
Projection-Free Algorithms for Nonsmooth Stochastic Convex-Concave Saddle-Point Problems

Khanh-Hung Giang-Tran
Cornell University
tg452@cornell.edu

Soroosh Shafiee
Cornell University
shafiee@cornell.edu

Abstract

We study nonsmooth convex-concave saddle-point problems over compact convex sets, assuming access to stochastic subgradients of the payoff function. We develop *single-loop projection-free* algorithms that use linear minimization oracles over the primal and dual domains. Unlike prior projection-free approaches that rely on smoothing, our methods are purely subgradient-based and handle nonsmoothness directly. This design makes the framework modular: when a linear minimization oracle is not used over a specific domain, the corresponding update can be replaced by a standard projection oracle without changing the single-loop structure. Thus, our framework covers both fully projection-free and hybrid oracle configurations. We prove *anytime* strong saddle gap guarantees under standard unbiased stochastic oracle assumptions with bounded variance. Our algorithms achieve an $O(\epsilon^{-2})$ iteration complexity, matching projected stochastic subgradient methods. We also provide matching lower bounds for the corresponding oracle models, establishing minimax optimality. Our results show that, for nonsmooth stochastic minimax problems, the computational advantages of linear minimization oracles need not come at the expense of statistical or oracle efficiency.

1 Introduction

Projected subgradient descent is a fundamental approach for constrained nonsmooth convex optimization [7, 62]. When projections onto the feasible set are computationally tractable, these methods are minimax optimal [51, 56], reaching an ϵ -approximate solution in $O(\epsilon^{-2})$ iterations. This optimality extends to the stochastic regime, where projected stochastic subgradient descent maintains the same iteration complexity [50, 61]. However, in many large-scale applications, the projection step becomes the primary computational bottleneck, necessitating the use of projection-free alternatives such as the Frank-Wolfe (FW) algorithm [9, 32]. While projection-free methods are well-established for smooth problems, their application to the nonsmooth setting is a more recent development. Notably, Thekumparampil et al. [65, 66] and Asgari and Neely [3, 4] have introduced nonsmooth FW variants that achieve an iteration complexity of $O(\epsilon^{-2})$, matching the minimax optimal rate [39].

This work addresses the nonsmooth convex-concave saddle-point problem of the form

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y), \quad (1)$$

where we assume access only to unbiased stochastic subgradients of f . For this class of problems, projected stochastic subgradient descent-ascent is known to achieve the optimal convergence rate of $O(\epsilon^{-2})$ [48, 50]. However, the landscape of projection-free methods for nonsmooth saddle-point problems is significantly less understood. This paper provides a comprehensive analysis of *single-loop*, linear minimization oracle (LMO)-based algorithms for (1), establishing anytime convergence guarantees. Our proposed framework is flexible, utilizing an LMO over $\mathcal{X}$, $\mathcal{Y}$, or both. In instances where an LMO is not employed for a specific domain, we utilize a standard projection oracle (PO).

1.1 Contributions

Our main contributions are as follows.

- ◊ Building upon the subgradient-based framework introduced in [3, 4], we propose a novel algorithmic update for saddle-point problems that employs a non-trivial mechanism to derive the LMO direction. This approach directly addresses objective nonsmoothness *without* requiring smoothing techniques. Furthermore, the update rule is designed for modularity, allowing it to be seamlessly integrated with a standard PO when an LMO is not employed for a specific variable. This allows us to rigorously analyze three distinct oracle configurations: LMO-LMO (fully projection-free), LMO-PO, and PO-LMO (hybrid). This flexibility enables practitioners to select the most efficient oracle for each constraint set $(\mathcal{X}, \mathcal{Y})$.
- ◊ We establish *anytime* convergence guarantees for all proposed algorithmic variants. This is of significant practical importance as it eliminates the need to pre-specify a target accuracy ϵ or a total iteration horizon T to tune parameters. We prove that our algorithms achieve the optimal $O(\epsilon^{-2})$ iteration complexity, matching the theoretical performance of standard projected stochastic subgradient descent-ascent methods [24, 51, 56].
- ◊ We complement our upper bounds with matching lower bounds. In particular, we prove an $\Omega(\epsilon^{-2})$ iteration complexity lower bound for any algorithm based on the LMO-LMO oracle model, as well as for any method operating under the LMO-PO oracle model. By strong duality, the same lower bound immediately extends to the PO-LMO setting. Hence, the rates achieved by our algorithms are *minimax optimal* in all three oracle regimes.
- ◊ Our results substantially strengthen the current theory of projection-free minimax optimization. In particular, even in the *smooth deterministic single-loop* setting, previous projection-free methods were only known to achieve slower rates [8, 18]. By contrast, we obtain the optimal $O(\epsilon^{-2})$ rate in the more general *nonsmooth stochastic* setting. This closes the convergence-rate gap between projection-free and projection-based methods for minimax optimization, showing that the computational advantages of LMO-based updates need not come at the expense of statistical or oracle complexity.

1.2 Related Works

Nonsmooth Frank-Wolfe Algorithms. Constrained minimax problems of the form (1) can be viewed as instances of nonsmooth convex optimization. This is because the primal function $p(x) := \max_{y \in \mathcal{Y}} f(x, y)$ is generally nonsmooth, even when f is smooth. The classic FW algorithm, designed for smooth objectives, cannot be directly applied here. White [68] first addressed nonsmooth optimization using an LMO. Since then, only a few studies have followed, and earlier works often rely on strict assumptions [13, 60]. More recently, Hazan and Kale [28] introduced the Online Frank-Wolfe algorithm, obtaining an $O(T^{-3/4})$ regret bound for general online convex optimization. This translates to an $O(\epsilon^{-4})$ rate for the offline nonsmooth convex setting. Under an additional smoothness assumption, Hazan and Minasyan [29] improved this complexity to $O(\epsilon^{-3})$ by using a Follow-the-Perturbed-Leader approach as a smoothing technique. Lan [39] used randomized smoothing in the deterministic setting to achieve an $O(\epsilon^{-2})$ rate.

A complementary approach addresses nonsmoothness through the composite structure of the objective. Lan and Zhou [42] and Lan [40] proposed the conditional gradient sliding algorithm, which decouples gradient and LMO computations. Using this sliding framework with Moreau-Yosida smoothing, Thekumparampil et al. [66] proposed a *double-loop* algorithm that achieves $\tilde{O}(\epsilon^{-2})$ LMO and subgradient oracle calls. Thekumparampil et al. [65] then extended this framework to a *single-loop* procedure for nonsmooth minimization. However, both algorithms rely on smoothing techniques that require parameters to be selected a priori. Consequently, these algorithms are not anytime procedures. Rather than relying on smoothing, Asgari and Neely [3] recently proposed a subgradient-based projection-free algorithm for nonsmooth convex optimization. This method attains an $O(\epsilon^{-2})$ convergence rate using a single LMO call per iteration, serving as a projection-free counterpart to projected subgradient descent. This was extended by Asgari and Neely [4] to handle general convex functional inequality constraints while preserving the $O(\epsilon^{-2})$ complexity, consistent with existing lower bounds [39]. A key algorithmic feature we share with the nonsmooth literature [3, 4, 21, 22, 42, 44, 64–66] is the use of subgradients *outside the feasible set*. Besides, similar to the approach in [3, 4], our analysis is subgradient-based rather than dependent on smoothing techniques.

LMO-Based Methods for Minimax Problems. Earlier works focused on special primal-dual Lagrangian-based formulations and established $O(\epsilon^{-2})$ rates in the smooth setting [27, 41, 43, 70]. Subsequent work extended this to smooth convex-strongly concave problems via stochastic *double-loop* methods [11] and to smooth convex-concave problems via *multi-loop* Follow-the-Perturbed-Leader approach [63], both recovering the $O(\epsilon^{-2})$ rate. Gidel et al. [19] pioneered the study of *single-loop* algorithms with linear convergence for smooth strongly convex-strongly concave problems with interior solutions, as well as for smooth convex-concave problems over strongly convex sets with lower-bounded gradients or polytope feasible set. Kolmogorov and Pock [35] developed a one-sided method for the bilinear case, achieving $O(\epsilon^{-1})$ rate.

A separate line of work considers the smooth nonconvex-concave setting, whose guarantees translate to the convex-concave case but at inferior rates since convexity is not exploited. Nouiehed et al. [57] proposed a multi-loop one-sided projection-free method achieving an $O(\epsilon^{-3.5})$ iterations. Boroun et al. [8] and Giang-Tran et al. [18] subsequently developed single-loop projection-free analyses for nonconvex-concave problems under various conditions, with rates ranging from $O(\epsilon^{-6})$ to $O(\epsilon^{-2})$ depending on the setting. To the best of our knowledge, this is the first work to address nonsmooth stochastic convex-concave minimax problems while achieving the optimal $O(\epsilon^{-2})$ rate.

LMO-Based Methods for Variational Inequalities. Smooth convex-concave minimax problems are an instance of monotone variational inequality problems. Hammond [26] showed that a single-loop FW algorithm with an open-loop stepsize converges *asymptotically* for monotone variational inequalities over strongly convex sets, and this was recently extended to general convex compact sets in [31]. While several works have established non-asymptotic complexity bounds via *nested multi-loop* schemes, such as semi-proximal mirror-prox methods [30] and extra gradient methods [5], single-loop algorithms have required either specific geometries or stronger oracles to achieve competitive rates. Specifically, Cox et al. [14] and Juditsky and Nemirovski [33] proved an $O(\epsilon^{-2})$ ergodic convergence rate under the accessibility of Fenchel-type representation of the payoff function and its corresponding first-order information. Moreover, Chen and Mazumdar [12] proved an $O(\epsilon^{-2})$ last-iterate convergence rate for a single-loop *generalized* FW method, but their approach relies on a stronger oracle rather than a pure LMO. In contrast, this paper considers nonsmooth stochastic minimax problems and develops single-loop purely LMO-based algorithms, enjoying the same rate.

PO-Based Methods. Many algorithms exist for solving monotone variational inequalities and smooth convex-concave saddle-point problems of the form (1). In the deterministic setting, the extragradient method [36], Popov’s algorithm [59], and Tseng’s forward-backward-forward scheme [67] all achieve an $O(\epsilon^{-1})$ rate, as does the mirror-prox algorithm of Nemirovski [49]. This rate is optimal [58] and is also attained by modern variants including reflected gradient methods [45, 47], optimistic gradient descent-ascent [15], dual extrapolation [54], and adaptive methods [2, 46]. For structured problems such as Lagrangian-based payoff functions for affinely constrained problems, specialized methods can exploit additional structure to obtain faster rates [10, 53]. Moving to the nonsmooth setting, Nedić and Ozdaglar [48] established an $O(\epsilon^{-2})$ rate for projected subgradient descent-ascent, and Nemirovski et al. [50] achieved the same rate under additional stochasticity. We develop algorithms based on various combinations of LMO and PO for nonsmooth stochastic convex-concave problems and establish the optimal $O(\epsilon^{-2})$ rate under this more challenging setting.

Lower Bound Complexities. Lower bounds for first-order methods have been extensively studied in the optimization literature. For nonsmooth convex minimization, the $O(\epsilon^{-2})$ complexity of subgradient methods is known to be optimal [24, 51, 56]. For smooth convex minimization, the matching lower bound for gradient-based methods was established by Nesterov [52], while Lan [39] derived analogous lower bounds for conditional gradient methods. More recently, lower bounds under additional strong convexity of the feasible set were studied by Halbey et al. [25] and Grimmer and Liu [23]. In the stochastic setting, optimal lower bounds for convex and nonconvex problems have been established in [1, 17]. For our minimax optimization problem, we show that our $O(\epsilon^{-2})$ upper bounds match their corresponding lower bounds, establishing that these rates are minimax optimal.

1.3 Notation and Outline

For a compact convex set $\mathcal{U}$, the LMO at a given direction d and the PO at a point u are defined as

$$\text{LMO}_{\mathcal{U}}(d) \in \arg \min_{v \in \mathcal{U}} d^{\top} v \quad \text{and} \quad \text{PO}_{\mathcal{U}}(u) := \arg \min_{v \in \mathcal{U}} \|u - v\|_2^2.$$

Convergence results in the main text are reported using standard big- O and $\tilde{O}$ notations, where the latter suppresses logarithmic factors for clarity. Detailed proofs, including explicit convergence constants, as well as implementation details are provided in the Appendix. Section 2 presents our proposed algorithms and derives the key recursions underlying each variant. Section 3 establishes their convergence guarantees, and Section 4 provides lower bounds for any algorithm relying on combinations of LMO and PO updates. Section 5 reports numerical experiments comparing our methods against state-of-the-art baselines.

2 Proposed Algorithms

We begin by stating our assumptions on the problem structure and the stochastic subgradient oracle.

Assumption 1. *The payoff function $f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ and the feasible sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ satisfy the followings.*

(i) *The domain $\mathcal{X}$ is a compact convex set and there exists another convex compact set $\bar{\mathcal{X}} \supseteq \mathcal{X}$ with*

$$R_{\mathcal{X}} := \max_{x \in \mathcal{X}, \bar{x} \in \bar{\mathcal{X}}} \|x - \bar{x}\|_2 > 0.$$

(ii) *The domain $\mathcal{Y}$ is a compact convex set and there exists another convex compact set $\bar{\mathcal{Y}} \supseteq \mathcal{Y}$ with*

$$R_{\mathcal{Y}} := \max_{y \in \mathcal{Y}, \bar{y} \in \bar{\mathcal{Y}}} \|y - \bar{y}\|_2 > 0.$$

(iii) *The function $f_y(\cdot) := f(\cdot, y)$ is convex on $\bar{\mathcal{X}}$.*

(iv) *The function $(-f)_x(\cdot) := -f(x, \cdot)$ is convex on $\bar{\mathcal{Y}}$.*

Assumption 2. *There exists a stochastic oracle $\mathcal{O}_f$ such that given any $(x, y) \in \bar{\mathcal{X}} \times \bar{\mathcal{Y}}$ and $(g^x, g^y) = \mathcal{O}_f(x, y)$, it holds that*

(i) $\mathbb{E}[(g^x, g^y)] \in \partial f_y(x) \times \partial(-f)_x(y)$.

(ii) *There exist finite $G_x, G_y > 0$ independent of (x, y) such that $\mathbb{E}[\|g^x\|_2^2] \leq G_x^2$ and $\mathbb{E}[\|g^y\|_2^2] \leq G_y^2$.*

Assumption 2 is standard in stochastic programming where $f(x, y) := \mathbb{E}_z[F(x, y, z)]$, in which the oracle $\mathcal{O}_f$ can be implemented by sampling z and returning $(g^x, g^y) \in \partial_x F(x, y, z) \times \partial_y(-F(x, y, z))$. The validity of this construction follows from Assumption 1, which ensures the interchange of expectation and subdifferentiation [50, Section 3.2].

We now describe our three proposed algorithms, summarized in Algorithms 1–3. All three share a common subroutine for computing LMO-based directions, which we describe for the primal variable x ; the dual counterpart is symmetric.

