathbf{t}', T_k}^k, \hat{z}) \\ &\quad + \frac{D_k^{\mathbf{t}'}}{2} (\|z_{\mathbf{t}', T_k}^k - \hat{z}\|_{\mathbf{P}}^2 - \|z_{\mathbf{t}', 0}^k - \hat{z}\|_{\mathbf{P}}^2) \\ &\quad - \delta_k^{k-1,\mathbf{t}'} \sum_{t_k=0}^{T_k-1} \frac{1}{\alpha_{t_k}^2}. \end{aligned}$$

After multiplying both sides by $\alpha_{T_k-1}^2$, we obtain

$$\begin{aligned} \sum_{t_k=0}^{T_k-1} \frac{\alpha_{T_k-1}^2}{\alpha_{t_k}} (S_{\mathbf{t}}^k + R_{\mathbf{t}}^k) &\geq \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\bar{z}_{\mathbf{t}', T_k}^k, \hat{z}) \\ &\quad + \frac{C_k^{\mathbf{t}'}}{2} (\|z_{\mathbf{t}', T_k}^k - \hat{z}\|_{\mathbf{P}}^2 - \|z_{\mathbf{t}', 0}^k - \hat{z}\|_{\mathbf{P}}^2) \\ &\quad - \delta_k^{k-1,\mathbf{t}'} \sum_{t_k=0}^{T_k-1} \frac{\alpha_{T_k-1}^2}{\alpha_{t_k}^2}. \end{aligned} \quad (50)$$

By lines 25 and 34 of Algorithm 1, we have

$$\bar{z}_{\mathbf{t}'}^{k-1} = \bar{z}_{\mathbf{t}', T_k}^k.$$

Therefore, (50) becomes

$$\begin{aligned} \sum_{t_k=0}^{T_k-1} \frac{\alpha_{T_k-1}^2}{\alpha_{t_k}} (S_{\mathbf{t}}^k + R_{\mathbf{t}}^k) &\geq \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\hat{z}_{\mathbf{t}'}^{k-1}, \hat{z}) \\ &\quad + \frac{C_k^{\mathbf{t}'}}{2} (\|z_{\mathbf{t}', T_k}^k - \hat{z}\|_{\mathbf{P}}^2 - \|z_{\mathbf{t}', 0}^k - \hat{z}\|_{\mathbf{P}}^2) \\ &\quad - \delta_k^{k-1, \mathbf{t}'} \sum_{t_k=0}^{T_k-1} \frac{\alpha_{T_k-1}^2}{\alpha_{t_k}^2}. \end{aligned} \quad (51)$$

Rearranging (51), we get

$$\begin{aligned} \Phi_{\mathbf{t}', \hat{z}}^{k-1}(\hat{z}_{\mathbf{t}'}^{k-1}, \hat{z}) &\leq \sum_{t_k=0}^{T_k-1} \frac{\alpha_{T_k-1}^2}{\alpha_{t_k}} (S_{\mathbf{t}}^k + R_{\mathbf{t}}^k) \\ &\quad + \frac{C_k^{\mathbf{t}'}}{2} (\|z_{\mathbf{t}', 0}^k - \hat{z}\|_{\mathbf{P}}^2 - \|z_{\mathbf{t}', T_k}^k - \hat{z}\|_{\mathbf{P}}^2) \\ &\quad + \delta_k^{k-1, \mathbf{t}'} \sum_{t_k=0}^{T_k-1} \frac{\alpha_{T_k-1}^2}{\alpha_{t_k}^2}. \end{aligned} \quad (52)$$

By the definitions of $L_i^{k, \mathbf{t}}$, $M_i^{k, \mathbf{t}}$, and $S_{\mathbf{t}}^k$, and by the warm-start relation in line 30 of Algorithm 1, the terms involving squared distances telescope. Hence

$$\sum_{t_k=0}^{T_k-1} \frac{\alpha_{T_k-1}^2}{\alpha_{t_k}} S_{\mathbf{t}}^k + \frac{C_k^{\mathbf{t}'}}{2} (\|z_{\mathbf{t}', 0}^k - \hat{z}\|_{\mathbf{P}}^2 - \|z_{\mathbf{t}', T_k}^k - \hat{z}\|_{\mathbf{P}}^2) = S_{\mathbf{t}'}^{k-1}. \quad (53)$$

We have

$$\frac{1}{\alpha_t^2} - \frac{1}{\alpha_{t-1}^2} = \frac{1}{\alpha_t}, \quad t \geq 1.$$

Therefore,

$$\sum_{t=0}^{T-1} \frac{1}{\alpha_t} = \frac{1}{\alpha_{T-1}^2}.$$

Since $\{\alpha_t\}$ is nonincreasing, for $0 \leq t \leq T-1$,

$$\frac{\alpha_{T-1}^2}{\alpha_t^2} \leq \frac{\alpha_{T-1}}{\alpha_t}.$$

Hence

$$\sum_{t=0}^{T-1} \frac{\alpha_{T-1}^2}{\alpha_t^2} \leq \alpha_{T-1} \sum_{t=0}^{T-1} \frac{1}{\alpha_t} = \frac{1}{\alpha_{T-1}}.$$

Applying this with $T = T_k$, and using

$$\delta_i^{k, \mathbf{t}} = \frac{\delta_i^{k-1, \mathbf{t}'}}{\alpha_{t_k}}, \quad \mathbf{t} = (\mathbf{t}', t_k),$$

we obtain

$$\begin{aligned} &\sum_{t_k=0}^{T_k-1} \frac{\alpha_{T_k-1}^2}{\alpha_{t_k}} R_{\mathbf{t}}^k + \delta_k^{k-1, \mathbf{t}'} \sum_{t_k=0}^{T_k-1} \frac{\alpha_{T_k-1}^2}{\alpha_{t_k}^2} \\ &= \left(\sum_{t_k=0}^{T_k-1} \frac{\alpha_{T_k-1}^2}{\alpha_{t_k}^2} \right) \left[\delta_k^{k-1, \mathbf{t}'} + \sum_{i=k+1}^n \frac{\delta_i^{k-1, \mathbf{t}'}}{\prod_{j=k+1}^i \alpha_{T_j-1}} \right] \\ &\leq \frac{\delta_k^{k-1, \mathbf{t}'}}{\alpha_{T_k-1}} + \sum_{i=k+1}^n \frac{\delta_i^{k-1, \mathbf{t}'}}{\prod_{j=k}^i \alpha_{T_j-1}} = R_{\mathbf{t}'}^{k-1}. \end{aligned} \quad (54)$$

Combining (52), (53), and (54), we get

$$\Phi_{\mathbf{t}', \hat{z}}^{k-1}(z_{\mathbf{t}'-1}^{k-1}, \hat{z}) \leq S_{\mathbf{t}'}^{k-1} + R_{\mathbf{t}'}^{k-1}.$$

Since the induction step is valid for every $1 \leq k \leq n$, the claim holds for all $0 \leq k \leq n$. $\square$

Proof of Theorem 3.3. Fix an arbitrary $z \in \mathcal{Z}$ and set $\hat{z} = z$. We use the empty multi-index convention and write

$$\begin{aligned} f_{\emptyset}^0(u) &= p(u), \\ g_{\hat{z}}^0(u) &= \langle Q(\hat{z}), u \rangle, \\ h_{\emptyset, \hat{z}}^0(u) &= p(u) + \langle Q(\hat{z}), u \rangle. \end{aligned}$$

Since $H_{\Sigma}^0 = 0$, we have

$$\Phi_{\emptyset, \hat{z}}^0(u, \hat{z}) = \Delta h_{\emptyset, \hat{z}}^0(u, \hat{z}) = p(u) - p(\hat{z}) + \langle Q(\hat{z}), u - \hat{z} \rangle.$$

Let $z_{\text{out}} = z_{\emptyset}^0$ denote the output of Algorithm 1. Applying Lemma D.3 with $k = 0$ and $\hat{z} = z$, we get

$$\mathcal{G}(z) := p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle.$$

Thus

$$\mathcal{G}(z) = \Phi_{\emptyset, z}^0(z_{\text{out}}, z) \leq S_{\emptyset}^0 + R_{\emptyset}^0.$$

We first bound $R_{\emptyset}^0$. By the definition of $R_{\mathbf{t}}^k$,

$$R_{\emptyset}^0 = \sum_{i=1}^n \frac{\delta_i}{\prod_{j=1}^i \alpha_{T_j-1}}.$$

By Lemma D.1,

$$\alpha_{T_j-1} \geq \frac{1}{T_j}, \quad j = 1, \dots, n.$$

Hence

$$R_{\emptyset}^0 \leq \sum_{i=1}^n \delta_i \prod_{j=1}^i T_j.$$

Next, we bound $S_{\emptyset}^0$. By the definition of $S_{\mathbf{t}}^k$, the initialization $z_{\mathbf{0}_i}^i = z_{\text{in}}$, and the nonpositivity of the terminal squared-distance terms, we have

$$S_{\emptyset}^0 \leq \sum_{i=1}^n \frac{L_i \prod_{j=1}^i \alpha_{T_j-1}^2 + M_i \prod_{j=1}^i \alpha_{T_j-1}}{2} \|z_{\text{in}} - z\|_{\mathbf{P}}^2.$$

Using Lemma D.1 again,

$$\alpha_{T_j-1} \leq \frac{2}{T_j}, \quad j = 1, \dots, n,$$

and therefore

$$S_{\emptyset}^0 \leq \sum_{i=1}^n \left(\frac{4^i L_i}{\prod_{j=1}^i T_j^2} + \frac{2^i M_i}{\prod_{j=1}^i T_j} \right) \frac{1}{2} \|z_{\text{in}} - z\|_{\mathbf{P}}^2.$$

Combining the bounds on $S_{\emptyset}^0$ and $R_{\emptyset}^0$, we obtain

$$\mathcal{G}(z) \leq \sum_{i=1}^n \left[\left(\frac{2^{2i-1} L_i}{\prod_{j=1}^i T_j^2} + \frac{2^{i-1} M_i}{\prod_{j=1}^i T_j} \right) \|z_{\text{in}} - z\|_{\mathbf{P}}^2 + \delta_i \prod_{j=1}^i T_j \right].$$

Since $z \in \mathcal{Z}$ was arbitrary, the proof is complete. $\square$

E Proof of Theorem 3.5

Proof. Step 1: From the Hölder condition to an inexact oracle.

Lemma E.1 (Hölder-to-inexact oracle). *Let $\mathcal{U} \subseteq \mathbb{R}^{d_z}$ be a convex set and let $h: \mathcal{U} \rightarrow \mathbb{R}$ be a convex function which satisfies the (ν, H) -Hölder condition on $\mathcal{U}$ with respect to $\|\cdot\|_{\mathbf{P}}$ (Assumption 3.4), with $0 \leq \nu \leq 1$ and $H \geq 0$. Then for every $\delta > 0$, the pair*

$$\tilde{h}(v) = h(v), \quad \tilde{\nabla} h(v) = h'(v) \in \partial h(v)$$

defines a first-order $(\delta, L(\delta))$ -oracle for h on $\mathcal{U}$ with respect to $\|\cdot\|_{\mathbf{P}}$, where

$$L(\delta) = \left[\frac{1-\nu}{2(1+\nu)} \cdot \frac{1}{\delta} \right]^{\frac{1-\nu}{1+\nu}} H^{\frac{2}{1+\nu}}. \quad (55)$$

In the limit $\nu = 1$, the formula reduces to $L(\delta) = H$ for every $\delta \geq 0$ (a $(0, H)$ -oracle, i.e., exact gradient with H -Lipschitz smoothness in $\|\cdot\|_{\mathbf{P}}$).

Proof. Fix $u, v \in \mathcal{U}$, $h'(u) \in \partial h(u)$, and $h'(v) \in \partial h(v)$. The lower bound of Definition 3 is the standard convexity inequality

$$h(u) - h(v) - \langle h'(v), u - v \rangle \geq 0. \quad (56)$$

We turn to the upper bound. Define $\varphi: [0, 1] \rightarrow \mathbb{R}$ by

$$\varphi(s) = h(v + s(u - v)).$$

Since $\mathcal{U}$ is convex and h is convex, φ is finite, convex, and absolutely continuous on $[0, 1]$, with (sub)derivative

$$\varphi'(s) = \langle h'(v + s(u - v)), u - v \rangle \quad (57)$$

for almost every $s \in (0, 1)$. Hence

$$\begin{aligned} h(u) - h(v) - \langle h'(v), u - v \rangle &= \int_0^1 \langle h'(v + s(u - v)) - h'(v), u - v \rangle ds. \end{aligned} \quad (58)$$

By Cauchy–Schwarz with respect to the dual pairing $\|\cdot\|_{\mathbf{P}}, \|\cdot\|_{\mathbf{P}^{-1}}$ and the Hölder hypothesis applied to $(v + s(u - v), v) \in \mathcal{U} \times \mathcal{U}$,

$$\begin{aligned} &\langle h'(v + s(u - v)) - h'(v), u - v \rangle \\ &\leq \|h'(v + s(u - v)) - h'(v)\|_{\mathbf{P}^{-1}} \|u - v\|_{\mathbf{P}} \\ &\leq H \|s(u - v)\|_{\mathbf{P}}^{\nu} \|u - v\|_{\mathbf{P}} \\ &= H s^{\nu} \|u - v\|_{\mathbf{P}}^{1+\nu}. \end{aligned} \quad (59)$$

Integrating $s \in [0, 1]$ gives

$$\begin{aligned} h(u) - h(v) - \langle h'(v), u - v \rangle &\leq \frac{H}{1+\nu} \|u - v\|_{\mathbf{P}}^{1+\nu}. \end{aligned} \quad (60)$$

It remains to show that

$$\frac{H}{1+\nu} r^{1+\nu} \leq \frac{L(\delta)}{2} r^2 + \delta \quad (61)$$

for all $r = \|u - v\|_{\mathbf{P}} \geq 0$. If $\nu = 1$, then

$$\frac{H}{1+\nu} r^{1+\nu} = \frac{H}{2} r^2,$$

so the desired bound holds with $L = H$ and $\delta = 0$.

Now let $0 \leq \nu < 1$. If $H = 0$, the claim is immediate with $L(\delta) = 0$. Assume $H > 0$. For $L > 0$, consider

$$F(r) = \frac{H}{1+\nu} r^{1+\nu} - \frac{L}{2} r^2, \quad r \geq 0.$$

The maximum of F is attained at

$$r_\star = \left(\frac{H}{L}\right)^{1/(1-\nu)},$$

and direct computation gives

$$\sup_{r \geq 0} F(r) = \frac{1-\nu}{2(1+\nu)} H^{2/(1-\nu)} L^{-(1+\nu)/(1-\nu)}.$$

Therefore

$$\frac{H}{1+\nu} r^{1+\nu} \leq \frac{L}{2} r^2 + \delta \quad \forall r \geq 0$$

whenever

$$\frac{1-\nu}{2(1+\nu)} H^{2/(1-\nu)} L^{-(1+\nu)/(1-\nu)} \leq \delta.$$

Solving for L gives

$$L \geq \left[\frac{1-\nu}{2(1+\nu)} \cdot \frac{1}{\delta} \right]^{\frac{1-\nu}{1+\nu}} H^{\frac{2}{1+\nu}}.$$

Choosing equality gives $L = L(\delta)$. Combining with (60),

$$\begin{aligned} 0 &\leq h(u) - h(v) - \langle h'(v), u - v \rangle \\ &\leq \frac{L(\delta)}{2} \|u - v\|_{\mathbf{P}}^2 + \delta. \end{aligned} \tag{62}$$

This is the $(\delta, L(\delta))$ -oracle bound of Definition 3. $\square$

Step 2: Reduction to Theorem 3.3. Temporarily fix $\delta_1, \dots, \delta_n > 0$ and run the Sliding Algorithm with $L_k(\delta_k)$ in place of $L_k(\delta_k^\circ)$ in the step sizes (15). Lemma E.1 shows that the exact value and (sub)gradient of each p_i form a $(\delta_i, L_i(\delta_i))$ -oracle. Hence Theorem 3.3 applies. In Step 4 we set $\delta_i = \delta_i^\circ$, at which point these step sizes coincide with (15). For brevity, set

$$\mathcal{G}(z) := p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle.$$

Also write

$$D_z := \|z_{\text{in}} - z\|_{\mathbf{P}}, \quad N_i := \prod_{j=1}^i T_j.$$

Therefore, for every $z \in \mathcal{Z}$,

$$\mathcal{G}(z) \leq \sum_{i=1}^n \left[\left(\frac{2^{2i-1} L_i(\delta_i)}{N_i^2} + \frac{2^{i-1} M_i}{N_i} \right) D_z^2 + \delta_i N_i \right]. \tag{63}$$

Step 3: Choice of δ_i . Let

$$a_i := \frac{1-\nu_i}{1+\nu_i}.$$

For $0 \leq \nu_i < 1$, write

$$L_i(\delta) = K_i \delta^{-a_i}, \quad K_i = \left[\frac{1-\nu_i}{2(1+\nu_i)} \right]^{\frac{1-\nu_i}{1+\nu_i}} H_i^{\frac{2}{1+\nu_i}}.$$

The δ_i -dependent contribution in (63), with $D = \|z_{\text{in}} - z\|_{\mathbf{P}}$, is

$$g_i(\delta; D) = \frac{4^i L_i(\delta)}{2} \frac{D^2}{N_i^2} + \delta N_i = \frac{2^{2i-1} K_i D^2}{N_i^2} \delta^{-a_i} + N_i \delta. \tag{64}$$

If $H_i D > 0$, then $g_i(\cdot; D)$ is strictly convex on $(0, \infty)$, and its first-order condition is

$$-a_i \frac{2^{2i-1} K_i D^2}{N_i^2} \delta^{-a_i-1} + N_i = 0. \tag{65}$$

Equivalently,

$$-\frac{1-\nu_i}{1+\nu_i} \frac{2^{2i-1} K_i D^2}{N_i^2} \delta^{-2/(1+\nu_i)} + N_i = 0.$$

Thus the unique minimizer is

$$\delta_i^*(D) = \left[\frac{(1 - \nu_i) 2^{2i-1} K_i D^2}{(1 + \nu_i) N_i^3} \right]^{(1+\nu_i)/2}. \quad (66)$$

If $H_i = 0$ or $D = 0$, the corresponding function contribution is trivial and the same expression may be interpreted by continuity. At the minimum, the two summands satisfy

$$\frac{4^i L_i(\delta_i^*)}{N_i^2} \frac{D^2}{2} = \frac{1 + \nu_i}{1 - \nu_i} \delta_i^* N_i, \quad (67)$$

and therefore

$$g_i(\delta_i^*(D); D) = \frac{2}{1 - \nu_i} N_i \delta_i^*(D). \quad (68)$$

Using

$$K_i^{(1+\nu_i)/2} = \left[\frac{1 - \nu_i}{2(1 + \nu_i)} \right]^{(1-\nu_i)/2} H_i, \quad (69)$$

we obtain

$$g_i(\delta_i^*(D); D) = \frac{2^{i(1+\nu_i)} H_i D^{1+\nu_i}}{1 + \nu_i} \frac{1}{N_i^{(1+3\nu_i)/2}}. \quad (70)$$

Step 4: Choosing δ_i from $\Omega_{z_{\text{in}}}$ and taking the supremum over z .

The minimizer $\delta_i^*(D)$ in (66) depends on

$$D = \|z_{\text{in}} - z\|_{\mathbf{P}},$$

and hence generally depends on the comparator z . Since the algorithm must use a single value of δ_i throughout the run, we choose it using the worst-case bound

$$D^2 \leq \Omega_{z_{\text{in}}} := \sup_{z \in \mathcal{Z}} \|z_{\text{in}} - z\|_{\mathbf{P}}^2.$$

Thus, for $0 \leq \nu_i < 1$, we set

$$\delta_i^\circ = \delta_i^*(\sqrt{\Omega_{z_{\text{in}}}}) = \left[\frac{(1 - \nu_i) 2^{2i-1} K_i \Omega_{z_{\text{in}}}}{(1 + \nu_i) N_i^3} \right]^{(1+\nu_i)/2}.$$

If $H_i = 0$, then $K_i = 0$, and we interpret this choice as $\delta_i^\circ = 0$ and $L_i(\delta_i^\circ) = 0$. For $\nu_i = 1$, we set

$$\delta_i^\circ = 0, \quad L_i(\delta_i^\circ) = H_i.$$

Substituting $\delta_i = \delta_i^\circ$ into (63), we obtain, for every $z \in \mathcal{Z}$,

$$\mathcal{G}(z) \leq \sum_{i=1}^n \left[\left(\frac{2^{2i-1} L_i(\delta_i^\circ)}{N_i^2} + \frac{2^{i-1} M_i}{N_i} \right) D_z^2 + \delta_i^\circ N_i \right].$$

Since

$$\|z_{\text{in}} - z\|_{\mathbf{P}}^2 \leq \Omega_{z_{\text{in}}}, \quad \forall z \in \mathcal{Z},$$

we get

$$\sup_{z \in \mathcal{Z}} \mathcal{G}(z) \leq \sum_{i=1}^n \left[\left(\frac{2^{2i-1} L_i(\delta_i^\circ)}{N_i^2} + \frac{2^{i-1} M_i}{N_i} \right) \Omega_{z_{\text{in}}} + \delta_i^\circ N_i \right].$$

For $0 \leq \nu_i < 1$, the choice of δ_i° gives

$$\frac{4^i L_i(\delta_i^\circ)}{N_i^2} \frac{\Omega_{z_{\text{in}}}}{2} + \delta_i^\circ N_i = \frac{2^{i(1+\nu_i)} H_i \Omega_{z_{\text{in}}}^{(1+\nu_i)/2}}{(1 + \nu_i) N_i^{(1+3\nu_i)/2}}.$$

For $\nu_i = 1$, the same expression is obtained by taking $\delta_i^\circ = 0$ and $L_i(\delta_i^\circ) = H_i$, since

$$\frac{4^i H_i}{N_i^2} \frac{\Omega_{z_{\text{in}}}}{2} = \frac{2^{2i-1} H_i \Omega_{z_{\text{in}}}}{N_i^2} = \frac{2^{i(1+\nu_i)} H_i \Omega_{z_{\text{in}}}^{(1+\nu_i)/2}}{(1 + \nu_i) N_i^{(1+3\nu_i)/2}} \Big|_{\nu_i=1}.$$

Therefore,

$$\sup_{z \in \mathcal{Z}} \mathcal{G}(z) \leq \sum_{i=1}^n \left[\frac{2^{i(1+\nu_i)} H_i \Omega_{z_{\text{in}}}^{(1+\nu_i)/2}}{(1 + \nu_i) N_i^{(1+3\nu_i)/2}} + \frac{2^{i-1} M_i \Omega_{z_{\text{in}}}}{N_i} \right].$$

This proves (17). $\square$

F Proof of Corollary 3.6

Proof. Set $\Omega := \Omega_{z_{\text{in}}}$. Since the finite-sum representation

$$p = \sum_{i=1}^n p_i, \quad Q = \sum_{i=1}^n Q_i$$

is unchanged by a simultaneous relabeling of the pairs (p_i, Q_i) , we order the components so that

$$R_1 \leq R_2 \leq \dots \leq R_n.$$

We first compare R_i with the explicit lower bounds needed in Theorem 3.5. Define

$$\begin{aligned} \tilde{R}_{i,H} &:= \left(\frac{2^{i(1+\nu_i)+1} H_i \Omega^{(1+\nu_i)/2}}{(1+\nu_i)\varepsilon} \right)^{\frac{2}{1+3\nu_i}}, \\ \tilde{R}_{i,M} &:= \frac{2^i M_i \Omega}{\varepsilon}. \end{aligned}$$

Then set

$$\tilde{R}_i := \max \left\{ \tilde{R}_{i,H}, \tilde{R}_{i,M}, 1 \right\}.$$

We claim that there exists a constant $\kappa_n > 0$, depending only on n , such that

$$\tilde{R}_i \leq \kappa_n R_i, \quad i = 1, \dots, n.$$

Indeed, since $i \leq n$, $0 \leq \nu_i \leq 1$, and $1 + \nu_i \geq 1$,

$$\left(\frac{2^{i(1+\nu_i)+1}}{1+\nu_i} \right)^{\frac{2}{1+3\nu_i}} \leq 2^{2n+2}.$$

Also $2^i \leq 2^n \leq 2^{2n+2}$. Hence we may take

$$\kappa_n := 2^{2n+2},$$

and obtain $\tilde{R}_i \leq \kappa_n R_i$ for all i .

