FASTER PROJECTION-FREE ALGORITHMS FOR MINIMAX PROBLEMS OVER POLYTOPES

Anonymous authors
Paper under double-blind review

ABSTRACT

Projection-free methods are attractive for constrained minimax problems when linear minimization oracles (LMOs) are substantially cheaper than projection. However, existing fully projection-free methods can suffer from much worse iteration complexities than their projection-based counterparts. We develop a fully projection-free method for polytope-constrained minimax problems, which exploits polyhedral geometry through away-step Frank–Wolfe updates, requiring only one LMO call over each feasible set per iteration. For nonconvex–concave problems, our method finds an ϵ-stationary point in $\mathcal{O}(\epsilon^{-3})$ iterations, improving the previous projection-free bound of $\mathcal{O}(\epsilon^{-6})$. Under a one-sided Kurdyka–Łojasiewicz condition with exponent $\theta \in (0,1)$, imposed on either the primal or dual problem, we obtain an $\mathcal{O}\left(\epsilon^{-\max\{2,4\theta\}}\right)$ complexity. To the best of our knowledge, these are the first fully projection-free complexity guarantees for the corresponding nonconvex–nonconcave problems. We further obtain an $\mathcal{O}(\epsilon^{-3})$ complexity for convex–nonconcave problems. For convex–concave problems, we improve the best-known projection-free complexity from $\mathcal{O}(\epsilon^{-5})$ to $\mathcal{O}(\epsilon^{-2})$. Our numerical results further demonstrate the practical effectiveness of the proposed method.

1 INTRODUCTION

Projection-free methods are attractive for constrained optimization problems in which projections onto the feasible set are computationally expensive, while linear minimization remains cheap. This situation arises for many structured polytopes in machine learning and combinatorial optimization, where linear optimization over the polytope often admits a closed-form solution or reduces to a well-studied combinatorial optimization problem (Combettes & Pokutta, 2021). This computational distinction is particularly relevant in constrained minimax optimization, where both the primal and dual variables must be updated repeatedly and projection-based methods may require projections onto two separate feasible sets. Fully projection-free methods, which replace these projections by linear minimization oracles, therefore provide an appealing alternative. However, achieving strong convergence guarantees while retaining this fully projection-free structure remains challenging.

Motivated by this challenge, we study smooth minimax problems of the form

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y}), \tag{P}$$

where f is L-smooth, and $A \subseteq \mathbb{R}^n, B \subseteq \mathbb{R}^m$ are finite sets with $\mathcal{X} = \mathrm{conv}(A)$ and $\mathcal{Y} = \mathrm{conv}(B)$. We focus primarily on the nonconvex–concave (NC–C) and convex–concave (C–C) settings, while also providing guarantees for convex–nonconcave (C–NC) problems as well as problems satisfying a one-sided Kurdyka–Łojasiewicz (KŁ) condition on either the primal or dual problem.

Many existing first-order methods for constrained minimax optimization rely on projected primal–dual updates (Lin et al., 2020a;b; Xu et al., 2023). A natural approach to constructing a fully projection-free algorithm for Problem P is to replace these projected updates by Frank–Wolfe steps, requiring only linear minimization over the two feasible sets. Combining such updates with a quadratic perturbation of the dual objective, Boroun et al. (2023) develop a fully projection-free single-loop method for NC–C problems with an $\mathcal{O}(\epsilon^{-6})$ iteration complexity under the additional assumption that the dual feasible set is strongly convex. More recently, Giang-Tran et al. (2026)

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

Table 1: Summary of convergence guarantees for fully projection-free minimax methods. Certificates are listed as primal / dual. "FW" denotes Frank–Wolfe stationarity; "Primal gap" denotes the primal optimality gap of the outer minimax problem; and "Dual gap" denotes suboptimality of the dual maximization problem. Their precise definitions and relations are given in Section 2.1. SC $\mathcal{Y}$ denotes a strongly convex dual feasible set.

Setting	Method	Certificate	Iteration Complexity
NC–C	(Boroun et al., 2023) (SC $\mathcal{Y}$)	FW / FW	$\mathcal{O}(\epsilon^{-6})$
	(Giang-Tran et al., 2026) (SC $\mathcal{Y}$)	FW / Dual gap	$\mathcal{O}(\epsilon^{-5})$
	(Giang-Tran et al., 2026)	FW / Dual gap	$\mathcal{O}(\epsilon^{-6})$
	Ours	**FW / FW**	$\mathcal{O}(\epsilon^{-\mathbf{3}})$
C–C	(Giang-Tran et al., 2026)	Primal gap / Dual gap	$\mathcal{O}(\epsilon^{-5})$
	(Giang-Tran et al., 2026) (SC $\mathcal{Y}$)	Primal gap / Dual gap	$\mathcal{O}(\epsilon^{-4})$
	Ours	**FW / FW**	$\mathcal{O}(\epsilon^{-\mathbf{2}})$
NC–KŁ	**Ours**	**FW / FW**	$\mathcal{O}(\epsilon^{-\max\{\mathbf{2},4\theta\}})$
KŁ–NC	**Ours**	**FW / FW**	$\mathcal{O}(\epsilon^{-\max\{\mathbf{2},4\theta\}})$
C–NC	**Ours**	**FW / FW**	$\mathcal{O}(\epsilon^{-\mathbf{3}})$

establish fully projection-free single-loop guarantees over general compact convex sets, obtaining iteration complexities of $\mathcal{O}(\epsilon^{-6})$ for the NC–C setting, and $\mathcal{O}(\epsilon^{-5})$ for the C–C setting.

These results establish finite-time convergence guarantees for fully projection-free minimax optimization. It remains open, however, whether the additional polyhedral structure in Problem P can be exploited to obtain sharper complexity guarantees. In smooth convex minimization over polytopes, such geometry is known to play an important role in the convergence of Frank–Wolfe type methods. Vanilla Frank–Wolfe can exhibit zig-zagging behavior near optimal faces, whereas away-step variants allow weight to be removed from previously selected atoms and can exploit the facial structure of the feasible set to achieve faster convergence (Lacoste-Julien & Jaggi, 2015; Wirth et al., 2025). While these results are specific to minimization, they suggest that away-step mechanisms may also provide leverage in minimax optimization. This motivates the following question:

Can fully projection-free minimax methods achieve sharper complexity guarantees by exploiting the geometry of polytopes?

We answer this question in the affirmative by developing a fully projection-free away-step Frank–Wolfe method based on a doubly smoothed formulation of Problem P. Specifically, double smoothing is introduced through auxiliary proximal variables on both the primal and dual sides, yielding strongly–convex–strongly–concave regularized subproblems and providing the flexibility to balance the primal–dual updates. For NC–C problems, we further introduce a quadratic perturbation on the dual side, whose magnitude balances the additional curvature needed in the analysis against the bias introduced to the original problem. The primal and dual variables are then updated jointly using away-step Frank–Wolfe updates over the product polytope $\mathcal{X} \times \mathcal{Y}$, exploiting the facial geometry of both feasible sets with only one LMO call over each set per iteration.

Integrating away-step Frank–Wolfe updates into the doubly smoothed framework is technically nontrivial. Our analysis is built on the Lyapunov function associated with the doubly smoothed formulation, but two additional difficulties arise. First, the resulting one-step estimate contains a primal–dual coupling term that cannot be directly absorbed by the Frank–Wolfe and proximal progress terms and requires an appropriate proximal error bound. Second, the Frank–Wolfe or away step may be clipped by its maximal feasible stepsize, in which case the proximal variables remain fixed. While clipped Frank–Wolfe steps still yield a uniform amount of Lyapunov decrease, clipped away steps need not. The number of clipped away iterations must therefore be controlled separately, which we achieve through an active-set counting argument.

These ingredients lead to the convergence guarantees summarized below.

Contributions. First, we develop a fully projection-free algorithm for smooth minimax optimization over polytopes that integrates away-step Frank–Wolfe updates into a perturbed doubly smoothed framework, while requiring only one LMO call over each feasible set per iteration. For NC–C problems, our method finds an ϵ-Frank–Wolfe stationary point in $\mathcal{O}(\epsilon^{-3})$ iterations, improving the

existing fully projection-free complexity of $\mathcal{O}(\epsilon^{-6})$. For C–C problems, we obtain an $\mathcal{O}(\epsilon^{-2})$ complexity, improving the previous $\mathcal{O}(\epsilon^{-5})$ bound.

Second, we extend the fully projection-free guarantees beyond the standard NC–C and C–C settings. Under a one-sided KŁ condition with exponent $\theta \in (0, 1)$, imposed on either the primal or dual problem, we obtain $\mathcal{O}(\epsilon^{-\max\{2,4\theta\}})$ complexity, matching the best-known projection-based rate. To the best of our knowledge, these are the first fully projection-free convergence guarantees for the corresponding NC–KŁ and KŁ–NC problems. We further obtain an $\mathcal{O}(\epsilon^{-3})$ guarantee for C–NC problems. A summary of these guarantees and their comparison with existing fully projection-free results is provided in Table 1. Numerical experiments on graphical games and robust budget allocation further demonstrate the practical performance of the proposed method.

2 Problem Setup

In this section, we introduce the structural assumptions and convergence criteria used throughout the paper. We first impose the following standard smoothness condition.

Assumption 2.1. The function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable, and its partial gradients are jointly L-Lipschitz continuous on $\mathcal{X} \times \mathcal{Y}$, i.e.,

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}', \boldsymbol{y}')\| \le L\big(\|\boldsymbol{x} - \boldsymbol{x}'\| + \|\boldsymbol{y} - \boldsymbol{y}'\|\big),$$
$$\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) - \nabla_{\boldsymbol{y}} f(\boldsymbol{x}', \boldsymbol{y}')\| \le L\big(\|\boldsymbol{x} - \boldsymbol{x}'\| + \|\boldsymbol{y} - \boldsymbol{y}'\|\big)$$

for all $(\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{X} \times \mathcal{Y}$. □

Beyond smoothness, we consider two structural conditions on the dual problem: concavity and the KŁ property.

Assumption 2.2. One of the following conditions is satisfied by the dual problem:

(i) (Concavity): For all $\boldsymbol{x} \in \mathcal{X}$, the function $f(\boldsymbol{x}, \cdot)$ is concave on $\mathcal{Y}$.

(ii) (KŁ property): For all $\boldsymbol{x} \in \mathcal{X}$, the maximization problem $\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$ has a nonempty solution set and a finite optimal value. Moreover, there exist constants $\tau_{\boldsymbol{y}} > 0$ and $\theta \in (0, 1)$ such that

$$\left(\max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}') - f(\boldsymbol{x}, \boldsymbol{y})\right)^{\theta} \le \frac{1}{\tau_{\boldsymbol{y}}} \operatorname{dist}\left(\mathbf{0}, -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) + \partial \mathbf{1}_{\mathcal{Y}}(\boldsymbol{y})\right)$$

for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$.

□

Under Assumption 2.2(i), we refer to the problem as nonconvex–concave (NC–C). If, in addition, $f(\cdot, \boldsymbol{y})$ is convex on $\mathcal{X}$ for every $\boldsymbol{y} \in \mathcal{Y}$, the problem is convex–concave (C–C). Under Assumption 2.2(ii), we refer to such problems as NC–KŁ problems. The symmetric primal-side conditions for KŁ–NC and C–NC problems are stated in Appendix J.

2.1 Stationarity and convergence certificates

We measure convergence throughout the paper using two-sided Frank–Wolfe stationarity. For comparison with prior work, we also introduce the saddle-point gap and the primal and dual optimality gaps used in existing fully projection-free analyses.

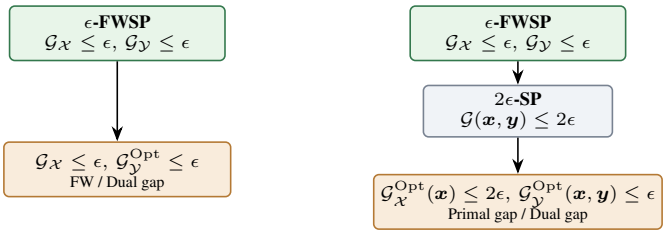
Definition 2.3 (Stationarity concepts)**.** Let $\epsilon \ge 0$ and $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$.

(i) The pair $(\boldsymbol{x}, \boldsymbol{y})$ is called an *ϵ-saddle point* (ϵ-SP) of Problem P if

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}') - \min_{\boldsymbol{x}' \in \mathcal{X}} f(\boldsymbol{x}', \boldsymbol{y}) \le \epsilon.$$

(ii) The pair $(\boldsymbol{x}, \boldsymbol{y})$ is called an *ϵ-Frank–Wolfe stationary point* (ϵ-FWSP) of Problem P if

$$\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \sup_{\boldsymbol{x}' \in \mathcal{X}} \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{x}' \rangle \le \epsilon, \qquad \text{and}$$
$$\mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \sup_{\boldsymbol{y}' \in \mathcal{Y}} \langle -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y} - \boldsymbol{y}' \rangle \le \epsilon.$$



(a) NC–C **(b) C–C**

Figure 1: Relations between our ϵ-FWSP guarantee and the criteria in (Giang-Tran et al., 2026). Green, orange, and gray boxes denote our guarantee, prior criteria, and the 2ϵ-SP guarantee, respectively. Arrows indicate implications; see Proposition 2.4.

□

The quantities $\mathcal{G}_{\mathcal{X}}$ and $\mathcal{G}_{\mathcal{Y}}$ are the Frank–Wolfe gaps associated with the primal and dual problems, respectively. For constrained smooth optimization, the Frank–Wolfe gap provides a natural measure of first-order non-stationarity, playing a role analogous to the gradient norm in the unconstrained setting; see, e.g., (Lacoste-Julien, 2016; Boroun et al., 2023). Accordingly, $\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y})$ measures first-order stationarity of $\min_{\boldsymbol{x}' \in \mathcal{X}} f(\boldsymbol{x}', \boldsymbol{y})$, while $\mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y})$ measures first-order stationarity of $\min_{\boldsymbol{y}' \in \mathcal{Y}} -f(\boldsymbol{x}, \boldsymbol{y}')$.

To compare this two-sided Frank–Wolfe criterion with prior fully projection-free guarantees in Table 1, we recall the optimality-gap criteria used in prior work. Under dual concavity, (Giang-Tran et al., 2026, Section 1.2.1) considers the dual optimality gap

$$\mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(\boldsymbol{x}, \boldsymbol{y}) := \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}') - f(\boldsymbol{x}, \boldsymbol{y}).$$

In the convex–concave setting, we also consider the primal optimality gap

$$\mathcal{G}_{\mathcal{X}}^{\mathrm{Opt}}(\boldsymbol{x}) := \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}') - \min_{\boldsymbol{x}' \in \mathcal{X}} \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}', \boldsymbol{y}').$$

The following proposition summarizes the relations among these certificates. Its proof is deferred to Appendix E.

Proposition 2.4 (Relations among convergence certificates)**.** Suppose Assumption 2.2(i) holds. For every $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$, the dual optimality gap satisfies

$$\mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(\boldsymbol{x}, \boldsymbol{y}) \le \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y}).$$

Consequently, an ϵ-FWSP also satisfies $\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) \le \epsilon, \mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(\boldsymbol{x}, \boldsymbol{y}) \le \epsilon$. If, in addition, $f(\cdot, \boldsymbol{y})$ is convex on $\mathcal{X}$ for every $\boldsymbol{y} \in \mathcal{Y}$, then

$$\max\left\{\mathcal{G}_{\mathcal{X}}^{\mathrm{Opt}}(\boldsymbol{x}), \mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(\boldsymbol{x}, \boldsymbol{y})\right\} \le \mathcal{G}(\boldsymbol{x}, \boldsymbol{y}) \le \mathcal{G}_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y}).$$

Hence every ϵ-FWSP satisfies $\mathcal{G}_{\mathcal{X}}^{\mathrm{Opt}}(\boldsymbol{x}) \le 2\epsilon, \mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(\boldsymbol{x}, \boldsymbol{y}) \le \epsilon$ and is a 2ϵ-saddle point. □

Figure 1 summarizes these implications and shows how the convergence criteria in Table 1 are related.

Remark 2.5. The converse implication in the C–C setting generally fails. In particular, simultaneous vanishing of the primal and dual optimality gaps need not imply saddle-point optimality of the corresponding primal–dual pair. A bilinear counterexample is given in Appendix E. □

3 Main Results

3.1 Algorithm

We now present our fully projection-free method. Motivated by smoothing techniques for minimax optimization (Zhang et al., 2020; Zheng et al., 2023; Li et al., 2025), we introduce auxiliary variables

$(\boldsymbol{z}, \boldsymbol{v})$ and define the regularized function

$$F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) \coloneqq f(\boldsymbol{x}, \boldsymbol{y}) + \frac{r_{\boldsymbol{x}}}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 - \frac{r_{\boldsymbol{y}}}{2}\|\boldsymbol{y} - \boldsymbol{v}\|^2 - \frac{\delta}{2}\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^2, \tag{1}$$

where $r_{\boldsymbol{x}}, r_{\boldsymbol{y}} > L$, $\delta \geq 0$, and $\bar{\boldsymbol{y}} \in \mathcal{Y}$ is fixed. The quadratic terms centered at $(\boldsymbol{z}, \boldsymbol{v})$ smooth the primal and dual subproblems, making them strongly convex and strongly concave, respectively, while the last term introduces a dual perturbation whose magnitude δ is chosen according to the problem structure. The symmetric primal-side variant, obtained by interchanging the primal and dual problems, is given in Appendix J.