Algorithm 1 Nonsmooth LMO-LMO	Algorithm 2 Nonsmooth LMO-PO	Algorithm 3 Nonsmooth PO-LMO
Input: Parameters $\alpha_t, \beta_t, \eta_t, \tau_t > 0$	Input: Parameters $\alpha_t, \gamma_t, \eta_t > 0$	Input: Parameters $\rho_t, \beta_t, \tau_t > 0$
Initialize: $v_1 = x_1 \in \mathcal{X}, \bar{x}_0 = \lambda_1 = 0$ $u_1 = y_1 \in \mathcal{Y}, \bar{y}_0 = \mu_1 = 0$	Initialize: $v_1 = x_1 \in \mathcal{X}, \bar{x}_0 = \lambda_1 = 0$ $y_1 \in \mathcal{Y}, \bar{y}_0 = 0$	Initialize: $x_1 \in \mathcal{X}$ $u_1 = y_1 \in \mathcal{Y}, \mu_0 = 0$
for $t = 1, \dots, T$ do	for $t = 1, \dots, T$ do	for $t = 1, \dots, T$ do
$(g_t^x, g_t^y) = \mathcal{O}_f(v_t, u_t)$	$(g_t^x, g_t^y) = \mathcal{O}_f(v_t, u_t)$	$(g_t^x, g_t^y) = \mathcal{O}_f(x_t, u_t)$
$\bar{x}_t = \frac{t-1}{t} \bar{x}_{t-1} + \frac{1}{t} x_t$	$\bar{x}_t = \frac{t-1}{t} \bar{x}_{t-1} + \frac{1}{t} x_t$	$\bar{x}_t = \frac{t-1}{t} \bar{x}_{t-1} + \frac{1}{t} x_t$
$v_{t+1} = \text{PO}_{\bar{\mathcal{X}}}(v_t - \frac{\lambda_t + g_t^x}{\alpha_t})$	$v_{t+1} = \text{PO}_{\bar{\mathcal{X}}}(v_t - \frac{\lambda_t + g_t^x}{\alpha_t})$	
$x_{t+1} = \text{LMO}_{\mathcal{X}}(-\lambda_t)$	$x_{t+1} = \text{LMO}_{\mathcal{X}}(-\lambda_t)$	$x_{t+1} = \text{PO}_{\mathcal{X}}(x_t - \rho_t g_t^x)$
$\lambda_{t+1} = \eta_{t+1}(\frac{\lambda_t}{\eta_t} + v_{t+1} - x_{t+1})$	$\lambda_{t+1} = \eta_{t+1}(\frac{\lambda_t}{\eta_t} + v_{t+1} - x_{t+1})$	$\bar{y}_t = \frac{t-1}{t} \bar{y}_{t-1} + \frac{1}{t} y_t$
$\bar{y}_t = \frac{t-1}{t} \bar{y}_{t-1} + \frac{1}{t} y_t$	$\bar{y}_t = \frac{t-1}{t} \bar{y}_{t-1} + \frac{1}{t} y_t$	$u_{t+1} = \text{PO}_{\bar{\mathcal{Y}}}(u_t - \frac{\mu_t + g_t^y}{\beta_t})$
$u_{t+1} = \text{PO}_{\bar{\mathcal{Y}}}(u_t - \frac{\mu_t + g_t^y}{\beta_t})$		$y_{t+1} = \text{LMO}_{\mathcal{Y}}(-\mu_t)$
$y_{t+1} = \text{LMO}_{\mathcal{Y}}(-\mu_t)$		$\mu_{t+1} = \tau_{t+1}(\frac{\mu_t}{\tau_t} + u_{t+1} - y_{t+1})$
$\mu_{t+1} = \tau_{t+1}(\frac{\mu_t}{\tau_t} + u_{t+1} - y_{t+1})$	$y_{t+1} = \text{PO}_{\mathcal{Y}}(y_t - \gamma_t g_t^y)$	
end for	end for	end for
Output: $\bar{x}_T, \bar{y}_T$	Output: $\bar{x}_T, \bar{y}_T$	Output: $\bar{x}_T, \bar{y}_T$

The method maintains a feasible point $x_t \in \mathcal{X}$ and an auxiliary point $v_t \in \bar{\mathcal{X}}$, where $\bar{\mathcal{X}}$ is a simple enclosing set of $\mathcal{X}$. This set is chosen so that projection onto $\bar{\mathcal{X}}$ is *computationally cheap*; typical choices include boxes and Euclidean balls. At iteration t , the algorithm uses the current accumulated direction λ_t together with the stochastic subgradient g_t^x to update the auxiliary point v_{t+1} . The next feasible point is then obtained through a linear minimization oracle over the original domain $x_{t+1} = \text{LMO}_{\mathcal{X}}(-\lambda_t)$. Finally, λ_{t+1} is updated from the discrepancy $v_{t+1} - x_{t+1}$, so that λ_t accumulates information about the mismatch between the auxiliary and feasible sequences. Since the algorithm avoids projection onto the potentially complicated feasible set $\mathcal{X}$ and instead relies on an LMO over $\mathcal{X}$, with only a cheap projection onto the simple enclosing set $\bar{\mathcal{X}}$, we refer to it as *projection-free*.

Algorithm 1 applies this procedure to both primal and dual variables. Algorithms 2 and 3 are hybrid variants. That is, Algorithm 2 uses the LMO-based update for x and a standard projected subgradient step for y , while Algorithm 3 does the reverse. All three algorithms output the running average $\bar{x}_T := \frac{1}{T} \sum_{t \in [T]} x_t$ and $\bar{y}_T := \frac{1}{T} \sum_{t \in [T]} y_t$ as the approximate solution, which is standard practice for nonsmooth stochastic methods to ensure convergence of gap functions.

We assess the quality of the output $(\bar{x}_T, \bar{y}_T)$ via the *strong saddle gap*, defined as

$$\text{Gap}(\bar{x}_T, \bar{y}_T) := \mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right].$$

By Jensen's inequality, the *weak saddle gap*, defined as

$$\text{gap}(\bar{x}_T, \bar{y}_T) := \max_{y \in \mathcal{Y}} \mathbb{E}[f(\bar{x}_T, y)] - \min_{x \in \mathcal{X}} \mathbb{E}[f(x, \bar{y}_T)],$$

satisfies

$$\text{gap}(\bar{x}_T, \bar{y}_T) \leq \text{Gap}(\bar{x}_T, \bar{y}_T).$$

Therefore, all subsequent guarantees carry over to the weak saddle gap as well, which is the metric used in [50]. We also note that the strong saddle gap admits a natural generalization to variational inequalities, as discussed in [34].

The following proposition isolates the key recursive inequalities underlying the LMO-based updates, which serve as the common building block in the convergence analyses of Algorithms 1–3.

Proposition 1. *Suppose Assumptions 1 and 2 hold.*

- (i) *Let $\{v_t, x_t, \lambda_t\}_{t \geq 1}$ be the primal sequences generated by Algorithms 1 or 2 with parameters satisfying $\alpha_t \leq \alpha_{t+1}$ and $\eta_t \geq \eta_{t+1}$ for all $t \geq 1$. Define the dual evaluation sequence $\{z_t\}_{t \geq 1} = \{u_t\}_{t \geq 1}$ for Algorithm 1 and $\{z_t\}_{t \geq 1} = \{y_t\}_{t \geq 1}$ for Algorithm 2. Then, for every $T \geq 2$, we have*

$$\begin{aligned} & \mathbb{E} \left[\frac{\eta_{T-1}}{2} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 + \max_{x \in \mathcal{X}} \sum_{t=1}^T (f(v_t, z_t) - f(x, z_t)) \right] \\ & \leq 2 \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 + \frac{G_{\mathcal{X}}^2}{2\alpha_{T-1}} + \sum_{t=1}^{T-1} \frac{G_{\mathcal{X}}^2}{2\alpha_t} + 2G_{\mathcal{X}}R_{\mathcal{X}}T^{1/2}. \end{aligned} \quad (2)$$

- (ii) *Let $\{u_t, y_t, \mu_t\}_{t \geq 1}$ be the dual sequences generated by Algorithms 1 or 3 with parameters satisfying $\beta_t \leq \beta_{t+1}$ and $\tau_t \geq \tau_{t+1}$ for all $t \geq 1$. Define the primal evaluation sequence $\{w_t\}_{t \geq 1} = \{v_t\}_{t \geq 1}$ for Algorithm 1 and $\{w_t\}_{t \geq 1} = \{x_t\}_{t \geq 1}$ for Algorithm 3. Then, for every $T \geq 2$, we have*

$$\begin{aligned} & \mathbb{E} \left[\frac{\tau_{T-1}}{2} \left\| \frac{\mu_T}{\tau_T} \right\|_2^2 + \max_{y \in \mathcal{Y}} \sum_{t=1}^T (f(w_t, y) - f(w_t, u_t)) \right] \\ & \leq 2 \left(\beta_{T-1} + \sum_{t=1}^{T-1} \tau_t \right) R_{\mathcal{Y}}^2 + \frac{G_{\mathcal{Y}}^2}{2\beta_{T-1}} + \sum_{t=1}^{T-1} \frac{G_{\mathcal{Y}}^2}{2\beta_t} + 2G_{\mathcal{Y}}R_{\mathcal{Y}}T^{1/2}. \end{aligned} \quad (3)$$

Although the proposition is stated for both primal and dual updates, inequalities (2) and (3) are *symmetric* in structure. This symmetry is by design. Specifically, strong duality of the minimax problem (1) allows us to treat the primal and dual updates on equal footing, leading to algorithms with an identical structure for both variables. We therefore only provide the primal proof in the appendix, as the dual proof follows by an analogous argument.

The next proposition isolates the key recursive inequalities underlying the PO-based updates, which serve as the common building block in the convergence analyses of Algorithms 2 and 3.

Proposition 2. *Suppose Assumptions 1 and 2 hold.*

- (i) *Let $\{x_t\}_{t \geq 1}$ be the primal sequence generated by Algorithm 3 with parameters satisfying $\rho_t \geq \rho_{t+1}$ for all $t \geq 1$. Then, for every $T \geq 1$,*

$$\mathbb{E} \left[\max_{x \in \mathcal{X}} \sum_{t=1}^T (f(x_t, u_t) - f(x, u_t)) \right] \leq \frac{2R_{\mathcal{X}}^2}{\rho_T} + \sum_{t=1}^T \frac{\rho_t G_{\mathcal{X}}^2}{2} + 2G_{\mathcal{X}} R_{\mathcal{X}} T^{1/2}. \quad (4)$$

- (ii) *Let $\{y_t\}_{t \geq 1}$ be the dual sequences generated by Algorithm 2 with parameters satisfying $\gamma_t \geq \gamma_{t+1}$ for all $t \geq 1$. Then, for every $T \geq 1$,*

$$\mathbb{E} \left[\max_{y \in \mathcal{Y}} \sum_{t=1}^T (f(v_t, y) - f(v_t, y_t)) \right] \leq \frac{2R_{\mathcal{Y}}^2}{\gamma_T} + \sum_{t=1}^T \frac{\gamma_t G_{\mathcal{Y}}^2}{2} + 2G_{\mathcal{Y}} R_{\mathcal{Y}} T^{1/2}. \quad (5)$$

Propositions 1 and 2 provide all the building blocks needed for the convergence analyses of the three algorithms, each of which will follow by combining the appropriate pair of inequalities. As before, the primal and dual inequalities (4) and (5) are symmetric in structure, so we provide only the proof of the primal part in the appendix. To keep the analysis unified, we also prove a slightly more general version of Proposition 2 that accommodates the PO-PO combination.

3 Convergence Guarantees

We now present the convergence guarantees of Algorithms 1–3.

Theorem 1. *Suppose Assumptions 1 and 2 hold.*

- (i) *Let $\{x_t, y_t\}_{t \geq 1}$ be the sequence generated by Algorithm 1 with parameters $\alpha_t = \Theta(t^{1/2})$, $\beta_t = \Theta(t^{1/2})$, $\eta_t = \Theta(t^{-1/2})$ and $\tau_t = \Theta(t^{-1/2})$ for any $t \geq 1$. Then, for every $T \geq 2$,*

$$\text{Gap}(\bar{x}_T, \bar{y}_T) = \mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \leq O(T^{-1/2}).$$

- (ii) *Let $\{x_t, y_t\}_{t \geq 1}$ be the sequence generated by Algorithm 2 with parameters $\alpha_t = \Theta(t^{1/2})$, $\gamma_t = \Theta(t^{-1/2})$ and $\eta_t = \Theta(t^{-1/2})$ for any $t \geq 1$. Then, for every $T \geq 2$,*

$$\text{Gap}(\bar{x}_T, \bar{y}_T) = \mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \leq O(T^{-1/2}).$$

- (iii) *Let $\{x_t, y_t\}_{t \geq 1}$ be the sequence generated by Algorithm 3 with parameters $\rho_t = \Theta(t^{-1/2})$, $\beta_t = \Theta(t^{1/2})$ and $\tau_t = \Theta(t^{-1/2})$ for any $t \geq 1$. Then, for every $T \geq 2$,*

$$\text{Gap}(\bar{x}_T, \bar{y}_T) = \mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \leq O(T^{-1/2}).$$

The proof of Theorem 1, with explicit constants and parameter settings, is provided in the appendix, where it is broken into three separate theorems each handling one oracles combination. To keep the analysis unified, we also include a proof for the PO-PO combination, which yields an anytime implementation of the method of Nemirovski et al. [50] with the strong saddle gap guarantee, as opposed to the weak guarantee established in the original work.

We now position our contributions within the literature on projection-free minimax optimization. Across all three oracle configurations, the same advantages apply: by avoiding dual smoothing and exploiting the natural symmetry of the convex-concave saddle-point structure, we achieve a uniform $\mathcal{O}(\epsilon^{-2})$ oracle complexity, substantially improving on prior work.

In the context of smooth deterministic minimax optimization problems, Boroun et al. [8] proposed the first single-loop LMO-LMO method for nonconvex-concave problems, and their rate easily translates to the convex-concave setting, yielding $\mathcal{O}(\epsilon^{-6})$ complexity. However, their implementation is not anytime, requiring the total number of iterations or the target precision to be fixed in advance. Moreover, the algorithm is only applicable when the feasible set $\mathcal{Y}$ is strongly convex, as the stepsize selection crucially relies on this knowledge. Giang-Tran et al. [18] relax this requirement and propose an anytime implementation achieving the same $\mathcal{O}(\epsilon^{-6})$ rate, which can be further improved to $\mathcal{O}(\epsilon^{-5})$ when $\mathcal{Y}$ is additionally strongly convex. Nonetheless, both approaches are restricted to smooth deterministic problems and rely on a dual smoothing framework, which breaks the inherent symmetry of convex-concave saddle-point problems. As a result, they achieve these inferior rates even in the smooth deterministic setting. In contrast, our algorithm handles the more general nonsmooth stochastic setting, avoids smoothing entirely, and admits a naturally symmetric implementation that, as we will see, leads to the improved rate of $\mathcal{O}(\epsilon^{-2})$.

Both Boroun et al. [8] and Giang-Tran et al. [18] also study the LMO-PO configuration, again restricted to smooth deterministic problems. The former achieves $\mathcal{O}(\epsilon^{-4})$ without an anytime guarantee, while the latter removes this limitation at the same rate. Notably, the LMO-PO configuration outperforms its fully projection-free counterpart, because a single PO-based dual update yields a more accurate estimate of the gradient of the regularized primal function than a single LMO-based update, leading to a tighter control of the smoothing error. Yet despite this advantage, both methods remain hampered by the dual smoothing framework and its symmetry-breaking effect. Our algorithm, which operates in more general setting and avoids smoothing altogether, achieves $\mathcal{O}(\epsilon^{-2})$ regardless.

The PO-LMO configuration was first studied in the context of smooth deterministic minimax optimization by Giang-Tran et al. [18], where an anytime implementation is proposed. Unlike the LMO-PO case, replacing the primal PO update with an LMO update leads to a coarser approximation of the primal gradient, which explains why the rate degrades back to $\mathcal{O}(\epsilon^{-6})$, matching the fully projection-free case. Our algorithm sidesteps this degradation entirely: by operating in the nonsmooth stochastic setting without smoothing and preserving the natural symmetry of the saddle-point structure, it achieves $\mathcal{O}(\epsilon^{-2})$ across all three oracle configurations.

4 Lower Bounds

We now show that the upper bounds in Theorem 1 are minimax optimal in their dependence on T , up to absolute constants. To this end, we introduce the generic LMO-LMO and LMO-PO procedures in Algorithms 4 and 5, respectively. Throughout this section, lower bounds are stated in a worst-case oracle model. Thus, the adversary may choose the feasible sets $\mathcal{X}, \mathcal{Y}$, the initial points x_1, y_1 , the payoff function f , and a valid tie-breaking rule for the LMO oracles. In particular, the adversary initializes x_1 and y_1 at extreme points and uses LMO oracles that always return extreme-point minimizers. For the LMO-PO class, only the primal variable is constrained by LMO calls, while the dual variable may be updated via any arbitrary map of historical first-order and value information, potentially incorporating projections. We omit the PO-LMO case, since the corresponding lower bound follows by swapping the primal and dual roles.