Next, we construct positive integers $T_1, \dots, T_n$. Let $N_0 := 1$, and define recursively

$$T_i := \max \left\{ 1, \left\lceil \frac{\kappa_n R_i}{N_{i-1}} \right\rceil \right\}.$$

Set

$$N_i := N_{i-1} T_i, \quad i = 1, \dots, n.$$

Then each T_i is a positive integer. We claim that

$$\kappa_n R_i \leq N_i \leq 2\kappa_n R_i, \quad i = 1, \dots, n.$$

The lower bound follows directly from the definition of T_i . For the upper bound, use induction. For $i = 1$,

$$N_1 = \lceil \kappa_n R_1 \rceil \leq \kappa_n R_1 + 1 \leq 2\kappa_n R_1,$$

because $R_1 \geq 1$. Suppose $N_{i-1} \leq 2\kappa_n R_{i-1}$. If $\kappa_n R_i \leq N_{i-1}$, then $T_i = 1$, and therefore

$$N_i = N_{i-1} \leq 2\kappa_n R_{i-1} \leq 2\kappa_n R_i.$$

If $\kappa_n R_i > N_{i-1}$, then

$$T_i = \left\lceil \frac{\kappa_n R_i}{N_{i-1}} \right\rceil \leq \frac{\kappa_n R_i}{N_{i-1}} + 1,$$

and hence

$$N_i = N_{i-1} T_i \leq \kappa_n R_i + N_{i-1} < 2\kappa_n R_i.$$

Thus $N_i \leq 2\kappa_n R_i$ for all i .

Since

$$N_i \geq \kappa_n R_i \geq \tilde{R}_i,$$

the choice of N_i satisfies the explicit lower bounds

$$N_i \geq \left(\frac{2^{i(1+\nu_i)+1} H_i \Omega^{(1+\nu_i)/2}}{(1+\nu_i)\varepsilon} \right)^{\frac{2}{1+3\nu_i}},$$

$$N_i \geq \frac{2^i M_i \Omega}{\varepsilon}.$$

Therefore, by Theorem 3.5,

$$\sup_{z \in \mathcal{Z}} \{p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle\} \leq \sum_{i=1}^n (A_i + B_i),$$

where

$$A_i := \frac{2^{i(1+\nu_i)} H_i \Omega^{(1+\nu_i)/2}}{(1+\nu_i) N_i^{(1+3\nu_i)/2}}, \quad B_i := \frac{2^{i-1} M_i \Omega}{N_i}.$$

The first lower bound on N_i gives $A_i \leq \varepsilon/2$, and the second gives $B_i \leq \varepsilon/2$. Hence

$$A_i + B_i \leq \varepsilon, \quad i = 1, \dots, n.$$

Summing over i yields

$$\sum_{i=1}^n (A_i + B_i) \leq n\varepsilon.$$

Thus

$$\sup_{z \in \mathcal{Z}} \{p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle\} \leq n\varepsilon.$$

Finally, since $N_i \leq 2\kappa_n R_i$, the required number of evaluations of p'_i and Q_i is at most $C_n R_i$, with $C_n = 2^{2n+4}$. This proves the corollary. $\square$

G Subspace Structure under Assumption 4.3

Throughout this section we write

$$V_x := \text{range}(\mathbf{B}^\top) = (\ker \mathbf{B})^\perp, \quad V_y := \text{range}(\mathbf{B}) = (\ker \mathbf{B}^\top)^\perp.$$

For the zero matrix, Assumption 4.3 is understood with the convention $\lambda_{\min}^+(\mathbf{0}) = 0$. Under R_x , Assumption 4.3 uses the Rayleigh-quotient bound $\lambda_{\min}^+(\mathbf{B}^\top \mathbf{B})$ on V_x . Otherwise it uses the ambient-space bound $\lambda_{\min}(\mathbf{B}^\top \mathbf{B})$. The lemmas below establish the required subspace invariance. The y -block is symmetric.

Lemma G.1 (Structure of the solution set). *Suppose R_x holds and $\mathcal{X} + \ker \mathbf{B} = \mathcal{X}$. Then:*

1. $f(x+u) = f(x)$ and $\partial f(x+u) = \partial f(x)$ for every $x \in \mathcal{X}$ and $u \in \ker \mathbf{B}$.
2. if $\mu_x > 0$, then $\ker \mathbf{B} = \{0\}$.
3. $F(x+u, y) = F(x, y)$ for every $(x, y) \in \mathcal{Z}$ and $u \in \ker \mathbf{B}$. Consequently $\mathcal{S} + (\ker \mathbf{B} \times \{0\}) = \mathcal{S}$.

The symmetric statements hold for g , $\mathcal{Y}$ and $\ker \mathbf{B}^\top$ under R_y . If $\mathcal{S} \neq \emptyset$, it contains a point z^* such that

$$x_{\text{in}} - x^* \in V_x.$$

If, in addition, R_y and $\mathcal{Y} + \ker \mathbf{B}^\top = \mathcal{Y}$ hold, z^* may be chosen so that

$$z_{\text{in}} - z^* \in V_x \times V_y.$$

Proof. Part (1): invariance of f and ∂f . Fix $x \in \mathcal{X}$ and $u \in \ker \mathbf{B}$. The cylinder condition gives $x+u \in \mathcal{X}$, while R_x gives

$$f'(x), f'(x+u) \in V_x = (\ker \mathbf{B})^\perp.$$

Applying convexity in both directions,

$$f(x+u) \geq f(x) + \langle f'(x), u \rangle = f(x),$$

and

$$f(x) \geq f(x+u) + \langle f'(x+u), -u \rangle = f(x+u).$$

Hence $f(x+u) = f(x)$.

Now take $q \in \partial f(x)$. Since $q \in V_x$, for every $w \in \mathbb{R}^{d_x}$,

$$f(w) \geq f(x) + \langle q, w - x \rangle = f(x+u) + \langle q, w - (x+u) \rangle.$$

Thus $q \in \partial f(x+u)$, proving $\partial f(x) \subseteq \partial f(x+u)$. Replacing u by $-u$ gives the reverse inclusion.

Part (2): strong convexity eliminates kernel directions. Fix $x \in \mathcal{X}$ and $u \in \ker \mathbf{B}$. By part (1) and R_x , $f(x+u) = f(x)$ and $\langle f'(x), u \rangle = 0$. Strong convexity therefore gives

$$0 = f(x+u) - f(x) \geq \frac{\mu_x}{2} \|u\|^2.$$

If $\mu_x > 0$, then $u = 0$. Hence $\ker \mathbf{B} = \{0\}$.

Part (3): invariance of F and $\mathcal{S}$. For $u \in \ker \mathbf{B}$, part (1) gives

$$F(x+u, y) = f(x) + \langle y, \mathbf{B}x + \mathbf{B}u \rangle - g(y) = F(x, y).$$

The cylinder condition preserves feasibility, so translating the x -component of any saddle point by u yields another saddle point. This proves

$$\mathcal{S} + (\ker \mathbf{B} \times \{0\}) = \mathcal{S}.$$

Interchanging $x, f, \mathbf{B}$ with $y, g, \mathbf{B}^\top$ proves the symmetric statements.

Choice of representative. Pick $\hat{z} = (\hat{x}, \hat{y}) \in \mathcal{S}$ and define

$$x^* := \hat{x} + \Pi_{\ker \mathbf{B}}(x_{\text{in}} - \hat{x}), \quad y^* := \hat{y}.$$

Part (3) shows that $z^* := (x^*, y^*) \in \mathcal{S}$, and

$$x_{\text{in}} - x^* = \Pi_{V_x}(x_{\text{in}} - \hat{x}) \in V_x.$$

If the symmetric hypotheses also hold, replace y^* by

$$y^* := \hat{y} + \Pi_{\ker \mathbf{B}^\top}(y_{\text{in}} - \hat{y}).$$

The symmetric version of part (3) preserves $z^* \in \mathcal{S}$, and

$$y_{\text{in}} - y^* = \Pi_{V_y}(y_{\text{in}} - \hat{y}) \in V_y.$$

□

Consequences for the Lyapunov function. Under the x -block hypotheses, kernel translations generate the solution orbit

$$\mathcal{K}_x(z^*) := \{(x^* + u, y^*) : u \in \ker \mathbf{B}\} \subseteq \mathcal{S}.$$

Using the same subgradient $f'(x^*) \in \partial f(x^* + u)$, part (1) of Lemma G.1 gives

$$D_f(x, x^* + u) = f(x) - f(x^*) - \langle f'(x^*), x - x^* \rangle + \langle f'(x^*), u \rangle = D_f(x, x^*).$$

Moreover, if $x - x^* \in V_x$, orthogonality gives

$$\delta_x \|x - x^*\|^2 = \inf_{u \in \ker \mathbf{B}} \delta_x \|x - (x^* + u)\|^2.$$

The symmetric statements hold for the y -block. In particular, when both sets of hypotheses hold and

$$\mathcal{K}(z^*) := z^* + (\ker \mathbf{B} \times \ker \mathbf{B}^\top) \subseteq \mathcal{S},$$

then, for $z - z^* \in V_x \times V_y$,

$$\|z - z^*\|_{\mathbf{P}}^2 = \text{dist}_{\mathbf{P}}(z, \mathcal{K}(z^*))^2.$$

Since $\mathcal{K}(z^*) \subseteq \mathcal{S}$, this quantity upper-bounds $\text{dist}_{\mathbf{P}}(z, \mathcal{S})^2$. The restart analysis uses the fixed aligned representative z^* .

Lemma G.2 (Subspace invariance). *Suppose R_x holds and $\mathcal{X} + \ker \mathbf{B} = \mathcal{X}$. Apply Algorithm 1 to the decomposition (25)–(26) with $\mathbf{P} = \text{diag}(\delta_x \mathbf{I}_{d_x}, \delta_y \mathbf{I}_{d_y})$ and input $z_{\text{in}} = (x_{\text{in}}, y_{\text{in}})$. Then every point $z = (x, y)$ generated by the algorithm satisfies $x - x_{\text{in}} \in V_x$. The symmetric statement holds for y under R_y and $\mathcal{Y} + \ker \mathbf{B}^\top = \mathcal{Y}$.*

Proof. Set

$$K_x := \ker \mathbf{B}, \quad \mathcal{X}_V := \mathcal{X} \cap V_x.$$

The cylinder condition implies

$$\mathcal{X} = \mathcal{X}_V \oplus K_x.$$

Indeed, if $x \in \mathcal{X}$, then $\Pi_{V_x} x = x - \Pi_{K_x} x \in \mathcal{X}_V$. Conversely, $\mathcal{X}_V + K_x \subseteq \mathcal{X}$.

For any $q = q_V + q_K \in V_x \oplus K_x$, projection onto this direct sum separates across the two subspaces:

$$\Pi_{\mathcal{X}}(q) = \Pi_{\mathcal{X}_V}(q_V) + q_K, \quad \text{hence} \quad \Pi_{\mathcal{X}}(q) - q \in V_x.$$

Moreover, because $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathbf{P} = \text{diag}(\delta_x \mathbf{I}, \delta_y \mathbf{I})$, the $\mathbf{P}$ -projection is blockwise:

$$\Pi_{\mathcal{Z}}^{\mathbf{P}}(q_x, q_y) = (\Pi_{\mathcal{X}}(q_x), \Pi_{\mathcal{Y}}(q_y)).$$

Thus, if $q_x \in x_{\text{in}} + V_x$, the projected x -block remains in $x_{\text{in}} + V_x$.

By Appendix C.1, every minimization in Algorithm 1 has the form

$$\Pi_{\mathcal{Z}}^{\mathbf{P}} \left(\frac{\sum_m \theta_m a_m}{\sum_m \theta_m} - \frac{1}{\sum_m \theta_m} \mathbf{P}^{-1} c \right),$$

where c is a sum of terms $\tilde{\nabla} \hat{p}_i(\cdot)$ and $Q_i(\cdot)$. Their x -blocks satisfy

$$\begin{aligned} p_1 : & \quad f'(\cdot) \in V_x & \text{by } R_x, \\ p_2 : & \quad 0, \\ p_3 : & \quad \beta_x \mathbf{B}^\top (\mathbf{B} \cdot -g'(y_{\text{in}})) \in V_x, \\ Q_3 : & \quad \mathbf{B}^\top y \in V_x, \quad Q_1 = Q_2 = 0. \end{aligned}$$

Lemma E.1 supplies exact subgradients. The affine rescaling used to construct $\hat{p}_i$ changes only their evaluation points and scalar factors, so the same inclusions hold for $\tilde{\nabla} \hat{p}_i$. Therefore the x -block of c lies in V_x . Since the x -block of $\mathbf{P}^{-1}$ is $\delta_x^{-1} \mathbf{I}$, the x -block of $\mathbf{P}^{-1} c$ also lies in V_x .

To complete the argument, let

$$\mathcal{A}_x := \{(x, y) : x \in x_{\text{in}} + V_x\}.$$

The initial point belongs to $\mathcal{A}_x$. This affine set is preserved by convex combinations, subtraction of $\mathbf{P}^{-1} c / \sum_m \theta_m$, and the final projection. Induction over the algorithm therefore gives $x - x_{\text{in}} \in V_x$ for every generated point.

The y -block statement follows by the same argument with $(\mathcal{Y}, V_y, \ker \mathbf{B}^\top)$ in place of $(\mathcal{X}, V_x, \ker \mathbf{B})$. $\square$

Remark 1 (Fixed representative across restarts). Choose z^* as in Section 4, aligned with z^0 in each block whose range condition is invoked. Lemma G.2 keeps $x^s \in x^0 + V_x$ when R_x is invoked and $y^s \in y^0 + V_y$ when R_y is invoked. Hence every restart remains aligned with the same z^* , and the contraction inequalities telescope with a fixed Ψ .

Corollary G.3. *Let $z^* \in \mathcal{S}$ be chosen as in Section 4: require $x_{\text{in}} - x^* \in V_x$ when R_x is invoked and $y_{\text{in}} - y^* \in V_y$ when R_y is invoked. Then the output z_{out} of Algorithm 1 satisfies*

$$\|\mathbf{B}(x_{\text{out}} - x^*)\|^2 \geq \mu_{xy}^2 \|x_{\text{out}} - x^*\|^2, \quad \|\mathbf{B}^\top(y_{\text{out}} - y^*)\|^2 \geq \mu_{yx}^2 \|y_{\text{out}} - y^*\|^2.$$

Proof. If R_x does not hold, Assumption 4.3 gives $\mu_{xy}^2 \leq \lambda_{\min}(\mathbf{B}^\top \mathbf{B})$ and the first inequality holds for every vector. Under R_x , Lemma G.2 and the choice of z^* give $x_{\text{out}} - x^* \in V_x$, where the Rayleigh bound $\lambda_{\min}^+(\mathbf{B}^\top \mathbf{B}) \geq \mu_{xy}^2$ applies. The second inequality is symmetric. $\square$

Remark 2 (Normal conditions at z^*). The conditions $\zeta = 0$ when $\beta_x > 0$ and $\xi = 0$ when $\beta_y > 0$ yield, respectively,

$$g'(y^*) = \mathbf{B}x^*, \quad f'(x^*) = -\mathbf{B}^\top y^*.$$

Interior and unconstrained blocks have zero normal vectors. When $\beta_x = \beta_y = 0$, equivalently $\nu_x < 1$ and $\nu_y < 1$, these conditions impose no restriction.

H Proof of Theorem 4.4

Proof. Let $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $z' = (x', y') \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. For brevity, set

$$d_x := x - x', \quad d_y := y - y',$$

and define

$$\mathcal{E} := p(z_{\text{out}}) - p(z^*) + \langle Q(z^*), z_{\text{out}} - z^* \rangle.$$

At the smooth endpoints, we choose

$$\delta_f = 0 \quad \text{if } \nu_x = 1, \quad \delta_g = 0 \quad \text{if } \nu_y = 1.$$

The tolerances for components with exponent below one are balanced later.

For $p_1(z) = f(x)$, Lemma E.1 gives

$$p_1(z) - \tilde{p}_1(z') - \langle \tilde{\nabla} p_1(z'), z - z' \rangle \leq \frac{L_x}{2} \|x - x'\|^2 + \delta_f.$$

Since

$$\|z - z'\|_{\mathbf{P}}^2 = \delta_x \|x - x'\|^2 + \delta_y \|y - y'\|^2,$$

we get

$$\begin{aligned} p_1(z) - \tilde{p}_1(z') - \langle \tilde{\nabla} p_1(z'), z - z' \rangle &\leq \frac{L_x}{2\delta_x} \|z - z'\|_{\mathbf{P}}^2 + \delta_f \\ &= \frac{\kappa_x}{2} \|z - z'\|_{\mathbf{P}}^2 + \delta_f. \end{aligned}$$

Similarly, p_2 admits a (δ_g, κ_y) -oracle with respect to $\|\cdot\|_{\mathbf{P}}$. Hence we may take $L_1 = \kappa_x, L_2 = \kappa_y$.

For p_3 , we have

$$\nabla p_3(z) - \nabla p_3(z') = \begin{pmatrix} \beta_x \mathbf{B}^\top \mathbf{B}(x - x') \\ \beta_y \mathbf{B} \mathbf{B}^\top (y - y') \end{pmatrix}.$$

Define

$$\gamma_B := \max\{\beta_x \delta_y, \beta_y \delta_x\}.$$

Therefore

$$\|\nabla p_3(z) - \nabla p_3(z')\|_{\mathbf{P}^{-1}} \leq \kappa_{xy} \gamma_B \|z - z'\|_{\mathbf{P}}.$$

By the assumptions of the theorem,

$$\beta_x \delta_y \leq \frac{1}{4}, \quad \beta_y \delta_x \leq \frac{1}{4}.$$

Hence

$$\|\nabla p_3(z) - \nabla p_3(z')\|_{\mathbf{P}^{-1}} \leq \frac{\kappa_{xy}}{4} \|z - z'\|_{\mathbf{P}} \leq \kappa_{xy} \|z - z'\|_{\mathbf{P}}.$$

Therefore p_3 is κ_{xy} -smooth with respect to $\|\cdot\|_{\mathbf{P}}$, and we take $L_3 = \kappa_{xy}$.

Since

$$Q_3(z) - Q_3(z') = \begin{pmatrix} \mathbf{B}^\top d_y \\ -\mathbf{B} d_x \end{pmatrix},$$

we can upper-bound $\|Q_3(z) - Q_3(z')\|_{\mathbf{P}^{-1}}$ as follows:

$$\begin{aligned} \|Q_3(z) - Q_3(z')\|_{\mathbf{P}^{-1}}^2 &= \frac{1}{\delta_x} \|\mathbf{B}^\top d_y\|^2 + \frac{1}{\delta_y} \|\mathbf{B} d_x\|^2 \\ &\leq \kappa_{xy} (\delta_x \|d_x\|^2 + \delta_y \|d_y\|^2) \\ &= \kappa_{xy} \|z - z'\|_{\mathbf{P}}^2. \end{aligned}$$

Hence we can choose $M_3 = \sqrt{\kappa_{xy}}$.