Smoothing and perturbation play different roles in the analysis. For NC–C problems, recent work has shown that a suitably chosen dual perturbation can sharpen the stationarity complexity of smoothed projection-based methods (Li et al., 2026). This motivates us to study whether the perturbation mechanism remains effective when the primal and dual updates are replaced by Frank–Wolfe steps over polytopes.

To obtain a fully projection-free method, we use Frank–Wolfe updates for both the primal and dual variables. For a compact convex set $\mathcal{U}$ and a vector $\boldsymbol{c}$, let

$$\mathrm{LMO}_{\mathcal{U}}(\boldsymbol{c}) \in \arg\min_{\boldsymbol{u} \in \mathcal{U}} \langle \boldsymbol{c}, \boldsymbol{u} \rangle$$

denote a linear minimization oracle over $\mathcal{U}$. Each Frank–Wolfe update therefore requires only one call to $\mathrm{LMO}_{\mathcal{X}}$ and one call to $\mathrm{LMO}_{\mathcal{Y}}$, rather than projections onto the two feasible sets.

The difficulty is that vanilla Frank–Wolfe can suffer from zig-zagging and slow convergence over polytopes (Lacoste-Julien & Jaggi, 2015; Wolfe, 1970); the strong convexity and strong concavity introduced by smoothing do not by themselves eliminate this geometric issue. We therefore employ away-step Frank–Wolfe updates, which augment the usual Frank–Wolfe direction with an away direction selected from the current active set. By allowing weight to be removed from previously selected atoms, away steps exploit the facial geometry of the polytope while preserving the projection-free nature of the method (Lacoste-Julien & Jaggi, 2015; Rinaldi & Zeffiro, 2023).

To apply this mechanism simultaneously to the primal and dual variables, we perform an away-step Frank–Wolfe update over the product polytope $\mathcal{X} \times \mathcal{Y}$. At iteration t, define

$$\boldsymbol{u}_{\mathrm{pd}}^t \coloneqq [\boldsymbol{x}^t; \boldsymbol{y}^t], \qquad \boldsymbol{r}^t \coloneqq \left[\nabla_{\boldsymbol{x}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t); -\nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t)\right].$$

Given $\boldsymbol{r}^t$, the Frank–Wolfe atom is obtained by minimizing the linearized objective over $\mathcal{X} \times \mathcal{Y}$, whereas the away atom is obtained by maximizing it over the current active-set product $\mathcal{S}_{\boldsymbol{x}}^t \times \mathcal{S}_{\boldsymbol{y}}^t$. The method then selects whichever of the Frank–Wolfe and away directions yields the larger decrease in the linearized objective and uses a gap-adaptive stepsize. Apart from the two LMO calls, the remaining operations require only searches over the current active sets. The complete procedure is given in Algorithm 1.

Algorithm 1 uses singleton initial active sets for simplicity. Our analysis also permits initialization from given atomic decompositions, with unchanged convergence rates, since Lemma 4.3 allows $m_0 > 2$.

3.2 Convergence Guarantees

We first present the guarantees for NC–C and NC–KŁ problems and then give the sharper rate available in the C–C setting.

Theorem 3.1 (Convergence rates for NC–C and NC–KŁ problems)**.** Let $T > 1$. For each of the following settings, the algorithmic parameters can be chosen so that there exists $t \in \{0, \ldots, T-1\}$ for which the corresponding guarantee holds.

(i) (NC–C) Suppose Assumption 2.2(i) holds. If $\delta = \Theta(T^{-1/3})$ and $\beta_{\boldsymbol{x}} = \Theta(T^{-1/3})$, then $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}(T^{-1/3})$-FWSP of Problem P.

(ii) (NC–KŁ, $\theta \in (0, 1/2]$) Suppose Assumption 2.2(ii) holds with $\theta \in (0, 1/2]$. Let $\delta = 0$ and choose $\beta_{\boldsymbol{x}} > 0$ independently of T such that $\beta_{\boldsymbol{x}} \leq \frac{r_{\boldsymbol{y}}}{48\beta_{\boldsymbol{y}} r_{\boldsymbol{x}} \omega_1 \operatorname{diam}(\mathcal{Y})^{1/\theta - 2}}$. Then $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}(T^{-1/2})$-FWSP of Problem P.

Algorithm 1 Perturbed Doubly Smoothed Frank–Wolfe Method with Away-step

Input: Initial atomic points $\boldsymbol{x}^0 \in A$, $\boldsymbol{y}^0 \in B$, $\boldsymbol{z}^0 = \boldsymbol{x}^0$, $\boldsymbol{v}^0 = \boldsymbol{y}^0$, perturbation parameter $\delta \geq 0$, anchor point $\bar{\boldsymbol{y}} \in \mathcal{Y}$, and parameters $r_{\boldsymbol{x}}, r_{\boldsymbol{y}}, \beta_{\boldsymbol{x}}, \beta_{\boldsymbol{y}}, \bar{C}$. The active sets are initialized as $S_{\boldsymbol{x}}^0 = \{\boldsymbol{x}^0\}$ and $S_{\boldsymbol{y}}^0 = \{\boldsymbol{y}^0\}$.
for $t = 0, \ldots, T-1$ **do**
  Compute $\boldsymbol{r}^t$ and $\boldsymbol{u}_{\mathrm{pd}}^t$.
  Let $\boldsymbol{s}_{\min}^t \in \operatorname{argmin}_{\boldsymbol{s} \in A \times B} \langle \boldsymbol{r}^t, \boldsymbol{s} \rangle$, $\boldsymbol{s}_{\max}^t \in \operatorname{argmax}_{\boldsymbol{s} \in S_{\boldsymbol{x}}^t \times S_{\boldsymbol{y}}^t} \langle \boldsymbol{r}^t, \boldsymbol{s} \rangle$, and define:
    $\boldsymbol{d}_{\mathrm{FW}}^t = \boldsymbol{s}_{\min}^t - \boldsymbol{u}_{\mathrm{pd}}^t, \quad \boldsymbol{d}_{\mathrm{A}}^t = \boldsymbol{u}_{\mathrm{pd}}^t - \boldsymbol{s}_{\max}^t$.
  Set $g_{\mathrm{FW}}^t = \langle -\boldsymbol{r}^t, \boldsymbol{d}_{\mathrm{FW}}^t \rangle$, $g_{\mathrm{A}}^t = \langle -\boldsymbol{r}^t, \boldsymbol{d}_{\mathrm{A}}^t \rangle$,
  if $g_{\mathrm{FW}}^t \geq g_{\mathrm{A}}^t$ **then**
    $\boldsymbol{d}^t = \boldsymbol{d}_{\mathrm{FW}}^t$, $\gamma_{\max}^t = 1$, and $g^t = g_{\mathrm{FW}}^t$ {FW direction}
  else
    $\boldsymbol{d}^t = \boldsymbol{d}_{\mathrm{A}}^t$, $\gamma_{\max}^t = \min\left\{ \frac{\alpha_{\boldsymbol{x}}^t}{1-\alpha_{\boldsymbol{x}}^t}, \frac{\alpha_{\boldsymbol{y}}^t}{1-\alpha_{\boldsymbol{y}}^t} \right\}$, where $\alpha_{\boldsymbol{x}}^t, \alpha_{\boldsymbol{y}}^t$ are coefficients of away atoms. Let $g^t = g_{\mathrm{A}}^t$. {Away direction}
  end if
  Set $\widehat{\gamma}_t = \frac{1}{4\bar{C}} g^t$, $\gamma_t = \min\{\gamma_{\max}^t, \widehat{\gamma}_t\}$, and update $\boldsymbol{u}_{\mathrm{pd}}^{t+1} = \boldsymbol{u}_{\mathrm{pd}}^t + \gamma_t \boldsymbol{d}^t$.
  Update the atomic coefficients and the active sets $S_{\boldsymbol{x}}^{t+1} := \{\boldsymbol{v}_{\boldsymbol{x}} \in A \mid \alpha_{\boldsymbol{v}_{\boldsymbol{x}}}^{t+1} > 0\}$ and $S_{\boldsymbol{y}}^{t+1} := \{\boldsymbol{v}_{\boldsymbol{y}} \in B \mid \alpha_{\boldsymbol{v}_{\boldsymbol{y}}}^{t+1} > 0\}$.
  if $\gamma_{\max}^t < \widehat{\gamma}_t$ **then**
    $\boldsymbol{z}^{t+1} = \boldsymbol{z}^t$ and $\boldsymbol{v}^{t+1} = \boldsymbol{v}^t$ {clipped step}
  else
    $\boldsymbol{z}^{t+1} = \boldsymbol{z}^t + \beta_{\boldsymbol{x}}(\boldsymbol{x}^t - \boldsymbol{z}^t)$ and $\boldsymbol{v}^{t+1} = \boldsymbol{v}^t + \beta_{\boldsymbol{y}}(\boldsymbol{y}^t - \boldsymbol{v}^t)$ {non-clipped step}
  end if
end for

(iii) (NC–KŁ, $\theta \in (1/2, 1)$) Suppose Assumption 2.2(ii) holds with $\theta \in (1/2, 1)$. Set $\delta = 0$ and choose $\beta_{\boldsymbol{x}} = \Theta\left(T^{-(2\theta-1)/(2\theta)}\right)$. Then $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}(T^{-1/(4\theta)})$-FWSP of Problem P.

□

The perturbation plays different roles in the NC–C and NC–KŁ settings. In the NC–C setting, the quadratic perturbation makes the dual problem strongly concave and, together with smoothing, yields the $\mathcal{O}(T^{-1/3})$ convergence rate. Under the one-sided KŁ condition, the rate is instead governed by the corresponding proximal error bound, while an additional quadratic perturbation would introduce unnecessary bias. We therefore set $\delta = 0$ in the NC–KŁ cases.

The corresponding primal-side guarantees follow from the symmetric variant of Algorithm 1, obtained by interchanging the primal and dual problems. In particular, KŁ–NC and C–NC problems attain the same rates as their corresponding NC–KŁ and NC–C settings in Theorem 3.1. The formal statements, primal-side perturbation, and proofs are given in Appendix J.

When the primal problem is also convex, Algorithm 1 admits a sharper convergence rate.

Theorem 3.2 (Iteration complexity for C–C problems)**.** Let $T > 1$. Suppose Assumption 2.2(i) holds and $f(\cdot, \boldsymbol{y})$ is convex on $\mathcal{X}$ for every $\boldsymbol{y} \in \mathcal{Y}$. The algorithmic parameters can be chosen so that, with $\delta = \Theta(T^{-1/2})$ and $\beta_{\boldsymbol{x}} = \Theta(\delta) = \Theta(T^{-1/2})$, there exists an index $t \in \{0, 1, \ldots, T-1\}$ such that $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}(T^{-1/2})$-FWSP of Problem P. Consequently, by Proposition 2.4, $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is also an $\mathcal{O}(T^{-1/2})$-SP. □

4 Convergence Analysis

This section establishes the convergence results stated in Section 3.2. The analysis builds on the doubly smoothed Lyapunov construction introduced in Zheng et al. (2023), which is used throughout both the nonconvex and convex–concave analyses. We first analyze the NC–C and NC–KŁ settings; the corresponding KŁ–NC and C–NC guarantees follow by interchanging the roles of the primal and dual problems. For the C–C setting, we use the same Lyapunov function. The additional primal

convexity allows the Lyapunov function to be bounded in terms of the Frank–Wolfe gap and the residuals associated with the two proximal updates, leading to a sharper convergence rate.

4.1 Nonconvex Problems

We first consider the NC–C and NC–KŁ settings. Following Zheng et al. (2023), we define the Lyapunov function $\Phi : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ as

$$\begin{aligned}
&\Phi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) \\
&:= \underbrace{F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) - d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})}_{\text{Primal descent}} + \underbrace{p_\delta(\boldsymbol{z}, \boldsymbol{v}) - d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})}_{\text{Dual ascent}} + \underbrace{q_\delta(\boldsymbol{z}) - p_\delta(\boldsymbol{z}, \boldsymbol{v})}_{\text{Proximal ascent}} + \underbrace{q_\delta(\boldsymbol{z}) - \underline{F}}_{\text{Proximal descent}} + \underline{F} \\
&= F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) - 2d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) + 2q_\delta(\boldsymbol{z}), \qquad (2)
\end{aligned}$$

where $d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) := \min_{\boldsymbol{x} \in \mathcal{X}} F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, $p_\delta(\boldsymbol{z}, \boldsymbol{v}) := \max_{\boldsymbol{y} \in \mathcal{Y}} d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, $q_\delta(\boldsymbol{z}) := \max_{\boldsymbol{v} \in \mathbb{R}^m} p_\delta(\boldsymbol{z}, \boldsymbol{v})$ and $\underline{F} := \min_{\boldsymbol{z} \in \mathbb{R}^n} q_\delta(\boldsymbol{z})$. The decomposition separates the Lyapunov function into four nonnegative components associated with the primal, dual, and proximal updates, allowing their one-step changes to be analyzed separately. In particular, $\Phi \geq \underline{F}$.

Recall that iteration t is non-clipped if $\widehat{\gamma}_t \leq \gamma_{\max}^t$, and clipped otherwise. The following proposition gives the one-step estimate for the Lyapunov function.

Proposition 4.1 (One-step Lyapunov estimate)**.** Suppose that $0 \leq \delta \leq \frac{L}{3}$, $r_{\boldsymbol{y}} \geq 2L$, $r_{\boldsymbol{x}} \geq \max\left\{2L,\, L + \frac{8L \operatorname{diam}(\mathcal{Y})}{\operatorname{PWidth}(A)}\right\}$ and define $\bar{C} := \frac{3r_{\boldsymbol{x}}}{2} \operatorname{diam}(\mathcal{X})^2 + 6r_{\boldsymbol{y}} \operatorname{diam}(\mathcal{Y})^2$. Let

$$0 < \beta_{\boldsymbol{y}} \leq \min\left\{\frac{1}{5},\, \frac{L^2 \operatorname{PWidth}(B)^2}{18432\, \bar{C} r_{\boldsymbol{y}}}\right\}, \qquad \text{and}$$

$$0 < \beta_{\boldsymbol{x}} \leq \min\left\{\frac{r_{\boldsymbol{y}}}{1536 r_{\boldsymbol{x}}},\, \frac{4L^2 r_{\boldsymbol{x}}}{80L^2 r_{\boldsymbol{x}} + 9r_{\boldsymbol{y}}(2r_{\boldsymbol{x}} + L)^2},\, \frac{L^2 \operatorname{PWidth}(B)^2}{221184\, \bar{C} r_{\boldsymbol{x}}}\right\}.$$

Then, at every non-clipped iteration t,

$$\begin{aligned}
\Phi^t - \Phi^{t+1} \geq \frac{1}{32\bar{C}}(g^t)^2 + \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{r_{\boldsymbol{y}}}{4\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 \\
- 6r_{\boldsymbol{x}}\beta_{\boldsymbol{x}}\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2, \qquad (3)
\end{aligned}$$

where $\boldsymbol{y}(\boldsymbol{z}, \boldsymbol{v}) \in \arg\max_{\boldsymbol{y} \in \mathcal{Y}} d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, $\boldsymbol{x}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, and $\boldsymbol{x}(\boldsymbol{z}, \boldsymbol{v}) := \boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}, \boldsymbol{v}), \boldsymbol{z}, \boldsymbol{v})$. Moreover, $\boldsymbol{v}(\boldsymbol{z}) \in \arg\max_{\boldsymbol{v} \in \mathbb{R}^m} p_\delta(\boldsymbol{z}, \boldsymbol{v})$ and $\boldsymbol{v}_+^t(\boldsymbol{z}) := \boldsymbol{v}^t + \beta_{\boldsymbol{y}}\,(\boldsymbol{y}(\boldsymbol{z}, \boldsymbol{v}^t) - \boldsymbol{v}^t)$.

At every clipped iteration t, we instead have

$$\Phi^t - \Phi^{t+1} > \frac{\gamma_t}{4} g^t \geq 0. \qquad (4)$$

□

The one-step estimates in Proposition 4.1 reveal two distinct difficulties in the convergence analysis.

At a non-clipped iteration, the estimate equation 3 follows the Lyapunov structure of the doubly smoothed analysis in Zheng et al. (2023), adapted to the away-step Frank–Wolfe updates considered here. The first three terms on the right-hand side quantify the progress of the Frank–Wolfe and proximal updates, while the last negative term arises from the primal–dual coupling. Therefore, equation 3 does not by itself guarantee Lyapunov descent, and the coupling term must be controlled using additional structure of the dual problem.