Algorithm 4 A generic LMO-LMO method

Input: $x_1 \in \mathcal{X}, y_1 \in \mathcal{Y}$
for $t = 1, \dots, T$ **do**
 Construct some $(p_t, q_t) \in \mathbb{R}^n \times \mathbb{R}^m$
 $x_{t+1} = \text{LMO}_{\mathcal{X}}(p_t)$
 $y_{t+1} = \text{LMO}_{\mathcal{Y}}(-q_t)$
end for
Output: $\bar{x}_T \in \text{conv}(\{x_1, \dots, x_{T+1}\})$,
 $\bar{y}_T \in \text{conv}(\{y_1, \dots, y_{T+1}\})$

Algorithm 5 A generic LMO-PO method

Input: $x_1 \in \mathcal{X}, y_1 \in \mathcal{Y}$
for $t = 1, \dots, T$ **do**
 Construct some $p_t \in \mathbb{R}^n$
 $x_{t+1} = \text{LMO}_{\mathcal{X}}(p_t)$
 Construct some $v_1, \dots, v_t \in \mathbb{R}^n$ and a map A_t
 $y_{t+1} = A_t(\{\partial f_{y_i}(v_i), \partial(-f)_{v_i}(y_i), f(v_i, y_i)\}_{1 \leq i \leq t})$
end for
Output: $\bar{x}_T \in \text{conv}(\{x_1, \dots, x_{T+1}\})$, $\bar{y}_T = y_{T+1}$

Given convex compact subsets $\mathcal{X} \subseteq \mathbb{R}^n, \mathcal{Y} \subseteq \mathbb{R}^m$ with radius $R_{\mathcal{X}} > 0, R_{\mathcal{Y}} > 0$, we denote by $\mathcal{F}_{G_x, G_y, \|\cdot\|_2}^0(\mathcal{X}, \mathcal{Y})$ the class of all saddle-point problems (1) whose payoff function f is convex-concave and for any $x, x' \in \mathcal{X}, y, y' \in \mathcal{Y}$

$$|f(x, y) - f(x', y)| \leq G_x \|x - x'\|_2 \quad \& \quad |f(x, y) - f(x, y')| \leq G_y \|y - y'\|_2.$$

Theorem 2. Let $G_x > 0, G_y > 0, R_{\mathcal{X}} > 0, R_{\mathcal{Y}} > 0$ be fixed. Then, for any generic LMO-LMO method of the form Algorithm 4 running for $T \geq 1$ iterations, there exist convex compact subsets $\mathcal{X} \subseteq \mathbb{R}^n, \mathcal{Y} \subseteq \mathbb{R}^m$ with radii $R_{\mathcal{X}}$ and $R_{\mathcal{Y}}$, respectively, and a function $f \in \mathcal{F}_{G_x, G_y, \|\cdot\|_2}^0(\mathcal{X}, \mathcal{Y})$ with $T \leq \min\{m, n\}/4 - 1$ such that

$$\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \geq \frac{G_x R_{\mathcal{X}}}{2(T+1)^{1/2}} + \frac{G_y R_{\mathcal{Y}}}{2(T+1)^{1/2}}.$$

Theorem 3. Let $G_x > 0, G_y > 0, R_{\mathcal{X}} > 0, R_{\mathcal{Y}} > 0$ be fixed. Then, for any generic LMO-PO method of the form Algorithm 5 running for $T \geq 1$ iterations, there exist convex compact subsets $\mathcal{X} \subseteq \mathbb{R}^n, \mathcal{Y} \subseteq \mathbb{R}^m$ with radii $R_{\mathcal{X}}$ and $R_{\mathcal{Y}}$, respectively, and a function $f \in \mathcal{F}_{G_x, G_y, \|\cdot\|_2}^0(\mathcal{X}, \mathcal{Y})$ with $T \leq \min\{m, n\}/4 - 1$ such that

$$\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \geq \frac{G_x R_{\mathcal{X}}}{2(T+1)^{1/2}} + \frac{G_y R_{\mathcal{Y}}}{2(T+1)^{1/2}}.$$

The proofs, presented in the appendix, closely follow the construction of Nesterov [56] and Lan [39], adapted to the saddle-point setting.

5 Numerical Results

We evaluate the proposed methods on two saddle-point problems: matrix completion with spectral norm fit and robust multiclass classification with hinge loss. We compare the three LMO/PO variants in Algorithms 1–3 with the fully projected baseline PO-PO, presented in Algorithm 6 in Appendix A. Throughout the experiments, the suffix *Deterministic* means that the oracle returns the full subgradient of the payoff function, while the suffix *Stochastic* means that the oracle returns an unbiased stochastic estimate. Thus, for instance, LMO-LMO (*Deterministic*) refers to Algorithm 1 implemented with the full oracle, whereas LMO-LMO (*Stochastic*) refers to the same method implemented with the stochastic oracle. For LMO-based updates, we use simple enclosing sets $\bar{\mathcal{X}} \supseteq \mathcal{X}$ and $\bar{\mathcal{Y}} \supseteq \mathcal{Y}$ chosen so that projection onto them is inexpensive. The explicit projection formulas for these covering sets, together with all remaining implementation details and parameter choices, are provided in Appendix B. In the following, $\|\cdot\|_2, \|\cdot\|_F, \|\cdot\|_{\text{op}}$ and $\|\cdot\|_*$ denote the Euclidean norm, Frobenius norm, spectral (operator) norm and nuclear norm, respectively.

Matrix Completion with Spectral Norm Fit. We first consider the bilinear saddle-point formulation of the spectral norm fit problem studied by Juditsky et al. [34]

$$\min_{\|X\|_* \leq 1} \max_{\|Y\|_* \leq 1} \text{Tr}[(\mathcal{A}(X) - B)^\top Y], \quad (6)$$

where $X \in \mathbb{R}^{n \times p}, Y \in \mathbb{R}^{m \times q}, B \in \mathbb{R}^{m \times q}$, and $\mathcal{A} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^{m \times q}$ is a linear map. In this experiment, we have

$$\mathcal{X} = \{X \in \mathbb{R}^{n \times p} : \|X\|_* \leq 1\} \quad \text{and} \quad \mathcal{Y} = \{Y \in \mathbb{R}^{m \times q} : \|Y\|_* \leq 1\}.$$

For the LMO-based variants, we choose the simple covering sets

$$\bar{\mathcal{X}} = \{X \in \mathbb{R}^{n \times p} : \|X\|_F \leq 1\} \quad \text{and} \quad \bar{\mathcal{Y}} = \{Y \in \mathbb{R}^{m \times q} : \|Y\|_F \leq 1\}.$$

Projection onto these Frobenius balls is just a *rescaling operation* and is therefore computationally cheap; see Section B for details. For this problem, we report the (strong) saddle gap, which admits a closed-form expression. Let $\mathcal{A}^*$ denote the adjoint of $\mathcal{A}$. Then, for any pair $(\bar{X}_T, \bar{Y}_T)$, the saddle gap

$$\begin{aligned} & \max_{\|Y\|_* \leq 1} \text{Tr}[(\mathcal{A}(\bar{X}_T) - B)^\top Y] - \min_{\|X\|_* \leq 1} \text{Tr}[(\mathcal{A}(X) - B)^\top \bar{Y}_T] \\ &= \|\mathcal{A}(\bar{X}_T) - B\|_{\text{op}} + \|\mathcal{A}^*(\bar{Y}_T)\|_{\text{op}} + \text{Tr}[B^\top \bar{Y}_T]. \end{aligned}$$

Since the payoff in (6) is smooth, we also compare against the smoothing-based projection-free methods of Giang-Tran et al. [18], the smoothing-based projection method of Xu et al. [69], and the classical extra-gradient and optimistic gradient descent-ascent methods. These additional comparisons are reported in Appendix B.

Robust Multiclass Classification with Hinge Loss. We next consider a robust multiclass classification problem. Given a training set $\mathcal{D}^{\text{tr}} = \{(a_i, b_i)\}_{i=1}^n$, with features $a_i \in \mathbb{R}^d$ and labels $b_i \in \{1, \dots, k\}$, the goal is to learn a predictor matrix $\Theta = [\theta_1, \dots, \theta_k]^\top \in \mathbb{R}^{k \times d}$. The prediction for a new point $\hat{a}$ is $\hat{b} = \arg \max_{j \in [k]} \theta_j^\top \hat{a}$. We use the multiclass hinge loss $\ell_i(\Theta) = \max_{j \in [k]} \{1(j \neq b_i) + \theta_j^\top a_i - \theta_{b_i}^\top a_i\}$. To promote low-rank structure, we impose the nuclear norm constraint $\mathcal{X} = \{\Theta \in \mathbb{R}^{k \times d} : \|\Theta\|_* \leq r\}$. Following the penalized distributionally robust optimization formulation in [6, 20, 37], we aim to solve

$$\min_{\Theta \in \mathcal{X}} \max_{y \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^n y_i \ell_i(\Theta) - \lambda D_\phi(y, \frac{1}{n} \mathbf{1}_n), \quad (7)$$

where $\mathcal{Y} = \{y \in \mathbb{R}^n : y \geq 0, \sum_{i=1}^n y_i = 1\}$ is the probability simplex. We use the Pearson χ^2 -divergence $D_\phi(y, \frac{1}{n} \mathbf{1}_n) := \|ny - \mathbf{1}_n\|_2^2$, which leads to a form of variance regularization [16, 20, 38]. For the LMO-based update, we choose

$$\bar{\mathcal{X}} = \{\Theta \in \mathbb{R}^{k \times d} : \|\Theta\|_F \leq r\} \quad \text{and} \quad \bar{\mathcal{Y}} = \{y \in \mathbb{R}^n : y \geq 0, \|y\|_2 \leq 1\}.$$

Projection onto $\bar{\mathcal{X}}$ is a Frobenius-norm rescaling, while projection onto $\bar{\mathcal{Y}}$ is obtained by nonnegative clipping followed by rescaling when necessary; see Section B for more details. For this experiment, we report the primal objective value as computing the full saddle gap is expensive at this scale, requiring solving both primal and dual problems. The primal objective is a tractable proxy and empirically reflects the same convergence behavior as the saddle-gap metric.

Summary of Results. Figure 1 reports the convergence of the LMO and PO variants in deterministic and stochastic settings. In the deterministic experiments, methods that avoid nuclear-norm projections perform best: LMO-LMO (Deterministic) achieves the fastest decrease and the best final value, with LMO-PO (Deterministic) typically close behind. This is consistent with the fact that an LMO over the nuclear-norm ball only requires a leading singular vector computation, whereas projection requires a substantially more expensive singular value decomposition. The gap is most visible in the classification experiment, where the primal nuclear-norm projection is the main computational bottleneck. The stochastic experiments show a similar pattern, with additional variability due to sampling. LMO-LMO (Stochastic) performs best on matrix completion, while LMO-PO (Stochastic) is especially competitive on classification. For the classification task, LMO and PO updates have comparable computational cost over $\mathcal{Y}$. When dual projections are feasible, they can therefore provide additional stability without introducing a significant computational overhead, which helps explain the strong performance of LMO-PO (Stochastic) in terms of the primal objective. By contrast, using a projection on the primal nuclear-norm constraint remains costly and leads to slower progress in wall-clock time. Overall, the results support the main message of the paper: replacing costly projections with LMOs can yield clear computational gains without sacrificing the optimal stochastic rate.

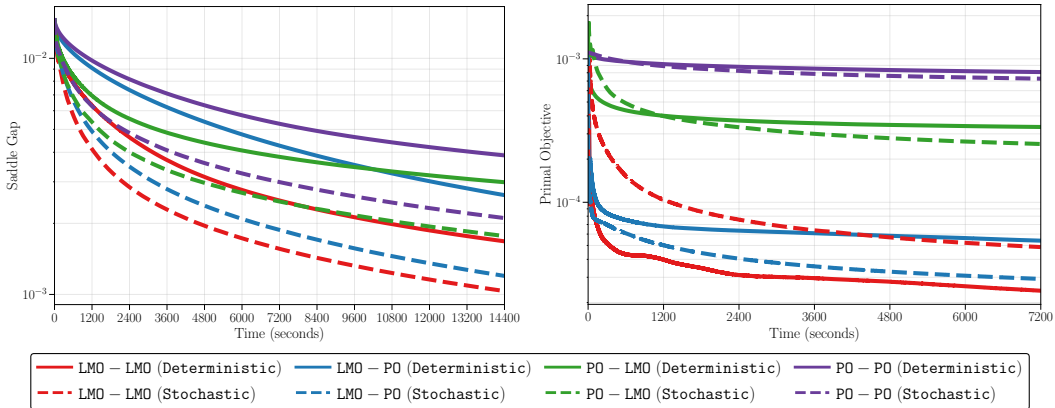


Figure 1: Convergence comparison of deterministic and stochastic LMO/PO variants.

References

- [1] A. Agarwal, P. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.
- [2] A. Alacaoglu, A. Böhm, and Y. Malitsky. Beyond the golden ratio for variational inequality algorithms. *Journal of Machine Learning Research*, 24(172):1–33, 2023.
- [3] K. Asgari and M. J. Neely. Projection-free non-smooth convex programming. *arXiv:2208.05127*, 2022.
- [4] K. Asgari and M. J. Neely. Nonsmooth projection-free optimization with functional constraints. *Computational Optimization and Applications*, 89(3):927–975, 2024.
- [5] R. R. Baghbadorani, P. M. Esfahani, and S. Grammatico. A Frank-Wolfe algorithm for strongly monotone variational inequalities. *Operations Research Letters*, page 107388, 2025.
- [6] A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.
- [7] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.
- [8] M. Boroun, E. Yazdandoost Hamedani, and A. Jalilzadeh. Projection-free methods for solving nonconvex-concave saddle point problems. In *Advances in Neural Information Processing Systems*, pages 53844–53856, 2023.
- [9] G. Braun, A. Carderera, C. W. Combettes, H. Hassani, A. Karbasi, A. Mokhtari, and S. Pokutta. *Conditional Gradient Methods: From Core Principles to AI Applications*. SIAM, 2025.
- [10] 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.
- [11] C. Chen, L. Luo, W. Zhang, and Y. Yu. Efficient projection-free algorithms for saddle point problems. In *Advances in Neural Information Processing Systems*, pages 10799–10808, 2020.
- [12] Z. Chen and E. Mazumdar. Last-iterate convergence for generalized Frank-Wolfe in monotone variational inequalities. In *Advances in Neural Information Processing Systems*, pages 115440–115467, 2024.
- [13] E. Cheung and Y. Li. Nonsmooth Frank-Wolfe using uniform affine approximations. *arXiv:1710.05776*, 2017.
- [14] B. Cox, A. Juditsky, and A. Nemirovski. Dual subgradient algorithms for large-scale nonsmooth learning problems. *Mathematical Programming*, 148(1):143–180, 2014.
- [15] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.
- [16] J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.
- [17] D. J. Foster, A. Sekhari, O. Shamir, N. Srebro, K. Sridharan, and B. Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*, pages 1319–1345. PMLR, 2019.
- [18] K.-H. Giang-Tran, S. Shafiee, and N. Ho-Nguyen. Projection-free algorithms for minimax problems. *arXiv:2603.29870*, 2026.
- [19] G. Gidel, T. Jebara, and S. Lacoste-Julien. Frank-Wolfe algorithms for saddle point problems. In *International Conference on Artificial Intelligence and Statistics*, pages 362–371, 2017.
- [20] J.-y. Gotoh, M. J. Kim, and A. E. Lim. Robust empirical optimization is almost the same as mean–variance optimization. *Operations Research Letters*, 46(4):448–452, 2018.