We have

$$\begin{aligned}\mathcal{E} &= f(x_{\text{out}}) - f(x^*) + \langle \mathbf{B}^\top y^*, x_{\text{out}} - x^* \rangle \\ &\quad + g(y_{\text{out}}) - g(y^*) - \langle \mathbf{B}x^*, y_{\text{out}} - y^* \rangle \\ &\quad + \frac{\beta_x}{2} (\|\mathbf{B}x_{\text{out}} - g'(y_{\text{in}})\|^2 - \|\mathbf{B}x^* - g'(y_{\text{in}})\|^2) \\ &\quad + \frac{\beta_y}{2} (\|\mathbf{B}^\top y_{\text{out}} + f'(x_{\text{in}})\|^2 - \|\mathbf{B}^\top y^* + f'(x_{\text{in}})\|^2).\end{aligned}$$

By the optimality conditions (28),

$$f'(x^*) + \mathbf{B}^\top y^* = -\xi, \quad g'(y^*) - \mathbf{B}x^* = -\zeta,$$

with $\xi \in N_{\mathcal{X}}(x^*)$ and $\zeta \in N_{\mathcal{Y}}(y^*)$. Since $x_{\text{out}} \in \mathcal{X}$ and $y_{\text{out}} \in \mathcal{Y}$, the definition of the normal cone gives $\langle \xi, x_{\text{out}} - x^* \rangle \leq 0$ and $\langle \zeta, y_{\text{out}} - y^* \rangle \leq 0$. Hence

$$\begin{aligned}f(x_{\text{out}}) - f(x^*) + \langle \mathbf{B}^\top y^*, x_{\text{out}} - x^* \rangle &= D_f(x_{\text{out}}, x^*) \\ &\quad - \langle \xi, x_{\text{out}} - x^* \rangle \\ &\geq D_f(x_{\text{out}}, x^*), \\ g(y_{\text{out}}) - g(y^*) - \langle \mathbf{B}x^*, y_{\text{out}} - y^* \rangle &= D_g(y_{\text{out}}, y^*) \\ &\quad - \langle \zeta, y_{\text{out}} - y^* \rangle \\ &\geq D_g(y_{\text{out}}, y^*).\end{aligned}$$

For brevity, write

$$\begin{aligned}D_{f,o} &:= D_f(x_{\text{out}}, x^*), & D_{g,o} &:= D_g(y_{\text{out}}, y^*), \\ D_{f,i} &:= D_f(x_{\text{in}}, x^*), & D_{g,i} &:= D_g(y_{\text{in}}, y^*).\end{aligned}$$

For the quadratic terms, we use the elementary inequality

$$\frac{1}{2}\|a + b\|^2 - \frac{1}{2}\|b\|^2 \geq \frac{1}{4}\|a\|^2 - \|b\|^2.$$

Multiplying this inequality by the nonnegative coefficients β_x and β_y , and applying it with $a = \mathbf{B}(x_{\text{out}} - x^*)$, $b = \mathbf{B}x^* - g'(y_{\text{in}})$, and then with $a = \mathbf{B}^\top(y_{\text{out}} - y^*)$, $b = \mathbf{B}^\top y^* + f'(x_{\text{in}})$, gives

$$\begin{aligned}\mathcal{E} &\geq D_{f,o} + D_{g,o} \\ &\quad + \frac{\beta_x}{4} \|\mathbf{B}(x_{\text{out}} - x^*)\|^2 \\ &\quad + \frac{\beta_y}{4} \|\mathbf{B}^\top(y_{\text{out}} - y^*)\|^2 \\ &\quad - \beta_x \|\mathbf{B}x^* - g'(y_{\text{in}})\|^2 - \beta_y \|\mathbf{B}^\top y^* + f'(x_{\text{in}})\|^2.\end{aligned}$$

For active coefficients, the normal conditions in Section 4 and (28) give

$$\mathbf{B}x^* = g'(y^*) \text{ in the } \beta_x\text{-term,} \quad \mathbf{B}^\top y^* = -f'(x^*) \text{ in the } \beta_y\text{-term}$$

(see Remark 2). We obtain

$$\begin{aligned}\mathcal{E} &\geq D_{f,o} + D_{g,o} \\ &\quad + \frac{\beta_x}{4} \|\mathbf{B}(x_{\text{out}} - x^*)\|^2 \\ &\quad + \frac{\beta_y}{4} \|\mathbf{B}^\top(y_{\text{out}} - y^*)\|^2 \\ &\quad - \beta_x \|g'(y_{\text{in}}) - g'(y^*)\|^2 - \beta_y \|f'(x_{\text{in}}) - f'(x^*)\|^2.\end{aligned}$$

If $\beta_x > 0$, then $\nu_y = 1$, for which we chose $\delta_g = 0$, and the ambient self-bounding inequality (20) gives

$$\|g'(y_{\text{in}}) - g'(y^*)\|^2 \leq 2L_y D_g(y_{\text{in}}, y^*).$$

Similarly, if $\beta_y > 0$, then $\nu_x = 1$, for which we chose $\delta_f = 0$, and (20) gives

$$\|f'(x_{\text{in}}) - f'(x^*)\|^2 \leq 2L_x D_f(x_{\text{in}}, x^*).$$

Therefore, using the definition of β_x and β_y in (21),

$$2\beta_x L_y \leq \frac{1}{2}, \quad 2\beta_y L_x \leq \frac{1}{2}.$$

Hence

$$\beta_x \|g'(y_{\text{in}}) - g'(y^*)\|^2 + \beta_y \|f'(x_{\text{in}}) - f'(x^*)\|^2 \leq \frac{1}{2}(D_{g,i} + D_{f,i}).$$

Substituting this into the previous lower bound gives

$$\begin{aligned} \mathcal{E} &\geq D_{f,o} + D_{g,o} \\ &\quad + \frac{\beta_x}{4} \|\mathbf{B}(x_{\text{out}} - x^*)\|^2 \\ &\quad + \frac{\beta_y}{4} \|\mathbf{B}^\top(y_{\text{out}} - y^*)\|^2 \\ &\quad - \frac{1}{2}(D_{f,i} + D_{g,i}). \end{aligned}$$

Using Corollary G.3, which supplies

$$\begin{aligned} \|\mathbf{B}(x_{\text{out}} - x^*)\|^2 &\geq \mu_{xy}^2 \|x_{\text{out}} - x^*\|^2, \\ \|\mathbf{B}^\top(y_{\text{out}} - y^*)\|^2 &\geq \mu_{yx}^2 \|y_{\text{out}} - y^*\|^2, \end{aligned}$$

and strong convexity,

$$\begin{aligned} D_f(x_{\text{out}}, x^*) &\geq \frac{\mu_x}{2} \|x_{\text{out}} - x^*\|^2, \\ D_g(y_{\text{out}}, y^*) &\geq \frac{\mu_y}{2} \|y_{\text{out}} - y^*\|^2, \end{aligned}$$

we obtain

$$\begin{aligned} c_x &:= \frac{\mu_x}{8} + \frac{\beta_x \mu_{xy}^2}{4}, \\ c_y &:= \frac{\mu_y}{8} + \frac{\beta_y \mu_{yx}^2}{4}. \end{aligned}$$

Then

$$\mathcal{E} \geq \frac{3}{4}(D_{f,o} + D_{g,o}) - \frac{1}{2}(D_{f,i} + D_{g,i}) + c_x \|x_{\text{out}} - x^*\|^2 + c_y \|y_{\text{out}} - y^*\|^2.$$

By the definitions of the effective strong convexity parameters, equivalently,

$$\begin{aligned} \delta_x &= \mu_x + 4\beta_x \mu_{xy}^2, \\ \delta_y &= \mu_y + 4\beta_y \mu_{yx}^2, \end{aligned}$$

we have

$$c_x \geq \frac{\delta_x}{16}, \quad c_y \geq \frac{\delta_y}{16}.$$

Therefore,

$$\mathcal{E} \geq \frac{3}{4}(D_{f,o} + D_{g,o}) - \frac{1}{2}(D_{f,i} + D_{g,i}) + \frac{1}{16} \|z_{\text{out}} - z^*\|_{\mathbf{P}}^2.$$

Let

$$R_{\text{in}}^2 := \|z_{\text{in}} - z^*\|_{\mathbf{P}}^2, \quad A_i := \prod_{j=1}^i T_j,$$

so that A_i is the number of evaluations of the component placed at level i , and set

$$\bar{N}_f := A_{\pi^{-1}(1)}, \quad \bar{N}_g := A_{\pi^{-1}(2)}, \quad \bar{N}_B := A_{\pi^{-1}(3)}.$$

Thus $\bar{N}_f, \bar{N}_g, \bar{N}_B$ are the component-oracle call budgets, irrespective of the levels to which π assigns the components. Applying Theorem 3.3 with comparator $z = z^*$, we obtain

$$\Theta := \frac{\kappa_x}{\bar{N}_f^2} + \frac{\kappa_y}{\bar{N}_g^2} + \frac{\kappa_{xy}}{\bar{N}_B^2} + \frac{\sqrt{\kappa_{xy}}}{\bar{N}_B}.$$

$$\mathcal{E} \leq C_{\text{VI}} \Theta R_{\text{in}}^2 + \bar{N}_f \delta_f + \bar{N}_g \delta_g, \quad (71)$$

where $C_{\text{VI}} > 0$ absorbs the level-dependent factors 2^{2i-1} and 2^{i-1} from Theorem 3.3. These factors are at most 2^5 for $n = 3$, uniformly over π , so Θ depends only on the component budgets $\bar{N}_f, \bar{N}_g, \bar{N}_B$.

Combining the upper bound (71) with the lower bound obtained above, the oracle-error terms are

$$\bar{N}_f \delta_f + \bar{N}_g \delta_g.$$

We now choose $\bar{N}_f, \bar{N}_g, \bar{N}_B$ and δ_f, δ_g so that, for a numerical constant $\rho > 0$,

$$C_{\text{VI}} \Theta \leq \rho$$

and

$$\bar{N}_f \delta_f \leq \rho \Omega_{\text{in}}, \quad \bar{N}_g \delta_g \leq \rho \Omega_{\text{in}}.$$

Since $R_{\text{in}}^2 \leq \Psi(z_{\text{in}}) \leq \Omega_{\text{in}}$, we get

$$\mathcal{E} \leq \rho R_{\text{in}}^2 + 2\rho \Omega_{\text{in}}.$$

Hence, combining this upper bound with the lower bound obtained above,

$$\begin{aligned} & \frac{1}{16} \|z_{\text{out}} - z^*\|_{\mathbf{P}}^2 + \frac{3}{4} (\text{D}_f(x_{\text{out}}, x^*) + \text{D}_g(y_{\text{out}}, y^*)) \\ & \leq \rho R_{\text{in}}^2 + 2\rho \Omega_{\text{in}} + \frac{1}{2} (\text{D}_f(x_{\text{in}}, x^*) + \text{D}_g(y_{\text{in}}, y^*)). \end{aligned}$$

Multiplying both sides by 16 gives

$$\begin{aligned} & \|z_{\text{out}} - z^*\|_{\mathbf{P}}^2 + 12 (\text{D}_f(x_{\text{out}}, x^*) + \text{D}_g(y_{\text{out}}, y^*)) \\ & \leq 16\rho R_{\text{in}}^2 + 32\rho \Omega_{\text{in}} + 8 (\text{D}_f(x_{\text{in}}, x^*) + \text{D}_g(y_{\text{in}}, y^*)). \end{aligned}$$

Since

$$\Psi(z_{\text{in}}) = R_{\text{in}}^2 + 12 \text{D}_f(x_{\text{in}}, x^*) + 12 \text{D}_g(y_{\text{in}}, y^*) \leq \Omega_{\text{in}},$$

and with $16\rho \leq 2/3$, we have

$$\begin{aligned} 16\rho R_{\text{in}}^2 + 8(D_{f,i} + D_{g,i}) & \leq \frac{2}{3} [R_{\text{in}}^2 + 12(D_{f,i} + D_{g,i})] \\ & = \frac{2}{3} \Psi(z_{\text{in}}) \leq \frac{2}{3} \Omega_{\text{in}}. \end{aligned}$$

Therefore,

$$\Psi(z_{\text{out}}) \leq \left(\frac{2}{3} + 32\rho \right) \Omega_{\text{in}}.$$

Choosing

$$\rho = \frac{1}{384},$$

we obtain

$$\frac{2}{3} + 32\rho = \frac{2}{3} + \frac{1}{12} = \frac{3}{4}.$$

Consequently,

$$\Psi(z_{\text{out}}) \leq \frac{3}{4} \Omega_{\text{in}}.$$

It remains to verify the claimed evaluation bounds. We use the same balancing argument as in the proof of Theorem 3.5.

We first consider the x -component. Suppose $0 \leq \nu_x < 1$, and define

$$a_x := \frac{1 - \nu_x}{1 + \nu_x}, \quad L_x(\delta_f) = K_x \delta_f^{-a_x},$$

where

$$K_x = \left[\frac{1 - \nu_x}{2(1 + \nu_x)} \right]^{\frac{1 - \nu_x}{1 + \nu_x}} H_x^{\frac{2}{1 + \nu_x}}.$$

The two conditions that need to be satisfied are

$$C_{\text{VI}} \frac{\kappa_x}{\bar{N}_f^2} = C_{\text{VI}} \frac{L_x(\delta_f)}{\delta_x \bar{N}_f^2} \leq \frac{\rho}{4}, \quad \bar{N}_f \delta_f \leq \rho \Omega_{\text{in}}.$$

Choose

$$\delta_f := \frac{\rho \Omega_{\text{in}}}{\bar{N}_f}.$$

Then the oracle-error condition $\bar{N}_f \delta_f \leq \rho \Omega_{\text{in}}$ holds with equality. Moreover,

$$L_x(\delta_f) = K_x \left(\frac{\rho \Omega_{\text{in}}}{\bar{N}_f} \right)^{-a_x} = K_x \rho^{-a_x} \Omega_{\text{in}}^{-a_x} \bar{N}_f^{a_x}.$$

Therefore, the first condition becomes

$$C_{\text{VI}} \frac{K_x \rho^{-a_x} \Omega_{\text{in}}^{-a_x} \bar{N}_f^{a_x}}{\delta_x \bar{N}_f^2} \leq \frac{\rho}{4}.$$

Equivalently,

$$\bar{N}_f^{2-a_x} \geq \frac{4C_{\text{VI}}}{\rho^{1+a_x}} \frac{K_x}{\delta_x} \Omega_{\text{in}}^{-a_x}.$$

Thus it is enough to choose

$$\bar{N}_f \geq \left(\frac{4C_{\text{VI}}}{\rho^{1+a_x}} \frac{K_x}{\delta_x} \Omega_{\text{in}}^{-a_x} \right)^{\frac{1}{2-a_x}}.$$

Since

$$2 - a_x = 2 - \frac{1 - \nu_x}{1 + \nu_x} = \frac{1 + 3\nu_x}{1 + \nu_x},$$

we obtain

$$\bar{N}_f = \mathcal{O} \left(K_x^{\frac{1+\nu_x}{1+3\nu_x}} \delta_x^{-\frac{1+\nu_x}{1+3\nu_x}} \Omega_{\text{in}}^{\frac{\nu_x-1}{1+3\nu_x}} \right).$$

Using the definition of K_x , this becomes

$$\bar{N}_f = \mathcal{O} \left(H_x^{\frac{2}{1+3\nu_x}} \delta_x^{-\frac{1+\nu_x}{1+3\nu_x}} \Omega_{\text{in}}^{\frac{\nu_x-1}{1+3\nu_x}} \right).$$

Equivalently, with

$$\tilde{H}_x := H_x \delta_x^{-(1+\nu_x)/2},$$

we have

$$\bar{N}_f = \mathcal{O} \left(\tilde{H}_x^{\frac{2}{1+3\nu_x}} \Omega_{\text{in}}^{\frac{\nu_x-1}{1+3\nu_x}} \right).$$

When $\nu_x = 1$, we take $\delta_f = 0$ and $L_x = H_x$. Then the only condition involving $\bar{N}_f$ is

$$C_{\text{VI}} \frac{\kappa_x}{\bar{N}_f^2} = C_{\text{VI}} \frac{L_x}{\delta_x \bar{N}_f^2} \leq \frac{\rho}{4},$$

which is satisfied by

$$\bar{N}_f = \mathcal{O} \left(\sqrt{\frac{L_x}{\delta_x}} \right) = \mathcal{O}(\sqrt{\kappa_x}).$$

The same two cases, with $C_{\text{VI}}\kappa_y/\bar{N}_g^2 \leq \rho/4$, give

$$\bar{N}_g = \mathcal{O}\left(\tilde{H}_y^{\frac{2}{1+3\nu_y}} \Omega_{\text{in}}^{\frac{\nu_y-1}{1+3\nu_y}}\right), \quad \tilde{H}_y := H_y \delta_y^{-(1+\nu_y)/2}.$$

Finally, the B -dependent terms require

$$C_{\text{VI}} \frac{\kappa_{xy}}{\bar{N}_B^2} \leq \frac{\rho}{4}, \quad C_{\text{VI}} \frac{\sqrt{\kappa_{xy}}}{\bar{N}_B} \leq \frac{\rho}{4}.$$

These are satisfied by

$$\bar{N}_B = \mathcal{O}\left(\sqrt{\kappa_{xy}}\right).$$

The preceding estimates hold once component m receives cR_m evaluations for some constant $c \geq 1$. Let π order the budgets as $R_{\pi(1)} \leq R_{\pi(2)} \leq R_{\pi(3)}$, matching the nondecreasing cumulative counts $A_i := \prod_{j \leq i} T_j$. Set $A_0 := 1$ and

$$T_i := \max\left\{1, \left\lceil \frac{cR_{\pi(i)}}{A_{i-1}} \right\rceil\right\}, \quad A_i := A_{i-1}T_i.$$

The induction in Appendix F gives

$$cR_{\pi(i)} \leq A_i \leq 2cR_{\pi(i)}, \quad i = 1, 2, 3,$$

because every $R_m \geq 1$. Hence

$$\bar{N}_f = \mathcal{O}(R_1), \quad \bar{N}_g = \mathcal{O}(R_2), \quad \bar{N}_B = \mathcal{O}(R_3).$$

Accounting for one anchor-gradient call of each type in p_3 and a constant number of $\mathbf{B}, \mathbf{B}^\top$ products per p_3, Q_3 call, the evaluation counts satisfy, for an absolute constant C_B ,

$$N_f \leq \bar{N}_f + 1, \quad N_g \leq \bar{N}_g + 1, \quad N_B \leq C_B(\bar{N}_B + 1).$$

Since $R_1, R_2, R_3 \geq 1$, the claimed bounds for N_f, N_g, N_B follow. $\square$

I Proof of Corollary 4.5

Proof. By Remark 1, every restart uses the function Ψ from Section 4 and the same representative z^* , so the one-step inequalities share a common reference point.

The case $S = 0$ is immediate. Assume $S \geq 1$.

Let z^s be the point at the beginning of restart s , and write

$$\Psi_s := \Psi(z^s), \quad \Psi_0 := \Psi(z^0), \quad \Omega_s := \left(\frac{3}{4}\right)^s \Psi_0.$$

We prove by induction that

$$\Psi_s \leq \Omega_s, \quad s = 0, 1, \dots, S.$$

The claim is true for $s = 0$. If $\Psi_s \leq \Omega_s$, then Theorem 4.4, applied with $\Omega_{\text{in}} = \Omega_s$, gives

$$\Psi_{s+1} \leq \frac{3}{4}\Omega_s = \Omega_{s+1}.$$

Therefore,

$$\Psi_S \leq \Omega_S = \left(\frac{3}{4}\right)^S \Psi_0 \leq \varepsilon$$

by the definition of S .

We now bound the total number of evaluations. For fixed problem parameters and every $\alpha > 0$,

$$1 + \log(\Psi_0/\varepsilon) = \mathcal{O}(\varepsilon^{-\alpha}) \quad \text{as } \varepsilon \downarrow 0.$$

First consider the f -component. If $0 \leq \nu_x < 1$, define

$$\alpha_x := \frac{1 - \nu_x}{1 + 3\nu_x} > 0.$$

At restart s , Theorem 4.4 gives

$$N_{f,s} = \mathcal{O} \left(\max \left\{ \tilde{H}_x^{\frac{2}{1+3\nu_x}} \Omega_s^{-\alpha_x}, 1 \right\} \right).$$

Hence

$$\begin{aligned} N_f^{\text{tot}} &= \sum_{s=0}^{S-1} N_{f,s} \\ &= \mathcal{O} \left(\tilde{H}_x^{\frac{2}{1+3\nu_x}} \sum_{s=0}^{S-1} \Omega_s^{-\alpha_x} + S \right) \\ &= \mathcal{O} \left(\tilde{H}_x^{\frac{2}{1+3\nu_x}} \Psi_0^{-\alpha_x} \sum_{s=0}^{S-1} \left(\frac{3}{4} \right)^{-\alpha_x s} + S \right). \end{aligned}$$

By the minimality of S ,

$$\varepsilon < \Omega_{S-1} \leq \frac{4}{3} \varepsilon.$$

The geometric sum is therefore bounded by

$$\sum_{s=0}^{S-1} \Omega_s^{-\alpha_x} \leq \Omega_{S-1}^{-\alpha_x} \sum_{j=0}^{\infty} \left(\frac{3}{4} \right)^{\alpha_x j} = \mathcal{O}(\varepsilon^{-\alpha_x}).$$

Thus

$$N_f^{\text{tot}} = \mathcal{O} \left(\tilde{H}_x^{\frac{2}{1+3\nu_x}} \varepsilon^{-\alpha_x} + S \right).$$

Because $\alpha_x > 0$, this becomes

$$N_f^{\text{tot}} = \mathcal{O} \left(\tilde{H}_x^{\frac{2}{1+3\nu_x}} \varepsilon^{\frac{\nu_x-1}{1+3\nu_x}} \right).$$

If $\nu_x = 1$, then each restart requires

$$N_{f,s} = \mathcal{O}(\sqrt{\kappa_x}),$$

and hence

$$N_f^{\text{tot}} = \sum_{s=0}^{S-1} N_{f,s} = \mathcal{O} \left(\sqrt{\kappa_x} \left(1 + \log \frac{\Psi_0}{\varepsilon} \right) \right).$$

The proof for the g -component is identical. If $0 \leq \nu_y < 1$, with

$$\alpha_y := \frac{1 - \nu_y}{1 + 3\nu_y},$$

we obtain

$$N_g^{\text{tot}} = \mathcal{O} \left(\tilde{H}_y^{\frac{2}{1+3\nu_y}} \varepsilon^{-\alpha_y} + S \right).$$

Because $\alpha_y > 0$, this becomes

$$N_g^{\text{tot}} = \mathcal{O} \left(\tilde{H}_y^{\frac{2}{1+3\nu_y}} \varepsilon^{\frac{\nu_y-1}{1+3\nu_y}} \right).$$

If $\nu_y = 1$, then

$$N_g^{\text{tot}} = \mathcal{O} \left(\sqrt{\kappa_y} \left(1 + \log \frac{\Psi_0}{\varepsilon} \right) \right).$$

Finally, at each restart,

$$N_{B,s} = \mathcal{O}(\max\{\sqrt{\kappa_{xy}}, 1\}).$$

Therefore

$$N_B^{\text{tot}} = \sum_{s=0}^{S-1} N_{B,s} = \mathcal{O} \left(\max\{\sqrt{\kappa_{xy}}, 1\} \left(1 + \log \frac{\Psi_0}{\varepsilon} \right) \right).$$

This proves the corollary. $\square$

J The Degenerate and Mixed Cases

The degenerate regime $\delta_x = \delta_y = 0$ of Section 4 yields the complexities collected in Table 4.