Clipping creates a separate difficulty. If the trial stepsize exceeds the maximal feasible stepsize, the step is clipped and the proximal variables are kept fixed. Although equation 4 shows that such an iteration does not increase the Lyapunov function, the two types of clipped iterations behave differently: a clipped Frank–Wolfe step yields a decrease uniformly bounded away from zero, whereas a clipped away step need not. The clipped Frank–Wolfe iterations can therefore be controlled directly through the total Lyapunov decrease, while the clipped away iterations require a separate argument based on the evolution of the active sets.

We first control the coupling term in equation 3. Specifically, we bound the primal–dual coupling term in terms of $\|\boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}) - \boldsymbol{v}^t\|$. Under the KŁ condition, the required bound follows from Zheng et al. (2023, Proposition 2). For a concave dual problem, we further exploit the strong concavity induced by the quadratic perturbation to obtain a stronger quadratic bound.

Proposition 4.2 (Proximal error bound)**.** At every non-clipped iteration t, the following bounds hold.

(i) Suppose that Assumption 2.2(i) holds and $\delta > 0$. Then

$$\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1}))\|^2 \le \omega_0 \|\boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}) - \boldsymbol{v}^t\|^2,$$

where $\omega_0 := \frac{2r_{\boldsymbol{y}}(r_{\boldsymbol{y}}+\delta)}{(r_{\boldsymbol{x}}-L)\delta}\left(\frac{1-\beta_{\boldsymbol{y}}}{\beta_{\boldsymbol{y}}} + \frac{r_{\boldsymbol{y}}}{r_{\boldsymbol{y}}-L+\delta}\right)^2$. In particular, for $0 < \delta \le \bar{\delta}$, we have $\omega_0 \le \frac{C_{\mathrm{sc}}}{\delta}$, where $\bar{\delta} > 0$ and $C_{\mathrm{sc}} > 0$ are independent of δ and T.

(ii) Suppose that Assumption 2.2(ii) holds and $\delta = 0$. Then

$$\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1}))\|^2 \le \omega_1 \|\boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}) - \boldsymbol{v}^t\|^{1/\theta},$$

where $\omega_1 := \frac{2}{(r_{\boldsymbol{x}}-L)\tau_{\boldsymbol{y}}^{1/\theta}}\left(\frac{r_{\boldsymbol{y}}(1-\beta_{\boldsymbol{y}})}{\beta_{\boldsymbol{y}}} + \frac{r_{\boldsymbol{y}}^2}{r_{\boldsymbol{y}}-L}\right)^{1/\theta}$.

□

The two bounds are used differently in the subsequent rate analysis. Under dual concavity, Proposition 4.2(i) with $\omega_0 = \Theta(1/\delta)$ allows the coupling term in equation 3 to be absorbed by taking $\beta_{\boldsymbol{x}} \le \mathcal{O}(\delta)$. Under the KŁ condition, Proposition 4.2(ii) gives a Hölder-type bound. When $\theta \le 1/2$, boundedness of $\mathcal{Y}$ converts this bound into a quadratic one, so the coupling term can again be absorbed. The case $\theta > 1/2$ requires a different argument and is treated in Appendix H.3.

We next control the second difficulty, namely the clipped away iterations. They are controlled by the following active-set counting bound; its proof is deferred to Appendix H.2.

Lemma 4.3 (Active-set counting)**.** Let $N_{\mathrm{nc}}(T)$, $N_{\mathrm{cFW}}(T)$, and $N_{\mathrm{cA}}(T)$ denote the numbers of non-clipped, clipped Frank–Wolfe, and clipped away iterations, respectively, among the first T iterations, and let $m_0 := |S_{\boldsymbol{x}}^0| + |S_{\boldsymbol{y}}^0|$. Then, $N_{\mathrm{cA}}(T) \le 2\big(N_{\mathrm{nc}}(T) + N_{\mathrm{cFW}}(T)\big) + m_0 - 2$. Consequently,

$$N_{\mathrm{nc}}(T) + N_{\mathrm{cFW}}(T) \ge \frac{T - m_0 + 2}{3}.$$

□

A clipped Frank–Wolfe iteration has $\gamma_t = 1$ and $g^t > 4\bar{C}$, so equation 4 gives $\Phi^t - \Phi^{t+1} > \bar{C}$. Once the coupling term is controlled, all non-clipped and clipped away iterations are non-increasing in Φ; thus $N_{\mathrm{cFW}}(T) \le (\Phi^0 - \underline{F})/\bar{C} = O(1)$, and Lemma 4.3 gives $N_{\mathrm{nc}}(T) = \Omega(T)$. Combining these estimates with the stationarity transfer argument in Appendix E.1 yields Theorem 3.1; the complete rate derivation is given in Appendix H.3.

4.2 Convex–Concave Problems

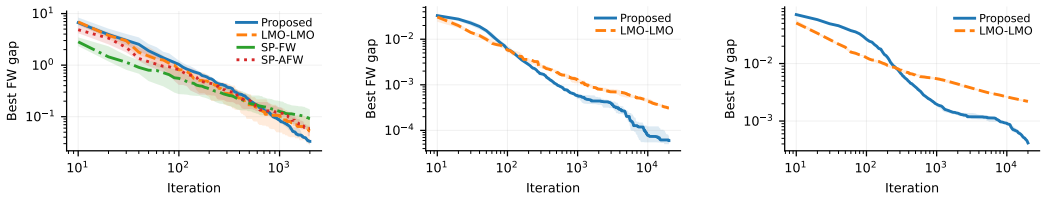
We now consider the C–C setting. Since it is a special case of the NC–C setting, the one-step Lyapunov estimates, the control of the primal–dual coupling term, and the active-set counting argument developed above remain valid. Here, the additional primal convexity allows the Lyapunov function to be bounded above by the Frank–Wolfe gap and the two proximal-update terms in the one-step estimate, leading to a sharper rate.

Define the shifted Lyapunov function $\widetilde{\Phi}^t := \Phi^t - \underline{F}$. Since $\Phi^t \ge \underline{F}$, we have $\widetilde{\Phi}^t \ge 0$, and all one-step estimates for Φ^t hold unchanged for $\widetilde{\Phi}^t$.

Lemma 4.4 (Lyapunov control under primal convexity)**.** Suppose that $f(\cdot, \boldsymbol{y})$ is convex on $\mathcal{X}$ for every $\boldsymbol{y} \in \mathcal{Y}$ and $f(\boldsymbol{x}, \cdot)$ is concave on $\mathcal{Y}$ for every $\boldsymbol{x} \in \mathcal{X}$. At every non-clipped iteration t, there exists a constant $C_{\mathrm{cc}} > 0$, independent of T, such that

$$\widetilde{\Phi}^t \le C_{\mathrm{cc}}\left(g^t + \frac{\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|}{\beta_{\boldsymbol{x}}} + \frac{\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|}{\beta_{\boldsymbol{y}}}\right).$$

□



(a) Graphical matching. (b) Budget allocation 50x50. (c) Budget allocation 100x100.

Figure 2: Best two-sided Frank–Wolfe stationarity gap, $\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y})$, versus iteration count. Solid lines show the median over 10 instances, with shaded interquartile ranges.

The proof is given in Appendix I. The first three terms in the Lyapunov decomposition are controlled using properties already available in the NC–C analysis. The additional primal convexity makes $p_\delta(\cdot, \boldsymbol{v})$ convex, which allows the remaining term $q_\delta(\boldsymbol{z}^t) - \underline{F}$ to be controlled by the Frank–Wolfe gap and the two proximal updates.

Taking $\beta_{\boldsymbol{x}} = c_{\boldsymbol{x}}\delta$ with $c_{\boldsymbol{x}} > 0$ sufficiently small, the coupling term is absorbed as in the NC–C analysis. Combining the resulting descent estimate with Lemma 4.4 gives

$$\widetilde{\Phi}^t - \widetilde{\Phi}^{t+1} \geq c\,\beta_{\boldsymbol{x}}(\widetilde{\Phi}^t)^2$$

at every non-clipped iteration, for some constant $c > 0$ independent of T. Together with the clipped-step bounds and Lemma 4.3, the standard reciprocal argument and the stationarity transfer argument in Appendix E.1 yield, for some $t < T$,

$$\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}^t, \boldsymbol{y}^t) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}^t, \boldsymbol{y}^t) = O\left(\frac{1}{\beta_{\boldsymbol{x}} T} + \delta\right).$$

With $\beta_{\boldsymbol{x}} = \Theta(\delta)$, choosing $\delta = \Theta(T^{-1/2})$ gives the $O(T^{-1/2})$ rate in Theorem 3.2.

5 Numerical Experiments

We evaluate the proposed algorithm on graphical matching and robust budget allocation problems, representing the convex–concave and nonconvex–concave settings, respectively. Performance is measured by the best two-sided Frank–Wolfe stationarity gap $G_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) + G_{\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y})$. For graphical matching, we compare with `LMO-LMO` (Giang-Tran et al., 2026), `SP-FW`, and `SP-AFW` (Gidel et al., 2017). For robust budget allocation, we compare only with `LMO-LMO`, since the methods of Gidel et al. (2017) do not provide guarantees for nonconvex–concave problems. As shown in Figure 2, our method is competitive on graphical matching and achieves smaller stationarity gaps than `LMO-LMO` on both robust budget allocation problems. Additional experimental details are provided in Appendix K.

6 Conclusion

In this paper, we presented a fully projection-free algorithm for minimax problems over polytopes that combines smoothing, quadratic perturbation, and away-step Frank–Wolfe updates. By exploiting the polyhedral geometry of the feasible sets, our algorithm achieves sharper iteration complexity guarantees. In particular, we show iteration complexities of $\mathcal{O}(\epsilon^{-3})$ for NC–C and C–NC problems and $\mathcal{O}(\epsilon^{-2})$ for C–C problems. Under a one-sided Kurdyka–Łojasiewicz condition with exponent θ, we obtain an $\mathcal{O}(\epsilon^{-\max\{2,4\theta\}})$ complexity. Experiments on graphical matching and robust budget allocation problems illustrate the method's performance in practice.

Two natural directions for future work are as follows. First, our current guarantees use horizon-dependent parameter choices, and it would be interesting to develop anytime variants that retain the improved fully projection-free rates. Second, although the proposed framework matches the corresponding projection-based complexity orders in several settings considered here, it remains open whether the NC–C and C–C rates can be further sharpened toward the best projection-based guarantees.

AI USE STATEMENT

In this work, we used generative AI tools for checking mathematical derivations and for assisting with code debugging and modification. We have not used generative AI tools to develop the core algorithmic ideas, theoretical results, or proofs, which were developed by the authors, and the remaining tasks requiring disclosure are not applicable to this work. Additionally, we used generative AI tools for language editing and for improving the presentation and organization of the manuscript. We have reviewed all AI-assisted work. In particular, all AI-assisted mathematical suggestions and code modifications were independently checked and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical and methodological. The numerical experiments use synthetically generated problem instances and do not involve human subjects, personal data, or sensitive datasets. We are not aware of any specific ethical concerns associated with this work.

REPRODUCIBILITY STATEMENT

The proposed algorithm and its parameter choices are fully specified in the main text. Complete proofs of all theoretical results are provided in the appendices. Detailed descriptions of the benchmark problems, instance generation, initialization, parameter settings, and implementation details for the numerical experiments are provided in Appendix K.

REFERENCES

Morteza Boroun, Erfan Yazdandoost Hamedani, and Afrooz Jalilzadeh. Projection-free methods for solving nonconvex-concave saddle point problems. *Advances in Neural Information Processing Systems*, 36:53844–53856, 2023.

Cyrille W Combettes and Sebastian Pokutta. Complexity of linear minimization and projection on some sets. *Operations Research Letters*, 49(4):565–571, 2021.

Jack Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965. doi: 10.4153/CJM-1965-045-4.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

Khanh-Hung Giang-Tran, Soroosh Shafiee, and Nam Ho-Nguyen. Projection-free algorithms for minimax problems. In *Forty-third International Conference on Machine Learning*, 2026.

Gauthier Gidel, Tony Jebara, and Simon Lacoste-Julien. Frank-wolfe algorithms for saddle point problems. In *Artificial Intelligence and Statistics*, pp. 362–371. PMLR, 2017.

Jacques GuéLat and Patrice Marcotte. Some comments on wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119, May 1986. ISSN 1436-4646.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman (eds.), *Proceedings of the 7th Python in Science Conference*, pp. 11 – 15, Pasadena, CA USA, 2008.

Janice H Hammond. *Solving asymmetric variational inequality problems and systems of equations with generalized nonlinear programming algorithms*. PhD thesis, Massachusetts Institute of Technology, 1984.

Matthew Hough. Convergence of the frank-wolfe algorithm for monotone variational inequalities, 2026. URL https://arxiv.org/abs/2603.12104.

Simon Lacoste-Julien. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. *Advances in neural information processing systems*, 28, 2015.

Jiajin Li, Linglingzhi Zhu, and Anthony Man-Cho So. Nonsmooth nonconvex–nonconcave minimax optimization: Primal–dual balancing and iteration complexity analysis. *Mathematical Programming*, 214(1):591–641, 2025.

Jiajin Li, Mahesh Nagarajan, Siyu Pan, and Nanxi Zhang. Smoothing meets perturbation: Unified and tight analysis for nonconvex-concave minimax optimization. *arXiv preprint arXiv:2602.14185*, 2026.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International conference on machine learning*, pp. 6083–6093. PMLR, 2020a.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on learning theory*, pp. 2738–2779. PMLR, 2020b.

Marien Renaud, Arthur Leclaire, and Nicolas Papadakis. On the moreau envelope properties of weakly convex functions. *arXiv preprint arXiv:2509.13960*, 2025.

Francesco Rinaldi and Damiano Zeffiro. Avoiding bad steps in frank-wolfe variants. *Computational Optimization and Applications*, 84(1):225–264, 2023.

Bryan Wilder. Equilibrium computation and robust optimization in zero sum games with submodular structure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, 2018.

Elias Wirth, Javier Pena, and Sebastian Pokutta. Fast convergence of frank-wolfe algorithms on polytopes, 2025. URL `https://arxiv.org/abs/2406.18789`.

Philip Wolfe. Convergence theory in nonlinear programming. In J. Abadie (ed.), *Integer and Nonlinear Programming*, pp. 1–36. North-Holland, Amsterdam, 1970.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex–concave and convex–nonconcave minimax problems. *Mathematical Programming*, 201(1):635–706, 2023.

Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in neural information processing systems*, 33:7377–7389, 2020.

Taoli Zheng, Linglingzhi Zhu, Anthony Man-Cho So, José Blanchet, and Jiajin Li. Universal gradient descent ascent method for nonconvex-nonconcave minimax optimization. *Advances in neural information processing systems*, 36:54075–54110, 2023.

A Appendix

Organization of the appendix. The appendix is organized as follows.

- Appendix B provides additional discussion of related work.
- Appendix C collects the Frank–Wolfe geometric tools used throughout the analysis, including the relations among the Frank–Wolfe, away, and pairwise Frank–Wolfe gaps and the pyramidal-width error bound.
- Appendix D summarizes the value functions and solution mappings associated with the regularized minimax problem and establishes the Lipschitz and Frank–Wolfe error bounds used later in the analysis.
- Appendix E develops the stationarity relations used in the main text, including the transfer from the algorithmic residual to Frank–Wolfe stationarity of the original problem, and proves Proposition 2.4.
- Appendix F derives the one-step Lyapunov estimates for non-clipped and clipped iterations and proves Proposition 4.1.

- Appendix G establishes the proximal error bounds used to control the primal–dual coupling term and proves Proposition 4.2.
- Appendix H combines these estimates with the clipped-iteration analysis and active-set counting argument. It proves Lemma 4.3 and then Theorem 3.1.
- Appendix I establishes the additional Lyapunov–residual relation available in the C–C setting and uses it to prove Theorem 3.2.
- Appendix J derives the corresponding KŁ–NC and C–NC guarantees by interchanging the primal and dual problems and proves Theorem J.1.
- Appendix K provides additional details for the numerical experiments.

B Additional Related work

Away-step variants of the classical Frank–Wolfe algorithm (Frank & Wolfe, 1956) were developed to mitigate zig-zagging in convex minimization over polytopes. The modification was first proposed by Wolfe (1970), subsequently corrected and analyzed by GuéLat & Marcotte (1986), and later given modern global linear convergence guarantees by Lacoste-Julien & Jaggi (2015).

Extending the Frank–Wolfe algorithm to minimax problems and more general variational inequalities has proved difficult. Hammond (1984) proves asymptotic convergence of the Frank–Wolfe algorithm for monotone variational inequalities over strongly convex sets on which the operator does not vanish, and conjectures convergence over polytopes under strong monotonicity. Gidel et al. (2017) consider Hammond's algorithm for the C–C setting, establishing convergence rates when the feasible set is strongly convex and the gradient does not vanish over the feasible set. They also partially resolve Hammond's conjecture by establishing convergence of Hammond's algorithm and an away-step variant for strongly convex–strongly concave objectives over polytopes under a restrictive condition involving the pyramidal widths of the feasible sets. Hough (2026) resolves Hammond's conjecture by proving asymptotic convergence of the Frank–Wolfe algorithm for monotone variational inequalities over general compact convex sets. This also establishes convergence of the corresponding algorithm studied by Gidel et al. (2017) in the C–C setting without requiring the additional geometric assumptions used in their polytope analysis.