- [21] B. Grimmer. Radial duality part I: Foundations. *Mathematical Programming*, 205(1):33–68, 2024.
- [22] B. Grimmer. Radial duality part II: Applications and algorithms. *Mathematical Programming*, 205(1):69–105, 2024.
- [23] B. Grimmer and N. Liu. Lower bounds for linear minimization oracle methods optimizing over strongly convex sets. *arXiv:2602.22608*, 2026.
- [24] C. Guzmán and A. Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015.
- [25] J. Halbey, D. Deza, M. Zimmer, C. Roux, B. Stellato, and S. Pokutta. Lower bounds for Frank-Wolfe on strongly convex sets. *arXiv:2602.04378*, 2026.
- [26] J. H. Hammond. *Solving Asymmetric Variational Inequality Problems and Systems of Equations with Generalized Nonlinear Programming Algorithms*. PhD thesis, Massachusetts Institute of Technology, 1984.
- [27] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1):75–112, 2015.
- [28] E. Hazan and S. Kale. Projection-free online learning. In *International Conference on Machine Learning*, pages 1843–1850, 2012.
- [29] E. Hazan and E. Minasyan. Faster projection-free online learning. In *Conference on Learning Theory*, pages 1877–1893, 2020.
- [30] N. He and Z. Harchaoui. Semi-proximal mirror-prox for nonsmooth composite minimization. In *Advances in Neural Information Processing Systems*, pages 3411–3419, 2015.
- [31] M. Hough. Asymptotic convergence of the Frank-Wolfe algorithm for monotone variational inequalities. *arXiv:2603.12104*, 2026.
- [32] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435, 2013.
- [33] 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.
- [34] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.
- [35] V. Kolmogorov and T. Pock. One-sided Frank-Wolfe algorithms for saddle problems. In *International Conference on Machine Learning*, pages 5665–5675, 2021.
- [36] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.
- [37] D. Kuhn, S. Shafiee, and W. Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.
- [38] H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.
- [39] G. Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv:1309.5550*, 2013.
- [40] G. Lan. Gradient sliding for composite optimization. *Mathematical Programming*, 159(1):201–235, 2016.
- [41] G. Lan, E. Romeijn, and Z. Zhou. Conditional gradient methods for convex optimization with general affine and nonlinear constraints. *SIAM Journal on Optimization*, 31(3):2307–2339, 2021.

- [42] G. Lan and Y. Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.
- [43] F. Locatello, A. Yurtsever, O. Fercoq, and V. Cevher. Stochastic Frank-Wolfe for composite convex minimization. In *Advances in Neural Information Processing Systems*, pages 14291–14301, 2019.
- [44] Z. Lu, N. Brukhim, P. Gradu, and E. Hazan. Projection-free adaptive regret with membership oracles. In *Conference on Algorithmic Learning Theory*, pages 1055–1073, 2023.
- [45] Y. Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.
- [46] Y. Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184(1):383–410, 2020.
- [47] Y. Malitsky and M. K. Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.
- [48] A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.
- [49] 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.
- [50] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- [51] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.
- [52] Y. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl akad nauk Sssr*, 27(2):372–376, 1983.
- [53] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.
- [54] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.
- [55] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.
- [56] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2013.
- [57] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14934–14942, 2019.
- [58] Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.
- [59] L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.
- [60] S. N. Ravi, M. D. Collins, and V. Singh. A deterministic nonsmooth Frank Wolfe algorithm with coresets guarantees. *INFORMS Journal on Optimization*, 1(2):120–142, 2019.
- [61] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.
- [62] N. Z. Shor. *Minimization Methods for Non-Differentiable Functions*. Springer, 1985.

- [63] A. Suggala and P. Netrapalli. Follow the perturbed leader: Optimism and fast parallel algorithms for smooth minimax games. In *Advances in Neural Information Processing Systems*, pages 22316–22326, 2020.
- [64] W. Tao, Z. Pan, G. Wu, and Q. Tao. The strength of Nesterov’s extrapolation in the individual convergence of nonsmooth optimization. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2557–2568, 2019.
- [65] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Optimal nonsmooth Frank-Wolfe method for stochastic regret minimization. In *Optimization for Machine Learning Workshop*, 2020.
- [66] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Projection efficient subgradient method and optimal nonsmooth Frank-Wolfe method. In *Advances in Neural Information Processing Systems*, pages 12211–12224, 2020.
- [67] P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.
- [68] D. J. White. Extension of the Frank-Wolfe algorithm to concave nondifferentiable objective functions. *Journal of Optimization Theory and Applications*, 78(2):283–301, 1993.
- [69] Z. Xu, H. Zhang, Y. Xu, and G. Lan. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, 201(1):635–706, 2023.
- [70] A. Yurtsever, O. Fercoq, and V. Cevher. A conditional-gradient-based augmented Lagrangian framework. In *International Conference on Machine Learning*, pages 7272–7281, 2019.

A Proofs

A.1 Proofs of Proposition 1

For the convenience, we repeat the statement of Proposition 1 below.

Proposition A.1 (Restatement of Proposition 1). *Suppose Assumptions 1 and 2 hold.*

- (i) *Let $\{v_t, x_t, \lambda_t\}_{t \geq 1}$ be the primal sequences generated by Algorithms 1 or 2 with parameters satisfying $\alpha_t \leq \alpha_{t+1}$ and $\eta_t \geq \eta_{t+1}$ for all $t \geq 1$. Define the dual evaluation sequence $\{z_t\}_{t \geq 1} = \{u_t\}_{t \geq 1}$ for Algorithm 1 and $\{z_t\}_{t \geq 1} = \{y_t\}_{t \geq 1}$ for Algorithm 2. Then, for every $T \geq 2$, we have*

$$\begin{aligned} & \mathbb{E} \left[\frac{\eta_{T-1}}{2} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 + \max_{x \in \mathcal{X}} \sum_{t=1}^T (f(v_t, z_t) - f(x, z_t)) \right] \\ & \leq 2 \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 + \frac{G_{\mathbf{x}}^2}{2\alpha_{T-1}} + \sum_{t=1}^{T-1} \frac{G_{\mathbf{x}}^2}{2\alpha_t} + 2G_{\mathbf{x}}R_{\mathcal{X}}T^{1/2}. \end{aligned} \quad (\text{A.1})$$

- (ii) *Let $\{u_t, y_t, \mu_t\}_{t \geq 1}$ be the dual sequences generated by Algorithms 1 or 3 with parameters satisfying $\beta_t \leq \beta_{t+1}$ and $\tau_t \geq \tau_{t+1}$ for all $t \geq 1$. Define the primal evaluation sequence $\{w_t\}_{t \geq 1} = \{v_t\}_{t \geq 1}$ for Algorithm 1 and $\{w_t\}_{t \geq 1} = \{x_t\}_{t \geq 1}$ for Algorithm 3. Then, for every $T \geq 2$, we have*

$$\begin{aligned} & \mathbb{E} \left[\frac{\tau_{T-1}}{2} \left\| \frac{\mu_T}{\tau_T} \right\|_2^2 + \max_{y \in \mathcal{Y}} \sum_{t=1}^T (f(w_t, y) - f(w_t, u_t)) \right] \\ & \leq 2 \left(\beta_{T-1} + \sum_{t=1}^{T-1} \tau_t \right) R_{\mathcal{Y}}^2 + \frac{G_{\mathbf{y}}^2}{2\beta_{T-1}} + \sum_{t=1}^{T-1} \frac{G_{\mathbf{y}}^2}{2\beta_t} + 2G_{\mathbf{y}}R_{\mathcal{Y}}T^{1/2}. \end{aligned} \quad (\text{A.2})$$

The proof uses the following standard Euclidean three-point inequality; see also [55, Lemma 6].

Lemma A.1. *Let h be a convex function defined on a nonempty convex set $\mathcal{C} \subseteq \mathbb{R}^d$. Given $\alpha > 0$ and $\bar{x} \in \mathcal{C}$, let*

$$x' = \arg \min_{x \in \mathcal{C}} \{h(x) + \alpha \|x - \bar{x}\|_2^2\}.$$

Then, for any $x \in \mathcal{C}$,

$$h(x') + \alpha \|x' - \bar{x}\|_2^2 \leq h(x) + \alpha \|x - \bar{x}\|_2^2 - \alpha \|x - x'\|_2^2.$$

Proof of Lemma A.1. We provide a proof for completeness. Define the function

$$\phi(x) := h(x) + \alpha \|x - \bar{x}\|_2^2, \quad x \in \mathcal{C}.$$

By 2α -strong convexity of ϕ on $\mathcal{C}$, for every $x \in \mathcal{C}$ and every $\theta \in [0, 1]$, we have

$$\phi((1-\theta)x' + \theta x) \leq (1-\theta)\phi(x') + \theta\phi(x) - \alpha\theta(1-\theta)\|x - x'\|_2^2.$$

Since x' minimizes ϕ over $\mathcal{C}$ and $(1-\theta)x' + \theta x \in \mathcal{C}$, we also have

$$\phi(x') \leq \phi((1-\theta)x' + \theta x).$$

Combining the last two displays yields

$$\phi(x') \leq (1-\theta)\phi(x') + \theta\phi(x) - \alpha\theta(1-\theta)\|x - x'\|_2^2,$$

and therefore, for every $\theta \in (0, 1]$, $\phi(x') \leq \phi(x) - \alpha(1-\theta)\|x - x'\|_2^2$. Letting $\theta \downarrow 0$ gives

$$\phi(x') \leq \phi(x) - \alpha\|x - x'\|_2^2.$$

Expanding the definition of ϕ completes the proof. $\square$

We are now ready to present the proof.

Proof of Proposition A.1. By symmetry, it suffices to prove part (i). The proof of part (ii) follows by exchanging

$$(v_t, x_t, \lambda_t, \alpha_t, \eta_t, \mathcal{X}, R_{\mathcal{X}}, G_{\mathcal{X}}, z_t) \quad \text{with} \quad (u_t, y_t, \mu_t, \beta_t, \tau_t, \mathcal{Y}, R_{\mathcal{Y}}, G_{\mathcal{Y}}, w_t),$$

and replacing $f_{z_t}(\cdot)$ by $(-f)_{w_t}(\cdot)$. We therefore focus on the primal inequality.

Fix any $x \in \mathcal{X}$. The update of v_{t+1} can be written as

$$v_{t+1} = \arg \min_{v \in \bar{\mathcal{Y}}} \left\{ (\lambda_t + g_t^x)^\top v + \frac{\alpha_t}{2} \|v - v_t\|_2^2 \right\}.$$

Since $x \in \mathcal{X} \subseteq \bar{\mathcal{X}}$, invoking Lemma A.1 with $h(v) := (\lambda_t + g_t^x)^\top v$, $\mathcal{C} := \bar{\mathcal{Y}}$, $\bar{x} := v_t$ and $\alpha := \alpha_t/2$ gives

$$(\lambda_t + g_t^x)^\top v_{t+1} + \frac{\alpha_t}{2} \|v_{t+1} - v_t\|_2^2 \leq (\lambda_t + g_t^x)^\top x + \frac{\alpha_t}{2} \|x - v_t\|_2^2 - \frac{\alpha_t}{2} \|x - v_{t+1}\|_2^2.$$

Subtracting $\lambda_t^\top x_{t+1}$ and $(g_t^x)^\top v_t$ from both sides yields

$$\begin{aligned} & \lambda_t^\top (v_{t+1} - x_{t+1}) + (g_t^x)^\top (v_{t+1} - v_t) + \frac{\alpha_t}{2} \|v_{t+1} - v_t\|_2^2 \\ & \leq \lambda_t^\top (x - x_{t+1}) + (g_t^x)^\top (x - v_t) + \frac{\alpha_t}{2} \|x - v_t\|_2^2 - \frac{\alpha_t}{2} \|x - v_{t+1}\|_2^2. \end{aligned}$$

Since $x_{t+1} = \text{LMO}_{\mathcal{X}}(-\lambda_t)$, we have $\lambda_t^\top (x - x_{t+1}) \leq 0$. Hence,

$$\begin{aligned} & \lambda_t^\top (v_{t+1} - x_{t+1}) + (g_t^x)^\top (v_{t+1} - v_t) + \frac{\alpha_t}{2} \|v_{t+1} - v_t\|_2^2 \\ & \leq (g_t^x)^\top (x - v_t) + \frac{\alpha_t}{2} \|x - v_t\|_2^2 - \frac{\alpha_t}{2} \|x - v_{t+1}\|_2^2. \end{aligned}$$

The update of λ_{t+1} implies $\lambda_{t+1}/\eta_{t+1} = \lambda_t/\eta_t + v_{t+1} - x_{t+1}$. Therefore,

$$\lambda_t^\top (v_{t+1} - x_{t+1}) = \eta_t \left(\frac{1}{\eta_t} \lambda_t^\top (v_{t+1} - x_{t+1}) \right) = \frac{\eta_t}{2} \left(\left\| \frac{\lambda_{t+1}}{\eta_{t+1}} \right\|_2^2 - \left\| \frac{\lambda_t}{\eta_t} \right\|_2^2 - \|v_{t+1} - x_{t+1}\|_2^2 \right).$$

Substituting this identity into the previous inequality gives

$$\begin{aligned} & \frac{\eta_t}{2} \left(\left\| \frac{\lambda_{t+1}}{\eta_{t+1}} \right\|_2^2 - \left\| \frac{\lambda_t}{\eta_t} \right\|_2^2 - \|v_{t+1} - x_{t+1}\|_2^2 \right) + (g_t^x)^\top (v_{t+1} - v_t) + \frac{\alpha_t}{2} \|v_{t+1} - v_t\|_2^2 \\ & \leq (g_t^x)^\top (x - v_t) + \frac{\alpha_t}{2} \|x - v_t\|_2^2 - \frac{\alpha_t}{2} \|x - v_{t+1}\|_2^2. \end{aligned}$$

By Young's inequality,

$$(g_t^x)^\top (v_{t+1} - v_t) + \frac{\alpha_t}{2} \|v_{t+1} - v_t\|_2^2 \geq -\frac{\|g_t^x\|_2^2}{2\alpha_t}.$$

Also, since $v_{t+1} \in \bar{\mathcal{X}}$ and $x_{t+1} \in \mathcal{X}$, we use the crude bound $\|v_{t+1} - x_{t+1}\|_2 \leq 2R_{\mathcal{X}}$. Thus,

$$\frac{\eta_t}{2} \left(\left\| \frac{\lambda_{t+1}}{\eta_{t+1}} \right\|_2^2 - \left\| \frac{\lambda_t}{\eta_t} \right\|_2^2 \right) - \frac{\|g_t^x\|_2^2}{2\alpha_t} \leq (g_t^x)^\top (x - v_t) + \frac{\alpha_t}{2} \|x - v_t\|_2^2 - \frac{\alpha_t}{2} \|x - v_{t+1}\|_2^2 + 2\eta_t R_{\mathcal{X}}^2.$$

Summing from $t = 1$ to $T - 1$ gives

$$\begin{aligned} & \frac{\eta_{T-1}}{2} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 + \frac{1}{2} \sum_{t=2}^{T-1} (\eta_{t-1} - \eta_t) \left\| \frac{\lambda_t}{\eta_t} \right\|_2^2 - \frac{\eta_1}{2} \left\| \frac{\lambda_1}{\eta_1} \right\|_2^2 - \sum_{t=1}^{T-1} \frac{\|g_t^x\|_2^2}{2\alpha_t} \\ & \leq \sum_{t=1}^{T-1} (g_t^x)^\top (x - v_t) + \frac{\alpha_1}{2} \|x - v_1\|_2^2 + \frac{1}{2} \sum_{t=2}^{T-1} (\alpha_t - \alpha_{t-1}) \|x - v_t\|_2^2 - \frac{\alpha_{T-1}}{2} \|x - v_T\|_2^2 + 2 \sum_{t=1}^{T-1} \eta_t R_{\mathcal{X}}^2. \end{aligned}$$

Since $\lambda_1 = 0$, $\eta_t \geq \eta_{t+1}$, $\alpha_t \leq \alpha_{t+1}$, and $\|x - v_t\|_2 \leq 2R_{\mathcal{X}}$ for all $x \in \mathcal{X}$ and $v_t \in \bar{\mathcal{X}}$, we obtain

$$\frac{\eta_{T-1}}{2} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 + \sum_{t=1}^{T-1} (g_t^x)^\top (v_t - x) \leq 2 \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 - \frac{\alpha_{T-1}}{2} \|x - v_T\|_2^2 + \sum_{t=1}^{T-1} \frac{\|g_t^x\|_2^2}{2\alpha_t}.$$

Adding $(g_T^x)^\top (v_T - x)$ to both sides and applying Young's inequality once more yields

$$\begin{aligned} & \frac{\eta_{T-1}}{2} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 + \sum_{t=1}^T (g_t^x)^\top (v_t - x) \\ & \leq 2 \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 + (g_T^x)^\top (v_T - x) - \frac{\alpha_{T-1}}{2} \|x - v_T\|_2^2 + \sum_{t=1}^{T-1} \frac{\|g_t^x\|_2^2}{2\alpha_t} \\ & \leq 2 \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 + \frac{\|g_T^x\|_2^2}{2\alpha_{T-1}} + \sum_{t=1}^{T-1} \frac{\|g_t^x\|_2^2}{2\alpha_t}. \end{aligned}$$