Oracle	Complexity
Evaluations of f'	$\mathcal{O}\left(\left(\frac{H_x \Omega^{(1+\nu_x)/2}}{\varepsilon}\right)^{\frac{2}{1+3\nu_x}}\right)$
Evaluations of g'	$\mathcal{O}\left(\left(\frac{H_y \Omega^{(1+\nu_y)/2}}{\varepsilon}\right)^{\frac{2}{1+3\nu_y}}\right)$
Products with $\mathbf{B}, \mathbf{B}^\top$	$\mathcal{O}\left(\frac{L_{xy} \Omega}{\varepsilon}\right)$

Table 4: Evaluation complexity for the bilinear saddle-point problem in the degenerate case $\delta_x = \delta_y = 0$. Here $\Omega := \sup_{z \in \mathcal{Z}} \|z^0 - z\|^2$, and the bounds hold for $0 \leq \nu_x, \nu_y \leq 1$.

The mixed cases $\delta_x > 0, \delta_y = 0$ and $\delta_x = 0, \delta_y > 0$. In the mixed regimes, exactly one effective strong-convexity parameter is positive. We regularize the zero-curvature block and rebuild the active part of p_3 with the regularized constant and anchor.

By symmetry, it suffices to consider

$$\delta_x > 0, \quad \delta_y = 0.$$

Because $\delta_y = \mu_y + 4\beta_y \mu_{yx}^2 = 0$, nonnegativity implies

$$\mu_y = 0, \quad \beta_y \mu_{yx}^2 = 0.$$

Assume that $\mathcal{Y}$ is bounded, fix $y^0 \in \mathbb{R}^{d_y}$, and define

$$\Omega_y := \sup_{y \in \mathcal{Y}} \|y - y^0\|^2 < \infty, \quad D_{\mathcal{Y}} := \sup_{u, v \in \mathcal{Y}} \|u - v\| \leq 2\sqrt{\Omega_y}.$$

If $\Omega_y = 0$, then $\mathcal{Y} = \{y^0\}$ and the y -block can be eliminated. Hence assume $\Omega_y > 0$. For a target accuracy $\varepsilon > 0$, set

$$\lambda_y := \frac{\varepsilon}{\Omega_y},$$

and replace g by

$$g_\lambda(y) := g(y) + \frac{\lambda_y}{2} \|y - y^0\|^2.$$

For $u, v \in \mathcal{Y}$,

$$\lambda_y \|u - v\| \leq \lambda_y D_{\mathcal{Y}}^{1-\nu_y} \|u - v\|^{\nu_y}.$$

Consequently, g_λ is λ_y -strongly convex and satisfies the $(\nu_y, H_{y,\lambda})$ -Hölder condition with

$$H_{y,\lambda} := H_y + \lambda_y D_{\mathcal{Y}}^{1-\nu_y}. \quad (72)$$

For $\nu_y = 1$, the function g_λ is smooth on the ambient space with constant $L_{y,\lambda} = L_y + \lambda_y$. In particular, (20) holds for g_λ with this constant. The coefficient of the x -part of the coupling regularizer must therefore be recomputed:

$$\beta_x^\lambda := \begin{cases} \frac{1}{4(L_y + \lambda_y)}, & \nu_y = 1, \\ 0, & 0 \leq \nu_y < 1, \end{cases} \quad \beta_y^\lambda := \beta_y. \quad (73)$$

The effective strong convexity parameters of the regularized problem are

$$\begin{aligned} \delta_x^\lambda &:= \mu_x + 4\beta_x^\lambda \mu_{xy}^2, \\ \delta_y^\lambda &:= \mu_y + \lambda_y + 4\beta_y^\lambda \mu_{yx}^2 = \lambda_y. \end{aligned} \quad (74)$$

The last equality uses $\delta_y = 0$. Thus $\delta_x^\lambda = \delta_x$ when $\nu_y < 1$, but generally not when $\nu_y = 1$.

At every restart with input $z_{\text{in}} = (x_{\text{in}}, y_{\text{in}})$, use

$$g'_\lambda(y_{\text{in}}) = g'(y_{\text{in}}) + \lambda_y(y_{\text{in}} - y^0)$$

and replace (25) by

$$\begin{aligned} p_{3,\lambda}(x, y) := & \frac{\beta_x^\lambda}{2} \|\mathbf{B}x - g'_\lambda(y_{\text{in}})\|^2 \\ & + \frac{\beta_y^\lambda}{2} \|\mathbf{B}^\top y + f'(x_{\text{in}})\|^2. \end{aligned} \quad (75)$$

The other components are $p_1 = f$, $p_2 = g_\lambda$, and the same bilinear operator Q_3 as in (26). The norm and condition parameters used by Theorem 4.4, which requires positive effective strong convexity parameters, are

$$\begin{aligned} \mathbf{P}_\lambda &= \text{diag}(\delta_x^\lambda \mathbf{I}_{d_x}, \delta_y^\lambda \mathbf{I}_{d_y}), \\ \kappa_x^\lambda &= \frac{L_x}{\delta_x^\lambda}, \quad \kappa_{xy}^\lambda = \frac{L_{xy}^2}{\delta_x^\lambda \lambda_y}. \end{aligned}$$

Assume that the regularized problem has a saddle point $z_\lambda^* = (x_\lambda^*, y_\lambda^*)$, and choose subgradients and normal vectors satisfying

$$f'(x_\lambda^*) + \mathbf{B}^\top y_\lambda^* + \xi_\lambda = 0, \quad g'_\lambda(y_\lambda^*) - \mathbf{B}x_\lambda^* + \zeta_\lambda = 0.$$

Applying Theorem 4.4 requires

$$\begin{aligned} \zeta_\lambda &= 0 \quad \text{if } \beta_x^\lambda > 0, & \xi_\lambda &= 0 \quad \text{if } \beta_y^\lambda > 0, \\ \beta_x^\lambda \delta_y^\lambda &\leq \frac{1}{4}, & \beta_y^\lambda \delta_x^\lambda &\leq \frac{1}{4}. \end{aligned} \quad (76)$$

The first inequality follows from $\delta_y^\lambda = \lambda_y$ and $\beta_x^\lambda \lambda_y \leq 1/4$. The second is automatic when either component is nonsmooth. In the fully smooth case it is retained as a compatibility assumption. We set $\mu_{yx} = 0$ on the regularized y -block.

Let Ψ_λ be the Lyapunov function (30) formed with $\mathbf{P}_\lambda$, g_λ , and z_λ^* . Apply Corollary 4.5 to $(f, g_\lambda, p_{3,\lambda}, Q_3)$, with terminal target $\varepsilon/2$. Its output z_λ^S satisfies

$$\Psi_\lambda(z_\lambda^S) \leq \frac{\varepsilon}{2}.$$

Define

$$\tilde{H}_x^\lambda := H_x(\delta_x^\lambda)^{-(1+\nu_x)/2}.$$

The direct bound for the f -oracle is

$$N_f = \begin{cases} \mathcal{O}\left((\tilde{H}_x^\lambda)^{\frac{2}{1+3\nu_x}} \varepsilon^{-\frac{1-\nu_x}{1+3\nu_x}}\right), & 0 \leq \nu_x < 1, \\ \tilde{\mathcal{O}}\left(\sqrt{\kappa_x^\lambda}\right), & \nu_x = 1, \end{cases}.$$

For the regularized g -oracle and the bilinear operator, the direct bounds are

$$N_g = \begin{cases} \mathcal{O}\left(\left(\frac{H_{y,\lambda} \Omega_y^{(1+\nu_y)/2}}{\varepsilon}\right)^{\frac{2}{1+3\nu_y}}\right), & 0 \leq \nu_y < 1, \\ \tilde{\mathcal{O}}\left(\sqrt{1 + \frac{L_y \Omega_y}{\varepsilon}}\right), & \nu_y = 1, \end{cases} \quad (77)$$

and

$$N_B = \tilde{\mathcal{O}}\left(L_{xy} \sqrt{\frac{\Omega_y}{\delta_x^\lambda \varepsilon}}\right). \quad (78)$$

For any saddle function H on $\mathcal{X} \times \mathcal{Y}$, define

$$\text{Gap}_H(x, y) := \max_{v \in \mathcal{Y}} H(x, v) - \min_{u \in \mathcal{X}} H(u, y).$$

Write $\text{Gap} := \text{Gap}_F$ and $\text{Gap}_{\lambda_y} := \text{Gap}_{F_{\lambda_y}}$, where

$$F_{\lambda_y}(x, y) := F(x, y) - \frac{\lambda_y}{2} \|y - y^0\|^2.$$

The exact regularized saddle has $\text{Gap}_{\lambda_y}(z_\lambda^*) = 0$ and satisfies

$$\text{Gap}(z_\lambda^*) \leq \text{Gap}_{\lambda_y}(z_\lambda^*) + \frac{\lambda_y}{2} \Omega_y = \frac{\varepsilon}{2}.$$

Thus z_λ^* is an $\varepsilon/2$ -weak-gap solution of the original problem. It remains to transfer the Lyapunov guarantee for z_λ^S to a weak-gap bound for the original problem (1).

Recall

$$F_{\lambda_y}(x, y) = F(x, y) - \frac{\lambda_y}{2} \|y - y^0\|^2 = f(x) + \langle y, \mathbf{B}x \rangle - g_\lambda(y),$$

and set

$$\varphi_\lambda(x) := \max_{v \in \mathcal{Y}} F_{\lambda_y}(x, v), \quad \psi_\lambda(y) := \min_{u \in \mathcal{X}} F_{\lambda_y}(u, y),$$

so that $\text{Gap}_{\lambda_y}(z) = \varphi_\lambda(x) - \psi_\lambda(y)$ and $\varphi_\lambda(x_\lambda^*) = \psi_\lambda(y_\lambda^*) = F_{\lambda_y}(z_\lambda^*)$.

Define

$$h^*(w) := \max_{v \in \mathcal{Y}} \{\langle w, v \rangle - g_\lambda(v)\}, \quad q^*(w') := \max_{u \in \mathcal{X}} \{\langle w', u \rangle - f(u)\}, \quad (79)$$

so that $\varphi_\lambda(x) = f(x) + h^*(\mathbf{B}x)$ and $\psi_\lambda(y) = -q^*(-\mathbf{B}^\top y) - g_\lambda(y)$. The first maximum is finite because $\mathcal{Y}$ is bounded, the second by Assumption J.1(ii), which is in force from here on. We write

$$D_{h^*}(w, w') := h^*(w) - h^*(w') - \langle \nabla h^*(w'), w - w' \rangle,$$

and, for the fixed choice $x_\lambda^* \in \partial q^*(-\mathbf{B}^\top y_\lambda^*)$ supplied by Lemma J.2,

$$D_{q^*}(w, -\mathbf{B}^\top y_\lambda^*) := q^*(w) - q^*(-\mathbf{B}^\top y_\lambda^*) - \langle x_\lambda^*, w + \mathbf{B}^\top y_\lambda^* \rangle.$$

Assumption J.1. In addition to (76):

(i) $\xi_\lambda = 0$ and $\zeta_\lambda = 0$.

(ii) either $\mu_x > 0$, or $\mathcal{X}$ is bounded, in which case we set $D_{\mathcal{X}} := \sup_{u, u' \in \mathcal{X}} \|u - u'\| < \infty$.

Lemma J.2. Let Assumption J.1(ii) hold, so that q^* is finite. The function h^* is differentiable on $\mathbb{R}^{d_y}$, with $\nabla h^*(w) = \arg \max_{v \in \mathcal{Y}} \{\langle w, v \rangle - g_\lambda(v)\}$ and $1/\lambda_y$ -Lipschitz gradient. In particular

$$D_{h^*}(w, w') \leq \frac{\|w - w'\|^2}{2\lambda_y}, \quad \nabla h^*(\mathbf{B}x_\lambda^*) = y_\lambda^*. \quad (80)$$

Moreover $x_\lambda^* \in \partial q^*(-\mathbf{B}^\top y_\lambda^*)$, and for every $y \in \mathcal{Y}$,

$$D_{q^*}(-\mathbf{B}^\top y, -\mathbf{B}^\top y_\lambda^*) \leq \max_{u \in \mathcal{X}} \left\{ \langle -\mathbf{B}^\top (y - y_\lambda^*), u - x_\lambda^* \rangle - D_f(u, x_\lambda^*) \right\}. \quad (81)$$

Proof. The function equal to g_λ on $\mathcal{Y}$ and to $+\infty$ elsewhere is proper, closed and λ_y -strongly convex, and h^* is its Fenchel conjugate. Hence h^* is differentiable with $1/\lambda_y$ -Lipschitz gradient given by the unique maximizer, and the first part of (80) is the descent lemma for h^* . Since z_λ^* is a saddle point of F_{λ_y} , the point y_λ^* maximizes $v \mapsto \langle \mathbf{B}x_\lambda^*, v \rangle - g_\lambda(v)$ over $\mathcal{Y}$, which is the second part. Likewise x_λ^* minimizes $u \mapsto f(u) + \langle \mathbf{B}^\top y_\lambda^*, u \rangle$ over $\mathcal{X}$, i.e. $x_\lambda^* \in \partial q^*(-\mathbf{B}^\top y_\lambda^*)$.

For (81), write $w := -\mathbf{B}^\top y$ and $w' := -\mathbf{B}^\top y_\lambda^*$. Using $q^*(w') = \langle w', x_\lambda^* \rangle - f(x_\lambda^*)$,

$$D_{q^*}(w, w') = \max_{u \in \mathcal{X}} \{ \langle w, u - x_\lambda^* \rangle - (f(u) - f(x_\lambda^*)) \}.$$

By the regularized optimality conditions, $f'(x_\lambda^*) = w' - \xi_\lambda$ with $\xi_\lambda \in N_{\mathcal{X}}(x_\lambda^*)$, so for every $u \in \mathcal{X}$,

$$f(u) - f(x_\lambda^*) = D_f(u, x_\lambda^*) + \langle w', u - x_\lambda^* \rangle - \langle \xi_\lambda, u - x_\lambda^* \rangle \geq D_f(u, x_\lambda^*) + \langle w', u - x_\lambda^* \rangle,$$

because $\langle \xi_\lambda, u - x_\lambda^* \rangle \leq 0$. Substituting gives the claim. $\square$

Lemma J.3. *Let Assumption J.1(ii) hold. For every $z = (x, y) \in \mathcal{Z}$,*

$$\begin{aligned} \text{Gap}_{\lambda_y}(z) &= D_f(x, x_\lambda^*) + D_{h^*}(\mathbf{B}x, \mathbf{B}x_\lambda^*) - \langle \xi_\lambda, x - x_\lambda^* \rangle \\ &\quad + D_{g_\lambda}(y, y_\lambda^*) + D_{q^*}(-\mathbf{B}^\top y, -\mathbf{B}^\top y_\lambda^*) - \langle \zeta_\lambda, y - y_\lambda^* \rangle. \end{aligned} \quad (82)$$

Proof. Since $\varphi_\lambda = f + h^* \circ \mathbf{B}$, Lemma J.2 gives

$$\begin{aligned} \varphi_\lambda(x) - \varphi_\lambda(x_\lambda^*) &= D_f(x, x_\lambda^*) + \langle f'(x_\lambda^*), x - x_\lambda^* \rangle \\ &\quad + D_{h^*}(\mathbf{B}x, \mathbf{B}x_\lambda^*) + \langle y_\lambda^*, \mathbf{B}(x - x_\lambda^*) \rangle, \end{aligned}$$

and the two linear terms combine into $\langle f'(x_\lambda^*) + \mathbf{B}^\top y_\lambda^*, x - x_\lambda^* \rangle = -\langle \xi_\lambda, x - x_\lambda^* \rangle$. Similarly,

$$\begin{aligned} \psi_\lambda(y_\lambda^*) - \psi_\lambda(y) &= q^*(-\mathbf{B}^\top y) - q^*(-\mathbf{B}^\top y_\lambda^*) + g_\lambda(y) - g_\lambda(y_\lambda^*) \\ &= D_{q^*}(-\mathbf{B}^\top y, -\mathbf{B}^\top y_\lambda^*) - \langle \mathbf{B}x_\lambda^*, y - y_\lambda^* \rangle \\ &\quad + D_{g_\lambda}(y, y_\lambda^*) + \langle g'_\lambda(y_\lambda^*), y - y_\lambda^* \rangle, \end{aligned}$$

whose linear terms combine into $\langle g'_\lambda(y_\lambda^*) - \mathbf{B}x_\lambda^*, y - y_\lambda^* \rangle = -\langle \zeta_\lambda, y - y_\lambda^* \rangle$. Adding the two displays and using $\varphi_\lambda(x_\lambda^*) = \psi_\lambda(y_\lambda^*)$ yields (82). The Bregman terms are nonnegative by convexity, and the two remaining terms are nonnegative because $\xi_\lambda \in N_{\mathcal{X}}(x_\lambda^*)$ with $x \in \mathcal{X}$, and $\zeta_\lambda \in N_{\mathcal{Y}}(y_\lambda^*)$ with $y \in \mathcal{Y}$. $\square$

Theorem J.4. *Let Assumption J.1 hold and set*

$$\hat{\kappa} := \frac{L_{xy}^2}{\delta_x^\lambda \lambda_y} = \frac{L_{xy}^2 \Omega_y}{\delta_x^\lambda \varepsilon}, \quad \hat{\kappa}_\mu := \frac{L_{xy}^2}{\mu_x \lambda_y} = \frac{L_{xy}^2 \Omega_y}{\mu_x \varepsilon} \quad (\mu_x > 0). \quad (83)$$

If $z \in \mathcal{Z}$ satisfies $\Psi_\lambda(z) \leq \varepsilon'$, then

$$\text{Gap}(z) \leq \frac{\varepsilon}{2} + \frac{\varepsilon'}{6} + \frac{\hat{\kappa}}{2} \varepsilon' + \begin{cases} \frac{\hat{\kappa}_\mu}{2} \varepsilon', & \mu_x > 0, \\ L_{xy} D_{\mathcal{X}} \sqrt{\frac{\varepsilon'}{\lambda_y}}, & \mu_x = 0. \end{cases} \quad (84)$$

In particular $\text{Gap}(z) \leq \varepsilon$ whenever

$$\varepsilon' \leq \varepsilon'_* := \begin{cases} \frac{\varepsilon}{2 \left(\frac{1}{6} + \frac{1}{2} \hat{\kappa} + \frac{1}{2} \hat{\kappa}_\mu \right)}, & \mu_x > 0, \\ \min \left\{ \frac{\varepsilon}{4 \left(\frac{1}{6} + \frac{1}{2} \hat{\kappa} \right)}, \frac{\lambda_y \varepsilon^2}{16 L_{xy}^2 D_{\mathcal{X}}^2} \right\}, & \mu_x = 0, \end{cases} \quad (85)$$

where the second entry of the minimum is read as $+\infty$ when $D_{\mathcal{X}} = 0$.