C Frank–Wolfe Geometry over Polytopes

In this section, we collect several useful inequalities for Frank–Wolfe methods applied to the general constrained minimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}),$$

where $\mathcal{X} = \operatorname{conv}(A)$ and A is a finite set. We assume that f is continuously differentiable. These results will later be applied to the minimization objective induced by our minimax formulation.

Let $\mathbf{x}^{(t)}$ be the iterate at step t, and let $\mathcal{S}^{(t)} \subseteq A$ denote the active set maintained by the algorithm, i.e., the subset of atoms used to represent $\mathbf{x}^{(t)}$ as a convex combination.

We first recall the definitions of the standard Frank–Wolfe vertex $\mathbf{s}^{(t)}$ and the away vertex $\mathbf{v}^{(t)}$:

$$\mathbf{s}^{(t)} \in \arg\min_{\mathbf{s}\in A}\langle \mathbf{s}, \nabla f(\mathbf{x}^{(t)})\rangle \quad \text{and} \quad \mathbf{v}^{(t)} \in \arg\max_{\mathbf{v}\in\mathcal{S}^{(t)}} \langle \mathbf{v}, \nabla f(\mathbf{x}^{(t)})\rangle.$$

We define the corresponding search directions as

$$\boldsymbol{d}_{\mathrm{FW}}^{(t)} \coloneqq \mathbf{s}^{(t)} - \mathbf{x}^{(t)}, \quad \boldsymbol{d}_{\mathrm{A}}^{(t)} \coloneqq \mathbf{x}^{(t)} - \mathbf{v}^{(t)}, \quad \text{and} \quad \boldsymbol{d}_{\mathrm{PFW}}^{(t)} \coloneqq \mathbf{s}^{(t)} - \mathbf{v}^{(t)}.$$

The associated Frank–Wolfe, away, and pairwise Frank–Wolfe gaps are defined by

$$g_{\mathrm{FW}}^{t} \coloneqq \langle \boldsymbol{d}_{\mathrm{FW}}^{(t)}, -\nabla f(\mathbf{x}^{(t)})\rangle, \quad g_{\mathrm{A}}^{t} \coloneqq \langle \boldsymbol{d}_{\mathrm{A}}^{(t)}, -\nabla f(\mathbf{x}^{(t)})\rangle,$$

and

$$g_{\mathrm{PFW}}^{t} \coloneqq \langle \boldsymbol{d}_{\mathrm{PFW}}^{(t)}, -\nabla f(\mathbf{x}^{(t)})\rangle.$$

In the Frank–Wolfe algorithm with away step (AFW), the search direction is chosen between the Frank–Wolfe and away directions according to the larger linearized descent. Accordingly, we define the AFW gap as $g_t \coloneqq \max\{g_{\text{FW}}^t, g_{\text{A}}^t\}$. The AFW gap and the pairwise Frank–Wolfe gap are equivalent up to a factor of two.

Lemma C.1 (Away-step gap bounds)**.** For the AFW algorithm, we have $\frac{1}{2} g_{\text{PFW}}^t \leq g_t \leq g_{\text{PFW}}^t$. □

Proof. By the definitions of the three directions, $\boldsymbol{d}_{\text{PFW}}^{(t)} = \boldsymbol{d}_{\text{FW}}^{(t)} + \boldsymbol{d}_{\text{A}}^{(t)}$. Hence, $g_{\text{PFW}}^t = g_{\text{FW}}^t + g_{\text{A}}^t$. Moreover, both g_{FW}^t and g_{A}^t are nonnegative. Therefore,

$$\max\left\{g_{\text{FW}}^t, g_{\text{A}}^t\right\} \leq g_{\text{FW}}^t + g_{\text{A}}^t \leq 2\max\left\{g_{\text{FW}}^t, g_{\text{A}}^t\right\},$$

which finishes the proof. □

The key geometric quantity in our analysis is the pyramidal width $\mathrm{PWidth}(A)$ introduced in (Lacoste-Julien & Jaggi, 2015). Informally, it is obtained by taking the worst-case directional width over all faces of the polytope, all feasible directions within each face, and all admissible active-set representations; we refer to (Lacoste-Julien & Jaggi, 2015) for the formal definition.

Geometrically, $\mathrm{PWidth}(A)$ captures the worst-case directional geometry induced by the faces and active sets of the polytope. In contrast to the diameter, which only measures the global size of the feasible set, the pyramidal width captures additional geometric information about its faces and the corresponding away and pairwise Frank–Wolfe directions.

For a finite atom set A generating a polytope in its affine hull, $\mathrm{PWidth}(A)$ is strictly positive. Moreover, (Lacoste-Julien & Jaggi, 2015, Lemma 4) shows that for the unit cube $\mathcal{X} = [0,1]^d$ with vertex set $A = \{0,1\}^d$,

$$\mathrm{PWidth}(A) = \frac{1}{\sqrt{d}}.$$

C.1 Error Bound Conditions

We next relate the Frank–Wolfe gap to a first-order optimality residual. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex, and define $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$. Under strong convexity, the distance from an iterate to the unique minimizer can be controlled by the first-order optimality residual:

$$\|\boldsymbol{x} - \boldsymbol{x}^\star\| \leq \frac{1}{\mu}\operatorname{dist}\left(\mathbf{0},\, \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) + \partial \boldsymbol{I}_{\mathcal{X}}(\boldsymbol{x})\right).$$

The first-order optimality residual on the right-hand side is not directly available from a Frank–Wolfe update. We therefore relate it to the pairwise Frank–Wolfe gap, which is available from the Frank–Wolfe and away directions. The required connection is provided by the pyramidal width of the polytope.

Lemma C.2 ((Rinaldi & Zeffiro, 2023, Lemma 3.1))**.** Let $A \subset \mathbb{R}^n$ be a finite set and let $\boldsymbol{x} \in \mathrm{conv}(A)$. Suppose that f is continuously differentiable at $\boldsymbol{x}$. Then

$$\operatorname{dist}\left(\mathbf{0}, \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) + \partial \boldsymbol{I}_{\mathrm{conv}(A)}(\boldsymbol{x})\right) \leq \frac{g_{\text{PFW}}}{\mathrm{PWidth}(A)}.$$

□

D Properties of the Regularized Problem

For reference, Table 2 summarizes the value functions and solution mappings associated with F_δ that are used throughout the analysis.

Since $r_{\boldsymbol{x}} > L$ and $r_{\boldsymbol{y}} > L$, for every fixed $(\boldsymbol{z}, \boldsymbol{v})$, the function $F_\delta(\cdot, \cdot, \boldsymbol{z}, \boldsymbol{v})$ is strongly convex in $\boldsymbol{x}$ and strongly concave in $\boldsymbol{y}$. Hence, the two representations of $p_\delta(\boldsymbol{z}, \boldsymbol{v})$ in Table 2 coincide.

We first collect several Lipschitz properties of the solution mappings associated with the regularized problem. These properties will be used repeatedly in the subsequent convergence analysis.

Table 2: Notation for the regularized problems and their solution mappings.

Problem	Value	Solution
$\min_{\boldsymbol{x}\in\mathcal{X}} F_\delta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{v})$	$d_\delta(\boldsymbol{y},\boldsymbol{z},\boldsymbol{v})$	$\boldsymbol{x}(\boldsymbol{y},\boldsymbol{z},\boldsymbol{v})$
$\max_{\boldsymbol{y}\in\mathcal{Y}} F_\delta(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{v})$	$h_\delta(\boldsymbol{x},\boldsymbol{z},\boldsymbol{v})$	$\boldsymbol{y}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{v})$
$\max_{\boldsymbol{y}\in\mathcal{Y}} d_\delta(\boldsymbol{y},\boldsymbol{z},\boldsymbol{v})$	$p_\delta(\boldsymbol{z},\boldsymbol{v})$	$\boldsymbol{y}(\boldsymbol{z},\boldsymbol{v})$
$\min_{\boldsymbol{x}\in\mathcal{X}} h_\delta(\boldsymbol{x},\boldsymbol{z},\boldsymbol{v})$	$p_\delta(\boldsymbol{z},\boldsymbol{v})$	$\boldsymbol{x}(\boldsymbol{z},\boldsymbol{v})$
$\max_{\boldsymbol{v}\in\mathbb{R}^m} p_\delta(\boldsymbol{z},\boldsymbol{v})$	$q_\delta(\boldsymbol{z})$	$\boldsymbol{v}(\boldsymbol{z})$

Lemma D.1 ((Zheng et al., 2023)[Lemma 2])**.** Suppose $r_{\boldsymbol{x}} > L, r_{\boldsymbol{y}} > L$. Then, for any $\boldsymbol{x},\boldsymbol{x}' \in \mathcal{X}$, $\boldsymbol{y},\boldsymbol{y}' \in \mathcal{Y}$, $\boldsymbol{z},\boldsymbol{z}' \in \mathbb{R}^n$ and $\boldsymbol{v},\boldsymbol{v}' \in \mathbb{R}^m$, the following inequalities hold:

(i) $\|\boldsymbol{x}(\boldsymbol{y}',\boldsymbol{z},\boldsymbol{v}) - \boldsymbol{x}(\boldsymbol{y},\boldsymbol{z},\boldsymbol{v})\| \le \sigma_1\|\boldsymbol{y}' - \boldsymbol{y}\|$,

(ii) $\|\boldsymbol{x}(\boldsymbol{y},\boldsymbol{z}',\boldsymbol{v}) - \boldsymbol{x}(\boldsymbol{y},\boldsymbol{z},\boldsymbol{v})\| \le \sigma_2\|\boldsymbol{z} - \boldsymbol{z}'\|$,

(iii) $\|\boldsymbol{x}(\boldsymbol{z}',\boldsymbol{v}) - \boldsymbol{x}(\boldsymbol{z},\boldsymbol{v})\| \le \sigma_2\|\boldsymbol{z} - \boldsymbol{z}'\|$,

(iv) $\|\boldsymbol{y}(\boldsymbol{z},\boldsymbol{v}) - \boldsymbol{y}(\boldsymbol{z}',\boldsymbol{v})\| \le \sigma_3\|\boldsymbol{z} - \boldsymbol{z}'\|$,

(v) $\|\boldsymbol{y}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}) - \boldsymbol{y}(\boldsymbol{x}',\boldsymbol{z},\boldsymbol{v})\| \le \sigma_4\|\boldsymbol{x} - \boldsymbol{x}'\|$,

(vi) $\|\boldsymbol{y}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}) - \boldsymbol{y}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{v}')\| \le \sigma_5\|\boldsymbol{v} - \boldsymbol{v}'\|$,

(vii) $\|\boldsymbol{y}(\boldsymbol{z},\boldsymbol{v}) - \boldsymbol{y}(\boldsymbol{z},\boldsymbol{v}')\| \le \sigma_5\|\boldsymbol{v} - \boldsymbol{v}'\|$,

where $\sigma_1 = \sigma_2 = \frac{r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}-L}$, $\sigma_3 = \frac{r_{\boldsymbol{x}}\sigma_1}{r_{\boldsymbol{y}}-L+\delta} + 1$, $\sigma_4 = \frac{r_{\boldsymbol{y}}+\delta}{r_{\boldsymbol{y}}-L+\delta}$, and $\sigma_5 = \frac{r_{\boldsymbol{y}}}{r_{\boldsymbol{y}}-L+\delta}$. □

The next result gives the smoothness of the dual function d_δ with respect to the $\boldsymbol{y}$ variable.

Lemma D.2 ((Zheng et al., 2023)[Lemma 3])**.** For any fixed $\boldsymbol{z} \in \mathbb{R}^n$, $\boldsymbol{v} \in \mathbb{R}^m$, the dual function $d_\delta(\cdot,\boldsymbol{z},\boldsymbol{v})$ is continuously differentiable with the gradient $\nabla_{\boldsymbol{y}} d_\delta(\boldsymbol{y},\boldsymbol{z},\boldsymbol{v}) = \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}(\boldsymbol{y},\boldsymbol{z},\boldsymbol{v}),\boldsymbol{y},\boldsymbol{z},\boldsymbol{v})$ and

$$\|\nabla_{\boldsymbol{y}} d_\delta(\boldsymbol{y},\boldsymbol{z},\boldsymbol{v}) - \nabla_{\boldsymbol{y}} d_\delta(\boldsymbol{y}',\boldsymbol{z},\boldsymbol{v})\| \le L_d\|\boldsymbol{y} - \boldsymbol{y}'\|,$$

where $L_d \coloneqq L\sigma_1 + L + r_{\boldsymbol{y}} + \delta$. □

We next relate the primal and dual Frank–Wolfe gaps to the corresponding optimization errors. We use the subscripts $\boldsymbol{x}$ and $\boldsymbol{y}$ for quantities associated with the primal and dual subproblems, respectively.

More precisely, $g^t_{\text{FW},\boldsymbol{x}}$, $g^t_{\text{A},\boldsymbol{x}}$, and $g^t_{\text{PFW},\boldsymbol{x}}$ denote the Frank–Wolfe, away, and pairwise Frank–Wolfe gaps associated with the minimization problem $\min_{\boldsymbol{x}\in\mathcal{X}} F_\delta(\boldsymbol{x},\boldsymbol{y}^t,\boldsymbol{z}^t,\boldsymbol{v}^t)$. Similarly, $g^t_{\text{FW},\boldsymbol{y}}$, $g^t_{\text{A},\boldsymbol{y}}$, and $g^t_{\text{PFW},\boldsymbol{y}}$ denote the corresponding gaps for the dual maximization problem $\max_{\boldsymbol{y}\in\mathcal{Y}} F_\delta(\boldsymbol{x}^t,\boldsymbol{y},\boldsymbol{z}^t,\boldsymbol{v}^t)$. In accordance with the notation of Section C,

$$g^t_{\text{PFW},\boldsymbol{x}} = g^t_{\text{FW},\boldsymbol{x}} + g^t_{\text{A},\boldsymbol{x}}, \qquad g^t_{\text{PFW},\boldsymbol{y}} = g^t_{\text{FW},\boldsymbol{y}} + g^t_{\text{A},\boldsymbol{y}}.$$

Here, $\text{PWidth}(A)$ and $\text{PWidth}(B)$ denote the pyramidal widths associated with $\mathcal{X} = \text{conv}(A)$ and $\mathcal{Y} = \text{conv}(B)$, respectively.

With this notation, the generic error bound from Section C can be applied separately to the primal and dual subproblems.

Lemma D.3. Suppose $r_{\boldsymbol{x}}, r_{\boldsymbol{y}} > L$. For any $t \ge 0$, the following inequalities hold:

(i) $\|\boldsymbol{x}^t - \boldsymbol{x}(\boldsymbol{y}^t,\boldsymbol{z}^t,\boldsymbol{v}^t)\| \le \sigma_6 g^t_{\text{PFW},\boldsymbol{x}}$,

(ii) $\|\boldsymbol{y}^t - \boldsymbol{y}(\boldsymbol{x}^t,\boldsymbol{z}^t,\boldsymbol{v}^t)\| \le \sigma_7 g^t_{\text{PFW},\boldsymbol{y}}$,

(iii) $\|\boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{y}^t\| \leq \sigma_8 g^t_{\text{PFW},\boldsymbol{y}} + \sigma_9 g^t_{\text{PFW},\boldsymbol{x}}$,

where $\sigma_6 = \frac{1}{(r_{\boldsymbol{x}} - L)\,\text{PWidth}(A)}, \sigma_7 = \frac{1}{(r_{\boldsymbol{y}} - L + \delta)\,\text{PWidth}(B)}, \sigma_8 = \frac{\sqrt{2}}{(r_{\boldsymbol{y}} - L + \delta)\,\text{PWidth}(B)}$, and $\sigma_9 = \frac{\sqrt{2} L \sigma_6}{r_{\boldsymbol{y}} - L + \delta}$. □

Proof. (i) By Assumption 2.1 and $r_{\boldsymbol{x}} > L$, we have $F_\delta(\cdot, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$ is $(r_{\boldsymbol{x}} - L)$-strongly convex and

$$\begin{aligned} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - \min_{\boldsymbol{x} \in \mathcal{X}} F_\delta(\boldsymbol{x}, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) &\leq \frac{1}{2(r_{\boldsymbol{x}} - L)} \operatorname{dist}(\boldsymbol{0}, \nabla_{\boldsymbol{x}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) + \partial \boldsymbol{I}_{\mathcal{X}}(\boldsymbol{x}^t))^2 \\ &\leq \frac{(g^t_{\text{PFW},\boldsymbol{x}})^2}{2(r_{\boldsymbol{x}} - L)\,\text{PWidth}(A)^2}, \end{aligned}$$

where the last inequality follows from Lemma C.2. In fact, applying Lemma C.2 with $F_\delta(\cdot, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t)$ yields

$$\operatorname{dist}\left(\boldsymbol{0},\, \nabla_{\boldsymbol{x}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) + \partial \boldsymbol{I}_{\mathcal{X}}(\boldsymbol{x}^t)\right) \leq \frac{g^t_{\text{PFW},\boldsymbol{x}}}{\text{PWidth}(A)}, \tag{5}$$

where $g^t_{\text{PFW},\boldsymbol{x}} \coloneqq g^t_{\text{FW},\boldsymbol{x}} + g^t_{\text{A},\boldsymbol{x}}$. On the other hand, we get

$$F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - \min_{\boldsymbol{x} \in \mathcal{X}} F_\delta(\boldsymbol{x}, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) \geq \frac{r_{\boldsymbol{x}} - L}{2} \|\boldsymbol{x}^t - \boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t)\|^2.$$

Combining them, we conclude that

$$\|\boldsymbol{x}^t - \boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t)\| \leq \frac{g^t_{\text{PFW},\boldsymbol{x}}}{(r_{\boldsymbol{x}} - L)\,\text{PWidth}(A)}.$$

Similarly, we can show (ii) holds. Now, we focus on the third bound.