Let $\mathcal{F}_t$ be the sigma algebra generated by all randomness up to the beginning of iteration t , and define $\bar{g}_t^x := \mathbb{E}[g_t^x \mid \mathcal{F}_t]$ and $\xi_t := \bar{g}_t^x - g_t^x$. By Assumption 2, $\bar{g}_t^x \in \partial f_{z_t}(v_t)$. Adding $\sum_{t=1}^T \xi_t^\top (v_t - x)$ to both sides of the last display and using convexity of $f_{z_t}(\cdot)$, we get

$$\begin{aligned} & \frac{\eta_{T-1}}{2} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 + \sum_{t=1}^T (f(v_t, z_t) - f(x, z_t)) \\ & \leq 2 \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 + \frac{\|g_T^x\|_2^2}{2\alpha_{T-1}} + \sum_{t=1}^{T-1} \frac{\|g_t^x\|_2^2}{2\alpha_t} + \sum_{t=1}^T \xi_t^\top (v_t - x). \end{aligned}$$

Taking the maximum over $x \in \mathcal{X}$ gives

$$\begin{aligned} & \frac{\eta_{T-1}}{2} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 + \max_{x \in \mathcal{X}} \sum_{t=1}^T (f(v_t, z_t) - f(x, z_t)) \\ & \leq 2 \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 + \frac{\|g_T^x\|_2^2}{2\alpha_{T-1}} + \sum_{t=1}^{T-1} \frac{\|g_t^x\|_2^2}{2\alpha_t} + \max_{x \in \mathcal{X}} \sum_{t=1}^T \xi_t^\top (v_t - x). \end{aligned}$$

It remains to control the martingale term. Since $x_1 \in \mathcal{X}$, for every $x \in \mathcal{X}$,

$$\sum_{t=1}^T \xi_t^\top (v_t - x) = \sum_{t=1}^T \xi_t^\top (v_t - x_1) + \sum_{t=1}^T \xi_t^\top (x_1 - x) \leq \sum_{t=1}^T \xi_t^\top (v_t - x_1) + 2R_{\mathcal{X}} \left\| \sum_{t=1}^T \xi_t \right\|_2.$$

Because $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$ and $v_t - x_1$ is $\mathcal{F}_t$ -measurable, $\mathbb{E} \left[\sum_{t=1}^T \xi_t^\top (v_t - x_1) \right] = 0$. Moreover,

$$\mathbb{E} \left[\left\| \sum_{t=1}^T \xi_t \right\|_2 \right] \leq \left(\mathbb{E} \left[\left\| \sum_{t=1}^T \xi_t \right\|_2^2 \right] \right)^{1/2} = \left(\sum_{t=1}^T \mathbb{E} \|\xi_t\|_2^2 \right)^{1/2} \leq G_x T^{1/2},$$

where the last inequality follows from $\mathbb{E} \|\xi_t\|_2^2 \leq \mathbb{E} \|g_t^x\|_2^2 \leq G_x^2$. Hence,

$$\mathbb{E} \left[\max_{x \in \mathcal{X}} \sum_{t=1}^T \xi_t^\top (v_t - x) \right] \leq 2G_x R_{\mathcal{X}} T^{1/2}. \quad (\text{A.3})$$

Taking expectations and using $\mathbb{E} \|g_t^x\|_2^2 \leq G_x^2$ for all t finally yields (A.1). This proves part (i), and the proof is complete by symmetry. $\square$

A.2 Proofs of Proposition 2

To provide a unified theoretical framework, we extend the scope of Proposition 2 to include the iterative structure of the PO-PO scheme in Algorithm 6.

This algorithm can be viewed as an anytime variant of the stochastic subgradient descent-ascent algorithm introduced by Nemirovski et al. [50]. Crucially, while the original results in [50] are limited to the weak saddle gap guarantees, our generalized proposition allows us to establish the stronger saddle gap guarantee. We note that the stronger saddle gap admits a natural generalization to variational inequalities, as discussed in [34]. We provide the full statement and its proof below for completeness.

Algorithm 6 Nonsmooth PO-PO

Input: $\rho_t, \gamma_t > 0$

1: **Init:** $x_1 \in \mathcal{X}, y_1 \in \mathcal{Y}$

2: **for** $t = 1, \dots, T$ **do**

3: $(g_t^x, g_t^y) = \mathcal{O}_f(x_t, y_t)$

4: $\bar{x}_t = \frac{(t-1)\bar{x}_{t-1} + x_t}{t}$

5: $x_{t+1} = \text{PO}_{\mathcal{X}}(x_t - \rho_t g_t^x)$

6: $\bar{y}_t = \frac{(t-1)\bar{y}_{t-1} + y_t}{t}$

7: $y_{t+1} = \text{PO}_{\mathcal{Y}}(y_t - \gamma_t g_t^y)$

8: **end for**

Output: $\bar{x}_T, \bar{y}_T$

Proposition A.2 (Generalization of Proposition 2). *Suppose Assumptions 1 and 2 hold.*

- (i) *Let $\{x_t\}_{t \geq 1}$ be the primal sequence generated by Algorithm 3 or 6 with parameters satisfying $\rho_t \geq \rho_{t+1}$ for all $t \geq 1$. Define the dual evaluation sequence $\{z_t\}_{t \geq 1} = \{u_t\}_{t \geq 1}$ for Algorithm 3 and $\{z_t\}_{t \geq 1} = \{y_t\}_{t \geq 1}$ for Algorithm 6. Then, for every $T \geq 1$,*

$$\mathbb{E} \left[\max_{x \in \mathcal{X}} \sum_{t=1}^T (f(x_t, z_t) - f(x, z_t)) \right] \leq \frac{2R_{\mathcal{X}}^2}{\rho_T} + \sum_{t=1}^T \frac{\rho_t G_x^2}{2} + 2G_x R_{\mathcal{X}} T^{1/2}. \quad (\text{A.4})$$

- (ii) *Let $\{y_t\}_{t \geq 1}$ be the dual sequences generated by Algorithm 2 or 6 with parameters satisfying $\gamma_t \geq \gamma_{t+1}$ for all $t \geq 1$. Define the primal evaluation sequence $\{w_t\}_{t \geq 1} = \{v_t\}_{t \geq 1}$ for Algorithm 2 and $\{w_t\}_{t \geq 1} = \{x_t\}_{t \geq 1}$ for Algorithm 6. Then, for every $T \geq 1$,*

$$\mathbb{E} \left[\max_{y \in \mathcal{Y}} \sum_{t=1}^T (f(w_t, y) - f(w_t, y_t)) \right] \leq \frac{2R_{\mathcal{Y}}^2}{\gamma_T} + \sum_{t=1}^T \frac{\gamma_t G_y^2}{2} + 2G_y R_{\mathcal{Y}} T^{1/2}. \quad (\text{A.5})$$

Proof of Proposition A.2. By symmetry, it suffices to prove part (i). The proof of part (ii) follows from the same argument after exchanging

$$(x_t, g_t^x, \rho_t, \mathcal{X}, R_{\mathcal{X}}, G_x, z_t) \quad \text{with} \quad (y_t, g_t^y, \gamma_t, \mathcal{Y}, R_{\mathcal{Y}}, G_y, w_t),$$

and replacing $f_{z_t}(\cdot)$ by $(-f)_{w_t}(\cdot)$. We therefore prove only the primal inequality.

Fix any $x \in \mathcal{X}$. The projected update gives

$$x_{t+1} = \text{PO}_{\mathcal{X}}(x_t - \rho_t g_t^x), \quad t = 1, \dots, T.$$

By nonexpansiveness of the Euclidean projection,

$$\|x_{t+1} - x\|_2^2 \leq \|x_t - \rho_t g_t^x - x\|_2^2 = \|x_t - x\|_2^2 - 2\rho_t (g_t^x)^\top (x_t - x) + \rho_t^2 \|g_t^x\|_2^2.$$

Rearranging yields

$$(g_t^x)^\top (x_t - x) \leq \frac{\|x_t - x\|_2^2}{2\rho_t} - \frac{\|x_{t+1} - x\|_2^2}{2\rho_t} + \frac{\rho_t \|g_t^x\|_2^2}{2}.$$

Summing this inequality from $t = 1$ to T gives

$$\sum_{t=1}^T (g_t^x)^\top (x_t - x) \leq \frac{\|x_1 - x\|_2^2}{2\rho_1} + \frac{1}{2} \sum_{t=2}^T \left(\frac{1}{\rho_t} - \frac{1}{\rho_{t-1}} \right) \|x_t - x\|_2^2 - \frac{\|x_{T+1} - x\|_2^2}{2\rho_T} + \sum_{t=1}^T \frac{\rho_t \|g_t^x\|_2^2}{2}.$$

Since $\rho_t \geq \rho_{t+1}$ and $\|x_t - x\|_2 \leq 2R_{\mathcal{X}}$ for all t , we may drop the final negative term and obtain

$$\sum_{t=1}^T (g_t^x)^\top (x_t - x) \leq \frac{2R_{\mathcal{X}}^2}{\rho_1} + 2R_{\mathcal{X}}^2 \sum_{t=2}^T \left(\frac{1}{\rho_t} - \frac{1}{\rho_{t-1}} \right) + \sum_{t=1}^T \frac{\rho_t \|g_t^x\|_2^2}{2} = \frac{2R_{\mathcal{X}}^2}{\rho_T} + \sum_{t=1}^T \frac{\rho_t \|g_t^x\|_2^2}{2}.$$

Let $\mathcal{F}_t$ denote the sigma algebra generated by all randomness up to the beginning of iteration t , so that x_t and z_t are $\mathcal{F}_t$ -measurable. Define $\bar{g}_t^x := \mathbb{E}[g_t^x \mid \mathcal{F}_t]$ and $\xi_t := \bar{g}_t^x - g_t^x$. By Assumption 2, $\bar{g}_t^x \in \partial f_{z_t}(x_t)$. Adding $\sum_{t=1}^T \xi_t^\top (x_t - x)$ to both sides of the last display and using convexity of $f_{z_t}(\cdot)$ gives, for every $x \in \mathcal{X}$,

$$\sum_{t=1}^T (f(x_t, z_t) - f(x, z_t)) \leq \sum_{t=1}^T (\bar{g}_t^x)^\top (x_t - x) \leq \frac{2R_{\mathcal{X}}^2}{\rho_T} + \sum_{t=1}^T \frac{\rho_t \|g_t^x\|_2^2}{2} + \sum_{t=1}^T \xi_t^\top (x_t - x).$$

Taking the maximum over $x \in \mathcal{X}$ yields

$$\max_{x \in \mathcal{X}} \sum_{t=1}^T (f(x_t, z_t) - f(x, z_t)) \leq \frac{2R_{\mathcal{X}}^2}{\rho_T} + \sum_{t=1}^T \frac{\rho_t \|g_t^x\|_2^2}{2} + \max_{x \in \mathcal{X}} \sum_{t=1}^T \xi_t^\top (x_t - x).$$

The martingale term is controlled by the same argument used to derive (A.3). Namely, replacing v_t by x_t gives

$$\mathbb{E} \left[\max_{x \in \mathcal{X}} \sum_{t=1}^T \xi_t^\top (x_t - x) \right] \leq 2G_x R_{\mathcal{X}} T^{1/2}.$$

Taking expectations and using $\mathbb{E}\|g_t^x\|_2^2 \leq G_x^2$ for all t gives (A.4). This proves part (i), and the proof is complete by symmetry. $\square$

A.3 Proofs of Theorem 1 (i)

We first present the detailed version of Theorem 1 (i), including explicit constants and parameter choices.

Theorem A.1 (Detailed version of Theorem 1 (i)). *Suppose Assumptions 1 and 2 hold. Let $\{x_t, y_t\}_{t \geq 1}$ be the primal-dual sequence generated by Algorithm 1 with parameters*

$$\alpha_t = C_1(t+1)^{1/2}, \quad \beta_t = C_2(t+1)^{1/2}, \quad \eta_t = C_3(t+1)^{-1/2}, \quad \tau_t = C_4(t+1)^{-1/2}$$

for some $C_1, C_2, C_3, C_4 > 0$. Then, for every $T \geq 2$,

$$\mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \leq \frac{2(C_1 + 2C_3)R_{\mathcal{X}}^2 + \frac{3G_{\mathcal{X}}^2}{2C_1} + \frac{G_{\mathcal{X}}^2}{2C_3} + 2G_{\mathcal{X}}R_{\mathcal{X}}}{T^{1/2}} + \frac{2(C_2 + 2C_4)R_{\mathcal{Y}}^2 + \frac{3G_{\mathcal{Y}}^2}{2C_2} + \frac{G_{\mathcal{Y}}^2}{2C_4} + 2G_{\mathcal{Y}}R_{\mathcal{Y}}}{T^{1/2}}.$$

If $C_1 = C_3 = G_{\mathcal{X}}/R_{\mathcal{X}}$ and $C_2 = C_4 = G_{\mathcal{Y}}/R_{\mathcal{Y}}$ then the bound becomes

$$\mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \leq \frac{10(G_{\mathcal{X}}R_{\mathcal{X}} + G_{\mathcal{Y}}R_{\mathcal{Y}})}{T^{1/2}}$$

The proof relies on the following lemma, which bounds the strong saddle gap for arbitrary monotone parameter sequences using Proposition A.1.

Lemma A.2. *Suppose Assumptions 1 and 2 hold. Let $\{x_t, y_t\}_{t \geq 1}$ be the primal-dual sequence generated by Algorithm 1 with parameters satisfying $\alpha_t \leq \alpha_{t+1}$, $\beta_t \leq \beta_{t+1}$, $\eta_t \geq \eta_{t+1}$ and $\tau_t \geq \tau_{t+1}$ for all $t \geq 1$. Then, for every $T \geq 2$,*

$$\begin{aligned} & \mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \\ & \leq \frac{2}{T} \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 + \frac{G_{\mathcal{X}}^2}{2\alpha_{T-1}T} + \frac{1}{T} \sum_{t=1}^{T-1} \frac{G_{\mathcal{X}}^2}{2\alpha_t} + \frac{G_{\mathcal{X}}^2}{2T\eta_{T-1}} + \frac{2G_{\mathcal{X}}R_{\mathcal{X}}}{T^{1/2}} \\ & \quad + \frac{2}{T} \left(\beta_{T-1} + \sum_{t=1}^{T-1} \tau_t \right) R_{\mathcal{Y}}^2 + \frac{G_{\mathcal{Y}}^2}{2\beta_{T-1}T} + \frac{1}{T} \sum_{t=1}^{T-1} \frac{G_{\mathcal{Y}}^2}{2\beta_t} + \frac{G_{\mathcal{Y}}^2}{2T\tau_{T-1}} + \frac{2G_{\mathcal{Y}}R_{\mathcal{Y}}}{T^{1/2}}. \end{aligned}$$

Proof of Lemma A.2. Define $\bar{v}_T := \frac{1}{T} \sum_{t=1}^T v_t$ and $\bar{u}_T := \frac{1}{T} \sum_{t=1}^T u_t$. Since $v_1 = x_1$ and $u_1 = y_1$, the LMO recursions imply

$$\begin{cases} \frac{\lambda_T}{\eta_T} = \frac{\lambda_{T-1}}{\eta_{T-1}} + v_T - x_T = \frac{\lambda_0}{\eta_0} + (v_T - x_T) + \dots + (v_0 - x_0) = \sum_{t=1}^T (v_t - x_t) = T(\bar{v}_T - \bar{x}_T), \\ \frac{\mu_T}{\tau_T} = \frac{\mu_{T-1}}{\tau_{T-1}} + u_T - y_T = \frac{\mu_0}{\tau_0} + (u_T - y_T) + \dots + (u_1 - y_1) = \sum_{t=1}^T (u_t - y_t) = T(\bar{u}_T - \bar{y}_T). \end{cases}$$