Proof. For $v \in \mathcal{Y}$ we have $\|v - y^0\|^2 \leq \Omega_y$, hence $F(x, v) \leq F_{\lambda_y}(x, v) + \frac{\lambda_y}{2} \Omega_y$, while $F(u, y) \geq F_{\lambda_y}(u, y)$. Taking the maximum over $v \in \mathcal{Y}$ and the minimum over $u \in \mathcal{X}$,

$$\text{Gap}(z) \leq \text{Gap}_{\lambda_y}(z) + \frac{\lambda_y}{2} \Omega_y = \text{Gap}_{\lambda_y}(z) + \frac{\varepsilon}{2}.$$

We bound $\text{Gap}_{\lambda_y}(z)$ by (82), in which the two normal-cone terms vanish by Assumption J.1(i). Since $\Psi_\lambda(z) \leq \varepsilon'$ and every summand of Ψ_λ is nonnegative,

$$D_f(x, x_\lambda^*) \leq \frac{\varepsilon'}{12}, \quad D_{g_\lambda}(y, y_\lambda^*) \leq \frac{\varepsilon'}{12}, \quad \|x - x_\lambda^*\|^2 \leq \frac{\varepsilon'}{\delta_x^\lambda}, \quad \|y - y_\lambda^*\|^2 \leq \frac{\varepsilon'}{\lambda_y}. \quad (86)$$

By (80) and $\|\mathbf{B}(x - x_\lambda^*)\| \leq L_{xy} \|x - x_\lambda^*\|$,

$$D_{h^*}(\mathbf{B}x, \mathbf{B}x_\lambda^*) \leq \frac{L_{xy}^2 \|x - x_\lambda^*\|^2}{2\lambda_y} \leq \frac{\hat{\kappa}}{2} \varepsilon'.$$

For the dual Bregman term we use (81). If $\mu_x > 0$, then $D_{q^*}(u, x_\lambda^*) \geq \frac{\mu_x}{2} \|u - x_\lambda^*\|^2$, and maximizing the resulting concave quadratic in $u - x_\lambda^*$ over $\mathbb{R}^{d_x} \supseteq \mathcal{X} - x_\lambda^*$ gives

$$D_{q^*}(-\mathbf{B}^\top y, -\mathbf{B}^\top y_\lambda^*) \leq \frac{\|\mathbf{B}^\top (y - y_\lambda^*)\|^2}{2\mu_x} \leq \frac{\hat{\kappa}_\mu}{2} \varepsilon'.$$

If $\mu_x = 0$, we discard $D_f \geq 0$ and use $u, x_\lambda^* \in \mathcal{X}$:

$$D_{q^*}(-\mathbf{B}^\top y, -\mathbf{B}^\top y_\lambda^*) \leq L_{xy} \|y - y_\lambda^*\| D_{\mathcal{X}} \leq L_{xy} D_{\mathcal{X}} \sqrt{\varepsilon'/\lambda_y}.$$

Adding the four contributions gives (84), and (85) makes everything after $\varepsilon/2$ at most $\varepsilon/2$. $\square$

Corollary J.5. *Let Assumption J.1 hold and set*

$$\Psi_0 := \Psi_\lambda(z^0), \quad \Gamma := \frac{\varepsilon}{\varepsilon'_*}, \quad \alpha_x := \frac{1 - \nu_x}{1 + 3\nu_x}, \quad \alpha_y := \frac{1 - \nu_y}{1 + 3\nu_y}.$$

Let $z_\lambda^S := z^0$ if $\Psi_0 \leq \varepsilon'_*$, and otherwise let z_λ^S be the output of Corollary 4.5 applied to $(f, g_\lambda, p_{3,\lambda}, Q_3)$ with terminal target ε'_* of (85). Then

$$\text{Gap}(z_\lambda^S) \leq \varepsilon \quad \text{for (1),}$$

and, with N_f, N_g, N_B as in (77)–(78) and $S = \mathcal{O}(1 + \log_+(\Psi_0 \Gamma / \varepsilon))$,

$$N_f^{\text{tot}} = \mathcal{O}(\Gamma^{\alpha_x} N_f), \quad N_g^{\text{tot}} = \mathcal{O}(\Gamma^{\alpha_y} N_g), \quad N_B^{\text{tot}} = \mathcal{O}(N_B). \quad (87)$$

Moreover

$$\nu_x = \nu_y = 1 \implies \alpha_x = \alpha_y = 0, \quad \nu_y < 1 \implies \hat{\kappa} = \hat{\kappa}_\mu = \kappa_{xy}^\lambda, \quad \Gamma = \Theta(1 + \kappa_{xy}^\lambda).$$

Proof. If $\Psi_0 \leq \varepsilon'_*$, then $\Psi_\lambda(z_\lambda^S) = \Psi_0 \leq \varepsilon'_*$ at zero oracle cost. Otherwise $\varepsilon'_* \in (0, \Psi_0]$ and Corollary 4.5 returns z_λ^S with $\Psi_\lambda(z_\lambda^S) \leq \varepsilon'_*$. In both cases (85) gives $\text{Gap}(z_\lambda^S) \leq \varepsilon$.

By Appendix I, the terminal target ε' enters only through

$$N_f^{\text{tot}} = \mathcal{O}((\varepsilon')^{-\alpha_x}) (\nu_x < 1), \quad N_g^{\text{tot}} = \mathcal{O}((\varepsilon')^{-\alpha_y}) (\nu_y < 1),$$

and, in the remaining cases, through $S = \mathcal{O}(1 + \log_+(\Psi_0 / \varepsilon'))$, with $N_B^{\text{tot}} = \mathcal{O}(\max\{\sqrt{\kappa_{xy}^\lambda}, 1\} S)$. Setting $\varepsilon' = \varepsilon'_* = \varepsilon / \Gamma$,

$$(\varepsilon'_*)^{-\alpha} = \Gamma^\alpha \varepsilon^{-\alpha}, \quad S = \mathcal{O}(1 + \log_+(\Psi_0 \Gamma / \varepsilon)),$$

while $\lambda_y = \varepsilon / \Omega_y$, κ_x^λ and κ_{xy}^λ are unchanged, which gives (87).

Finally, $\nu_y < 1$ forces $\beta_x^\lambda = 0$ in (73), hence $\delta_x^\lambda = \mu_x = \delta_x > 0$ and

$$\hat{\kappa} = \hat{\kappa}_\mu = \frac{L_{xy}^2 \Omega_y}{\delta_x \varepsilon} = \kappa_{xy}^\lambda, \quad \Gamma = 2\left(\frac{1}{6} + \hat{\kappa}\right) = \Theta(1 + \kappa_{xy}^\lambda)$$

by (83) and (85). $\square$

K Stochastic Settings

We allow function components with arbitrary Hölder exponents in $[0, 1]$, accessed through conditionally unbiased stochastic (sub)gradient oracles; the monotone operators and bilinear products are evaluated exactly. The analysis retains both the Hölder modeling error and the martingale contribution. On bounded domains it gives a uniform expected gap. When f, g have positive intrinsic strong convexity on their feasible sets, a restarted decomposition with $p_3 = 0$ gives an expected Lyapunov guarantee without anchor samples. The two guarantees and their domain assumptions are stated separately below.

K.1 Stochastic Hölder Oracle Extension

The function components p_i are accessed through conditionally unbiased stochastic first-order oracles, while the operators Q_i are evaluated exactly. We allow each true function to have its own Hölder exponent $\nu_i \in [0, 1]$. The Hölder modeling error and the stochastic error enter the recursion additively and can be optimized independently. The smooth stochastic bounds follow by setting every $\nu_i = 1$.

Let $\|\cdot\|_{\mathbf{P},*}$ denote the dual norm of $\|\cdot\|_{\mathbf{P}}$,

$$\|s\|_{\mathbf{P},*} := \sqrt{\langle s, \mathbf{P}^{-1} s \rangle}.$$

The stochastic algorithm and its filtration. Replace each component-gradient call in Algorithm 1 by a stochastic (sub)gradient using a fresh sample, leaving the other steps unchanged. Only the sampled vectors are needed: the function expressions in the recursion specify affine changes of variables, and no function-value oracle is required. The oracle-call nodes are

$$\mathcal{T} := \{(k, \mathbf{t}) : 1 \leq k \leq n, \mathbf{t} = (t_1, \dots, t_k), 0 \leq t_\ell < T_\ell\}.$$

Enumerate them as $\rho_1, \dots, \rho_N$ in the algorithm's depth-first execution order, where $N = \sum_{k=1}^n \prod_{j=1}^k T_j$. If ξ_{ρ_m} is the sample drawn at ρ_m , set

$$\mathcal{F}_0 := \sigma(z_{\text{in}}), \quad \mathcal{F}_m := \mathcal{F}_0 \vee \sigma(\xi_{\rho_1}, \dots, \xi_{\rho_m}), \quad \mathcal{F}_{\rho_m} := \mathcal{F}_{m-1}.$$

Thus $\mathcal{F}_\rho$ is the history before the call at ρ . For $\rho = (k, \mathbf{t})$, the points $z_{\mathbf{t}}^k, \bar{z}_{\mathbf{t}}^k$ and the function $\hat{p}_k^{k, \mathbf{t}}$ are $\mathcal{F}_\rho$ -measurable, whereas $\bar{z}_{\mathbf{t}}^k$ generally depends on the current sample and later samples in the same subtree. All iteration budgets and step-size parameters below are deterministic. For a restarted invocation, $\mathcal{F}_0$ may instead contain the entire history preceding that invocation.

Assumption K.1 (Stochastic Hölder first-order oracle). *For every $1 \leq i \leq n$, the convex function p_i satisfies Assumption 3.4, with $0 \leq \nu_i \leq 1$ and $H_i \geq 0$. Let $s_i(v) \in \partial p_i(v)$ be a measurable selection; when $\nu_i > 0$, it is the true gradient $\nabla p_i(v)$. There is a measurable stochastic oracle G_i such that, writing*

$$\varepsilon_i(v, \xi) := G_i(v, \xi) - s_i(v),$$

for every call to component i , at node ρ , and every $\mathcal{F}_\rho$ -measurable query $v \in \mathcal{Z}$,

$$\mathbb{E}[\varepsilon_i(v, \xi_\rho) \mid \mathcal{F}_\rho] = 0, \quad \mathbb{E}[\|\varepsilon_i(v, \xi_\rho)\|_{\mathbf{P},*}^2 \mid \mathcal{F}_\rho] \leq \sigma_i^2 \quad \text{a.s.} \quad (88)$$

The input z_{in} has finite second moment.

The assumptions concern the true functions and the conditional moments of the oracle; individual samples need not be Hölder continuous. Hölder subgradients and Lipschitz operators have at most linear growth. Since the algorithm makes finitely many averaging and weighted projection steps with positive deterministic quadratic coefficients, its iterates have finite second moments. Thus the martingale products used below are integrable.

For $0 \leq \nu_i < 1$ and $H_i > 0$, choose a deterministic modeling tolerance $d_i > 0$, and use $L_i(d_i)$ from (14). For $\nu_i = 1$, set $d_i = 0$ and $L_i(d_i) = H_i$. If $H_i = 0$, the function is affine; set $d_i = L_i(d_i) = 0$. Lemma E.1 gives, for every $u, v \in \mathcal{Z}$,

$$0 \leq p_i(u) - p_i(v) - \langle s_i(v), u - v \rangle \leq \frac{L_i(d_i)}{2} \|u - v\|_{\mathbf{P}}^2 + d_i. \quad (89)$$

Here d_i is a modeling tolerance, not an oracle bias. The sampled vector need not satisfy the lower bound in (89) pathwise. We therefore use the following three-point inequality, which also permits the prox output to depend on the sample used to compute it.

Lemma K.2 (Stochastic three-point inequality). *Let $h: \mathcal{Z} \rightarrow \mathbb{R}$ be convex, let $s(v) \in \partial h(v)$, and suppose*

$$0 \leq h(w) - h(v) - \langle s(v), w - v \rangle \leq \frac{L}{2} \|w - v\|_{\mathbf{P}}^2 + d, \quad v, w \in \mathcal{Z}.$$

At a $\mathcal{F}_\rho$ -measurable query v , write $G = s(v) + \varepsilon$. For every $\tau > 0$, the following inequality holds pathwise for all $u, y \in \mathcal{Z}$, including random choices:

$$h(u) - h(y) \geq \langle G, u - y \rangle - \frac{L + \tau}{2} \|y - v\|_{\mathbf{P}}^2 - d - \frac{\|\varepsilon\|_{\mathbf{P},*}^2}{2\tau} - \langle \varepsilon, u - v \rangle. \quad (90)$$

If $\varepsilon = 0$ almost surely, the inequality holds with $\tau = 0$ and the last two terms omitted. If the conditional moments (88) hold with variance σ^2 , and u is $\mathcal{F}_\rho$ -measurable with u, v of finite second moment, then

$$\mathbb{E}[\langle \varepsilon, u - v \rangle \mid \mathcal{F}_\rho] = 0, \quad \mathbb{E}[\|\varepsilon\|_{\mathbf{P},*}^2 \mid \mathcal{F}_\rho] \leq \sigma^2. \quad (91)$$

Proof. Subtract the quadratic upper model at y from the convex lower bound at u , both based at v , to obtain

$$h(u) - h(y) \geq \langle s(v), u - y \rangle - \frac{L}{2} \|y - v\|_{\mathbf{P}}^2 - d.$$

Substitute $s(v) = G - \varepsilon$, and use

$$-\langle \varepsilon, u - y \rangle = -\langle \varepsilon, u - v \rangle + \langle \varepsilon, y - v \rangle, \quad \langle \varepsilon, y - v \rangle \geq -\frac{\tau}{2} \|y - v\|_{\mathbf{P}}^2 - \frac{\|\varepsilon\|_{\mathbf{P},*}^2}{2\tau}.$$

These are pathwise inequalities, so no measurability restriction on u, y is needed for (90). Conditional centering in (91) additionally uses the $\mathcal{F}_\rho$ -measurability of u, v . The zero-noise case follows directly from the first display. $\square$

Setting $\varepsilon = 0$ and $\tau = 0$ recovers (41) with deterministic tolerance d . For positive noise, the quadratic prox term absorbs the dependence on y ; only the comparator term must be centered conditionally.

Lemma K.3 (Rescaled stochastic three-point inequality). *Let Assumption K.1 hold, and suppose $0 < \alpha_t \leq 1$. Fix a node $\rho = (k, \mathbf{t}) \in \mathcal{T}$, and set $\Gamma_{\mathbf{t}} := \prod_{\ell=1}^k \alpha_{t_\ell}$. The induced oracle has a selected mean subgradient $\hat{s}_k^{k,\mathbf{t}}$. Write its error as*

$$\varepsilon^{k,\mathbf{t}} := G_k^{k,\mathbf{t}}(z_{\mathbf{t}}^k, \xi_\rho) - \hat{s}_k^{k,\mathbf{t}}(z_{\mathbf{t}}^k).$$

For every $\tau_k > 0$, almost surely and simultaneously for all $u, y \in \mathcal{Z}$,

$$\begin{aligned} \hat{p}_k^{k,\mathbf{t}}(u) - \hat{p}_k^{k,\mathbf{t}}(y) &\geq \langle G_k^{k,\mathbf{t}}(z_{\mathbf{t}}^k, \xi_\rho), u - y \rangle - \frac{L_k^{k,\mathbf{t}} + \tau_k^{k,\mathbf{t}}}{2} \|y - z_{\mathbf{t}}^k\|_{\mathbf{P}}^2 \\ &\quad - r_k^{k,\mathbf{t}} - \langle \varepsilon^{k,\mathbf{t}}, u - z_{\mathbf{t}}^k \rangle, \end{aligned} \quad (92)$$

where

$$L_k^{k,\mathbf{t}} = L_k(d_k)\Gamma_{\mathbf{t}}, \quad \tau_k^{k,\mathbf{t}} = \tau_k\Gamma_{\mathbf{t}}, \quad r_k^{k,\mathbf{t}} = \frac{d_k + \|\varepsilon^{k,\mathbf{t}}\|_{\mathbf{P},*}^2/(2\tau_k)}{\Gamma_{\mathbf{t}}}.$$

For a $\mathcal{F}_\rho$ -measurable comparator u with finite second moment,

$$\mathbb{E}[\langle \varepsilon^{k,\mathbf{t}}, u - z_{\mathbf{t}}^k \rangle \mid \mathcal{F}_\rho] = 0, \quad \mathbb{E}[r_k^{k,\mathbf{t}} \mid \mathcal{F}_\rho] \leq \bar{r}_k^{k,\mathbf{t}} := \frac{d_k + \sigma_k^2/(2\tau_k)}{\Gamma_{\mathbf{t}}}. \quad (93)$$

If $\sigma_k = 0$, then $\varepsilon^{k,\mathbf{t}} = 0$ almost surely, and the conclusions hold with $\tau_k = 0$ and $r_k^{k,\mathbf{t}} = \bar{r}_k^{k,\mathbf{t}} = d_k/\Gamma_{\mathbf{t}}$.

Proof. The affine transformation in line 20 does not depend on the type of oracle. As in Lemma D.2,

$$\hat{p}_k^{k,\mathbf{t}}(z) = \Gamma_{\mathbf{t}}^{-1} p_k(A_{\mathbf{t}}^k(z)),$$

where $A_{\mathbf{t}}^k$ is a $\mathcal{F}_\rho$ -measurable affine map with linear part $\Gamma_{\mathbf{t}} \text{Id}$. Its image lies in $\mathcal{Z}$ because the averaging points lie in the convex set $\mathcal{Z}$. The subgradient inequality gives the measurable selection and oracle

$$\hat{s}_k^{k,\mathbf{t}}(z) = s_k(A_{\mathbf{t}}^k(z)), \quad G_k^{k,\mathbf{t}}(z, \xi) = G_k(A_{\mathbf{t}}^k(z), \xi).$$

Apply Lemma K.2 to p_k at $u' = A_{\mathbf{t}}^k(u)$, $y' = A_{\mathbf{t}}^k(y)$, and $v' = A_{\mathbf{t}}^k(z_{\mathbf{t}}^k)$, using its model with parameters $(d_k, L_k(d_k))$. Divide by $\Gamma_{\mathbf{t}}$, and use

$$u' - y' = \Gamma_{\mathbf{t}}(u - y), \quad u' - v' = \Gamma_{\mathbf{t}}(u - z_{\mathbf{t}}^k), \quad \|y' - v'\|_{\mathbf{P}}^2 = \Gamma_{\mathbf{t}}^2 \|y - z_{\mathbf{t}}^k\|_{\mathbf{P}}^2.$$

This proves (92). The query v' is measurable before the call, so the original conditional oracle moments give (93). The zero-noise conclusion follows from the last clause of Lemma K.2. $\square$

The modeling and squared-noise errors both scale as $\Gamma_{\mathbf{t}}^{-1}$, while the martingale term carries no $\Gamma_{\mathbf{t}}$ factor in the rescaled coordinates. In particular, zero noise removes only the stochastic part of $r_k^{k,\mathbf{t}}$.

Lemma K.4 (Stochastic recursive gap bound). *Under Assumptions 3.1 and K.1, let $T_i \in \mathbb{N}$, and let $\{\alpha_t\}$ satisfy (11). Choose the modeling tolerances above and $\tau_i \geq 0$, with $\tau_i > 0$ whenever $\sigma_i > 0$, such that $L_i(d_i) + \tau_i + M_i > 0$. Run stochastic Algorithm 1 with*

$$\eta_{\mathbf{t}}^k = (L_k(d_k) + \tau_k) \prod_{\ell=1}^k \alpha_{t_\ell} + M_k \prod_{\ell=1}^k \frac{\alpha_{t_\ell}}{\alpha_{T_\ell-1}}. \quad (94)$$

For every deterministic $\hat{z} \in \mathcal{Z}$ and every admissible $(k, \mathbf{t})$, including $(0, \emptyset)$,

$$\mathbb{E}[\Phi_{\mathbf{t}, \hat{z}}^k(\hat{z}_{\mathbf{t}}^k, \hat{z})] \leq \mathbb{E}[S_{\mathbf{t}}^k] + R_{\mathbf{t}}^k.$$

Here $\Phi_{\mathbf{t}, \hat{z}}^k, S_{\mathbf{t}}^k$ are the stochastic analogues of the quantities in Lemma D.3, with sampled linear terms and with $L_i^{k,\mathbf{t}}$ replaced by $(L_i(d_i) + \tau_i)\Gamma_{\mathbf{t}}$, and

$$R_{\mathbf{t}}^k := \sum_{i=k+1}^n \frac{\bar{r}_i^{k,\mathbf{t}}}{\prod_{j=k+1}^i \alpha_{T_j-1}}, \quad \bar{r}_i^{k,\mathbf{t}} := \frac{d_i + \sigma_i^2/(2\tau_i)}{\Gamma_{\mathbf{t}}}.$$

The quantity $R_{\mathbf{t}}^k$ is deterministic. We set $\sigma_i^2/(2\tau_i) = 0$ when $\sigma_i = \tau_i = 0$, and use $\Gamma_{\emptyset} = 1$.

Proof. Repeat the downward induction of Lemma D.3. At $k = n$, the minimizing property gives $\Phi_{\mathbf{t}, \hat{z}}^n(\hat{z}_{\mathbf{t}}^n, \hat{z}) \leq 0$ pathwise, while $S_{\mathbf{t}}^n = R_{\mathbf{t}}^n = 0$.

At a node $(k, \mathbf{t})$, replace the deterministic three-point inequality in (42) by (92), with $u = \hat{z}$ and $y = \hat{z}_{\mathbf{t}}^k$. The prox output may depend on the sample, as permitted by that lemma. The deterministic derivation then holds pathwise with curvature $L_k^{k, \mathbf{t}} + \tau_k^{k, \mathbf{t}}$ and residual $-r_k^{k, \mathbf{t}} - \mathcal{E}_{\mathbf{t}}^k$, where

$$\mathcal{E}_{\mathbf{t}}^k := \langle \varepsilon^{k, \mathbf{t}}, \hat{z} - z_{\mathbf{t}}^k \rangle.$$

Indeed, with $\eta_{\mathbf{t}}^k = L_k^{k, \mathbf{t}} + \tau_k^{k, \mathbf{t}} + M_k^{k, \mathbf{t}}$, the coefficient discarded in (44) remains nonnegative:

$$\eta_{\mathbf{t}}^k - L_k^{k, \mathbf{t}} - \tau_k^{k, \mathbf{t}} - \frac{(M_k^{k, \mathbf{t}})^2}{\eta_{\mathbf{t}}^k} = \frac{M_k^{k, \mathbf{t}}(L_k^{k, \mathbf{t}} + \tau_k^{k, \mathbf{t}})}{\eta_{\mathbf{t}}^k} \geq 0.$$

Conditional centering and the remainder bound give

$$\mathbb{E}[\mathcal{E}_{\mathbf{t}}^k \mid \mathcal{F}_{(k, \mathbf{t})}] = 0, \quad \mathbb{E}[r_k^{k, \mathbf{t}} \mid \mathcal{F}_{(k, \mathbf{t})}] \leq \bar{r}_k^{k, \mathbf{t}}.$$

The remaining inequalities and telescoping identities in the deterministic proof use deterministic nonnegative weights, so they commute with expectation. Finally, $\bar{r}_i^{k, \mathbf{t}} = \bar{r}_i^{k-1, \mathbf{t}'} / \alpha_{t_k}$, exactly as for the deterministic oracle tolerances; (54) therefore gives $R_{\mathbf{t}'}^{k-1}$. This closes the induction.