(iii) By the strong concavity of $d_\delta(\cdot, \boldsymbol{z}, \boldsymbol{v}) = \min_{\boldsymbol{x} \in \mathcal{X}} F_\delta(\boldsymbol{x}, \cdot, \boldsymbol{z}, \boldsymbol{v})$, we have

$$\begin{aligned} &\max_{\boldsymbol{y} \in \mathcal{Y}} d_\delta(\boldsymbol{y}, \boldsymbol{z}^t, \boldsymbol{v}^t) - d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) \\ &\leq \frac{1}{2(r_{\boldsymbol{y}} - L + \delta)} \operatorname{dist}(\boldsymbol{0}, -\nabla_{\boldsymbol{y}} d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) + \partial \boldsymbol{I}_{\mathcal{Y}}(\boldsymbol{y}^t))^2 \\ &\leq \frac{1}{r_{\boldsymbol{y}} - L + \delta} \operatorname{dist}(\boldsymbol{0}, -\nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) + \partial \boldsymbol{I}_{\mathcal{Y}}(\boldsymbol{y}^t))^2 \\ &\quad + \frac{1}{r_{\boldsymbol{y}} - L + \delta} \|\nabla_{\boldsymbol{y}} d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t)\|^2 \\ &\leq \frac{(g^t_{\text{PFW},\boldsymbol{y}})^2}{(r_{\boldsymbol{y}} - L + \delta)\,\text{PWidth}(B)^2} + \frac{L^2}{r_{\boldsymbol{y}} - L + \delta} \|\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{x}^t\|^2 \\ &\leq \frac{(g^t_{\text{PFW},\boldsymbol{y}})^2}{(r_{\boldsymbol{y}} - L + \delta)\,\text{PWidth}(B)^2} + \frac{L^2 \sigma_6^2 (g^t_{\text{PFW},\boldsymbol{x}})^2}{r_{\boldsymbol{y}} - L + \delta}, \end{aligned}$$

where the first inequality follows from strong concavity. The second uses the Cauchy-Schwarz inequality, while the third follows from Lemma C.2 and Assumption 2.1:

$$\operatorname{dist}\left(\boldsymbol{0},\, -\nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) + \partial \boldsymbol{I}_{\mathcal{Y}}(\boldsymbol{y}^t)\right) \leq \frac{g^t_{\text{PFW},\boldsymbol{y}}}{\text{PWidth}(B)},$$

where $g^t_{\text{PFW},\boldsymbol{y}} \coloneqq g^t_{\text{FW},\boldsymbol{y}} + g^t_{\text{A},\boldsymbol{y}}$. On the other hand, we have

$$\max_{\boldsymbol{y} \in \mathcal{Y}} d_\delta(\boldsymbol{y}, \boldsymbol{z}^t, \boldsymbol{v}^t) - d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) \geq \frac{r_{\boldsymbol{y}} - L + \delta}{2} \|\boldsymbol{y}^t - \boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t)\|^2.$$

Combining them yields

$$\|\boldsymbol{y}^t - \boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t)\| \leq \frac{\sqrt{2} g^t_{\text{PFW},\boldsymbol{y}}}{(r_{\boldsymbol{y}} - L + \delta)\,\text{PWidth}(B)} + \frac{\sqrt{2} L \sigma_6 g^t_{\text{PFW},\boldsymbol{x}}}{r_{\boldsymbol{y}} - L + \delta}.$$

□

E STATIONARITY RELATIONS

This section establishes the stationarity relations used in the main text. We first relate the algorithmic residual to Frank–Wolfe stationarity of the original minimax problem by passing through the perturbed objective f_δ. We then prove Proposition 2.4, which provides the relations used in the comparison with prior fully projection-free results.

E.1 FROM ALGORITHMIC RESIDUALS TO FIRST-ORDER STATIONARITY

For every non-clipped iteration t, define

$$\mathcal{R}_t \coloneqq g^t + \frac{\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|}{\beta_{\boldsymbol{x}}} + \frac{\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|}{\beta_{\boldsymbol{y}}}. \tag{6}$$

The following lemma relates the quantities controlled by the algorithm to Frank–Wolfe stationarity of the original minimax problem.

Lemma E.1 (Residual-to-FW stationarity)**.** For every non-clipped iteration t, there exists a constant $C_{\mathrm{stat}} > 0$, independent of T, such that

$$\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}^t, \boldsymbol{y}^t) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}^t, \boldsymbol{y}^t) \le C_{\mathrm{stat}}\mathcal{R}_t + \delta \ \mathrm{diam}(\mathcal{Y})^2.$$

Consequently, if $\mathcal{R}_t \le \epsilon$, then $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}(\epsilon + \delta)$-FWSP of Problem P. □

To prove Lemma E.1, we first compare the Frank–Wolfe gaps of the perturbed and original problems. Recall that $f_\delta(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{y}) - \frac{\delta}{2}\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^2$. We define the Frank–Wolfe stationarity measures associated with the perturbed problem by

$$\mathcal{G}_{\mathcal{X}}^{\delta}(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \sup_{\boldsymbol{x}' \in \mathcal{X}} \langle \nabla_{\boldsymbol{x}} f_\delta(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{x}' \rangle,$$

and

$$\mathcal{G}_{\mathcal{Y}}^{\delta}(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \sup_{\boldsymbol{y}' \in \mathcal{Y}} \langle -\nabla_{\boldsymbol{y}} f_\delta(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y} - \boldsymbol{y}' \rangle.$$

The following lemma bounds the perturbation bias.

Lemma E.2 (Perturbation bias)**.** For any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y}) \le \mathcal{G}_{\mathcal{X}}^{\delta}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{G}_{\mathcal{Y}}^{\delta}(\boldsymbol{x}, \boldsymbol{y}) + \delta \, \mathrm{diam}(\mathcal{Y})^2.$$

□

Proof. Since the perturbation is independent of $\boldsymbol{x}$, this gives

$$\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{G}_{\mathcal{X}}^{\delta}(\boldsymbol{x}, \boldsymbol{y}).$$

Moreover, from the definition of f_δ, we have $-\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}) = -\nabla_{\boldsymbol{y}} f_\delta(\boldsymbol{x}, \boldsymbol{y}) - \delta(\boldsymbol{y} - \bar{\boldsymbol{y}})$. Hence,

$$\begin{aligned}
\mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y}) &= \sup_{\boldsymbol{y}' \in \mathcal{Y}} \langle -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y} - \boldsymbol{y}' \rangle \\
&\le \mathcal{G}_{\mathcal{Y}}^{\delta}(\boldsymbol{x}, \boldsymbol{y}) + \delta \sup_{\boldsymbol{y}' \in \mathcal{Y}} |\langle \boldsymbol{y} - \bar{\boldsymbol{y}}, \boldsymbol{y} - \boldsymbol{y}' \rangle| \\
&\le \mathcal{G}_{\mathcal{Y}}^{\delta}(\boldsymbol{x}, \boldsymbol{y}) + \delta \, \mathrm{diam}(\mathcal{Y})^2,
\end{aligned}$$

where the last inequality follows from $\boldsymbol{y}, \bar{\boldsymbol{y}}, \boldsymbol{y}' \in \mathcal{Y}$. □

We now prove Lemma E.1 by first controlling the perturbed Frank–Wolfe gaps and then applying Lemma E.2.

Proof of Lemma E.1. We first bound $\mathcal{G}_{\mathcal{X}}^{\delta}(\boldsymbol{x}^t, \boldsymbol{y}^t)$ and $\mathcal{G}_{\mathcal{Y}}^{\delta}(\boldsymbol{x}^t, \boldsymbol{y}^t)$ in terms of $\mathcal{R}_t$.

For the $\mathcal{X}$-stationarity measure, we have

$$\begin{aligned}\mathcal{G}^{\delta}_{\mathcal{X}}(\boldsymbol{x}^t,\boldsymbol{y}^t) &= \sup_{\boldsymbol{x}\in\mathcal{X}}\langle\nabla_{\boldsymbol{x}} f_\delta(\boldsymbol{x}^t,\boldsymbol{y}^t),\boldsymbol{x}^t-\boldsymbol{x}\rangle\\ &\le \sup_{\boldsymbol{x}\in\mathcal{X}}\langle\nabla_{\boldsymbol{x}} F_\delta(\boldsymbol{x}^t,\boldsymbol{y}^t,\boldsymbol{z}^t,\boldsymbol{v}^t),\boldsymbol{x}^t-\boldsymbol{x}\rangle + \sup_{\boldsymbol{x}\in\mathcal{X}} r_{\boldsymbol{x}}\langle\boldsymbol{z}^t-\boldsymbol{x}^t,\boldsymbol{x}^t-\boldsymbol{x}\rangle\\ &\le g^t_{\mathrm{FW},\boldsymbol{x}} + r_{\boldsymbol{x}}\operatorname{diam}(\mathcal{X})\|\boldsymbol{z}^t-\boldsymbol{x}^t\|\\ &= g^t_{\mathrm{FW},\boldsymbol{x}} + \frac{r_{\boldsymbol{x}}\operatorname{diam}(\mathcal{X})}{\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|.\end{aligned}$$

Here, the first inequality follows from the relation between $\nabla_{\boldsymbol{x}} f_\delta$ and $\nabla_{\boldsymbol{x}} F_\delta$, the second uses the boundedness of $\mathcal{X}$, and the last follows from the update of $\boldsymbol{z}^{t+1}$.

Similarly, for the $\mathcal{Y}$-stationarity measure,

$$\begin{aligned}\mathcal{G}^{\delta}_{\mathcal{Y}}(\boldsymbol{x}^t,\boldsymbol{y}^t) &= \sup_{\boldsymbol{y}\in\mathcal{Y}}\langle-\nabla_{\boldsymbol{y}} f_\delta(\boldsymbol{x}^t,\boldsymbol{y}^t),\boldsymbol{y}^t-\boldsymbol{y}\rangle\\ &\le g^t_{\mathrm{FW},\boldsymbol{y}} + \frac{r_{\boldsymbol{y}}\operatorname{diam}(\mathcal{Y})}{\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t-\boldsymbol{v}^{t+1}\|.\end{aligned}$$

It remains to control the last term by the residual used in the analysis. By the definition of $\boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})$, we have

$$\begin{aligned}\frac{1}{\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t-\boldsymbol{v}^{t+1}\| &\le \frac{1}{\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t-\boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\| + \|\boldsymbol{y}^t-\boldsymbol{y}(\boldsymbol{z}^{t+1},\boldsymbol{v}^t)\|\\ &\le \frac{1}{\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t-\boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\| + \|\boldsymbol{y}^t-\boldsymbol{y}(\boldsymbol{z}^{t},\boldsymbol{v}^t)\| + \sigma_3\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|\\ &\le \frac{1}{\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t-\boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\| + \sigma_8 g^t_{\mathrm{PFW},\boldsymbol{y}} + \sigma_9 g^t_{\mathrm{PFW},\boldsymbol{x}} + \sigma_3\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|,\end{aligned}$$

where the second inequality uses Lemma D.1(iv), and the last inequality follows from Lemma D.3(iii). Combining the bounds for the two stationarity measures gives

$$\begin{aligned}&\mathcal{G}^{\delta}_{\mathcal{X}}(\boldsymbol{x}^t,\boldsymbol{y}^t) + \mathcal{G}^{\delta}_{\mathcal{Y}}(\boldsymbol{x}^t,\boldsymbol{y}^t)\\ &\le g^t_{\mathrm{FW},\boldsymbol{x}} + g^t_{\mathrm{FW},\boldsymbol{y}} + \frac{r_{\boldsymbol{x}}\operatorname{diam}(\mathcal{X})}{\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\| + \frac{r_{\boldsymbol{y}}\operatorname{diam}(\mathcal{Y})}{\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t-\boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\|\\ &\quad + r_{\boldsymbol{y}}\operatorname{diam}(\mathcal{Y})\left(\sigma_8 g^t_{\mathrm{PFW},\boldsymbol{y}} + \sigma_9 g^t_{\mathrm{PFW},\boldsymbol{x}} + \sigma_3\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|\right)\\ &\le \left(1+2r_{\boldsymbol{y}}\operatorname{diam}(\mathcal{Y})\max\{\sigma_8,\sigma_9\}\right) g^t + \frac{r_{\boldsymbol{x}}\operatorname{diam}(\mathcal{X}) + r_{\boldsymbol{y}}\sigma_3\beta_{\boldsymbol{x}}\operatorname{diam}(\mathcal{Y})}{\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|\\ &\quad + \frac{r_{\boldsymbol{y}}\operatorname{diam}(\mathcal{Y})}{\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t-\boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\|.\end{aligned}$$

Consequently, if $g^t \le \epsilon$, $\frac{\|\boldsymbol{z}^t-\boldsymbol{z}^{t+1}\|}{\beta_{\boldsymbol{x}}} \le \epsilon$, $\frac{\|\boldsymbol{v}^t-\boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\|}{\beta_{\boldsymbol{y}}} \le \epsilon$, then

$$\mathcal{G}^{\delta}_{\mathcal{X}}(\boldsymbol{x}^t,\boldsymbol{y}^t) + \mathcal{G}^{\delta}_{\mathcal{Y}}(\boldsymbol{x}^t,\boldsymbol{y}^t) \le \mathcal{O}(\epsilon).$$

Finally, applying Lemma E.2 yields

$$\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}^t,\boldsymbol{y}^t) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}^t,\boldsymbol{y}^t) \le \mathcal{O}(\epsilon+\delta),$$

and hence $(\boldsymbol{x}^t,\boldsymbol{y}^t)$ is an $\mathcal{O}(\epsilon+\delta)$-FWSP of the original minimax problem. □

E.2 Proof of Proposition 2.4

We provide the proof of Proposition 2.4. The first relation, $\mathcal{G}^{\mathrm{Opt}}_{\mathcal{Y}} \le \mathcal{G}_{\mathcal{Y}}$, is also noted in (Giang-Tran et al., 2026); we include the short argument for completeness.

Proof of Proposition 2.4. Suppose first that Assumption 2.2(i) holds. By concavity of $f(\boldsymbol{x}, \cdot)$, for every $\boldsymbol{y}' \in \mathcal{Y}$,

$$f(\boldsymbol{x}, \boldsymbol{y}') - f(\boldsymbol{x}, \boldsymbol{y}) \leq \langle -\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y} - \boldsymbol{y}' \rangle. \tag{7}$$

Taking the supremum over $\boldsymbol{y}' \in \mathcal{Y}$ yields $\mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(\boldsymbol{x}, \boldsymbol{y}) \leq \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y})$. Thus, every ϵ-FWSP satisfies $\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) \leq \epsilon$ and $\mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(\boldsymbol{x}, \boldsymbol{y}) \leq \epsilon$.