We first convert the primal regret bound in Proposition A.1 into a bound with comparator $\bar{y}_T$. By convexity of $(-f)_x(\cdot)$, we have $\frac{1}{T} \sum_{t=1}^T -f(x, u_t) \geq -f(x, \bar{u}_T)$. Moreover, $(-f)_x(\cdot)$ is $G_{\mathcal{Y}}$ -Lipschitz on $\bar{\mathcal{Y}}$, so

$$-f(x, \bar{u}_T) \geq -f(x, \bar{y}_T) - G_{\mathcal{Y}} \|\bar{u}_T - \bar{y}_T\|_2 = -f(x, \bar{y}_T) - \frac{G_{\mathcal{Y}}}{T} \left\| \frac{\mu_T}{\tau_T} \right\|_2.$$

Therefore, applying Proposition A.1 (i) with $z_t = u_t$ gives

$$\begin{aligned} & \mathbb{E} \left[\frac{\eta_{T-1}}{2T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 - \frac{G_{\mathcal{Y}}}{T} \left\| \frac{\mu_T}{\tau_T} \right\|_2 + \frac{1}{T} \sum_{t=1}^T f(v_t, u_t) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \\ & \leq \frac{2}{T} \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 + \frac{G_{\mathcal{X}}^2}{2\alpha_{T-1}T} + \frac{1}{T} \sum_{t=1}^{T-1} \frac{G_{\mathcal{X}}^2}{2\alpha_t} + \frac{2G_{\mathcal{X}}R_{\mathcal{X}}}{T^{1/2}}. \end{aligned} \tag{A.6}$$

Similarly, by convexity of $f_y(\cdot)$, we have $\frac{1}{T} \sum_{t=1}^T f(v_t, y) \geq f(\bar{v}_T, y)$. Adding (A.6) and (A.7) cancels the average payoff term. It remains to lower bound the two quadratic expressions. For any $a > 0$, $b \geq 0$, and $s \geq 0$, we use $as^2/2 - bs \geq -b^2/(2a)$. Hence,

$$\begin{aligned} \frac{\eta_{T-1}}{2T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 - \frac{G_x}{T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2 &\geq -\frac{G_x^2}{2T\eta_{T-1}}, \\ \frac{\tau_{T-1}}{2T} \left\| \frac{\mu_T}{\tau_T} \right\|_2^2 - \frac{G_y}{T} \left\| \frac{\mu_T}{\tau_T} \right\|_2 &\geq -\frac{G_y^2}{2T\tau_{T-1}}. \end{aligned}$$

Combining these bounds with (A.6) and (A.7) proves the claim. Since $f_y(\cdot)$ is G_x -Lipschitz on $\bar{\mathcal{X}}$, we also have

$$f(\bar{v}_T, y) \geq f(\bar{x}_T, y) - G_x \|\bar{v}_T - \bar{x}_T\|_2 = f(\bar{x}_T, y) - \frac{G_x}{T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2.$$

Applying Proposition A.1 (ii) with $w_t = v_t$ gives

$$\begin{aligned} &\mathbb{E} \left[\frac{\tau_{T-1}}{2T} \left\| \frac{\mu_T}{\tau_T} \right\|_2^2 - \frac{G_x}{T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2 + \max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \frac{1}{T} \sum_{t=1}^T f(v_t, u_t) \right] \\ &\leq \frac{2}{T} \left(\beta_{T-1} + \sum_{t=1}^{T-1} \tau_t \right) R_y^2 + \frac{G_y^2}{2\beta_{T-1}T} + \frac{1}{T} \sum_{t=1}^{T-1} \frac{G_y^2}{2\beta_t} + \frac{2G_y R_y}{T^{1/2}}. \end{aligned} \quad (\text{A.7})$$

Adding (A.6) and (A.7) cancels the average payoff term. It remains to lower bound the two quadratic expressions. For any $a > 0$, $b \geq 0$, and $s \geq 0$, we use $as^2/2 - bs \geq -b^2/(2a)$. Hence,

$$\frac{\eta_{T-1}}{2T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 - \frac{G_x}{T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2 \geq -\frac{G_x^2}{2T\eta_{T-1}}, \quad \frac{\tau_{T-1}}{2T} \left\| \frac{\mu_T}{\tau_T} \right\|_2^2 - \frac{G_y}{T} \left\| \frac{\mu_T}{\tau_T} \right\|_2 \geq -\frac{G_y^2}{2T\tau_{T-1}}.$$

Combining these bounds with (A.6) and (A.7) proves the claim. $\square$

Proof of Theorem A.1. The parameter choices satisfy the monotonicity conditions in Lemma A.2. Moreover, for every $T \geq 2$,

$$\alpha_{T-1} = C_1 T^{1/2}, \quad \beta_{T-1} = C_2 T^{1/2}, \quad \eta_{T-1} = C_3 T^{-1/2}, \quad \tau_{T-1} = C_4 T^{-1/2}.$$

We also use

$$\sum_{t=1}^{T-1} \eta_t \leq 2C_3 T^{1/2}, \quad \sum_{t=1}^{T-1} \tau_t \leq 2C_4 T^{1/2}, \quad \sum_{t=1}^{T-1} \frac{1}{\alpha_t} \leq \frac{2T^{1/2}}{C_1}, \quad \sum_{t=1}^{T-1} \frac{1}{\beta_t} \leq \frac{2T^{1/2}}{C_2}.$$

Substituting these bounds into Lemma A.2 gives the desired bound. $\square$

A.4 Proofs of Theorem 1 (ii) and (iii)

We first present the detailed versions of Theorem 1 (ii) and (iii), including explicit constants and parameter choices.

Theorem A.2 (Detailed version of Theorem 1 (ii)). *Suppose Assumptions 1 and 2 hold. Let $\{x_t, y_t\}_{t \geq 1}$ be the primal-dual sequence generated by Algorithm 2 with parameters*

$$\alpha_t = C_1(t+1)^{1/2}, \quad \gamma_t = C_2 t^{-1/2}, \quad \eta_t = C_3(t+1)^{-1/2}$$

for some $C_1, C_2, C_3 > 0$. Then, for every $T \geq 2$,

$$\begin{aligned} \mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] &\leq \frac{2(C_1 + 2C_3)R_{\mathcal{X}}^2 + \frac{3G_{\mathcal{X}}^2}{2C_1} + \frac{G_{\mathcal{X}}^2}{2C_3} + 2G_{\mathcal{X}}R_{\mathcal{X}}}{T^{1/2}} \\ &\quad + \frac{\frac{2R_{\mathcal{Y}}^2}{C_2} + C_2 G_{\mathcal{Y}}^2 + 2G_{\mathcal{Y}}R_{\mathcal{Y}}}{T^{1/2}}. \end{aligned}$$

If $C_1 = C_3 = G_{\mathcal{X}}/R_{\mathcal{X}}$ and $C_2 = G_{\mathcal{Y}}/R_{\mathcal{Y}}$ then the bound becomes

$$\mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \leq \frac{10G_{\mathcal{X}}R_{\mathcal{X}} + 5G_{\mathcal{Y}}R_{\mathcal{Y}}}{T^{1/2}}$$

Theorem A.3 (Detailed version of Theorem 1 (iii)). *Suppose Assumptions 1 and 2 hold. Let $\{x_t, y_t\}_{t \geq 1}$ be the primal-dual sequence generated by Algorithm 3 with parameters*

$$\beta_t = C_1(t+1)^{1/2}, \quad \rho_t = C_2(t+1)^{-1/2}, \quad \tau_t = C_3(t+1)^{-1/2}$$

for some $C_1, C_2, C_3 > 0$. Then, for every $T \geq 2$,

$$\begin{aligned} \mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] &\leq \frac{\frac{2R_{\mathcal{X}}^2}{C_2} + C_2 G_{\mathcal{X}}^2 + 2G_{\mathcal{X}} R_{\mathcal{X}}}{T^{1/2}} \\ &\quad + \frac{2(C_1 + 2C_3)R_{\mathcal{Y}}^2 + \frac{3G_{\mathcal{Y}}^2}{2C_1} + \frac{G_{\mathcal{Y}}^2}{2C_3} + 2G_{\mathcal{Y}} R_{\mathcal{Y}}}{T^{1/2}}. \end{aligned}$$

If $C_1 = C_3 = G_{\mathcal{Y}}/R_{\mathcal{Y}}$ and $C_2 = G_{\mathcal{X}}/R_{\mathcal{X}}$ then the bound becomes

$$\mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \leq \frac{5G_{\mathcal{X}} R_{\mathcal{X}} + 10G_{\mathcal{Y}} R_{\mathcal{Y}}}{T^{1/2}}$$

The proof relies on the following lemma, which bounds the strong saddle gap for arbitrary monotone parameter sequences using Propositions A.1 and A.2.

Lemma A.3. *Suppose Assumptions 1 and 2 hold.*

- (i) *Let $\{x_t, y_t\}_{t \geq 1}$ be the primal-dual sequence generated by Algorithm 2 with parameters satisfying $\alpha_t \leq \alpha_{t+1}$, $\eta_t \geq \eta_{t+1}$, and $\gamma_t \geq \gamma_{t+1}$ for all $t \geq 1$. Then, for every $T \geq 2$,*

$$\begin{aligned} &\mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \\ &\leq \frac{2}{T} \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 + \frac{G_{\mathcal{X}}^2}{2\alpha_{T-1}T} + \frac{1}{T} \sum_{t=1}^{T-1} \frac{G_{\mathcal{X}}^2}{2\alpha_t} + \frac{G_{\mathcal{X}}^2}{2T\eta_{T-1}} + \frac{2G_{\mathcal{X}} R_{\mathcal{X}}}{T^{1/2}} \\ &\quad + \frac{2R_{\mathcal{Y}}^2}{\gamma_T T} + \frac{1}{T} \sum_{t=1}^T \frac{\gamma_t G_{\mathcal{Y}}^2}{2} + \frac{2G_{\mathcal{Y}} R_{\mathcal{Y}}}{T^{1/2}}. \end{aligned}$$

- (ii) *Let $\{x_t, y_t\}_{t \geq 1}$ be the primal-dual sequence generated by Algorithm 3 with parameters satisfying $\rho_t \geq \rho_{t+1}$, $\beta_t \leq \beta_{t+1}$, and $\tau_t \geq \tau_{t+1}$ for all $t \geq 1$. Then, for every $T \geq 2$,*

$$\begin{aligned} &\mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \\ &\leq \frac{2R_{\mathcal{X}}^2}{\rho_T T} + \frac{1}{T} \sum_{t=1}^T \frac{\rho_t G_{\mathcal{X}}^2}{2} + \frac{2G_{\mathcal{X}} R_{\mathcal{X}}}{T^{1/2}} \\ &\quad + \frac{2}{T} \left(\beta_{T-1} + \sum_{t=1}^{T-1} \tau_t \right) R_{\mathcal{Y}}^2 + \frac{G_{\mathcal{Y}}^2}{2\beta_{T-1}T} + \frac{1}{T} \sum_{t=1}^{T-1} \frac{G_{\mathcal{Y}}^2}{2\beta_t} + \frac{G_{\mathcal{Y}}^2}{2T\tau_{T-1}} + \frac{2G_{\mathcal{Y}} R_{\mathcal{Y}}}{T^{1/2}}. \end{aligned}$$

Proof of Lemma A.3. By symmetry, it suffices to prove part (i). The proof of part (ii) follows from the same argument after exchanging the primal and dual roles.

Define $\bar{v}_T := \frac{1}{T} \sum_{t=1}^T v_t$. Since $v_1 = x_1$, the LMO recursion implies

$$\frac{\lambda_T}{\eta_T} = \sum_{t=1}^T (v_t - x_t) = T(\bar{v}_T - \bar{x}_T).$$

We first control the primal side. By convexity of $(-f)_x(\cdot)$, for every $x \in \mathcal{X}$, we have $\frac{1}{T} \sum_{t=1}^T -f(x, y_t) \geq -f(x, \bar{y}_T)$. Applying Proposition A.1 (i) with $z_t = y_t$ gives

$$\begin{aligned} &\mathbb{E} \left[\frac{\eta_{T-1}}{2T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 + \frac{1}{T} \sum_{t=1}^T f(v_t, y_t) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \\ &\leq \frac{2}{T} \left(\alpha_{T-1} + \sum_{t=1}^{T-1} \eta_t \right) R_{\mathcal{X}}^2 + \frac{G_{\mathcal{X}}^2}{2\alpha_{T-1}T} + \frac{1}{T} \sum_{t=1}^{T-1} \frac{G_{\mathcal{X}}^2}{2\alpha_t} + \frac{2G_{\mathcal{X}} R_{\mathcal{X}}}{T^{1/2}}. \end{aligned} \tag{A.8}$$

We next control the dual side. By convexity of $f_y(\cdot)$, we have $\frac{1}{T} \sum_{t=1}^T f(v_t, y) \geq f(\bar{v}_T, y)$. Since $f_y(\cdot)$ is G_x -Lipschitz on $\bar{\mathcal{X}}$,

$$f(\bar{v}_T, y) \geq f(\bar{x}_T, y) - G_x \|\bar{v}_T - \bar{x}_T\|_2 = f(\bar{x}_T, y) - \frac{G_x}{T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2.$$

Applying Proposition A.2 (ii) with $w_t = v_t$ gives

$$\begin{aligned} & \mathbb{E} \left[-\frac{G_x}{T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2 + \max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \frac{1}{T} \sum_{t=1}^T f(v_t, y_t) \right] \\ & \leq \frac{2R_y^2}{\gamma_T T} + \frac{1}{T} \sum_{t=1}^T \frac{\gamma_t G_y^2}{2} + \frac{2G_y R_y}{T^{1/2}}. \end{aligned} \quad (\text{A.9})$$

Adding (A.8) and (A.9) cancels the average payoff term. It remains to lower bound the quadratic expression in λ_T . For any $a > 0$, $b \geq 0$, and $s \geq 0$, we use $as^2/2 - bs \geq -b^2/(2a)$. Hence,

$$\frac{\eta_{T-1}}{2T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2^2 - \frac{G_x}{T} \left\| \frac{\lambda_T}{\eta_T} \right\|_2 \geq -\frac{G_x^2}{2T\eta_{T-1}}.$$

Combining this bound with (A.8) and (A.9) proves part (i), and therefore completes the proof. $\square$

Proof of Theorems A.2 and A.3. We prove the two detailed theorems by substituting the stated parameter choices into Lemma A.3. For the LMO-PO case, the parameters satisfy

$$\alpha_{T-1} = C_1 T^{1/2}, \quad \eta_{T-1} = C_3 T^{-1/2}, \quad \gamma_T = C_2 T^{-1/2}.$$

Moreover,

$$\sum_{t=1}^{T-1} \eta_t \leq 2C_3 T^{1/2}, \quad \sum_{t=1}^{T-1} \frac{1}{\alpha_t} \leq \frac{2T^{1/2}}{C_1}, \quad \sum_{t=1}^T \gamma_t \leq 2C_2 T^{1/2}.$$

Substituting these estimates into Lemma A.3 (i) gives the desired bound in Theorem A.2.

The proof of Theorem A.3 is identical, using Lemma A.3 (ii) together with

$$\rho_T = C_2 T^{-1/2}, \quad \beta_{T-1} = C_1 T^{1/2}, \quad \tau_{T-1} = C_3 T^{-1/2},$$

and the analogous bounds

$$\sum_{t=1}^T \rho_t \leq 2C_2 T^{1/2}, \quad \sum_{t=1}^{T-1} \tau_t \leq 2C_3 T^{1/2}, \quad \sum_{t=1}^{T-1} \frac{1}{\beta_t} \leq \frac{2T^{1/2}}{C_1}.$$

gives the desired bound in Theorem A.3. $\square$

A.5 Strong Saddle-Gap Guarantee for the PO-PO Baseline

We first state the detailed convergence guarantee for Algorithm 6. This result is not part of Theorem 1, but it follows from the same framework and gives the standard PO-PO benchmark in the strong saddle-gap metric.

Theorem A.4. *Suppose Assumptions 1 and 2 hold. Let $\{x_t, y_t\}_{t \geq 1}$ be the primal-dual sequence generated by Algorithm 6 with parameters*

$$\rho_t = C_1 t^{-1/2}, \quad \gamma_t = C_2 t^{-1/2}$$

for some $C_1, C_2 > 0$. Then, for every $T \geq 1$,

$$\mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \leq \frac{\frac{2R_x^2}{C_1} + C_1 G_x^2 + 2G_x R_{\mathcal{X}}}{T^{1/2}} + \frac{\frac{2R_y^2}{C_2} + C_2 G_y^2 + 2G_y R_y}{T^{1/2}}.$$

If $C_1 = G_x/R_{\mathcal{X}}$ and $C_2 = G_y/R_y$ then the bound becomes

$$\mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \leq \frac{5(G_x R_{\mathcal{X}} + G_y R_y)}{T^{1/2}}$$

The proof relies on the following lemma, which bounds the strong saddle gap for arbitrary monotone parameter sequences using Proposition A.2.