The same induction may instead retain every error term pathwise. Each transition from level k to $k - 1$ multiplies that error by $\alpha_{T_{k-1}}^2 / \alpha_{t_k}$, as seen in (52). This observation will be used for the uniform gap below. $\square$

Theorem K.5 (Stochastic Hölder sliding). *Let the assumptions and parameter choices of Lemma K.4 hold. Define $A_i := \prod_{j=1}^i T_j$. For every deterministic comparator $z \in \mathcal{Z}$, the output satisfies*

$$\begin{aligned} \mathcal{G}_{\text{st}}(z) &:= \mathbb{E}[p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle] \\ &\leq \sum_{i=1}^n \left[\left(\frac{2^{2i-1}(L_i(d_i) + \tau_i)}{A_i^2} + \frac{2^{i-1}M_i}{A_i} \right) \mathbb{E}\|z_{\text{in}} - z\|_{\mathbf{P}}^2 + d_i A_i + \frac{\sigma_i^2 A_i}{2\tau_i} \right]. \end{aligned} \quad (95)$$

The zero-noise convention of Lemma K.4 applies.

Proof. At the root, the definitions in the proof of Theorem 3.3 give $H_{\Sigma}^0 = 0$, $z_{\text{out}} = \hat{z}_{\emptyset}^0$, and

$$\Phi_{\emptyset, z}^0(u, z) = p(u) - p(z) + \langle Q(z), u - z \rangle.$$

Thus Lemma K.4 yields $\mathcal{G}_{\text{st}}(z) \leq \mathbb{E}[S_{\emptyset}^0] + R_{\emptyset}^0$. By Lemma D.1, $1/T_j \leq \alpha_{T_{j-1}} \leq 2/T_j$. Consequently,

$$R_{\emptyset}^0 = \sum_{i=1}^n \frac{d_i + \sigma_i^2/(2\tau_i)}{\prod_{j=1}^i \alpha_{T_{j-1}}} \leq \sum_{i=1}^n \left(d_i A_i + \frac{\sigma_i^2 A_i}{2\tau_i} \right).$$

Use the initialization $z_{\emptyset}^i = z_{\text{in}}$, drop the nonpositive terminal squared distances in $S_{\emptyset}^0$, and then take expectations. This gives

$$\mathbb{E}[S_{\emptyset}^0] \leq \sum_{i=1}^n \left(\frac{2^{2i-1}(L_i(d_i) + \tau_i)}{A_i^2} + \frac{2^{i-1}M_i}{A_i} \right) \mathbb{E}\|z_{\text{in}} - z\|_{\mathbf{P}}^2.$$

Combining the two bounds proves (95). $\square$

Corollary K.6 (Optimized stochastic Hölder rate). *Let Assumptions 3.1 and K.1 hold, with $H_i + M_i + \sigma_i > 0$ for each retained component pair. Fix positive integers $T_1, \dots, T_n$, define $A_i := \prod_{j=1}^i T_j$, and use the sequence $\{\alpha_t\}$ in (11). Let $z \in \mathcal{Z}$ be deterministic and suppose that a deterministic $R > 0$ satisfies*

$$\mathbb{E}\|z_{\text{in}} - z\|_{\mathbf{P}}^2 \leq R^2. \quad (96)$$

For $0 \leq \nu_i < 1$ and $H_i > 0$, set

$$\begin{aligned} a_i &:= \frac{1 - \nu_i}{1 + \nu_i}, \quad K_i := \left[\frac{1 - \nu_i}{2(1 + \nu_i)} \right]^{a_i} H_i^{2/(1+\nu_i)}, \\ d_i^* &:= \left[\frac{a_i 2^{2i-1} K_i R^2}{A_i^3} \right]^{(1+\nu_i)/2}, \quad \tau_i^* := \frac{\sigma_i A_i^{3/2}}{2^i R}. \end{aligned} \quad (97)$$

At $\nu_i = 1$, use $d_i^* = 0$ and $L_i(d_i^*) = H_i$; at $H_i = 0$, use $d_i^* = L_i(d_i^*) = 0$. The formula for τ_i^* applies in every case, including $\sigma_i = 0$. With these choices and the step sizes (94), the stochastic version of Algorithm 1 satisfies

$$\mathcal{G}_{\text{st}}(z) \leq \sum_{i=1}^n \left[\frac{2^{i(1+\nu_i)} H_i R^{1+\nu_i}}{(1+\nu_i) A_i^{(1+3\nu_i)/2}} + \frac{2^{i-1} M_i R^2}{A_i} + \frac{2^i \sigma_i R}{\sqrt{A_i}} \right]. \quad (98)$$

Proof. For each component, the parameter-dependent terms in (95), after using (96), separate as

$$\left[\frac{2^{2i-1} L_i(d_i) R^2}{A_i^2} + d_i A_i \right] + \left[\frac{2^{2i-1} \tau_i R^2}{A_i^2} + \frac{\sigma_i^2 A_i}{2\tau_i} \right].$$

The first bracket is the deterministic Hölder optimization (64) with $D = R$ and $N_i = A_i$. Since $L_i(d) = K_i d^{-a_i}$, its first-order condition yields d_i^* in (97), and its minimum, by (70), is

$$\frac{2^{i(1+\nu_i)} H_i R^{1+\nu_i}}{(1+\nu_i) A_i^{(1+3\nu_i)/2}}.$$

For $\nu_i = 1$, the same expression is $2^{2i-1} H_i R^2 / A_i^2$; for $H_i = 0$, it is zero. For $\sigma_i > 0$, τ_i^* balances the two terms in the second bracket, each of which equals $2^{i-1} \sigma_i R / \sqrt{A_i}$. If $\sigma_i = 0$, both vanish by convention. This proves (98). $\square$

The expected radius bound permits random restart inputs while the parameters remain deterministic. A component with $H_i = M_i = \sigma_i = 0$ contributes only a deterministic affine function and a constant operator; it can be evaluated once and absorbed into another component's fixed linear term without changing that component's Hölder, Lipschitz, or variance constants. The optimized statements are written for the remaining nonzero pairs.

Corollary K.7 (Stochastic Hölder oracle complexity). *Let the assumptions of Corollary K.6 hold with comparator z^* , a solution of (8). For $\varepsilon > 0$, define the complete target budgets*

$$B_i := 1 + \left(\frac{H_i R^{1+\nu_i}}{\varepsilon} \right)^{2/(1+3\nu_i)} + \frac{M_i R^2}{\varepsilon} + \frac{\sigma_i^2 R^2}{\varepsilon^2}. \quad (99)$$

Relabel the component pairs so that $B_1 \leq \dots \leq B_n$. There are positive integers T_i with $A_i = \prod_{j=1}^i T_j = \mathcal{O}_n(B_i)$ such that, with the optimized parameters in (97),

$$\mathbb{E}[p(z_{\text{out}}) - p(z^*) + \langle Q(z^*), z_{\text{out}} - z^* \rangle] \leq \varepsilon.$$

The method uses A_i stochastic function-oracle calls and at most $2A_i$ evaluations of Q_i for the i -th component pair.

Proof. In (98), the three terms for component i are bounded by $\varepsilon/(3n)$ whenever $A_i \geq c_n B_i$, for a sufficiently large constant c_n depending only on n . The exponent $(1+3\nu_i)/2$ lies in $[1/2, 2]$, so the constant can be chosen uniformly for $\nu_i \in [0, 1]$. The integer construction of Appendix F is explicit: with $A_0 = 1$, choose

$$T_i := \max\{1, \lceil c_n B_i / A_{i-1} \rceil\}, \quad A_i := A_{i-1} T_i.$$

The nondecreasing order of the B_i gives $c_n B_i \leq A_i \leq 2c_n B_i$ by induction. Summing the error bounds proves the claim. Algorithm 1 makes one sampled function-oracle call and at most two operator evaluations per node at that level. $\square$

The ordering in (99) includes the variance contribution. Sorting only the deterministic costs can force an exact operator into a deeper level and make its evaluation count inherit an unnecessarily large stochastic function budget.

Remark 3 (Uniform and fixed-comparator residuals). For $p \equiv 0$, $Q(z) = \mathbf{S}z$ with $\mathbf{S}^\top = -\mathbf{S}$, and solution $z^* = 0$, the fixed-comparator residual is zero for every output. The uniform gap

$$\sup_{z \in \mathcal{Z}} \langle Q(z), z_{\text{out}} - z \rangle = \sup_{z \in \mathcal{Z}} \langle z, -\mathbf{S}z_{\text{out}} \rangle$$

can detect the remaining bilinear residual. Therefore the bounded convex-concave analysis uses the uniform expected gap below, while the strongly convex analysis uses a fixed comparator to control the Lyapunov function.

Theorem K.8 (Uniform stochastic Hölder gap). *Let the assumptions of Theorem K.5 hold. Assume that a deterministic constant $0 < \Omega < \infty$ satisfies*

$$\sup_{z \in \mathcal{Z}} \|z_{\text{in}} - z\|_{\mathbf{P}}^2 \leq \Omega \quad \text{almost surely.}$$

Then

$$\begin{aligned} & \mathbb{E} \left[\sup_{z \in \mathcal{Z}} \{p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle\} \right] \\ & \leq \sum_{i=1}^n \left[\frac{2^{2i-1}(L_i(d_i) + \tau_i)\Omega}{A_i^2} + \frac{2^{i-1}M_i\Omega}{A_i} + d_i A_i + \frac{\sigma_i^2 A_i}{2\tau_i} \right] + \sqrt{\Omega} \left(\sum_{i=1}^n \frac{2^i \sigma_i^2}{A_i} \right)^{1/2}. \end{aligned} \quad (100)$$

The same zero-noise convention applies.

Proof. The pathwise induction in Lemma K.4 holds simultaneously for all $\hat{z} \in \mathcal{Z}$: the three-point inequality is pathwise for every comparator, and the iterates do not depend on the comparator. A residual at node $(k, \mathbf{t})$ reaches the root with deterministic weight

$$w_{\mathbf{t}}^k := \prod_{\ell=1}^k \frac{\alpha_{T_\ell-1}^2}{\alpha_{t_\ell}}.$$

Therefore, almost surely and for all $\hat{z} \in \mathcal{Z}$,

$$\begin{aligned} & p(z_{\text{out}}) - p(\hat{z}) + \langle Q(\hat{z}), z_{\text{out}} - \hat{z} \rangle \\ & \leq S_{\emptyset}^0(\hat{z}) + \sum_{(k, \mathbf{t}) \in \mathcal{T}} w_{\mathbf{t}}^k r_k^{k, \mathbf{t}} + \sum_{(k, \mathbf{t}) \in \mathcal{T}} w_{\mathbf{t}}^k \langle \varepsilon^{k, \mathbf{t}}, \hat{z} - z_{\mathbf{t}}^k \rangle. \end{aligned} \quad (101)$$

By the conditional bound on $r_k^{k, \mathbf{t}}$ and Lemma D.1,

$$\begin{aligned} \mathbb{E} \sum_{\mathbf{t}} w_{\mathbf{t}}^k r_k^{k, \mathbf{t}} & \leq \left(d_k + \frac{\sigma_k^2}{2\tau_k} \right) \prod_{\ell=1}^k \left(\alpha_{T_\ell-1}^2 \sum_{t=0}^{T_\ell-1} \alpha_t^{-2} \right) \\ & \leq \frac{d_k + \sigma_k^2/(2\tau_k)}{\prod_{\ell=1}^k \alpha_{T_\ell-1}} \leq d_k A_k + \frac{\sigma_k^2 A_k}{2\tau_k}. \end{aligned}$$

This includes $\sigma_k = \tau_k = 0$, in which case the modeling term $d_k A_k$ is retained.

Let $V := \sum_{(k, \mathbf{t}) \in \mathcal{T}} w_{\mathbf{t}}^k \varepsilon^{k, \mathbf{t}}$. The last sum in (101) equals

$$\langle V, \hat{z} - z_{\text{in}} \rangle - \sum_{(k, \mathbf{t}) \in \mathcal{T}} w_{\mathbf{t}}^k \langle \varepsilon^{k, \mathbf{t}}, z_{\mathbf{t}}^k - z_{\text{in}} \rangle.$$

The second term has zero expectation because its query point and input are measurable before the corresponding call. The radius assumption gives the pathwise bound

$$\sup_{\hat{z} \in \mathcal{Z}} \langle V, \hat{z} - z_{\text{in}} \rangle \leq \sqrt{\Omega} \|V\|_{\mathbf{P},*}.$$

Conditional centering, in the depth-first sampling order, eliminates all cross terms in $\mathbb{E} \|V\|_{\mathbf{P},*}^2$. Jensen's inequality yields

$$\mathbb{E} \|V\|_{\mathbf{P},*} \leq \left(\sum_{k=1}^n \sigma_k^2 \sum_{\mathbf{t}} (w_{\mathbf{t}}^k)^2 \right)^{1/2}.$$

Again using Lemma D.1,

$$\sum_{\mathbf{t}} (w_{\mathbf{t}}^k)^2 = \prod_{\ell=1}^k \left(\alpha_{T_\ell-1}^4 \sum_{t=0}^{T_\ell-1} \alpha_t^{-2} \right) \leq \prod_{\ell=1}^k \alpha_{T_\ell-1} \leq \frac{2^k}{A_k}.$$

Take the supremum in (101), followed by expectation. Dropping the terminal distances in $S_{\emptyset}^0$ and using the radius bound gives the curvature and operator terms. The preceding bounds give the modeling, variance, and martingale terms in (100). $\square$

Corollary K.9 (Uniform stochastic Hölder complexity). *Under the assumptions of Theorem K.8, suppose $H_i + M_i + \sigma_i > 0$ for every retained pair and choose d_i, τ_i as in (97) with $R = \sqrt{\Omega}$, including the stated endpoint conventions. Then*

$$\begin{aligned} & \mathbb{E} \left[\sup_{z \in \mathcal{Z}} \{p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle\} \right] \\ & \leq \sum_{i=1}^n \left[\frac{2^{i(1+\nu_i)} H_i \Omega^{(1+\nu_i)/2}}{(1+\nu_i) A_i^{(1+3\nu_i)/2}} + \frac{2^{i-1} M_i \Omega}{A_i} + \frac{(2^i + 2^{i/2}) \sigma_i \sqrt{\Omega}}{\sqrt{A_i}} \right]. \end{aligned} \quad (102)$$

Consequently, the complete budgets (99), with $R^2 = \Omega$, sorted and rounded as in Corollary K.7, also suffice to make the uniform expected gap at most ε .

Proof. Optimize the first four terms of (100) exactly as in Corollary K.6. Bound the final square root by $\sum_i 2^{i/2} \sigma_i / \sqrt{A_i}$, and apply the same budget construction with a larger constant depending only on n . $\square$

At $\nu_i = 1$, the deterministic function contribution in (102) is $2^{2i-1} H_i \Omega / A_i^2$, recovering the smooth stochastic bound. At $\sigma_i = 0$ for every component, the bound becomes exactly the deterministic Hölder rate in Theorem 3.5.

K.2 Stochastic Bilinear Saddle-Point Problem

Consider Problem (1) on nonempty closed convex sets $\mathcal{X}, \mathcal{Y}$, with bounded product set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Assume that f, g are convex and satisfy the (ν_x, H_x) - and (ν_y, H_y) -Hölder conditions on their respective feasible sets, where $0 \leq \nu_x, \nu_y \leq 1$. Their stochastic oracles satisfy Assumption K.1 with Euclidean conditional variances σ_x^2, σ_y^2 . In particular, at exponent zero the conditional oracle mean is a selected ambient subgradient. Assume $\|\mathbf{B}\|_{\text{op}} \leq L_{xy}$, and evaluate products with $\mathbf{B}, \mathbf{B}^\top$ exactly. No positive curvature is needed in this subsection, so the result includes the regime $\delta_x = \delta_y = 0$.

For deterministic $z^0 \in \mathcal{Z}$, define

$$\Omega := \sup_{z \in \mathcal{Z}} \|z^0 - z\|^2 < \infty.$$

The case $\Omega = 0$ is immediate; hence assume $\Omega > 0$. Use Algorithm 1 in the Euclidean norm with

$$\begin{aligned} p_1(x, y) &= f(x), & p_2(x, y) &= g(y), & p_3 &\equiv 0, \\ Q_1 &= Q_2 = 0, & Q_3(x, y) &= \begin{pmatrix} \mathbf{B}^\top y \\ -\mathbf{B}x \end{pmatrix}. \end{aligned} \quad (103)$$

The operator Q_3 is monotone and L_{xy} -Lipschitz. Let N_f, N_g, N_B count, respectively, stochastic evaluations of the two function oracles and products with $\mathbf{B}, \mathbf{B}^\top$. A constant factor of two between a call to Q_3 and its two matrix-vector products is absorbed in the bounds below.

Theorem K.10 (Bounded-domain stochastic Hölder complexity). *Under the assumptions above, define*

$$\text{Gap}(x, y) := \sup_{v \in \mathcal{Y}} F(x, v) - \inf_{u \in \mathcal{X}} F(u, y).$$

With the modeling tolerances and noise parameters of Corollary K.9, the output $z_{\text{out}} = (x_{\text{out}}, y_{\text{out}})$ satisfies

$$\begin{aligned} \mathbb{E} \text{Gap}(z_{\text{out}}) &\leq C \left[\frac{H_x \Omega^{(1+\nu_x)/2}}{N_f^{(1+3\nu_x)/2}} + \frac{H_y \Omega^{(1+\nu_y)/2}}{N_g^{(1+3\nu_y)/2}} \right. \\ &\quad \left. + \frac{L_{xy} \Omega}{N_B} + \sigma_x \sqrt{\frac{\Omega}{N_f}} + \sigma_y \sqrt{\frac{\Omega}{N_g}} \right], \end{aligned} \quad (104)$$

where $C > 0$ is an absolute constant. In particular, for any $\varepsilon > 0$, there is a component ordering and a choice of integer iteration counts giving $\mathbb{E} \text{Gap}(z_{\text{out}}) \leq \varepsilon$ with

$$\begin{aligned} N_f &= \mathcal{O} \left(1 + \left(\frac{H_x \Omega^{(1+\nu_x)/2}}{\varepsilon} \right)^{\frac{2}{1+3\nu_x}} + \frac{\sigma_x^2 \Omega}{\varepsilon^2} \right), \\ N_g &= \mathcal{O} \left(1 + \left(\frac{H_y \Omega^{(1+\nu_y)/2}}{\varepsilon} \right)^{\frac{2}{1+3\nu_y}} + \frac{\sigma_y^2 \Omega}{\varepsilon^2} \right), \\ N_B &= \mathcal{O} \left(1 + \frac{L_{xy} \Omega}{\varepsilon} \right). \end{aligned} \quad (105)$$

Proof. For the decomposition (103), skew-symmetry gives the pointwise identity

$$p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle = F(x_{\text{out}}, y) - F(x, y_{\text{out}}).$$

Taking the supremum over $z = (x, y) \in \mathcal{Z}$ therefore gives precisely $\text{Gap}(z_{\text{out}})$. Apply Corollary K.9 with the component data

$$(H, \nu, M, \sigma) = (H_x, \nu_x, 0, \sigma_x), \quad (H_y, \nu_y, 0, \sigma_y), \quad (0, 1, L_{xy}, 0).$$

The constants for at most three recursive levels are absorbed into C , which proves (104).

To obtain (105), form the three target budgets on its right-hand side, multiply them by a sufficiently large absolute constant, and sort the complete budgets, including their variance terms, in nondecreasing order. The integer construction of Appendix F realizes them within absolute factors. Consequently, each summand in (104) is a sufficiently small multiple of ε . The ordering is part of the construction: a coupling component permanently assigned to the innermost level could otherwise inherit a larger stochastic function budget. $\square$

An exact function component with $H_q = \sigma_q = 0$ is affine on its feasible set. Its constant subgradient can be evaluated once and included as a fixed linear term in another retained component. An identically zero component pair can be omitted. If all components are exact and affine and $\mathbf{B} = 0$, one projection minimizing their linear sum plus $\varepsilon \|z - z^0\|^2 / (2\Omega)$ gives gap at most $\varepsilon/2$; thus the additive constants in (105) also cover this case.

At $\nu_x = \nu_y = 1$, the bounds recover the smooth stochastic rates. When $\sigma_x = \sigma_y = 0$, they recover the deterministic Hölder bounds. In both cases, the guarantee is the uniform expected primal–dual gap, with supremization performed before expectation.

K.3 Stochastic strongly convex–strongly concave regime

Assume that $\mathcal{X}, \mathcal{Y}$ are nonempty closed convex sets and that f, g are respectively μ_x - and μ_y -strongly convex on these sets, with $\mu_x, \mu_y > 0$. Suppose they satisfy the (ν_x, H_x) - and (ν_y, H_y) -Hölder conditions on these sets, for $0 \leq \nu_x, \nu_y \leq 1$. Assume a saddle point $z^* = (x^*, y^*)$ exists, and fix ambient subgradients and normal vectors satisfying (28). Define D_f, D_g using this fixed choice as in (29). The positive intrinsic curvatures make the saddle point unique.

The stochastic oracles have conditionally unbiased errors and Euclidean conditional second moments bounded by σ_x^2, σ_y^2 , as in Assumption K.1. These conditions hold relative to the history before every query, including all earlier restarts. Assume $\|\mathbf{B}\|_{\text{op}} \leq L_{xy}$, and evaluate products with $\mathbf{B}, \mathbf{B}^\top$ exactly.