Suppose now, in addition, that $f(\cdot, \boldsymbol{y})$ is convex on $\mathcal{X}$ for every $\boldsymbol{y} \in \mathcal{Y}$. Since $\min_{\boldsymbol{x}' \in \mathcal{X}} f(\boldsymbol{x}', \boldsymbol{y}) \leq \min_{\boldsymbol{x}' \in \mathcal{X}} \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}', \boldsymbol{y}')$, we have $\mathcal{G}_{\mathcal{X}}^{\mathrm{Opt}}(\boldsymbol{x}) \leq \mathcal{G}(\boldsymbol{x}, \boldsymbol{y})$. Moreover, $\mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(\boldsymbol{x}, \boldsymbol{y}) = \max_{\boldsymbol{y}' \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}') - f(\boldsymbol{x}, \boldsymbol{y}) \leq \mathcal{G}(\boldsymbol{x}, \boldsymbol{y})$, since $\min_{\boldsymbol{x}' \in \mathcal{X}} f(\boldsymbol{x}', \boldsymbol{y}) \leq f(\boldsymbol{x}, \boldsymbol{y})$. Therefore,

$$\max \left\{ \mathcal{G}_{\mathcal{X}}^{\mathrm{Opt}}(\boldsymbol{x}), \mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(\boldsymbol{x}, \boldsymbol{y}) \right\} \leq \mathcal{G}(\boldsymbol{x}, \boldsymbol{y}).$$

It remains to upper bound the saddle-point gap by the two Frank–Wolfe gaps. By convexity of $f(\cdot, \boldsymbol{y})$, for every $\boldsymbol{x}' \in \mathcal{X}$,

$$f(\boldsymbol{x}, \boldsymbol{y}) - f(\boldsymbol{x}', \boldsymbol{y}) \leq \langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{x} - \boldsymbol{x}' \rangle,$$

whereas concavity of $f(\boldsymbol{x}, \cdot)$ gives equation 7. Adding the two inequalities and then taking the supremum over $\boldsymbol{x}' \in \mathcal{X}$ and $\boldsymbol{y}' \in \mathcal{Y}$ yields

$$\mathcal{G}(\boldsymbol{x}, \boldsymbol{y}) \leq \mathcal{G}_{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y}).$$

Thus, if $(\boldsymbol{x}, \boldsymbol{y})$ is an ϵ-FWSP, then $\mathcal{G}(\boldsymbol{x}, \boldsymbol{y}) \leq 2\epsilon$, so $(\boldsymbol{x}, \boldsymbol{y})$ is a 2ϵ-SP. Furthermore, $\mathcal{G}_{\mathcal{X}}^{\mathrm{Opt}}(\boldsymbol{x}) \leq 2\epsilon$ and $\mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(\boldsymbol{x}, \boldsymbol{y}) \leq \epsilon$, where the sharper dual bound follows from $\mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}} \leq \mathcal{G}_{\mathcal{Y}}$. □

Remark E.3 (Failure of the reverse implication)**.** The converse of the C–C implication in Proposition 2.4 does not hold in general: small primal and dual optimality gaps do not necessarily imply a small saddle-point gap, nor two-sided Frank–Wolfe stationarity of the same primal–dual pair.

To see this, consider the convex–concave bilinear problem

$$\min_{x \in [-1,1]} \max_{y \in [-1,1]} xy$$

and the point $(x, y) = (0, 1)$. The primal value function is $\phi(x) = \max_{y \in [-1,1]} xy = |x|$, so $\mathcal{G}_{\mathcal{X}}^{\mathrm{Opt}}(0) = 0$. Moreover, $\mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(0, 1) = 0$, since $f(0, y) = 0$ for every $y \in [-1, 1]$.

However, the saddle-point gap is

$$\mathcal{G}(0, 1) = \max_{y' \in [-1,1]} f(0, y') - \min_{x' \in [-1,1]} f(x', 1) = 1.$$

Likewise, since $\nabla_x f(0, 1) = 1$ and $\nabla_y f(0, 1) = 0$, we have $\mathcal{G}_{\mathcal{X}}(0, 1) = 1$ and $\mathcal{G}_{\mathcal{Y}}(0, 1) = 0$. Thus,

$$\mathcal{G}_{\mathcal{X}}^{\mathrm{Opt}}(0) = \mathcal{G}_{\mathcal{Y}}^{\mathrm{Opt}}(0, 1) = 0, \qquad \mathcal{G}(0, 1) = \mathcal{G}_{\mathcal{X}}(0, 1) + \mathcal{G}_{\mathcal{Y}}(0, 1) = 1.$$

Hence, even vanishing primal and dual optimality gaps do not imply saddle-point optimality or two-sided Frank–Wolfe stationarity of the same primal–dual pair. □

F Proof of Proposition 4.1

We prove Proposition 4.1 by first establishing separate one-step estimates for the primal function F_δ, the dual function d_δ, and the proximal value function q_δ. We then combine these estimates to derive the Lyapunov bounds for non-clipped and clipped iterations. The distinction between the two cases arises because the proximal variables $\boldsymbol{z}$ and $\boldsymbol{v}$ are updated at non-clipped iterations but remain unchanged at clipped iterations. Finally, the two Lyapunov bounds are combined with the stepsize rule in Algorithm 1 to obtain the estimates stated in Proposition 4.1.

F.1 BASIC ONE-STEP ESTIMATES

Throughout this subsection, let t be a non-clipped iteration. In this case,

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t + \beta_{\boldsymbol{x}}(\boldsymbol{x}^t - \boldsymbol{z}^t), \qquad \boldsymbol{v}^{t+1} = \boldsymbol{v}^t + \beta_{\boldsymbol{y}}(\boldsymbol{y}^t - \boldsymbol{v}^t).$$

We establish the corresponding one-step estimates for the primal, dual, and proximal components.

Lemma F.1 (Primal descent)**.** For any non-clipped iteration $t \geq 0$, the following inequality holds:

$$\begin{aligned} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) \geq F_\delta(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) + \gamma_t g_{\boldsymbol{x}}^t - \gamma_t^2 \left(L + r_{\boldsymbol{x}}\right) \|\boldsymbol{d}_{\boldsymbol{x}}^t\|^2 \\ + \langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^t - \boldsymbol{y}^{t+1} \rangle - \gamma_t^2 \left(L + \frac{\delta}{2}\right) \|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2 \\ + \left(\frac{(2-\beta_{\boldsymbol{x}})r_{\boldsymbol{x}}}{2\beta_{\boldsymbol{x}}} - \frac{r_{\boldsymbol{x}}}{2}\right) \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \left(\frac{(\beta_{\boldsymbol{y}}-2)r_{\boldsymbol{y}}}{2\beta_{\boldsymbol{y}}} - \frac{r_{\boldsymbol{y}}}{2}\right) \|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2. \end{aligned}$$

□

Proof. We decompose the difference into four terms:

$$\begin{aligned} & F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - F_\delta(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) \\ = & \underbrace{F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t)}_{T_1} + \underbrace{F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t) - F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1})}_{T_2} \\ & + \underbrace{F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) - F_\delta(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1})}_{T_3} \\ & + \underbrace{F_\delta(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) - F_\delta(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1})}_{T_4}. \end{aligned}$$

For T_1, on top of the update of $\boldsymbol{z}^{t+1}$, i.e., $\boldsymbol{z}^{t+1} = \boldsymbol{z}^t + \beta_{\boldsymbol{x}}(\boldsymbol{x}^t - \boldsymbol{z}^t)$, we obtain

$$F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t) = \frac{(2-\beta_{\boldsymbol{x}})r_{\boldsymbol{x}}}{2\beta_{\boldsymbol{x}}} \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2.$$

Similarly, as for T_2, following the update of $\boldsymbol{v}^{t+1}$, i.e, $\boldsymbol{v}^{t+1} = \boldsymbol{v}^t + \beta_{\boldsymbol{y}}(\boldsymbol{y}^t - \boldsymbol{v}^t)$, we can verify that

$$F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t) - F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) = \frac{(\beta_{\boldsymbol{y}}-2)r_{\boldsymbol{y}}}{2\beta_{\boldsymbol{y}}} \|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2.$$

As for T_3, we bound it as follows:

$$\begin{aligned} & F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) - F_\delta(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) \\ \geq & \langle \nabla_{\boldsymbol{x}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t), \boldsymbol{x}^t - \boldsymbol{x}^{t+1} \rangle - \frac{L + r_{\boldsymbol{x}}}{2} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 \\ \geq & \gamma_t \langle -\nabla_{\boldsymbol{x}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{d}_{\boldsymbol{x}}^t \rangle - \frac{r_{\boldsymbol{x}}}{2} \|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2 - \frac{\gamma_t^2 \left(L + 2r_{\boldsymbol{x}}\right)}{2} \|\boldsymbol{d}_{\boldsymbol{x}}^t\|^2 \\ = & \gamma_t g_{\boldsymbol{x}}^t - \frac{r_{\boldsymbol{x}}}{2} \|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2 - \frac{\gamma_t^2 \left(L + 2r_{\boldsymbol{x}}\right)}{2} \|\boldsymbol{d}_{\boldsymbol{x}}^t\|^2, \end{aligned}$$

where the first inequality follows from the Lipschitz gradient property (see Assumption 2.1) and the second is due to the update of $\boldsymbol{x}^{t+1}$, Cauchy-Schwarz inequality, and AM-GM inequality. Next, by Assumption 2.1, the part T_4 is bounded as follows:

$$\begin{aligned} & F_\delta(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) - F_\delta(\boldsymbol{x}^{t+1}, \boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) \\ \geq & \langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^{t+1}, \boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}), \boldsymbol{y}^t - \boldsymbol{y}^{t+1} \rangle + \frac{r_{\boldsymbol{y}} - L + \delta}{2} \|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\|^2 \\ \geq & \langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^t - \boldsymbol{y}^{t+1} \rangle - \frac{L}{2} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|^2 - \frac{r_{\boldsymbol{y}}}{2} \|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2 - \left(L + \frac{\delta}{2}\right) \|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\|^2 \\ = & \langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^t - \boldsymbol{y}^{t+1} \rangle - \frac{\gamma_t^2 L}{2} \|\boldsymbol{d}_{\boldsymbol{x}}^t\|^2 - \frac{r_{\boldsymbol{y}}}{2} \|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2 - \gamma_t^2 \left(L + \frac{\delta}{2}\right) \|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2, \end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz and AM-GM inequalities.

Combining all the above bounds leads to the conclusion. □

The next estimate controls the variation of the dual function. Its proof adapts the argument of (Zheng et al., 2023, Lemma 6) to the updates used in our method.

Lemma F.2 (Dual ascent)**.** For any non-clipped iteration $t \geq 0$, the following inequality holds:

$$\begin{aligned} d_\delta(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) \geq d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) + \left(\frac{(2-\beta_{\boldsymbol{y}})r_{\boldsymbol{y}}}{2\beta_{\boldsymbol{y}}} - \frac{r_{\boldsymbol{y}}}{2}\right)\|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2 \\ + \frac{r_{\boldsymbol{x}}}{2}\langle \boldsymbol{z}^{t+1} + \boldsymbol{z}^t - 2\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t\rangle \\ + \langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle - \frac{\gamma_t^2 (L_d + r_{\boldsymbol{y}})}{2}\|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2. \end{aligned}$$

□

Proof. The difference of the update for the dual function is controlled by the following three parts:

$$\begin{aligned} & d_\delta(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) - d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) \\ = & \underbrace{d_\delta(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) - d_\delta(\boldsymbol{y}^{t+1}, \boldsymbol{z}^t, \boldsymbol{v}^{t+1})}_{T_1} + \underbrace{d_\delta(\boldsymbol{y}^{t+1}, \boldsymbol{z}^t, \boldsymbol{v}^{t+1}) - d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^{t+1})}_{T_2} \\ & + \underbrace{d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^{t+1}) - d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t)}_{T_3}. \end{aligned}$$

For the first part, by the definition of $\boldsymbol{x}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, we have

$$\begin{aligned} T_1 &= F_\delta(\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t), \boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) - F_\delta(\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^{t+1}, \boldsymbol{z}^t, \boldsymbol{v}^{t+1}) \\ &\geq F_\delta(\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t), \boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^{t+1}) - F_\delta(\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t), \boldsymbol{y}^{t+1}, \boldsymbol{z}^t, \boldsymbol{v}^{t+1}) \\ &= \frac{r_{\boldsymbol{x}}}{2}\langle \boldsymbol{z}^{t+1} + \boldsymbol{z}^t - 2\boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t\rangle, \end{aligned}$$

where the first inequality uses optimality condition of $\boldsymbol{x}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$ and the fact that $\boldsymbol{x}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) = \boldsymbol{x}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}')$. For the second part, using Lemma D.2 gives

$$\begin{aligned} T_2 &\geq \langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^{t+1}), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle - \frac{L_d}{2}\|\boldsymbol{y}^t - \boldsymbol{y}^{t+1}\|^2 \\ &\geq \langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle - \frac{r_{\boldsymbol{y}}}{2}\|\boldsymbol{v}^t - \boldsymbol{v}^{t+1}\|^2 - \frac{L_d + r_{\boldsymbol{y}}}{2}\|\boldsymbol{y}^t - \boldsymbol{y}^{t+1}\|^2. \end{aligned}$$

Consider the last part, following the update of $\boldsymbol{v}^{t+1}$, we obtain

$$T_3 = \frac{r_{\boldsymbol{y}}}{2}\left(\|\boldsymbol{y}^t - \boldsymbol{v}^t\|^2 - \|\boldsymbol{y}^t - \boldsymbol{v}^{t+1}\|^2\right) = \frac{(2-\beta_{\boldsymbol{y}})r_{\boldsymbol{y}}}{2\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2.$$

Combining the three estimates proves the result. □

Lemma F.3 (Proximal descent)**.** (Zheng et al., 2023, Lemma 7) For any non-clipped iteration $t \geq 0$, the following inequality holds:

$$q_\delta(\boldsymbol{z}^t) \geq q_\delta(\boldsymbol{z}^{t+1}) + \frac{r_{\boldsymbol{x}}}{2}\langle \boldsymbol{z}^t + \boldsymbol{z}^{t+1} - 2\boldsymbol{x}(\boldsymbol{z}^t, \boldsymbol{v}(\boldsymbol{z}^{t+1})), \boldsymbol{z}^t - \boldsymbol{z}^{t+1}\rangle.$$

□

F.2 Non-clipped Lyapunov estimate

Recall the Lyapunov function $\Phi : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ as follows:

$$\Phi(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) = F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) - 2d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) + 2q_\delta(\boldsymbol{z}),$$

where $d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) := \min_{\boldsymbol{x} \in \mathcal{X}} F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, $p_\delta(\boldsymbol{z}, \boldsymbol{v}) := \max_{\boldsymbol{y} \in \mathcal{Y}} d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, and $q_\delta(\boldsymbol{z}) := \max_{\boldsymbol{v} \in \mathbb{R}^m} p_\delta(\boldsymbol{z}, \boldsymbol{v})$.

We now combine the preceding primal, dual, and proximal estimates to derive a Lyapunov bound for non-clipped iterations.

Lemma F.4 (Non-clipped Lyapunov bound)**.** Under the assumptions and parameter conditions of Proposition 4.1, every non-clipped iteration t satisfies

$$\Phi^t - \Phi^{t+1} \geq \frac{\gamma_t}{2} g^t + \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}} \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{r_{\boldsymbol{y}}}{4\beta_{\boldsymbol{y}}} \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 - \bar{C}\gamma_t^2 - \frac{1}{32\bar{C}}(g^t)^2 - 6r_{\boldsymbol{x}}\beta_{\boldsymbol{x}} \left\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\right\|^2.$$

□

Proof of Lemma F.4. From the results in Lemmas F.1, F.2, and F.3, we know that

$$\begin{aligned}\Phi^t \geq \Phi^{t+1} &+ \gamma_t g^t - \gamma_t^2 (L + r_{\boldsymbol{x}}) \|\boldsymbol{d}_{\boldsymbol{x}}^t\|^2 - \gamma_t^2 \left(r_{\boldsymbol{y}} + \frac{\delta}{2} + L + L_d\right) \|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2 \\ &+ \frac{(1-\beta_{\boldsymbol{x}})r_{\boldsymbol{x}}}{\beta_{\boldsymbol{x}}} \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{(1-2\beta_{\boldsymbol{y}})r_{\boldsymbol{y}}}{\beta_{\boldsymbol{y}}} \|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2 \\ &+ \underbrace{2\langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t \rangle}_{T_1} \\ &+ \underbrace{2r_{\boldsymbol{x}} \langle \boldsymbol{x}(\boldsymbol{z}^t, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t \rangle}_{T_2}.\end{aligned}$$

We bound each term separately. For term T_1, using Lipschitz gradient property (see Assumption 2.1) and Young's inequality, we obtain

$$\begin{aligned}&2\langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t \rangle \\ \geq\ & -2L\|\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{x}^t\| \|\boldsymbol{y}^{t+1} - \boldsymbol{y}^t\| \geq -2\gamma_t L \sigma_6 g_{\text{PFW},\boldsymbol{x}}^t \|\boldsymbol{d}_{\boldsymbol{y}}^t\| \\ \geq\ & -2\gamma_t L \sigma_6 \operatorname{diam}(\mathcal{Y}) g_{\text{PFW}}^t \geq -4\gamma_t L \sigma_6 \operatorname{diam}(\mathcal{Y}) g^t, \end{aligned} \tag{8}$$

where the second inequality follows from the Lemma D.3 (i) and the last two inequalities use the non-negativity of $g_{\text{PFW},\boldsymbol{y}}^t$ and the boundedness of $\|\boldsymbol{d}_{\boldsymbol{y}}^t\|$.