Lemma A.4. *Suppose Assumptions 1 and 2 hold. Let $\{x_t, y_t\}_{t \geq 1}$ be the primal-dual sequence generated by Algorithm 6 with parameters satisfying $\rho_t \geq \rho_{t+1}$ and $\gamma_t \geq \gamma_{t+1}$ for all $t \geq 1$. Then, for every $T \geq 1$,*

$$\begin{aligned} & \mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \\ & \leq \frac{2R_{\mathcal{X}}^2}{\rho_T T} + \frac{1}{T} \sum_{t=1}^T \frac{\rho_t G_x^2}{2} + \frac{2G_x R_{\mathcal{X}}}{T^{1/2}} + \frac{2R_{\mathcal{Y}}^2}{\gamma_T T} + \frac{1}{T} \sum_{t=1}^T \frac{\gamma_t G_y^2}{2} + \frac{2G_y R_{\mathcal{Y}}}{T^{1/2}}. \end{aligned}$$

Proof of Lemma A.3. We first control the primal side. By convexity of $(-f)_x(\cdot)$, for every $x \in \mathcal{X}$, we have $\frac{1}{T} \sum_{t=1}^T f(x, y_t) \leq f(x, \bar{y}_T)$. Therefore,

$$\max_{x \in \mathcal{X}} \frac{1}{T} \sum_{t=1}^T f(x, y_t) - f(x, \bar{y}_T) \leq \frac{1}{T} \max_{x \in \mathcal{X}} \sum_{t=1}^T (f(x, y_t) - f(x, \bar{y}_T)).$$

Applying Proposition A.2 (i) with $z_t = y_t$ gives

$$\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T f(x_t, y_t) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \right] \leq \frac{2R_{\mathcal{X}}^2}{\rho_T T} + \frac{1}{T} \sum_{t=1}^T \frac{\rho_t G_x^2}{2} + \frac{2G_x R_{\mathcal{X}}}{T^{1/2}}. \quad (\text{A.10})$$

We next control the dual side. By convexity of $f(\cdot, y)$, for every $y \in \mathcal{Y}$, we have $f(\bar{x}_T, y) \leq \frac{1}{T} \sum_{t=1}^T f(x_t, y)$. Hence,

$$\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \frac{1}{T} \sum_{t=1}^T f(x_t, y) \leq \frac{1}{T} \max_{y \in \mathcal{Y}} \sum_{t=1}^T (f(x_t, y) - f(x_t, \bar{y}_T)).$$

Applying Proposition A.2 (ii) with $w_t = x_t$ gives

$$\mathbb{E} \left[\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \frac{1}{T} \sum_{t=1}^T f(x_t, y) \right] \leq \frac{2R_{\mathcal{Y}}^2}{\gamma_T T} + \frac{1}{T} \sum_{t=1}^T \frac{\gamma_t G_y^2}{2} + \frac{2G_y R_{\mathcal{Y}}}{T^{1/2}}. \quad (\text{A.11})$$

Adding (A.10) and (A.11) cancels the average payoff term and proves the lemma. $\square$

Proof of Theorem A.4. The parameter choices satisfy the monotonicity conditions in Lemma A.4. Moreover,

$$\rho_T = C_1 T^{-1/2}, \quad \gamma_T = C_2 T^{-1/2}.$$

We also use

$$\sum_{t=1}^T \rho_t \leq 2C_1 T^{1/2}, \quad \sum_{t=1}^T \gamma_t \leq 2C_2 T^{1/2}.$$

Substituting these estimates into Lemma A.4 gives the desired bound. $\square$

A.6 Proofs of Theorem 2

For the convenience, we repeat the statement of Theorem 2.

Theorem A.5 (Restatement of Theorem 2). *Let $G_x > 0, G_y > 0, R_{\mathcal{X}} > 0, R_{\mathcal{Y}} > 0$ be fixed. Then, for any generic LMO-LMO method of the form Algorithm 4 running for $T \geq 1$ iterations, there exist convex compact subsets $\mathcal{X} \subseteq \mathbb{R}^n, \mathcal{Y} \subseteq \mathbb{R}^m$ with radii $R_{\mathcal{X}}$ and $R_{\mathcal{Y}}$, respectively, and a function $f \in \mathcal{F}_{G_x, G_y, \|\cdot\|_2}^0(\mathcal{X}, \mathcal{Y})$ with $T \leq \min\{m, n\}/4 - 1$ such that*

$$\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \geq \frac{G_x R_{\mathcal{X}}}{2(T+1)^{1/2}} + \frac{G_y R_{\mathcal{Y}}}{2(T+1)^{1/2}}.$$

Proof of Theorem A.5. Fix $T \geq 1$, and choose dimensions n, m such that $n, m \geq 4(T + 1)$. Let

$$\mathcal{X} := R_{\mathcal{X}} \cdot \Delta_n, \quad \mathcal{Y} := R_{\mathcal{Y}} \cdot \Delta_m,$$

where $\Delta_k := \{z \in \mathbb{R}_+^k : \sum_{i=1}^k z_i = 1\}$. The adversary initializes the method at the extreme points $x_1 = R_{\mathcal{X}} \cdot e_1$ and $y_1 = R_{\mathcal{Y}} \cdot e_1$, and fixes LMO tie-breaking rules that always return extreme-point minimizers. This is valid because a linear objective over a simplex always admits an extreme-point minimizer. Consider the payoff function

$$f(x, y) := G_{\mathbf{x}} \|x\|_2 - G_{\mathbf{y}} \|y\|_2.$$

The function f is convex in x and concave in y . It is also easy to verify that $f \in \mathcal{F}_{G_{\mathbf{x}}, G_{\mathbf{y}}, \|\cdot\|_2}^0(\mathcal{X}, \mathcal{Y})$.

Since x_1 is an extreme point of $\mathcal{X}$ and each primal LMO call returns an extreme point of $\mathcal{X}$, all points $x_1, \dots, x_{T+1}$ are vertices of $R_{\mathcal{X}} \cdot \Delta_n$. By Algorithm 4, the output satisfies

$$\bar{x}_T \in \text{conv}\{x_1, \dots, x_{T+1}\}.$$

Therefore, there exist distinct indices $i_1, \dots, i_{k_T} \in [n]$ with $k_T \leq T + 1$ such that

$$\bar{x}_T \in R_{\mathcal{X}} \cdot \text{conv}\{e_{i_1}, \dots, e_{i_{k_T}}\}.$$

Consequently, we have

$$\|\bar{x}_T\|_2 \geq \min \left\{ \|x\|_2 : x \in R_{\mathcal{X}} \cdot \text{conv}\{e_{i_1}, \dots, e_{i_{k_T}}\} \right\} = \frac{R_{\mathcal{X}}}{k_T^{1/2}} \geq \frac{R_{\mathcal{X}}}{(T + 1)^{1/2}}.$$

Similarly, since y_1 is an extreme point of $\mathcal{Y}$ and each dual LMO call returns an extreme point of $\mathcal{Y}$, all points $y_1, \dots, y_{T+1}$ are vertices of $R_{\mathcal{Y}} \cdot \Delta_m$, we have

$$\|\bar{y}_T\|_2 \geq \frac{R_{\mathcal{Y}}}{(T + 1)^{1/2}}.$$

On the other hand, over the full simplices,

$$\min_{x \in \mathcal{X}} \|x\|_2 = \frac{R_{\mathcal{X}}}{n^{1/2}}, \quad \min_{y \in \mathcal{Y}} \|y\|_2 = \frac{R_{\mathcal{Y}}}{m^{1/2}}.$$

Therefore,

$$\begin{aligned} \max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) &= G_{\mathbf{x}} \|\bar{x}_T\|_2 - G_{\mathbf{y}} \min_{y \in \mathcal{Y}} \|y\|_2 - G_{\mathbf{x}} \min_{x \in \mathcal{X}} \|x\|_2 + G_{\mathbf{y}} \|\bar{y}_T\|_2 \\ &\geq \frac{G_{\mathbf{x}} R_{\mathcal{X}}}{(T + 1)^{1/2}} - \frac{G_{\mathbf{x}} R_{\mathcal{X}}}{n^{1/2}} + \frac{G_{\mathbf{y}} R_{\mathcal{Y}}}{(T + 1)^{1/2}} - \frac{G_{\mathbf{y}} R_{\mathcal{Y}}}{m^{1/2}}. \end{aligned}$$

Since $n, m \geq 4(T + 1)$, we have

$$\frac{1}{n^{1/2}} \leq \frac{1}{2(T + 1)^{1/2}}, \quad \frac{1}{m^{1/2}} \leq \frac{1}{2(T + 1)^{1/2}}.$$

Combining the last two displays gives the desired bound. $\square$

A.7 Proofs of Theorem 3

For the convenience, we repeat the statement of Theorem 3.

Theorem A.6 (Restatement of Theorem 3). *Let $G_{\mathbf{x}} > 0, G_{\mathbf{y}} > 0, R_{\mathcal{X}} > 0, R_{\mathcal{Y}} > 0$ be fixed. Then, for any generic LMO-PO method of the form Algorithm 5 running for $T \geq 1$ iterations, there exist convex compact subsets $\mathcal{X} \subseteq \mathbb{R}^n, \mathcal{Y} \subseteq \mathbb{R}^m$ with radii $R_{\mathcal{X}}$ and $R_{\mathcal{Y}}$, respectively, and a function $f \in \mathcal{F}_{G_{\mathbf{x}}, G_{\mathbf{y}}, \|\cdot\|_2}^0(\mathcal{X}, \mathcal{Y})$ with $T \leq \min\{m, n\}/4 - 1$ such that*

$$\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \geq \frac{G_{\mathbf{x}} R_{\mathcal{X}}}{2(T + 1)^{1/2}} + \frac{G_{\mathbf{y}} R_{\mathcal{Y}}}{2(T + 1)^{1/2}}.$$

Proof of Theorem A.6. Fix $T \geq 1$, and choose dimensions n, m such that $n, m \geq 4(T + 1)$. Let

$$\mathcal{X} := R_{\mathcal{X}} \cdot \Delta_n.$$

The adversary initializes the primal variable at the extreme point $x_1 = R_{\mathcal{X}} \cdot e_1$ and fixes an LMO tie-breaking rule that always returns an extreme-point minimizer. For the dual variable, we invoke the classical nonsmooth first-order lower bound over the Euclidean ball. Namely, for any adaptive first-order method making T oracle queries and returning $\bar{y}_T$, there exist a compact convex set $\mathcal{Y} := R_{\mathcal{Y}} \cdot \mathbb{B}_2^m$ and a convex $G_{\mathcal{Y}}$ -Lipschitz function $h_{T+1} : \mathcal{Y} \rightarrow \mathbb{R}$ such that

$$h_{T+1}(\bar{y}_T) - \min_{y \in \mathcal{Y}} h_{T+1}(y) \geq \frac{G_{\mathcal{Y}} R_{\mathcal{Y}}}{2(T+1)^{1/2}}. \quad (\text{A.12})$$

This is the standard nonsmooth lower-bound construction of Nesterov [56]. For completeness, we repeat the construction here. Set

$$\delta := \frac{R_{\mathcal{Y}}}{4T(T+1)^{1/2}}.$$

We inductively construct signs $a_1, \dots, a_{T+1} \in \{\pm 1\}$ and distinct coordinates $\sigma(1), \dots, \sigma(T+1) \in \{1, \dots, m\}$. After y_k has been determined, choose

$$\sigma(k) \in \arg \max_{j \in [m] \setminus \{\sigma(1), \dots, \sigma(k-1)\}} |(y_k)_j|,$$

and choose $a_k \in \{\pm 1\}$ so that

$$a_k (y_k)_{\sigma(k)} = |(y_k)_{\sigma(k)}|.$$

For $k = 1, \dots, T+1$, define

$$h_k(y) := G_{\mathcal{Y}} \max_{1 \leq i \leq k} \{a_i y_{\sigma(i)} - 2(i-1)\delta\}.$$

We claim that the final function h_{T+1} is consistent with all information revealed to the algorithm during the first T iterations. To see this, fix $1 \leq s < k \leq T+1$ and any $y \in \mathcal{Y}$ such that $\|y - y_s\|_2 \leq \delta$. For every $i > s$, the coordinate $\sigma(i)$ was still unused when $\sigma(s)$ was chosen. Hence

$$|(y_s)_{\sigma(s)}| \geq |(y_s)_{\sigma(i)}| \geq a_i (y_s)_{\sigma(i)}.$$

Therefore,

$$\begin{aligned} a_i y_{\sigma(i)} - 2(i-1)\delta &\leq a_i (y_s)_{\sigma(i)} + \delta - 2(i-1)\delta \\ &\leq |(y_s)_{\sigma(s)}| + \delta - 2s\delta \\ &= |(y_s)_{\sigma(s)}| - (2s-1)\delta \\ &\leq a_s y_{\sigma(s)} - 2(s-1)\delta. \end{aligned}$$

Thus every affine piece added after step s is dominated on the ball $\{y : \|y - y_s\|_2 \leq \delta\}$ by the s -th affine piece, which already appears in h_s . Consequently,

$$h_k(y) = h_s(y), \quad \forall y \in \mathcal{Y} \text{ such that } \|y - y_s\|_2 \leq \delta.$$

In particular, the values and subdifferentials observed at each previous query y_s , $s \leq T$, are identical for h_s and for the final function h_{T+1} . The function h_{T+1} is convex and $G_{\mathcal{Y}}$ -Lipschitz on $\mathcal{Y}$, since it is the maximum of affine functions whose slopes have Euclidean norm $G_{\mathcal{Y}}$. Moreover,

$$\min_{y \in \mathcal{Y}} h_{T+1}(y) \leq G_{\mathcal{Y}} \min_{\|y\|_2 \leq R_{\mathcal{Y}}} \max_{1 \leq i \leq T+1} a_i y_{\sigma(i)} = -\frac{G_{\mathcal{Y}} R_{\mathcal{Y}}}{(T+1)^{1/2}}.$$

On the other hand, since $\bar{y}_T = y_{T+1}$ in Algorithm 5, we have

$$h_{T+1}(y_{T+1}) \geq G_{\mathcal{Y}} (a_{T+1}(y_{T+1})_{\sigma(T+1)} - 2T\delta) = G_{\mathcal{Y}} (|(y_{T+1})_{\sigma(T+1)}| - 2T\delta) \geq -2G_{\mathcal{Y}} T\delta.$$

Therefore, by the choice of δ ,

$$h(\bar{y}_T) - \min_{y \in \mathcal{Y}} h(y) \geq \frac{G_{\mathcal{Y}} R_{\mathcal{Y}}}{(T+1)^{1/2}} - 2G_{\mathcal{Y}} T\delta = \frac{G_{\mathcal{Y}} R_{\mathcal{Y}}}{2(T+1)^{1/2}}.$$

Define now the saddle payoff

$$f(x, y) := G_x \|x\|_2 - h_{T+1}(y).$$

Since x_1 is an extreme point of $\mathcal{X}$ and each primal LMO call returns an extreme point of $\mathcal{X}$, all points $x_1, \dots, x_{T+1}$ are vertices of $R_{\mathcal{X}} \Delta_n$. By Algorithm 5, the output satisfies

$$\bar{x}_T \in R_{\mathcal{X}} \cdot \text{conv}\{e_{i_1}, \dots, e_{i_{k_T}}\}.$$

Consequently, we have

$$\|\bar{x}_T\|_2 \geq \min \left\{ \|x\|_2 : x \in R_{\mathcal{X}} \cdot \text{conv}\{e_{i_1}, \dots, e_{i_{k_T}}\} \right\} = \frac{R_{\mathcal{X}}}{k_T^{1/2}} \geq \frac{R_{\mathcal{X}}}{(T+1)^{1/2}}.$$

Since $n \geq 4(T+1)$, we may conclude that

$$G_x \|\bar{x}_T\|_2 - G_x \min_{x \in \mathcal{X}} \|x\|_2 \geq \frac{G_x R_{\mathcal{X}}}{2(T+1)^{1/2}}. \quad (\text{A.13})$$

Finally, by the separable form of f ,

$$\begin{aligned} \max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) &= G_x \|\bar{x}_T\|_2 - \min_{y \in \mathcal{Y}} h_{T+1}(y) - \left(G_x \min_{x \in \mathcal{X}} \|x\|_2 - h_{T+1}(\bar{y}_T) \right) \\ &= G_x \|\bar{x}_T\|_2 - G_x \min_{x \in \mathcal{X}} \|x\|_2 + h_{T+1}(\bar{y}_T) - \min_{y \in \mathcal{Y}} h_{T+1}(y). \end{aligned}$$

Combining this identity with (A.12) and (A.13) yields the desired bound. $\square$

B Implementation Details

This appendix collects the implementation details used in Section 5. In all experiments, the suffix **Deterministic** means that the method uses the full subgradient oracle, while the suffix **Stochastic** means that the method uses an unbiased stochastic estimate of the same oracle. The algorithm names LMO-LMO, LMO-PO, PO-LMO, and PO-PO refer to the oracle used for the primal and dual domains, respectively. They correspond to Algorithms 1 to 3 and 6, respectively. For LMO-based updates, we use simple enclosing sets $\bar{\mathcal{X}} \supseteq \mathcal{X}$ and $\bar{\mathcal{Y}} \supseteq \mathcal{Y}$. These sets are chosen so that projection onto them is computationally cheap, while the actual feasible updates over $\mathcal{X}$ and $\mathcal{Y}$ are performed using LMOs whenever the corresponding method uses an LMO. For the robust classification problem, we rely on the following set $\bar{\mathcal{Y}}$ for the inexpensive projection step.