Use the decomposition (103), including $p_3 \equiv 0$ when either function is smooth, with norm matrix

$$\mathbf{P} = \text{diag}(\mu_x \mathbf{I}_{d_x}, \mu_y \mathbf{I}_{d_y}).$$

Thus the curvature used here is supplied entirely by f, g ; in the effective-curvature notation we take $\mu_{xy} = \mu_{yx} = 0$ and $\delta_x = \mu_x, \delta_y = \mu_y$. The component constants in this norm are

$$\begin{aligned} \tilde{H}_x &= H_x \mu_x^{-(1+\nu_x)/2}, & \tilde{H}_y &= H_y \mu_y^{-(1+\nu_y)/2}, \\ \bar{\sigma}_x^2 &= \frac{\sigma_x^2}{\mu_x}, & \bar{\sigma}_y^2 &= \frac{\sigma_y^2}{\mu_y}. \end{aligned} \tag{106}$$

The operator constant is

$$M_B = \frac{L_{xy}}{\sqrt{\mu_x \mu_y}}.$$

Thus $H_3 = 0, M_1 = M_2 = 0, M_3 = M_B$, and $\sigma_3 = 0$.

A fixed residual controls the Lyapunov function. For $z = (x, y) \in \mathcal{Z}$, set

$$\mathcal{R}_*(z) := F(x, y^*) - F(x^*, y) = p(z) - p(z^*) + \langle Q(z^*), z - z^* \rangle.$$

By (28) and the definition of the normal cones,

$$\begin{aligned} \mathcal{R}_*(z) &= D_f(x, x^*) + D_g(y, y^*) - \langle \xi, x - x^* \rangle - \langle \zeta, y - y^* \rangle \\ &\geq D_f(x, x^*) + D_g(y, y^*) \geq \tfrac{1}{2} \|z - z^*\|_{\mathbf{P}}^2. \end{aligned} \tag{107}$$

Consequently, for the Lyapunov function (30) with this $\mathbf{P}$,

$$\begin{aligned} \Psi(z) &= \|z - z^*\|_{\mathbf{P}}^2 + 12 D_f(x, x^*) + 12 D_g(y, y^*) \\ &\leq 14 (D_f(x, x^*) + D_g(y, y^*)) \leq 14 \mathcal{R}_*(z). \end{aligned} \tag{108}$$

In particular, the normal vectors need not vanish. This estimate uses strong convexity on the feasible sets and does not require the ambient self-bounding inequality (20).

One restart. Let $\Gamma_s > 0$ be deterministic with $\mathbb{E}\Psi(z^s) \leq \Gamma_s$. For $q \in \{x, y\}$, define

$$\theta_q := \frac{1 - \nu_q}{1 + 3\nu_q}, \quad \mathcal{K}_q := \tilde{H}_q^{\frac{2}{1+3\nu_q}} = H_q^{\frac{2}{1+3\nu_q}} \mu_q^{-\frac{1+\nu_q}{1+3\nu_q}}. \quad (109)$$

Choose the stage oracle budgets, up to sufficiently large absolute factors, as

$$\begin{aligned} N_{f,s} &\asymp 1 + \mathcal{K}_x \Gamma_s^{-\theta_x} + \frac{\sigma_x^2}{\mu_x \Gamma_s}, \\ N_{g,s} &\asymp 1 + \mathcal{K}_y \Gamma_s^{-\theta_y} + \frac{\sigma_y^2}{\mu_y \Gamma_s}, \\ N_{B,s} &\asymp 1 + M_B. \end{aligned} \quad (110)$$

Sort these complete budgets at each restart and use the integer construction of Appendix F. Choose the modeling tolerances and noise parameters from Corollary K.6 with $R^2 = \Gamma_s$ and the actual cumulative iteration counts after sorting. They are deterministic even though z^s is random.

Since $\mathbb{E}\|z^s - z^*\|_{\mathbf{P}}^2 \leq \Gamma_s$, the fixed-comparator bound of Corollary K.6 gives

$$\begin{aligned} \mathbb{E}\mathcal{R}_*(z^{s+1}) &\leq C \left[\frac{\tilde{H}_x \Gamma_s^{(1+\nu_x)/2}}{N_{f,s}^{(1+3\nu_x)/2}} + \frac{\tilde{H}_y \Gamma_s^{(1+\nu_y)/2}}{N_{g,s}^{(1+3\nu_y)/2}} \right. \\ &\quad \left. + \frac{M_B \Gamma_s}{N_{B,s}} + \bar{\sigma}_x \sqrt{\frac{\Gamma_s}{N_{f,s}}} + \bar{\sigma}_y \sqrt{\frac{\Gamma_s}{N_{g,s}}} \right] \leq \frac{\Gamma_s}{28}. \end{aligned}$$

The final inequality follows by taking the constants in (110) sufficiently large. Combining this with (108) yields

$$\mathbb{E}\Psi(z^{s+1}) \leq \frac{1}{2} \Gamma_s. \quad (111)$$

Theorem K.11 (Restarted stochastic Hölder complexity). *Under the assumptions of this subsection, let $z^0 \in \mathcal{Z}$ be deterministic and let $\Gamma_0 > 0$ be a known upper bound on $\Psi(z^0)$. For $0 < \varepsilon < \Gamma_0$, set*

$$\Gamma_s := 2^{-s} \Gamma_0, \quad S := \left\lceil \log_2 \frac{\Gamma_0}{\varepsilon} \right\rceil, \quad \ell_\varepsilon := 1 + \log \frac{\Gamma_0}{\varepsilon}.$$

Run S restarts with (110) and the parameters specified above. Then

$$\mathbb{E}\Psi(z^S) \leq \varepsilon.$$

For $q \in \{x, y\}$, write $N_x^{\text{tot}} := N_f^{\text{tot}}$ and $N_y^{\text{tot}} := N_g^{\text{tot}}$. Define

$$G_\theta(S) := \begin{cases} \frac{2^{\theta S} - 1}{2^\theta - 1}, & \theta > 0, \\ S, & \theta = 0. \end{cases}$$

The total oracle counts satisfy

$$N_q^{\text{tot}} = \mathcal{O} \left(S + \mathcal{K}_q \Gamma_0^{-\theta_q} G_{\theta_q}(S) + \frac{\sigma_q^2}{\mu_q \Gamma_0} (2^S - 1) \right), \quad q \in \{x, y\}, \quad (112)$$

and

$$N_B^{\text{tot}} = \mathcal{O} \left(\left(1 + \frac{L_{xy}}{\sqrt{\mu_x \mu_y}} \right) \ell_\varepsilon \right). \quad (113)$$

In particular, for each function block,

$$N_q^{\text{tot}} = \begin{cases} \mathcal{O}_{\nu_q} \left(\ell_\varepsilon + \left(\frac{H_q^2}{\mu_q^{1+\nu_q} \varepsilon^{1-\nu_q}} \right)^{\frac{1}{1+3\nu_q}} + \frac{\sigma_q^2}{\mu_q \varepsilon} \right), & 0 \leq \nu_q < 1, \\ \mathcal{O} \left(\left(1 + \sqrt{\frac{H_q}{\mu_q}} \right) \ell_\varepsilon + \frac{\sigma_q^2}{\mu_q \varepsilon} \right), & \nu_q = 1. \end{cases} \quad (114)$$

The constants in (112) are absolute. The notation $\mathcal{O}_{\nu_q}$ in (114) permits dependence on the fixed Hölder exponent.

Proof. Induction using (111) gives $\mathbb{E}\Psi(z^s) \leq \Gamma_s$, and $\Gamma_S \leq \varepsilon$. Summing (110) yields (112), since

$$\sum_{s=0}^{S-1} \Gamma_s^{-\theta} = \Gamma_0^{-\theta} G_\theta(S), \quad \sum_{s=0}^{S-1} \Gamma_s^{-1} = \frac{2^S - 1}{\Gamma_0} < \frac{2}{\varepsilon}.$$

For $\theta_q > 0$, the first sum is at most

$$\frac{2^{\theta_q}}{2^{\theta_q} - 1} \varepsilon^{-\theta_q}.$$

This gives the first case of (114). For $\nu_q = 1$, $\theta_q = 0$ and $\mathcal{K}_q = \sqrt{H_q/\mu_q}$, giving its second case. The coupling budget is constant across stages, so its sum is (113). $\square$

If a known upper bound already satisfies $\Gamma_0 \leq \varepsilon$, including the case $\Gamma_0 = 0$, return z^0 . Formula (112) displays the smooth limit: $G_\theta(S) \rightarrow S$ as $\theta \downarrow 0$. In contrast, the constant in the simplified polynomial bound is not uniform as $\nu_q \uparrow 1$. The variance contribution has no restart logarithm, because its stage budgets form a geometric sum. The construction uses no anchor samples and gives ε -accuracy without an anchor error term.

The guarantee is in the Lyapunov metric Ψ , and implies

$$\mathbb{E}\|x^S - x^*\|^2 \leq \frac{\varepsilon}{\mu_x}, \quad \mathbb{E}\|y^S - y^*\|^2 \leq \frac{\varepsilon}{\mu_y}.$$

A uniform primal–dual gap guarantee requires a separate conversion.

Remark 4 (Scope of the strongly convex extension). The Hölder and strong-convexity assumptions above are imposed on the feasible sets. Indeed, if a function is μ -strongly convex and satisfies a (ν, H) -Hölder condition with $\mu > 0$ and $\nu < 1$, then for any two distinct feasible points u, v ,

$$\mu\|u - v\|^2 \leq \langle h'(u) - h'(v), u - v \rangle \leq H\|u - v\|^{1+\nu}.$$

Hence their distance is at most $(H/\mu)^{1/(1-\nu)}$. In particular, these assumptions cannot hold globally on a positive-dimensional full space.

Theorem K.11 uses $\mu_x, \mu_y > 0$ and $p_3 = 0$. It therefore needs neither active-anchor ambient smoothness nor samplewise range invariance. Exploiting coupling-induced curvature, or treating a block with zero intrinsic curvature, requires additional analysis: active anchor terms must retain the appropriate genuine smoothness assumptions, and any subspace invariance argument must also control the stochastic oracle samples. The inexact-model constants $L_i(d_i)$ do not by themselves imply (20). The uniform gap conversions for regularized partially strongly convex problems in Appendix J also require their own stochastic accuracy accounting.

L Experimental Details

All experiments are CPU-only and single-threaded, run on an Intel Core i5-8250U under x86-64 Linux 6.8 with Python 3.12, NumPy 2.5, SciPy 1.18, and Matplotlib 3.11.

L.1 Synthetic experiments

Instances. We use separable primal and dual functions with a linear tilt,

$$f(x) = \frac{1}{2} \sum_i a_i x_i^2 + \frac{\alpha_x}{1+\nu_x} \sum_i |x_i|^{1+\nu_x} - \langle c_x, x \rangle,$$

$$g(y) = \frac{1}{2} \sum_j b_j y_j^2 + \frac{\alpha_y}{1+\nu_y} \sum_j |y_j|^{1+\nu_y} - \langle c_y, y \rangle,$$

on boxes $\mathcal{X} = [-R_x, R_x]^{d_x}$, $\mathcal{Y} = [-R_y, R_y]^{d_y}$, with $d_x = d_y = 60$ and $R_x = R_y = 1$, so $\Omega = R_x^2 d_x + R_y^2 d_y = 120$. Separability is what makes the exact primal–dual gap computable, and the tilts $c_x, c_y \sim \mathcal{N}(0, I)$ move the saddle point away from the origin — without them $z^0 = 0$ is already optimal and every method looks identical. Tilts change no constant, since H, μ and L depend only on the even part.

The coupling is $\mathbf{B} = \mathbf{U}\Sigma\mathbf{V}^\top$ with $\mathbf{U}, \mathbf{V}$ Haar-distributed (from the QR factorization of Gaussians) and singular values log-spaced in $[\sigma_{\min}, \sigma_{\max}]$. Hence $\lambda_{\max}(\mathbf{B}^\top \mathbf{B}) = \sigma_{\max}^2$, $\lambda_{\min}^+(\mathbf{B}^\top \mathbf{B}) = \sigma_{\min}^2$, and $\mathbf{B}$ has full column rank, so the compatibility conditions in Assumption 4.3 hold automatically.

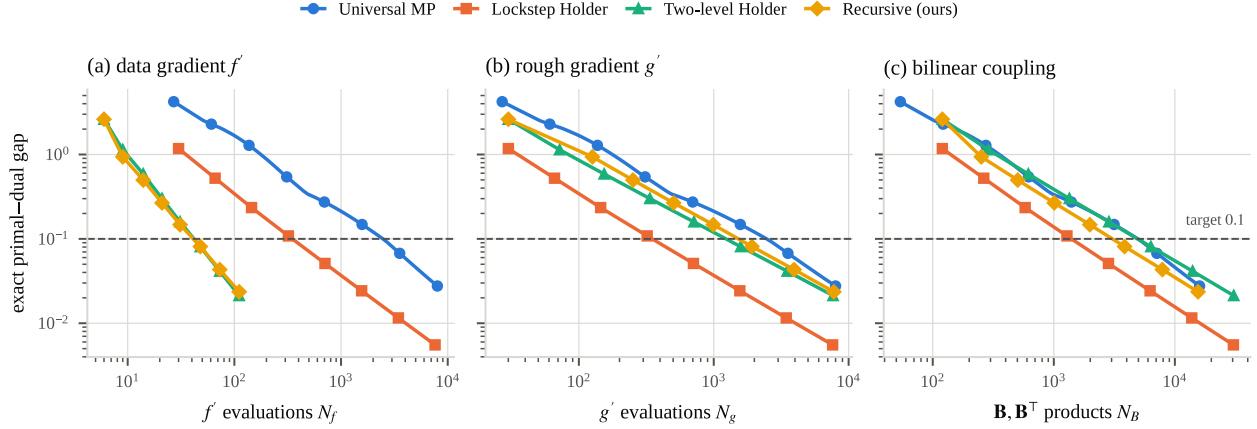


Figure 2: Oracle-wise convergence for $(\nu_x, \nu_y) = (0.75, 0.25)$: exact gap versus f' , g' , and $\mathbf{B}, \mathbf{B}^\top$ evaluations (median, seeds 0, 1, 2). Universal Mirror-Prox is the external baseline. Lockstep and two-level Hölder are no- and partial-separation ablations. Settings are frozen after tuning on seed 1729.

Constants used in the runs. Because Assumptions 4.1–4.3 allow any valid strong convexity lower bounds, we set

$$\mu_x = \mu_y = 0, \quad \mu_{xy} = \mu_{yx} = 0,$$

so $\delta_x = \delta_y = 0$ and Table 4 applies. We use $p_3 \equiv 0$, $\mathbf{P} = \mathbf{I}$, and no restarts, and set the available curvature lower bounds to zero to isolate the Hölder exponents.

The Hölder constant of ∇f on $\mathcal{X}$ is

$$H_x = 2^{1-\nu_x} \alpha_x d_x^{(1-\nu_x)/2} + \left(\max_i a_i \right) \text{diam}(\mathcal{X})^{1-\nu_x}.$$

The first term follows from the scalar constant $2^{1-\nu_x}$ and coordinate aggregation, with equality for equally spread increments. The second converts the quadratic part’s Lipschitz constant into a Hölder bound on the bounded set.

Metric and oracle accounting. We report

$$\text{Gap}(x, y) := \max_{y' \in \mathcal{Y}} F(x, y') - \min_{x' \in \mathcal{X}} F(x', y).$$

Separability and box constraints reduce both extrema to scalar problems, solved by vectorized bisection to machine precision. For $p = f + g$ and the skew coupling Q , this is precisely the residual in Theorem 3.5:

$$\text{Gap}(x_{\text{out}}, y_{\text{out}}) = \sup_z \{p(z_{\text{out}}) - p(z) + \langle Q(z), z_{\text{out}} - z \rangle\}.$$

Instrumented wrappers count each f' , g' , $\mathbf{B}v$, and $\mathbf{B}^\top u$ evaluation. Gap computations are excluded. With the coupling innermost, the three-level scheme uses $N_f = T_1$, $N_g = T_1 T_2$, and $N_B = 4T_1 T_2 T_3$.

Baselines and tuning. Figure 2 compares Generalized Universal Mirror-Prox (Stonyakin et al. 2022) with lockstep and two-level ablations. Generalized Universal Mirror-Prox covers Hölder-continuous operators. Lockstep assigns every component the largest required budget, while two-level groups g' with the coupling. Both schedules follow Corollary 3.6. The universal and lockstep methods query f' and g' equally often, so $N_g/N_f = 1$.

The smooth cost study uses Mirror-Prox (Nemirovski 2004) under the same oracle budget. The application also includes Accelerated Mirror-Prox (Chen et al. 2017) and two-level Mirror-Prox Sliding (Lan and Ouyang 2021), with f outside and $g + \mathbf{D}$ inside. Each of the four methods selects among twelve configurations on one validation instance per regime, then freezes them for five test instances. The sliding method uses the constants supplied by the synthetic generator, whereas the universal baseline estimates its parameters adaptively.

Sorting the recursive levels is essential when $R_c \ll R_g < R_f$. As ε decreases from 8 to 1, the sorted order uses 4–40 coupling products, compared with 28,804–1,843,204 for the fixed order, while both attain the requested targets.

ν_x	predicted	measured	per-seed range	decades	gap/bound
0.00	2.000	1.922	[1.897, 1.931]	1.8	[0.172, 0.270]
0.25	1.143	1.103	[1.099, 1.111]	2.7	[0.157, 0.196]
0.50	0.800	0.819	[0.817, 0.819]	3.6	[0.135, 0.210]
0.75	0.615	0.642	[0.642, 0.645]	3.7	[0.119, 0.210]
1.00	0.500	0.509	[0.506, 0.510]	2.9	[0.080, 0.112]

Table 5: Measured and predicted exponent of N_f in $1/\varepsilon$ (median and per-seed slope range over three seeds) and the measured gap-to-bound range over all budgets. “Decades” is the fitted gap range. Using the unrounded fits, the largest relative deviation from the predicted exponent is 4.28%, attained at $\nu_x = 0.75$.

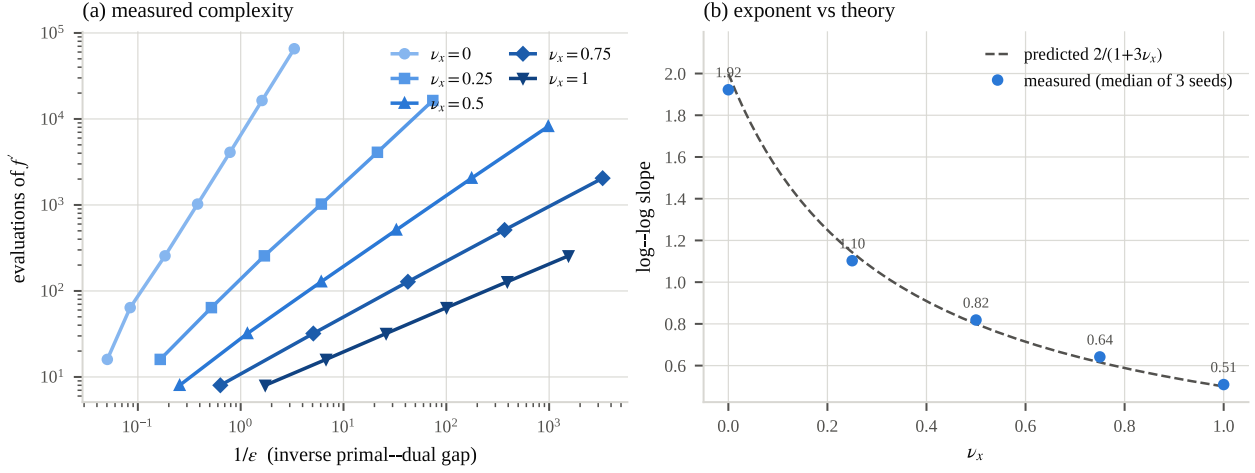


Figure 3: f' -complexity across the Hölder range in the degenerate regime: N_f versus inverse exact gap (left) and fitted versus predicted slope $2/(1+3\nu_x)$ (right), medians over three seeds.

(ν_x, ν_y)	levels	N_f	N_g	N_g/N_f	N_B	gap	seed range
(1, 0.5)	fBg	16	96	6	1920	0.152	[0.108, 0.196]
(1, 0)	fBg	16	115680	7230	1920	0.417	[0.332, 0.419]
(0.75, 0.25)	fBg	31	992	32	1984	0.147	[0.094, 0.185]
(0.5, 0.5)	fBg	93	93	1	2232	0.127	[0.073, 0.159]
(0, 0)	Bfg	115680	115680	1	1920	0.532	[0.456, 0.572]

Table 6: Component budgets at $\varepsilon = 0.5$ (median, three seeds). Levels are sorted outermost first. The last columns give the attained gap and range. Reversing unequal exponent pairs exchanges the roles of f and g , so mirrored rows are omitted. All displayed rows meet the target except the fully nonsmooth case, which is 6% above it with $\kappa_n = 1$.

The cost trade-off. Table 7 reports, at matched accuracy on a smooth instance, the gradient and matrix-vector counts of the sliding method and of Mirror-Prox, the resulting ratios, and the crossover $\rho^* = \text{cost}(f')/\text{cost}(\mathbf{B}v)$ above which sliding attains the lower total cost $\rho(N_f + N_g) + N_B$. The saving in gradient calls grows as ε decreases (from $4.7\times$ to $8.5\times$) and so does the extra spend on products (from $18\times$ to $40\times$), leaving ρ^* in the range 22–44. Sliding is therefore the cheaper method exactly when $\rho > \rho^*$, that is, when one (sub)gradient evaluation costs more than a few dozen matrix-vector products. Below that threshold Mirror-Prox is cheaper.

ε	sliding			Mirror-Prox			ratio		
	grad	B	gap	grad	B	gap	grad saved	B extra	ρ^*
$1e-01$	48	4096	$5.3e-02$	224	224	$9.5e-02$	$4.7\times$	$18.3\times$	22
$3e-02$	96	16384	$1.4e-02$	592	592	$2.9e-02$	$6.2\times$	$27.7\times$	32
$1e-02$	192	65536	$3.7e-03$	1632	1632	$1.0e-02$	$8.5\times$	$40.2\times$	44

Table 7: Oracle counts at matched accuracy and the cost crossover ρ^* . The entries are medians over three seeds. “grad” is $N_f + N_g$. For each method the reported counts are those at which its gap first crosses ε , so the accuracy actually attained is at most ε by construction.