For term T_2, for any $\kappa > 0$, it follows that

$$\begin{aligned}&2r_{\boldsymbol{x}} \langle \boldsymbol{x}(\boldsymbol{z}^t, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t \rangle \\ =\ & 2r_{\boldsymbol{x}} \langle \boldsymbol{x}(\boldsymbol{z}^t, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t \rangle \\ &+ 2r_{\boldsymbol{x}} \langle \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t), \boldsymbol{z}^{t+1} - \boldsymbol{z}^t \rangle \\ \geq\ & -2r_{\boldsymbol{x}}\sigma_2 \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 - \frac{r_{\boldsymbol{x}}}{\kappa} \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 - r_{\boldsymbol{x}}\kappa \|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t)\|^2,\end{aligned}$$

where the inequality is from the Cauchy-Schwarz inequality, Young's inequality, and error bound (iii) in Lemma D.1. Hence,

$$\begin{aligned}\Phi^t \geq \Phi^{t+1} &+ \gamma_t (1 - 4L\sigma_6 \operatorname{diam}(\mathcal{Y})) g^t + r_{\boldsymbol{x}} \left(\frac{(1-\beta_{\boldsymbol{x}})}{\beta_{\boldsymbol{x}}} - 2\sigma_2 - \frac{1}{\kappa}\right) \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 \\ &+ \frac{(1-2\beta_{\boldsymbol{y}})r_{\boldsymbol{y}}}{\beta_{\boldsymbol{y}}} \|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2 - \gamma_t^2 (L + r_{\boldsymbol{x}}) \|\boldsymbol{d}_{\boldsymbol{x}}^t\|^2 - \gamma_t^2 \left(r_{\boldsymbol{y}} + \frac{\delta}{2} + L + L_d\right) \|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2 \\ &- r_{\boldsymbol{x}}\kappa \underbrace{\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t)\|^2}_{T_3}.\end{aligned}$$

Next, we focus on the negative term. From the fact $\boldsymbol{x}(\boldsymbol{z}, \boldsymbol{v}) = \boldsymbol{x}(\boldsymbol{y}(\boldsymbol{z}, \boldsymbol{v}), \boldsymbol{z}, \boldsymbol{v})$ and $\boldsymbol{x}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) = \boldsymbol{x}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}')$, the term T_3 is bounded as follows:

$$\begin{aligned}&\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t)\|^2 \\ \leq\ & 3\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2 + 3\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t)\|^2 \\ &+ 3\|\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t)\|^2 \\ \leq\ & 3\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2 + 3\sigma_1^2 \left(\|\boldsymbol{y}^t - \boldsymbol{y}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2 + \|\boldsymbol{y}^t - \boldsymbol{y}^{t+1}\|^2\right).\end{aligned} \tag{9}$$

Here, the second inequality is due to the error bound (i) in Lemma D.1. We can further simplify the second error term using Lemma D.3(iii) and Lemma D.1(iv), (vii):

$$\begin{aligned}
&\|\boldsymbol{y}^t - \boldsymbol{y}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2 \\
&\leq 2\|\boldsymbol{y}^t - \boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t)\|^2 + 2\|\boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{y}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2 \\
&\leq 4\sigma_8^2 (g_{\text{PFW},\boldsymbol{y}}^t)^2 + 4\sigma_9^2 (g_{\text{PFW},\boldsymbol{x}}^t)^2 + 4\sigma_5^2 \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 + 4\sigma_3^2 \|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2. \qquad (10)
\end{aligned}$$

Plugging equation 10 into equation 9, we have

$$\begin{aligned}
&\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{y}^{t+1}, \boldsymbol{z}^{t+1}, \boldsymbol{v}^t)\|^2 \\
&\leq 3\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2 + 12\sigma_1^2\sigma_5^2 \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 \\
&\quad + 12\sigma_1^2\sigma_3^2 \|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2 + 12\sigma_1^2\sigma_8^2 (g_{\text{PFW},\boldsymbol{y}}^t)^2 + 12\sigma_1^2\sigma_9^2 (g_{\text{PFW},\boldsymbol{x}}^t)^2 + 3\sigma_1^2\gamma_t^2 \|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2 \\
&\leq 3\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2 + 12\sigma_1^2\sigma_5^2 \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 \\
&\quad + 12\sigma_1^2\sigma_3^2 \|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2 + 48\sigma_1^2 \max\{\sigma_8^2, \sigma_9^2\} (g^t)^2 + 3\sigma_1^2\gamma_t^2 \|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2,
\end{aligned}$$

where the last inequality uses $(g_{\text{PFW},\boldsymbol{y}}^t)^2 + (g_{\text{PFW},\boldsymbol{x}}^t)^2 \leq (g_{\text{PFW}}^t)^2 \leq 4(g^t)^2$.

Summing the above inequalities and letting $c_{\boldsymbol{z}} \coloneqq r_{\boldsymbol{x}} \left(\frac{(1-\beta_{\boldsymbol{x}})}{\beta_{\boldsymbol{x}}} - 2\sigma_2 - \frac{1}{\kappa} \right)$, we have

$$\begin{aligned}
\Phi^t \geq \Phi^{t+1} &+ \gamma_t \left(1 - 4L\sigma_6 \operatorname{diam}(\mathcal{Y})\right) g^t - 48\sigma_1^2 \max\{\sigma_8^2, \sigma_9^2\} r_{\boldsymbol{x}} \kappa (g^t)^2 \\
&+ \left(c_{\boldsymbol{z}} - 12\sigma_1^2\sigma_3^2 r_{\boldsymbol{x}} \kappa\right) \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{(1 - 2\beta_{\boldsymbol{y}}) r_{\boldsymbol{y}}}{\beta_{\boldsymbol{y}}} \|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2 \\
&- \gamma_t^2 \left(L + r_{\boldsymbol{x}}\right) \|\boldsymbol{d}_{\boldsymbol{x}}^t\|^2 - \gamma_t^2 \left(r_{\boldsymbol{y}} + \frac{\delta}{2} + L + L_d + 3\sigma_1^2 r_{\boldsymbol{x}} \kappa \right) \|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2 \\
&- 12\sigma_1^2\sigma_5^2 r_{\boldsymbol{x}} \kappa \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 - 3 r_{\boldsymbol{x}} \kappa \|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2. \qquad (11)
\end{aligned}$$

To express the positive $\boldsymbol{v}$-update term in terms of the proximal residual $\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|$, we have

$$\begin{aligned}
\|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2 &\geq \frac{1}{2} \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 - \|\boldsymbol{v}^{t+1} - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 \\
&\geq \frac{1}{2} \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 - \beta_{\boldsymbol{y}}^2 \left(16 \max\{\sigma_8^2, \sigma_9^2\} (g^t)^2 + 2\sigma_3^2 \|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2\right). \qquad (12)
\end{aligned}$$

Recalling that $\boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}) = \boldsymbol{v}^t + \beta_{\boldsymbol{y}}(\boldsymbol{y}(\boldsymbol{z}^{t+1}, \boldsymbol{v}^t) - \boldsymbol{v}^t)$, the last inequality in equation 12 can be obtained by using error bounds (iv) in Lemma D.1 and (iii) in Lemma D.3, i.e.,

$$\begin{aligned}
\|\boldsymbol{v}^{t+1} - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 &= \beta_{\boldsymbol{y}}^2 \|\boldsymbol{y}^t - \boldsymbol{y}(\boldsymbol{z}^{t+1}, \boldsymbol{v}^t)\|^2 \\
&\leq \beta_{\boldsymbol{y}}^2 \left(2\|\boldsymbol{y}^t - \boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t)\|^2 + 2\|\boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{y}(\boldsymbol{z}^{t+1}, \boldsymbol{v}^t)\|^2\right) \\
&\leq \beta_{\boldsymbol{y}}^2 \left(4\sigma_8^2 (g_{\text{PFW},\boldsymbol{y}}^t)^2 + 4\sigma_9^2 (g_{\text{PFW},\boldsymbol{x}}^t)^2 + 2\sigma_3^2 \|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2\right) \\
&\leq \beta_{\boldsymbol{y}}^2 \left(16 \max\{\sigma_8^2, \sigma_9^2\} (g^t)^2 + 2\sigma_3^2 \|\boldsymbol{z}^t - \boldsymbol{z}^{t+1}\|^2\right),
\end{aligned}$$

where the last inequality uses $(g_{\text{PFW},\boldsymbol{y}}^t)^2 + (g_{\text{PFW},\boldsymbol{x}}^t)^2 \leq (g_{\text{PFW}}^t)^2 \leq 4(g^t)^2$.

Substituting equation 12 into equation 11 yields

$$\begin{aligned}
\Phi^t \geq \Phi^{t+1} &+ \gamma_t \left(1 - 4L\sigma_6 \operatorname{diam}(\mathcal{Y})\right) g^t - 16 \max\{\sigma_8^2, \sigma_9^2\} \left(3\sigma_1^2 r_{\boldsymbol{x}} \kappa + \beta_{\boldsymbol{y}}(1 - 2\beta_{\boldsymbol{y}}) r_{\boldsymbol{y}}\right) (g^t)^2 \\
&+ \left(c_{\boldsymbol{z}} - 12\sigma_1^2\sigma_3^2 r_{\boldsymbol{x}} \kappa - 2\beta_{\boldsymbol{y}}(1 - 2\beta_{\boldsymbol{y}}) r_{\boldsymbol{y}} \sigma_3^2\right) \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 \\
&+ \left(\frac{(1 - 2\beta_{\boldsymbol{y}}) r_{\boldsymbol{y}}}{2\beta_{\boldsymbol{y}}} - 12\sigma_1^2\sigma_5^2 r_{\boldsymbol{x}} \kappa \right) \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2 \\
&- \gamma_t^2 \left(L + r_{\boldsymbol{x}}\right) \|\boldsymbol{d}_{\boldsymbol{x}}^t\|^2 - \gamma_t^2 \left(r_{\boldsymbol{y}} + \frac{\delta}{2} + L + L_d + 3\sigma_1^2 r_{\boldsymbol{x}} \kappa \right) \|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2 \\
&- 3 r_{\boldsymbol{x}} \kappa \|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2,
\end{aligned}$$

where $\kappa > 0$ is any positive constant, and the coefficients are defined as follows:

$$\sigma_1 = \sigma_2 = \frac{r_{\boldsymbol{x}}}{r_{\boldsymbol{x}} - L}, \qquad \nu = 1 - 4L\sigma_6 \operatorname{diam}(\mathcal{Y}),$$

$$\sigma_3 = \frac{r_{\boldsymbol{x}}\sigma_1}{r_{\boldsymbol{y}} - L + \delta} + 1, \qquad B_\beta = 16 \max\{\sigma_8^2, \sigma_9^2\} \left(3\sigma_1^2 r_{\boldsymbol{x}}\kappa + \beta_{\boldsymbol{y}}(1 - 2\beta_{\boldsymbol{y}})r_{\boldsymbol{y}}\right),$$

$$\sigma_4 = \frac{r_{\boldsymbol{y}} + \delta}{r_{\boldsymbol{y}} - L + \delta}, \qquad c_{\boldsymbol{z}} = r_{\boldsymbol{x}} \left(\frac{(1 - \beta_{\boldsymbol{x}})}{\beta_{\boldsymbol{x}}} - 2\sigma_2 - \frac{1}{\kappa}\right),$$

$$\sigma_5 = \frac{r_{\boldsymbol{y}}}{r_{\boldsymbol{y}} - L + \delta}, \qquad s^{\boldsymbol{z}} = c_{\boldsymbol{z}} - 12 r_{\boldsymbol{x}}\kappa\sigma_1^2\sigma_3^2 - 2r_{\boldsymbol{y}}\sigma_3^2\beta_{\boldsymbol{y}}(1 - 2\beta_{\boldsymbol{y}}),$$

$$\sigma_6 = \frac{1}{(r_{\boldsymbol{x}} - L)\operatorname{PWidth}(A)}, \qquad s^{\boldsymbol{v}} = \frac{(1 - 2\beta_{\boldsymbol{y}})r_{\boldsymbol{y}}}{2\beta_{\boldsymbol{y}}} - 12 r_{\boldsymbol{x}}\kappa\sigma_1^2\sigma_5^2,$$

$$\sigma_8 = \frac{\sqrt{2}}{(r_{\boldsymbol{y}} - L + \delta)\operatorname{PWidth}(B)}, \quad s_2^{\boldsymbol{x}} = L + r_{\boldsymbol{x}},$$

$$\sigma_9 = \frac{\sqrt{2}L\sigma_6}{r_{\boldsymbol{y}} - L + \delta}, \qquad s_2^{\boldsymbol{y}} = r_{\boldsymbol{y}} + \frac{\delta}{2} + L + L_d + 3\sigma_1^2 r_{\boldsymbol{x}}\kappa.$$

Choice of parameters. We next verify that the coefficients in the preceding estimate satisfy the bounds stated in Proposition 4.1.

- As $r_{\boldsymbol{x}} \geq 2L$, we have $\sigma_1 = \sigma_2 \leq 2$ and $s_2^{\boldsymbol{x}} \leq \frac{3r_{\boldsymbol{x}}}{2}$. Similarly, $r_{\boldsymbol{y}} \geq 2L$ yields

$$\sigma_3 \leq \frac{2r_{\boldsymbol{x}}}{L} + 1, \sigma_5 \leq 2, \text{and } \sigma_8 \leq \frac{3}{L\operatorname{PWidth}(B)}.$$

  With these bounds, setting $\kappa \coloneqq 2\beta_{\boldsymbol{x}}$, $\beta_{\boldsymbol{x}} \leq \min\{\frac{r_{\boldsymbol{y}}}{1536 r_{\boldsymbol{x}}}, \frac{4L^2 r_{\boldsymbol{x}}}{80L^2 r_{\boldsymbol{x}} + 9r_{\boldsymbol{y}}(2r_{\boldsymbol{x}} + L)^2}\}$, $\beta_{\boldsymbol{y}} \leq \frac{1}{5}$, and $\delta \leq \frac{L}{3}$, we can further bound other coefficients:

$$s^{\boldsymbol{v}} = \frac{(1 - 2\beta_{\boldsymbol{y}})r_{\boldsymbol{y}}}{2\beta_{\boldsymbol{y}}} - 12 r_{\boldsymbol{x}}\kappa\sigma_1^2\sigma_5^2 \geq \frac{(1 - 2\beta_{\boldsymbol{y}})r_{\boldsymbol{y}}}{2\beta_{\boldsymbol{y}}} - 384 r_{\boldsymbol{x}}\beta_{\boldsymbol{x}} \geq \frac{r_{\boldsymbol{y}}}{\beta_{\boldsymbol{y}}}\left(\frac{1}{2} - \frac{5\beta_{\boldsymbol{y}}}{4}\right) \geq \frac{r_{\boldsymbol{y}}}{4\beta_{\boldsymbol{y}}},$$

$$s_2^{\boldsymbol{y}} = r_{\boldsymbol{y}} + \frac{\delta}{2} + L + L_d + 3\sigma_1^2 r_{\boldsymbol{x}}\kappa \leq \frac{3\delta}{2} + 4L + 2r_{\boldsymbol{y}} + 24 r_{\boldsymbol{x}}\beta_{\boldsymbol{x}} \leq \frac{3\delta}{2} + 4L + 3r_{\boldsymbol{y}} \leq 6r_{\boldsymbol{y}}, \text{ and}$$

$$\begin{aligned} s^{\boldsymbol{z}} &= r_{\boldsymbol{x}} \left(\frac{(1 - \beta_{\boldsymbol{x}})}{\beta_{\boldsymbol{x}}} - 2\sigma_2 - \frac{1}{\kappa}\right) - 12 r_{\boldsymbol{x}}\kappa\sigma_1^2\sigma_3^2 - 2r_{\boldsymbol{y}}\sigma_3^2\beta_{\boldsymbol{y}}(1 - 2\beta_{\boldsymbol{y}}) \\ &\geq r_{\boldsymbol{x}} \left(\frac{1}{2\beta_{\boldsymbol{x}}} - 5\right) - 2\sigma_3^2 \left(48 r_{\boldsymbol{x}}\beta_{\boldsymbol{x}} + \beta_{\boldsymbol{y}}(1 - 2\beta_{\boldsymbol{y}})r_{\boldsymbol{y}}\right) \\ &\geq r_{\boldsymbol{x}} \left(\frac{1}{2\beta_{\boldsymbol{x}}} - 5\right) - \frac{9r_{\boldsymbol{y}}}{16} \left(\frac{4r_{\boldsymbol{x}}^2}{L^2} + 1 + \frac{4r_{\boldsymbol{x}}}{L}\right) \\ &\geq \frac{r_{\boldsymbol{x}}}{\beta_{\boldsymbol{x}}} \left(\frac{1}{2} - \left(5 + \frac{9r_{\boldsymbol{x}}r_{\boldsymbol{y}}}{4L^2} + \frac{9r_{\boldsymbol{y}}}{16r_{\boldsymbol{x}}} + \frac{9r_{\boldsymbol{y}}}{4L}\right)\beta_{\boldsymbol{x}}\right) \geq \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}}. \end{aligned}$$

  Combining the above estimates with $\|\boldsymbol{d}_{\boldsymbol{x}}^t\| \leq \operatorname{diam}(\mathcal{X})$ and $\|\boldsymbol{d}_{\boldsymbol{y}}^t\| \leq \operatorname{diam}(\mathcal{Y})$, we obtain the following uniform bound on the quadratic curvature term:

$$(L + r_{\boldsymbol{x}})\|\boldsymbol{d}_{\boldsymbol{x}}^t\|^2 + \left(r_{\boldsymbol{y}} + \frac{\delta}{2} + L + L_d + 3\sigma_1^2 r_{\boldsymbol{x}}\kappa\right)\|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2 \leq \frac{3r_{\boldsymbol{x}}}{2}\operatorname{diam}(\mathcal{X})^2 + 6r_{\boldsymbol{y}}\operatorname{diam}(\mathcal{Y})^2.$$