Lemma B.1. *Given any $c \in \mathbb{R}^n$, let $[c]_+ = \max\{c, 0\}$ componentwise. For*

$$\bar{\mathcal{Y}} := \{y \in \mathbb{R}^n : y \geq 0, \|y\|_2 \leq 1\},$$

it holds that

$$\text{PO}_{\bar{\mathcal{Y}}}(c) = \frac{[c]_+}{\max\{\|[c]_+\|_2, 1\}}.$$

Proof of Lemma B.1. Computing $\text{PO}_{\bar{\mathcal{Y}}}(c)$ is equivalent to solving

$$\min_{y \in \mathbb{R}^n} \{\|y - c\|_2^2 : y \geq 0, \|y\|_2 \leq 1\}.$$

Since the feasible set is the intersection of the nonnegative orthant and the Euclidean unit ball, the negative coordinates of c are first clipped to zero. Thus the problem reduces to projecting $[c]_+$ onto the Euclidean unit ball. If $\|[c]_+\|_2 \leq 1$, then $[c]_+$ is feasible and is the projection. If $\|[c]_+\|_2 > 1$, the Euclidean projection onto the unit ball is $[c]_+ / \|[c]_+\|_2$. Combining the two cases gives

$$\text{PO}_{\bar{\mathcal{Y}}}(c) = \frac{[c]_+}{\max\{\|[c]_+\|_2, 1\}}.$$

This proves the claim. $\square$

The projection and LMO formulas used in the experiments are summarized in Table 1, where the input is the vector c or the matrix C depending on the context.

Table 1: Summary of LMOs and POs.

Set	LMO	PO
Frobenius Ball ($\ Y\ _F \leq r$)	$-r \frac{C}{\ C\ _2}$	If $\ C\ _2 > r$, $Y = r \frac{C}{\ C\ _F}$. Else $Y = C$.
Nuclear Ball ($\ Y\ _* \leq r$)	$-r(u_1 v_1^T)$ (Top singular triplet)	1. Full SVD: $C = U\Sigma V^T$. 2. Project $\text{diag}(\Sigma)$ onto r -Simplex. 3. Reconstruct: $Y = U\Sigma_{\text{proj}}V^T$.
Nonnegative Ball ($\ y\ _2 \leq r, y \geq 0$)	—	1. Nonnegative clip: $[c]_+ = \max(0, c)$ 2. If $\ [c]_+\ _2 > r$, $Y = r \frac{[c]_+}{\ [c]_+\ _2}$. Else $y = [c]_+$.
r-Simplex ($\sum y_i = r, y \geq 0$)	$-r \text{sign}(c_i) \cdot e_i$ where $i = \text{argmin}(c)$	1. Sort C in descending order. 2. Find τ s.t. $\sum \max(c_i - \tau, 0) = r$. 3. Output: $y_i = \max(c_i - \tau, 0)$.

B.1 Matrix Completion with Spectral Norm Fit

Juditsky and Nemirovski [33] considered the spectral norm fit problem

$$\min_{\|X\|_* \leq 1} \|\mathcal{A}(X) - B\|_{\text{op}}, \quad (\text{B.1})$$

where $X \in \mathbb{R}^{n \times p}$ is the decision variable, $B \in \mathbb{R}^{m \times q}$, and $\mathcal{A} : \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^{m \times q}$ is a linear map. Problem (B.1) admits the saddle-point reformulation

$$\min_{\|X\|_* \leq 1} \max_{\|Y\|_* \leq 1} \text{Tr}[(\mathcal{A}(X) - B)^\top Y].$$

For the matrix completion with spectral norm fit problem, we generate a synthetic instance as suggested by [33] as follows: We set $p = n$ and $q = m = 2n$ with $n = 200$ and the map $\mathcal{A}$ given by

$$\mathcal{A}(X) = \frac{1}{k} \sum_{i=1}^k L_i X R_i^\top,$$

where $L_i \in \mathbb{R}^{m \times n}$ and $R_i \in \mathbb{R}^{q \times p}$ with $L_i = \tilde{L}_i / \|\tilde{L}_i\|_{\text{op}}$ and $R_i = \tilde{R}_i / \|\tilde{R}_i\|_{\text{op}}$ and $\tilde{L}_i, \tilde{R}_i$ are randomly generated by standard normal distributions. In our experiment, we set $k = 2$. The construction of the observation matrix B begins with the generation of a ground-truth signal $V_{\text{true}} \in \mathbb{R}^{n \times n}$. To ensure this matrix is low-rank, we first define the rank as $r = \lfloor \sqrt{n} \rfloor$. We then sample two latent factor matrices $U, V \in \mathbb{R}^{n \times r}$ from a standard normal distribution and form the base signal matrix through the outer product $V_{\text{base}} = UV^\top$.

To construct B , we first compute $V_{\text{true}} = V_{\text{base}} / \|V_{\text{base}}\|_*$, which ensures that the resulting ground-truth matrix has a unit nuclear norm. This normalized signal is then passed through a linear observation operator $\mathcal{A}(\cdot)$ to produce the clean measurement matrix $B_{\text{clean}} = \mathcal{A}(V_{\text{true}}) \in \mathbb{R}^{m \times m}$. We then generate a raw noise matrix $E_{\text{raw}} \in \mathbb{R}^{m \times m}$ where each entry is sampled from standard normal distribution. This matrix is then rescaled by its spectral norm and multiplied by a specified noise level δ to form the error matrix $E = \delta \cdot (E_{\text{raw}} / \|E_{\text{raw}}\|_{\text{op}})$. The observed matrix B is then obtained by the summation of the clean measurements and this spectrally-scaled noise, $B = B_{\text{clean}} + E$.

For initialization, we start with $X_0 = \text{PO}_{\mathcal{X}}(\tilde{X}_0)$ and $Y_0 = \text{PO}_{\mathcal{Y}}(\tilde{Y}_0)$ where entries of $\tilde{X}_0$ and $\tilde{Y}_0$ are randomly generated by uniform distribution over $[-1000, 1000]$. For Algorithm 1, Algorithm 2, Algorithm 3 and Algorithm 6, the stepsizes are given by

$$\alpha_t = \beta_t = \left(\frac{1}{t+1} \right)^{-1/2} \quad \& \quad \eta_t = \tau_t = \frac{0.1}{(t+1)^{1/2}} \quad \& \quad \gamma_t = \rho_t = \left(\frac{1}{t+1} \right)^{1/2},$$

which align with Theorem 1 and Theorem A.4. For $\mathcal{X} = \{X \in \mathbb{R}^{n \times p} \mid \|X\|_* \leq 1\}$, we choose the covering $\bar{\mathcal{X}} = \{X \in \mathbb{R}^{n \times p} \mid \|X\|_F \leq 1\}$. For $\mathcal{Y} = \{X \in \mathbb{R}^{m \times q} \mid \|Y\|_* \leq 1\}$, we choose the covering $\bar{\mathcal{Y}} = \{Y \in \mathbb{R}^{m \times q} \mid \|Y\|_F \leq 1\}$.

Moreover, since the payoff function in this experiment is smooth, we also implement LMO-LMO, LMO-PO and PO-LMO methods [18], Alternating Gradient Projection (AGP) method [69], Extra-Gradient (EG) method [36], Optimistic Gradient Descent-Ascent (OGDA) method [59]. To describe the parameter choices, we follow the notation adopted in the corresponding paper. For LMO-LMO method [18, Algorithm 1], we choose $\tau_t = (t+1)^{-1}$ and $\beta_t = (t+1)^{-1/5}$ according to [18, Theorem 2.2]. For LMO-PO method [18, Algorithm 2], we choose $\tau_t = (t+1)^{-1}$ and $\beta_t = (t+1)^{-1/3}$ according to [18, Theorem 3.2]. For PO-LMO method [18, Algorithm 3], we choose $\tau_t = (t+1)^{-2/3}$ and $\beta_t = (t+1)^{-1/3}$ according to [18, Theorem 4.2]. For AGP method [69], we implement both parameter choices in [69, Theorem 3.2], i.e., $b_t = 0, c_t = (t+1)^{-1/4}, \beta_t = (t+1)^{1/2}, \gamma_t = 10$ and in [69, Theorem 4.2], i.e., $b_t = (t+1)^{-1/4}, c_t = 0, \beta_t = 10, \gamma_k = (t+1)^{1/2}$ which correspond to parameter choices that ensure convergence in nonconvex-concave and convex-nonconcave settings, respectively. For EG method [36], we choose the stepsize to be $\frac{1}{2L}$ and for OGDA method [59], we choose the stepsize to be $\frac{1}{2.05L}$ where $L = \frac{1}{k} \sum_{i=1}^k \|L_i\|_{\text{op}} \|R_i\|_{\text{op}}$ is the smoothness constant of the payoff function.

Figure 2 provides the extended comparison on the smooth matrix completion instance. As expected for this bilinear problem, the classical smooth methods, especially OGDA and EG, perform very well when equipped with tuned stepsizes. Among the proposed nonsmoothing methods, LMO-LMO (Det) and LMO-LMO (Sto) are the strongest variants, and they substantially outperform the smoothing-based projection-free methods. The smoothing-based LMO variants decrease much more slowly and remain close to their initial saddle-gap values over the plotted time horizon, reflecting the cost of introducing and tuning a smoothing parameter. The AGP variants also improve more slowly than EG/OGDA and the proposed LMO-based methods. Overall, the figure illustrates that, even on a smooth problem where methods designed specifically for smooth saddle-point problems are highly competitive, the proposed subgradient-based LMO methods remain effective and avoid the deterioration observed for smoothing-based projection-free schemes.

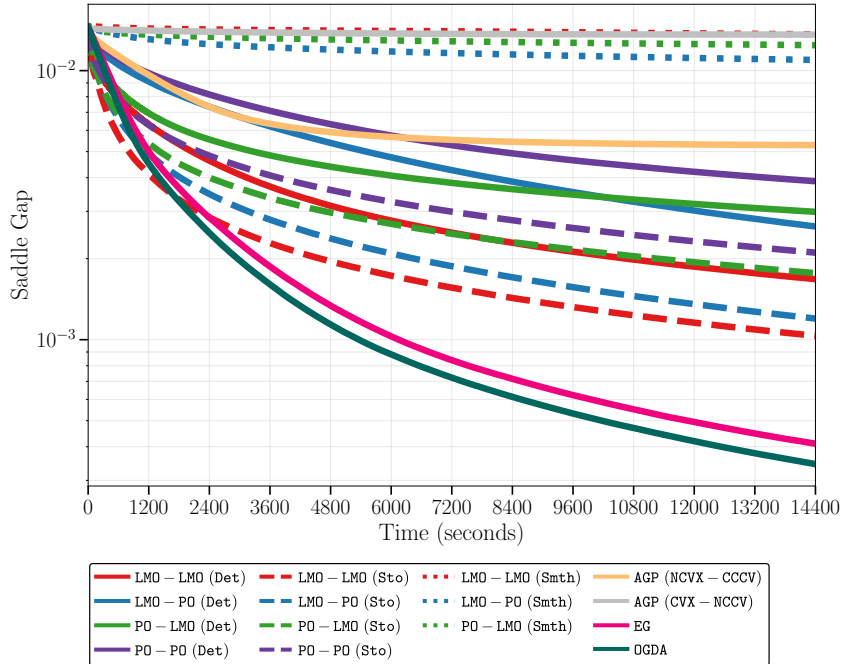


Figure 2: Additional comparison on the matrix completion problem with spectral norm fit.

B.2 Robust Multiclass Classification with Hinge Loss

We use the robust multiclass classification model described in Section 5. In this experiment, we set

$$r = 10, \quad \lambda = \frac{\lambda'}{n^2}, \quad \lambda' = 1.$$

The primal and dual feasible sets are

$$\mathcal{X} = \{\Theta \in \mathbb{R}^{k \times d} : \|\Theta\|_* \leq r\} \quad \text{and} \quad \mathcal{Y} = \left\{ y \in \mathbb{R}^n : y \geq 0, \sum_{i=1}^n y_i = 1 \right\}.$$

For the LMO-based variants, we use the covering sets

$$\bar{\mathcal{X}} = \{\Theta \in \mathbb{R}^{k \times d} : \|\Theta\|_F \leq r\}, \quad \text{and} \quad \bar{\mathcal{Y}} = \{y \in \mathbb{R}^n : y \geq 0, \|y\|_2 \leq 1\}.$$

The inclusion $\mathcal{Y} \subseteq \bar{\mathcal{Y}}$ follows because every probability vector has Euclidean norm at most one. Projection onto $\bar{\mathcal{X}}$ is a Frobenius-norm rescaling, while projection onto $\bar{\mathcal{Y}}$ is given in Lemma B.1. We initialize $\Theta_1 = \text{PO}_{\mathcal{X}}(\tilde{\Theta}_1)$ and $y_1 = e_1$, where the entries of $\tilde{\Theta}_1$ are sampled independently from the uniform distribution on $[-1000, 1000]$. The stepsizes are

$$\alpha_t = \beta_t = \left(\frac{0.01}{t+1} \right)^{-1/2}, \quad \eta_t = \tau_t = \left(\frac{10}{t+1} \right)^{1/2}, \quad \gamma_t = \rho_t = \left(\frac{0.01}{t+1} \right)^{1/2}.$$

These choices follow the scaling in Theorems 1 and A.4. Let

$$\ell_i(\Theta) = \max_{j \in [k]} \{ \mathbf{1}(j \neq b_i) + \theta_j^\top a_i - \theta_{b_i}^\top a_i \}.$$

The full primal subgradient is computed from $\frac{1}{n} \sum_{i=1}^n y_i \partial \ell_i(\Theta)$. The dual oracle uses the subgradient of $(-f)_\Theta(\cdot)$, namely $-\frac{1}{n} \ell(\Theta) + 2\lambda n(ny - \mathbf{1}_n)$, where $\ell(\Theta) = (\ell_1(\Theta), \dots, \ell_n(\Theta))^\top$. The stochastic variants replace the full sums by unbiased single-sample estimates. For this experiment, we report the primal objective

$$p(\Theta) := \max_{y \in \mathcal{Y}} \left\{ \frac{1}{n} \sum_{i=1}^n y_i \ell_i(\Theta) - \lambda \|ny - \mathbf{1}_n\|_2^2 \right\}.$$

Computing the full strong saddle gap is expensive at this scale because it requires solving both the dual maximization problem and the primal minimization problem over the nuclear-norm ball. The primal objective is substantially cheaper to compute and empirically tracks the same convergence behavior.