Parameters and seeds. The swept values and selection criteria are:

- **Loop counts.** T_1 is swept geometrically, with $T_2 = 2$ and $T_3 = \lceil 4T_1^{(3\nu_x-1)/2} \rceil$, the rule that keeps the g - and coupling terms of (17) below the f -term so that the measured exponent is the one belonging to f . In the ordering experiment the T_i are instead those the theory prescribes, from the budgets R_i of Corollary 3.6.
- **Oracle tolerances.** δ_i and the induced L_i are computed from (16). No parameter is tuned.
- **Baseline parameters.** For the smooth cost experiment, the Mirror-Prox step-size grid is $\{2^k\}_{k=-6}^6$ times $1/(L_{xy} + H_x + H_y)$, and the run reaching the smallest gap within the shared oracle budget is reported. For Figure 2, Universal Mirror-Prox uses a target of 0.0625 in its backtracking test and selects $L_0 \in \{0.1, 1, 10\}$ on validation seed 1729. The selected value $L_0 = 1$ is then frozen.
- **Instance constants.** $a_i \equiv 0$ and $b_j \equiv 0$ (so the quadratic part is absent and the power term alone shapes f, g), $\alpha_x = \alpha_y = 1$, $\sigma_{\min} = 0.5$, and $\sigma_{\max} = 2$. Two experiments deviate: the exponent sweep of Table 5 fixes $\nu_y = 1$ and takes $\alpha_y = 0.05$, $b_j \in [0, 0.05]$ so that the g -term never binds while ν_x is swept over $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$. The convergence comparison of Figure 2 uses the genuinely Hölder pair $(\nu_x, \nu_y) = (\frac{3}{4}, \frac{1}{4})$. Reversed exponent pairs were also run and exchange the f - and g -budgets as expected. The ordering experiment uses $\nu_x = 0$, $\nu_y = 1$, $\alpha_x = 4$, $\alpha_y = 0.1$, $\sigma \in [0.02, 0.05]$ to place $R_c \ll R_g < R_f$.
- **Loop-count constant.** The construction of Corollary 3.6 is applied with $\kappa_n = 1$, using the raw budgets R_i . The theorem’s sufficient value is $2^{2n+2} = 256$. Four of five displayed rows in Table 6 attain the target. The fully nonsmooth row is 6% above it.
- **Synthetic seeds.** Results for seeds 0, 1, 2 are reported as medians, with the per-seed range added in the exponent table. Runs are deterministic given the seed.

Validity of the bound. The last column of Table 5 reports the measured gap as a fraction of the right-hand side of (17) over the whole grid of Section 5. There are no violations in 90 runs. The ratio drifts upward slowly with the budget, which is why a log–log slope fitted over too narrow a range of accuracies overestimates the exponent. The fits therefore use the largest four budgets, over which the gap spans 1.8–3.7 decades.

L.2 Tomographic reconstruction

Objective, domains, and scaling. For $n = 64$, let $d = n^2$, $m = 180 \cdot 91 = 16,380$, and consider

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F_{\tau,c}(x, y) = \frac{1}{2} \|\mathbf{A}x - b\|^2 + \frac{\mu}{2} \|x\|^2 + \langle y, \mathbf{D}x \rangle - \frac{\beta}{2} \|y\|^2 - \frac{\tau}{2} \langle y, \mathbf{L}_{\text{nl}} y \rangle - \frac{c}{1+\nu_y} \sum_{i=1}^{2d} |y_i|^{1+\nu_y}, \quad (115)$$

that is, in the notation of (1),

$$f(x) = \frac{1}{2} \|\mathbf{A}x - b\|^2 + \frac{\mu}{2} \|x\|^2, \quad g_{\tau,c}(y) = \frac{\beta}{2} \|y\|^2 + \frac{\tau}{2} \langle y, \mathbf{L}_{\text{nl}} y \rangle + \frac{c}{1+\nu_y} \sum_{i=1}^{2d} |y_i|^{1+\nu_y}, \quad \mathbf{B} = \mathbf{D},$$

$$\mathcal{X} = \{x : \|x\| \leq R\}, \quad \mathcal{Y} = [-\lambda, \lambda]^{2d}, \quad R = 1.05(\|\hat{x}(0)\| + \|\mathbf{D}\| \lambda \sqrt{2d}/\mu), \quad \hat{x}(0) = (\mathbf{A}^\top \mathbf{A} + \mu \mathbf{I})^{-1} \mathbf{A}^\top b.$$

The definition of R bounds $\hat{x}(y) = \hat{x}(0) - (\mathbf{A}^\top \mathbf{A} + \mu \mathbf{I})^{-1} \mathbf{D}^\top y$ uniformly over $\mathcal{Y}$. Thus the ball leaves the primal minimizers unchanged and gives the finite domain constant $\Omega = R^2 + 2d\lambda^2$. In the reported runs, $R \approx 29.0$, $\|x^*\| \approx 14.9$, and the projection is inactive.

The projector $\mathbf{A} \in \mathbb{R}^{m \times d}$ is sparse pixel-driven parallel-beam, splatting each pixel, of width $2/(n-1)$, onto its two neighboring detector bins. Its stored transpose is the exact adjoint, and $\mathbf{A}$ is divided by its spectral norm. The target intensity lies in $[0, 1]$, and

$$b = \mathbf{A}x^\dagger + \xi, \quad \xi_j \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \quad \sigma = 0.01 \|\mathbf{A}x^\dagger\| / \sqrt{m}.$$

The map $\mathbf{D}$ concatenates the unscaled forward differences $(D_h x)_{i,j} = x_{i,j+1} - x_{i,j}$ and $(D_v x)_{i,j} = x_{i+1,j} - x_{i,j}$, with zero in the last column or row and the corresponding exact transpose implemented.

The three regimes. All share

$$\mu = 10^{-3}, \quad \beta = 0.05, \quad \lambda = 5 \cdot 10^{-5},$$

and differ only in the dual parameters (τ, c, ν_y) .

Standard ($\tau = c = 0$). Eliminating y in (115) gives the primal problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{A}x - b\|^2 + \frac{\mu}{2} \|x\|^2 + \lambda \sum_{i=1}^{2d} h_{\beta\lambda}((\mathbf{D}x)_i), \quad h_{\delta}(t) = \begin{cases} t^2/(2\delta), & |t| \leq \delta, \\ |t| - \delta/2, & |t| > \delta, \end{cases} \quad (116)$$

so the Huber threshold is $\beta\lambda$, not β .

Nonlocal ($\tau = 0.05, c = 0$), which tests the third level. Here $\mathbf{L}_{\text{nl}} = \text{diag}(L, L)$, where L is the graph Laplacian connecting all pixels at offsets $|\Delta i|, |\Delta j| \leq 5$ with weights $\exp(-(\Delta i^2 + \Delta j^2)/(2 \cdot 2.5^2))$, divided by the maximum weighted degree. This normalization gives $\|\mathbf{L}_{\text{nl}}\|_2 \leq 2$, with 908,552 stored entries at 64^2 . Here g' applies the sparse graph while $\mathbf{D}$ remains a local stencil, and eliminating y gives the spatially correlated penalty $g_{\tau,0}^*(\mathbf{D}x)$.

Hölder ($\tau = 0.05, \nu_y = \frac{1}{2}, c = 7.955 \cdot 10^{-3}$) adds a power term whose gradient $c \text{sign}(y)|y|^{\nu_y}$ is (ν_y, H_y) -Hölder but not Lipschitz. We set the coefficient using the dimensionless boundary-gradient ratio $r_c = 22.5$:

$$c = \text{round}_{10^{-6}}(r_c \beta \lambda^{1-\nu_y}) = 7.955 \cdot 10^{-3},$$

which makes the power-gradient magnitude 22.5 times the diagonal quadratic gradient at $|y_i| = \lambda$. The Hölder constant is the two-part bound of Appendix L.1, $H_y = c^{2^{1-\nu_y}}(2d)^{(1-\nu_y)/2} + (\beta + 2\tau) \text{diam}(\mathcal{Y})^{1-\nu_y}$, and the sliding curvature is $L_2(\delta_2^c)$ of (16) rather than a Lipschitz constant.

Hölder diagnostics. At a reference saddle point, 3.10% of coordinates satisfy $|y_i| \leq 0.01\lambda$, and the 99th percentile of the coordinatewise power-gradient Lipschitz quotient is 85.8, compared with the quadratic constant 0.15. The observed global Hölder quotient is $0.0616 < H_y = 0.1213$. Varying the coefficient over $c/2, c, 2c$ leaves the weighted-cost ordering unchanged and all median gaps below $3.1 \cdot 10^{-4}$.

The values (μ, β, λ) were selected before method tuning using a validation-only pilot over eight candidate pairs. Each received the same 250-step L-BFGS screen, and the selected pair maximized validation PSNR subject to $\mu \geq 10^{-3}$.

Metric and timing. Let $\text{Gap}_{\tau,c}$ be the gap of Appendix L.1 formed with $F_{\tau,c}$ on $\mathcal{X} \times \mathcal{Y}$. We report

$$\text{RelGap}_{\tau,c}(x, y) = \frac{\text{Gap}_{\tau,c}(x, y)}{\text{Gap}_{\tau,c}(0, 0)},$$

with the denominator recomputed on every instance. The primal infimum is computed by diagonally preconditioned CG to relative tolerance $2 \cdot 10^{-11}$, and the dual maximum by proximal gradient with coordinatewise bisection in the proximal map. Strong-convexity residual bounds certify both solves. On the recorded Hölder run, the resulting gap-value error bound is $9.73 \cdot 10^{-18}$.

The weights c_f, c_g, c_B are median isolated latencies of the actual f', g' , and one $\mathbf{D}$ or $\mathbf{D}^\top$ kernel, and $C_{\text{weighted}} = c_f N_f + c_g N_g + c_B N_B$. Median seedwise ratios are $(c_f/c_B, c_g/c_B) = (157, 0.118)$ in the diagonal regime and $(160, 49.4)$ in the nonlocal regime. End-to-end wall time includes all algorithm and projection overhead but excludes construction, validation, cost calibration, and gap evaluation.

Tuning protocol. The configurations of Appendix L.1, run on validation phantom/noise pair $(0, 0)$, are as follows. Mirror-Prox uses

$$\eta/\eta_0 \in \{0.75, 1, 1.25\}, \quad N \in \{1536, 1792, 2048, 2304\}, \quad \eta_0 = (H_x + H_y + L_{xy})^{-1},$$

Accelerated MP uses the same iteration set with curvature scales $\{0.75, 1, 1.25\}$. Each sliding grid contains twelve configurations. The MPS grid consists of $(T_1, T_2) = (96, 32), (112, 32)$ and every pair in $\{128, 144, 160, 176, 192\} \times \{24, 32\}$. The recursive grid consists of $(T_1, T_2, T_3) = (96, 4, 8), (112, 4, 8)$ and, for each $T_1 \in \{128, 144, 160, 176, 192\}$, both $(T_2, T_3) = (3, 10)$ and $(4, 8)$. Selection minimizes C_{weighted} subject to validation $\text{RelGap} \leq 3.75 \cdot 10^{-4}$, 75% of the test target, with no restarts and no test-time stopping. For $\tau = 0$, the selected fixed configurations are Mirror-Prox $(\eta, N) = (0.321384, 1792)$, Accelerated MP $(s, N) = (0.75, 1536)$, MPS $T = (128, 32)$, and recursive $T = (160, 3, 10)$. For $\tau = 0.05$, they are respectively $(0.313328, 1792)$, $(0.75, 1536)$, $(160, 24)$, and $(160, 3, 10)$. In the Hölder regime, Generalized Universal Mirror-Prox (Stonyakin et al. 2022) is the external baseline. The fixed-step methods are reported as ablations. Universal Mirror-Prox sweeps the controlled inexactness $\varepsilon \in \{2 \cdot 10^{-3}, 5 \cdot 10^{-4}, 10^{-4}\}$ of its backtracking test against $N \in \{384, 448, 512, 576\}$ at $L_0 = 1$. Validation selects $(\varepsilon, N) = (2 \cdot 10^{-3}, 576)$, whose test relative gaps, $2.32 \cdot 10^{-4}$ to $3.21 \cdot 10^{-4}$, are comparable to those of the other methods.

Test evaluation. The frozen counts are evaluated on five unseen pairs $(\text{seed}, \text{phantom}) = (1, 0), (2, 1), (3, 2), (4, 0), (5, 1)$, so variability includes three object geometries as well as independent measurement noise. Each method is timed three times per pair. The reported wall time is the median of the five per-pair medians, with their full range. All sixty frozen method/regime runs attain the common $\text{RelGap} \leq 5 \cdot 10^{-4}$ target.

Scope. At the reported accuracies, the component budgets are $R_f \approx 1.7 \cdot 10^3$, $R_g \approx 6.9 \cdot 10^3$, and $R_c \approx 7.9 \cdot 10^6$, so coupling calls determine the rate. Tomography therefore demonstrates oracle separation with an expensive, non-Lipschitz g' , but not the ν_y -dependent exponent. The synthetic experiments test that exponent.

method	N_f	N_g	N_B	weighted ms	wall ms [range]	rel. gap	PSNR [range]
<i>Huber-TV</i>							
Mirror-Prox	3584	3584	7168	16236	17540 [17484, 17660]	2.55×10^{-4}	25.67 [23.84, 25.82]
Accelerated MP	1536	1536	6144	7052	7899 [7882, 7937]	3.14×10^{-4}	25.75 [23.99, 26.05]
MP Sliding	128	4096	16384	1086	2433 [2420, 2461]	2.80×10^{-4}	25.77 [24.00, 26.08]
Recursive (ours)	160	480	19200	1302	3177 [3120, 3194]	2.98×10^{-4}	25.68 [23.91, 25.98]
<i>Nonlocal Huber-TV</i>							
Mirror-Prox	3584	3584	7168	21145	23070 [22996, 23365]	2.66×10^{-4}	25.59 [23.77, 25.74]
Accelerated MP	1536	1536	6144	9157	10348 [10195, 10506]	3.14×10^{-4}	25.75 [23.99, 26.05]
MP Sliding	160	3840	15360	6498	8387 [8227, 8445]	2.94×10^{-4}	25.71 [23.94, 26.01]
Recursive (ours)	160	480	19200	1968	3957 [3935, 4000]	2.98×10^{-4}	25.67 [23.91, 25.98]
<i>Nonlocal Hölder-TV</i>							
Universal MP	1728	1728	3456	10508	10780 [10698, 11331]	2.59×10^{-4}	25.64 [23.82, 25.80]
Accelerated MP	1536	1536	6144	9426	10429 [9824, 10458]	3.15×10^{-4}	25.74 [23.98, 26.05]
MP Sliding	160	3840	15360	6889	8562 [8322, 8618]	2.94×10^{-4}	25.70 [23.94, 26.01]
Recursive (ours)	160	480	19200	2029	3981 [3922, 4022]	3.03×10^{-4}	25.63 [23.87, 25.94]

Table 8: Application comparison at the common relative primal–dual-gap target $5 \cdot 10^{-4}$. Weighted cost uses measured per-component latency. Wall time is end-to-end optimization time. Entries are medians over five unseen phantom/noise instances. Brackets give the observed instance range for wall time and PSNR.

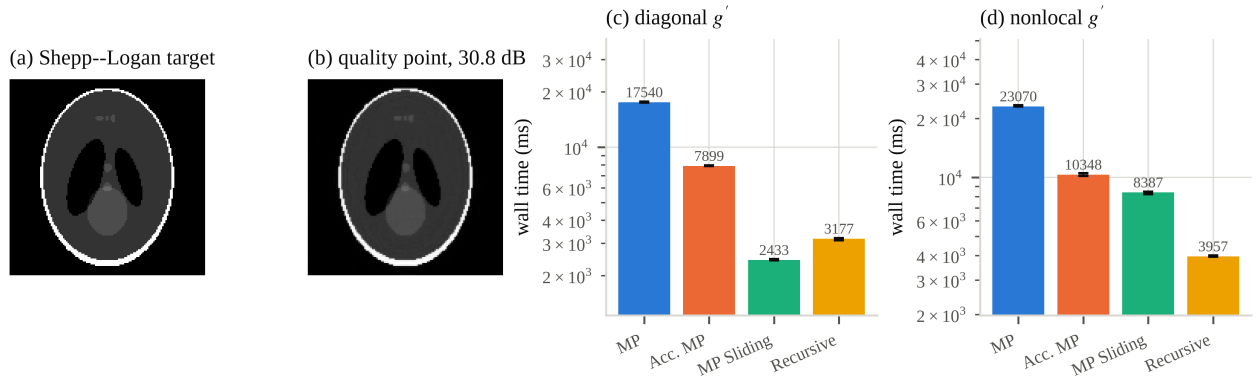


Figure 4: Tomographic reconstruction and timing. (a) 128^2 modified Shepp–Logan target. (b) Recursive result at $T = (768, 4, 8)$, reaching 30.81 dB at relative gap $3.26 \cdot 10^{-5}$. (c–d) Median end-to-end time at the common $5 \cdot 10^{-4}$ target, with bars spanning the five test-instance medians. The standard diagonal and nonlocal regimes use the counts and times in Table 8.

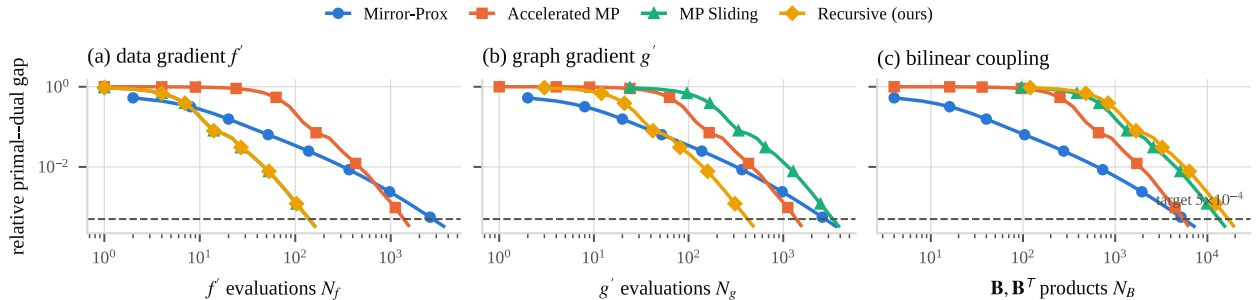


Figure 5: Oracle-wise convergence on the first unseen nonlocal tomography instance. The three panels plot the relative primal–dual gap against, respectively, evaluations of the data gradient f' , evaluations of the graph gradient g' , and products with B, B^T . The dashed line is the common target $5 \cdot 10^{-4}$.

M Broader Complexity-Separation Results

This appendix expands the positioning summarized in Section 1.2 by comparing the oracle models and assumptions of the closest complexity-separation results.

Joint versus separated oracle counts. Kovalev et al. (2022) solve the smooth bilinear problem (1) with the Accelerated Primal-Dual Gradient method, which is linearly convergent and optimal in the strongly-convex–strongly-concave regime. Its oracle model allows $\mathcal{O}(1)$ evaluations of ∇f , of ∇g , and of products with $\mathbf{B}, \mathbf{B}^\top$ at every iteration, and its complexity is a single count $\mathcal{O}(\min\{T_a, T_b, T_c, T_d\} \log(1/\varepsilon))$ whose leading term is

$$T_a = \max\left\{\sqrt{\kappa_x}, \sqrt{\kappa_y}, \sqrt{\kappa_{xy}}\right\}, \quad \sqrt{\kappa_{xy}} = \frac{L_{xy}}{\sqrt{\delta_x \delta_y}},$$

charged simultaneously to all three oracles. Separation replaces this maximum by the three terms individually, one per oracle. Borodich et al. (2025) obtain such separate counts with optimal dependence on $\sqrt{\kappa_x}$, $\sqrt{\kappa_y}$, and $\sqrt{\kappa_{xy}}$ for the smooth bilinear problem.

Separation for general coupling. Tominin et al. (2021) consider $\min_x \max_y \{f(x) + G(x, y) - h(y)\}$ with h of finite-sum form, strongly convex in x and strongly concave in y , and all components smooth. They estimate separately the numbers of evaluations of ∇f , $\nabla_x G$, $\nabla_y G$ and ∇h_i , obtaining budgets such as $\tilde{\mathcal{O}}(\sqrt{\kappa_x^{(f)} \kappa_y^{(G)}})$ for ∇f and $\tilde{\mathcal{O}}(\sqrt{\kappa_x^{(G)} \kappa_y^{(G)}})$ for the coupling. Their coupling budgets depend on products of primal and dual condition numbers of G , whereas the bilinear specialization considered here uses the spectral quantity $\sqrt{\kappa_{xy}}$. Their separation is obtained through a two-level accelerated construction for $\min_x F(x)$, with $F(x) = f(x) + \max_y \{G(x, y) - h(y)\}$, driven by inexact proximal steps and probabilistic inexact oracles. The resulting bounds carry polylogarithmic factors in ε^{-1} and the confidence level, and assume the strongly-convex–strongly-concave regime.

Universal Hölder primal–dual methods. Affine-constrained models have bilinear Lagrangians. Yurtsever et al. (2015) adapt to unknown Hölder regularity in the dual formulation, while Luo (2024) handle Lipschitz and Hölder gradients universally and prove an optimal mixed-type rate. These methods use joint iteration or line-search counts, while the present analysis tracks function-gradient and $A, A^\top$ evaluations separately.

Parameter-free composite separation. For composite minimization, the universal, parameter-free PFUGS method of Wu et al. (2026) separately counts a Hölder f' -oracle and a smooth g' -oracle. Its model has two function-oracle families and no coupling oracle. The model studied here tracks f', g' , and $\mathbf{B}, \mathbf{B}^\top$, with separate Hölder exponents for the two function components.