  Accordingly, we define

$$\bar{C} \coloneqq \frac{3r_{\boldsymbol{x}}}{2}\operatorname{diam}(\mathcal{X})^2 + 6r_{\boldsymbol{y}}\operatorname{diam}(\mathcal{Y})^2.$$

- We further impose $r_{\boldsymbol{x}} \geq L + \frac{8L\operatorname{diam}(\mathcal{Y})}{\operatorname{PWidth}(A)}$. Then

$$\sigma_6 \leq \frac{1}{8L\operatorname{diam}(\mathcal{Y})}, \sigma_8 \leq \frac{3}{L\operatorname{PWidth}(B)}, \text{and } \sigma_9 \leq \frac{1}{4L\operatorname{diam}(\mathcal{Y})}.$$

Since $\mathcal{Y} = \mathrm{conv}(B)$, the pyramidal width is bounded above by the diameter of $\mathcal{Y}$, i.e., $\mathrm{PWidth}(B) \le \mathrm{diam}(\mathcal{Y})$. Moreover, suppose that

$$\beta_{\boldsymbol{x}} \le \frac{L^2\,\mathrm{PWidth}(B)^2}{221184\,\bar{C}\,r_{\boldsymbol{x}}}, \beta_{\boldsymbol{y}} \le \frac{L^2\,\mathrm{PWidth}(B)^2}{18432\,\bar{C}\,r_{\boldsymbol{y}}}.$$

Then, recalling that $\kappa = 2\beta_{\boldsymbol{x}}$ and $\sigma_1 \le 2$, we can further bound the coefficients:

$$\begin{aligned}\nu &= 1 - 4L\sigma_6\,\mathrm{diam}(\mathcal{Y}) \ge \frac{1}{2}, \text{and}\\ B_\beta &= 16\max\{\sigma_8^2, \sigma_9^2\}\left(3\sigma_1^2 r_{\boldsymbol{x}}\kappa + \beta_{\boldsymbol{y}}(1-2\beta_{\boldsymbol{y}})r_{\boldsymbol{y}}\right)\\ &\le \frac{144}{L^2\,\mathrm{PWidth}(B)^2}\left(24 r_{\boldsymbol{x}}\beta_{\boldsymbol{x}} + 2 r_{\boldsymbol{y}}\beta_{\boldsymbol{y}}\right) \le \frac{1}{32\bar{C}}.\end{aligned}$$

Putting them together finishes the proof. □

F.3 Clipped Lyapunov estimate

We now consider a clipped iteration. By the update rule of Algorithm 1, the proximal variables remain unchanged:

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t, \qquad \boldsymbol{v}^{t+1} = \boldsymbol{v}^t.$$

The corresponding Lyapunov estimate is given below.

Lemma F.5 (Clipped-step Lyapunov bound)**.** Under the assumptions and parameter conditions of Proposition 4.1, suppose that $\boldsymbol{z}^{t+1} = \boldsymbol{z}^t$ and $\boldsymbol{v}^{t+1} = \boldsymbol{v}^t$. Then

$$\Phi^t - \Phi^{t+1} \ge \frac{\gamma_t}{2} g^t - \bar{C}\gamma_t^2.$$

□

Proof. Since $\boldsymbol{z}$ and $\boldsymbol{v}$ are fixed, all terms associated with the proximal updates vanish. Repeating the primal and dual descent arguments with $\boldsymbol{z}^{t+1} = \boldsymbol{z}^t$ and $\boldsymbol{v}^{t+1} = \boldsymbol{v}^t$ gives

$$\begin{aligned}\Phi^t - \Phi^{t+1} &\ge \gamma_t g^t + 2\langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^{t+1} - \boldsymbol{y}^t\rangle\\ &\quad - \gamma_t^2(L + r_{\boldsymbol{x}})\|\boldsymbol{d}_{\boldsymbol{x}}^t\|^2 - \gamma_t^2\left(r_{\boldsymbol{y}} + \frac{\delta}{2} + L + L_d\right)\|\boldsymbol{d}_{\boldsymbol{y}}^t\|^2\\ &\ge (1 - 4L\sigma_6\,\mathrm{diam}(\mathcal{Y}))\,\gamma_t g^t - \bar{C}\gamma_t^2,\end{aligned}$$

where the second inequality follows from the bound in equation 8 and the definition of $\bar{C}$. Using the choice of $r_{\boldsymbol{x}}$ finishes the proof. □

Proof of Proposition 4.1. Suppose first that iteration t is non-clipped. By Algorithm 1, $\gamma_t = \widehat{\gamma}_t = \frac{g^t}{4\bar{C}}$. Applying Lemma F.4 and using

$$\frac{\gamma_t}{2} g^t - \bar{C}\gamma_t^2 - \frac{1}{32\bar{C}}(g^t)^2 = \frac{1}{8\bar{C}}(g^t)^2 - \frac{1}{16\bar{C}}(g^t)^2 - \frac{1}{32\bar{C}}(g^t)^2 = \frac{1}{32\bar{C}}(g^t)^2,$$

we obtain

$$\begin{aligned}\Phi^t - \Phi^{t+1} \ge \frac{1}{32\bar{C}}(g^t)^2 &+ \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{r_{\boldsymbol{y}}}{4\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2\\ &- 6 r_{\boldsymbol{x}}\beta_{\boldsymbol{x}}\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2,\end{aligned} \tag{13}$$

which proves equation 3.

Suppose now that iteration t is clipped. Then $\gamma_t = \gamma_{\max}^t < \widehat{\gamma}_t = g^t/(4\bar{C})$, and the proximal variables remain unchanged. Lemma F.5 therefore gives

$$\Phi^t - \Phi^{t+1} \ge \frac{\gamma_t}{2} g^t - \bar{C}\gamma_t^2 > \frac{\gamma_t}{4} g^t \ge 0,$$

which proves equation 4. □

G PROOF OF PROPOSITION 4.2

The non-clipped Lyapunov estimate contains the coupling term $\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2$, which can be further controlled by the dual proximal residual $\|\boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}) - \boldsymbol{v}^t\|$. Proposition 4.2 provides the required proximal error bounds under the structural assumptions considered in our convergence analysis. Parts (ii) follows directly from (Zheng et al., 2023, Proposition 2), with the notation and KŁ normalization used in this paper. We prove part (i) below, exploiting the strong concavity induced by the quadratic perturbation.

Proof of Proposition 4.2. Suppose Assumption 2.2(i) holds and $\delta > 0$. For later use, define the proximal dual envelope $h_\delta(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}) := \max_{\boldsymbol{y}\in\mathcal{Y}} F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$, and let $\boldsymbol{y}_\delta(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}) := \operatorname{argmax}_{\boldsymbol{y}\in\mathcal{Y}} F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$.

Since $f(\boldsymbol{x}, \cdot)$ is concave and the perturbation $-\frac{\delta}{2}\|\cdot - \bar{\boldsymbol{y}}\|^2$ is δ-strongly concave, the maximizer $\boldsymbol{y}_\delta(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v})$ is unique. Moreover, by Danskin's theorem, $\nabla_{\boldsymbol{v}} h_\delta(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}) = r_{\boldsymbol{y}}\,(\boldsymbol{y}_\delta(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{v}) - \boldsymbol{v})$. Moreover, for $\delta > 0$, we have $h_\delta(\boldsymbol{x}, \boldsymbol{z}, \cdot)$ is $\frac{r_{\boldsymbol{y}}\delta}{r_{\boldsymbol{y}}+\delta}$-strongly concave (Renaud et al., 2025, Lemma 5.5).

We now prove the desired error bound. Since $F_\delta(\cdot, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v})$ is $(r_{\boldsymbol{x}} - L)$-strongly convex uniformly over $\boldsymbol{y} \in \mathcal{Y}$, the function $h_\delta(\cdot, \boldsymbol{z}, \boldsymbol{v})$ is also $(r_{\boldsymbol{x}} - L)$-strongly convex. Moreover, by the definition of $\boldsymbol{x}(\boldsymbol{z}, \boldsymbol{v})$, we have

$$\begin{aligned}
&\max_{\boldsymbol{v}\in\mathbb{R}^m} h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}) - h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) \\
&\geq h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) \\
&\geq \frac{r_{\boldsymbol{x}} - L}{2}\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1}))\|^2. \qquad (14)
\end{aligned}$$

On the other hand, by the definition of $\boldsymbol{v}(\boldsymbol{z}^{t+1})$, we obtain

$$\begin{aligned}
h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) &= \max_{\boldsymbol{v}\in\mathbb{R}^m} h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}), \boldsymbol{z}^{t+1}, \boldsymbol{v}) \\
&\geq h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})). \qquad (15)
\end{aligned}$$

Combining equation 14 and equation 15 yields

$$\begin{aligned}
&\frac{r_{\boldsymbol{x}} - L}{2}\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1}))\|^2 \\
&\leq \max_{\boldsymbol{v}\in\mathbb{R}^m} h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}) - h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})).
\end{aligned}$$

By $\frac{r_{\boldsymbol{y}}\delta}{r_{\boldsymbol{y}}+\delta}$-strong concavity of $h_\delta(\boldsymbol{x}, \boldsymbol{z}, \cdot)$, we obtain

$$\begin{aligned}
&\max_{\boldsymbol{v}\in\mathbb{R}^m} h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}) - h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})) \\
&\leq \frac{r_{\boldsymbol{y}} + \delta}{r_{\boldsymbol{y}}\delta}\|\nabla_{\boldsymbol{v}} h_\delta(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2 \\
&\leq \frac{r_{\boldsymbol{y}}^2(r_{\boldsymbol{y}} + \delta)}{r_{\boldsymbol{y}}\delta}\left(\frac{1 - \beta_{\boldsymbol{y}}}{\beta_{\boldsymbol{y}}} + \sigma_5\right)^2 \|\boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}) - \boldsymbol{v}^t\|^2.
\end{aligned}$$

where the last inequality follows from the same argument as (Zheng et al., 2023, Eq. (22)), namely,

$$\left\|\nabla_{\boldsymbol{v}} h_\delta\left(\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})), \boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\right)\right\| \leq r_{\boldsymbol{y}}\left(\frac{1 - \beta_{\boldsymbol{y}}}{\beta_{\boldsymbol{y}}} + \sigma_5\right)\|\boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}) - \boldsymbol{v}^t\|.$$

Combining the preceding inequalities gives

$$\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1}))\|^2 \leq \frac{2r_{\boldsymbol{y}}(r_{\boldsymbol{y}} + \delta)}{(r_{\boldsymbol{x}} - L)\delta}\left(\frac{1 - \beta_{\boldsymbol{y}}}{\beta_{\boldsymbol{y}}} + \sigma_5\right)^2 \|\boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}) - \boldsymbol{v}^t\|^2.$$

This finishes the proof. □

H PROOFS OF LEMMA 4.3 AND THEOREM 3.1

We now combine the one-step Lyapunov estimates in Proposition 4.1, the proximal error bounds in Proposition 4.2, and an active-set counting argument to prove Theorem 3.1.

For non-clipped iterations, Proposition 4.1 already provides the required one-step estimate; its remaining primal–dual coupling term will be controlled using Proposition 4.2 in the rate analysis below. Clipped iterations require additional consideration because clipped Frank–Wolfe and clipped away iterations have different implications for the global convergence analysis. We first record these properties, then establish the active-set counting bound, and finally derive the convergence rates.

H.1 PROPERTIES OF CLIPPED ITERATIONS

At every clipped iteration, Proposition 4.1 gives

$$\Phi^t - \Phi^{t+1} > \frac{\gamma_t}{4} g^t \geq 0. \tag{16}$$

Hence, every clipped iteration is Lyapunov non-increasing.

We next distinguish the two possible selected directions. If the joint Frank–Wolfe direction is selected, then $\gamma_{\max}^t = 1$. Since the iteration is clipped, $\widehat{\gamma}_t > 1$, and therefore $g^t > 4\bar{C}$. Using equation 16 with $\gamma_t = 1$ gives

$$\Phi^t - \Phi^{t+1} > \frac{1}{4} g^t > \bar{C}. \tag{17}$$

Thus, every clipped Frank–Wolfe iteration yields a Lyapunov decrease uniformly bounded away from zero.

If instead the joint away direction is selected, the iteration remains Lyapunov non-increasing by equation 16, but its decrease need not be uniformly bounded away from zero. The number of such iterations is therefore controlled through the evolution of the active sets.

H.2 PROOF OF LEMMA 4.3

We prove the active-set counting bound in Lemma 4.3. Following the product-domain away-step Frank–Wolfe construction of Gidel et al. (2017), we maintain separate convex-combination representations for the primal and dual iterates. At iteration t, write

$$\boldsymbol{x}^t = \sum_{\boldsymbol{a} \in S_{\boldsymbol{x}}^t} \alpha_{\boldsymbol{x},\boldsymbol{a}}^t \, \boldsymbol{a}, \qquad \boldsymbol{y}^t = \sum_{\boldsymbol{b} \in S_{\boldsymbol{y}}^t} \alpha_{\boldsymbol{y},\boldsymbol{b}}^t \, \boldsymbol{b},$$

where the coefficients are nonnegative and sum to one in each representation.

Let $\boldsymbol{s}_{\min}^t = [\boldsymbol{s}_{\min,\boldsymbol{x}}^t; \boldsymbol{s}_{\min,\boldsymbol{y}}^t]$ and $\boldsymbol{s}_{\max}^t = [\boldsymbol{s}_{\max,\boldsymbol{x}}^t; \boldsymbol{s}_{\max,\boldsymbol{y}}^t]$ denote the Frank–Wolfe and away atoms, respectively. Coefficients of atoms outside the current active sets are understood to be zero.

If the joint Frank–Wolfe direction is selected, the standard Frank–Wolfe coefficient update gives

$$\alpha_{\boldsymbol{x},\boldsymbol{a}}^{t+1} = (1-\gamma_t)\alpha_{\boldsymbol{x},\boldsymbol{a}}^t + \gamma_t \mathbf{1}_{\{\boldsymbol{a} = \boldsymbol{s}_{\min,\boldsymbol{x}}^t\}}, \qquad \boldsymbol{a} \in \mathcal{A},$$
$$\alpha_{\boldsymbol{y},\boldsymbol{b}}^{t+1} = (1-\gamma_t)\alpha_{\boldsymbol{y},\boldsymbol{b}}^t + \gamma_t \mathbf{1}_{\{\boldsymbol{b} = \boldsymbol{s}_{\min,\boldsymbol{y}}^t\}}, \qquad \boldsymbol{b} \in \mathcal{B}.$$

If the joint away direction is selected, then

$$\alpha_{\boldsymbol{x},\boldsymbol{a}}^{t+1} = (1+\gamma_t)\alpha_{\boldsymbol{x},\boldsymbol{a}}^t - \gamma_t \mathbf{1}_{\{\boldsymbol{a} = \boldsymbol{s}_{\max,\boldsymbol{x}}^t\}}, \qquad \boldsymbol{a} \in \mathcal{A},$$
$$\alpha_{\boldsymbol{y},\boldsymbol{b}}^{t+1} = (1+\gamma_t)\alpha_{\boldsymbol{y},\boldsymbol{b}}^t - \gamma_t \mathbf{1}_{\{\boldsymbol{b} = \boldsymbol{s}_{\max,\boldsymbol{y}}^t\}}, \qquad \boldsymbol{b} \in \mathcal{B}.$$

The active sets are therefore

$$S_{\boldsymbol{x}}^{t+1} = \{\boldsymbol{a} \in \mathcal{A} : \alpha_{\boldsymbol{x},\boldsymbol{a}}^{t+1} > 0\}, \qquad S_{\boldsymbol{y}}^{t+1} = \{\boldsymbol{b} \in \mathcal{B} : \alpha_{\boldsymbol{y},\boldsymbol{b}}^{t+1} > 0\}.$$

For an away step, preserving nonnegativity of the maintained convex-combination coefficients requires

$$\gamma_t \leq \gamma_{\max}^t := \min\left\{ \frac{\alpha_{\boldsymbol{x},\boldsymbol{s}_{\max,\boldsymbol{x}}^t}^t}{1 - \alpha_{\boldsymbol{x},\boldsymbol{s}_{\max,\boldsymbol{x}}^t}^t}, \frac{\alpha_{\boldsymbol{y},\boldsymbol{s}_{\max,\boldsymbol{y}}^t}^t}{1 - \alpha_{\boldsymbol{y},\boldsymbol{s}_{\max,\boldsymbol{y}}^t}^t} \right\}, \tag{18}$$

with the convention $\alpha/(1-\alpha) = +\infty$ when $\alpha = 1$. If an away iteration is clipped, then $\gamma_t = \gamma_{\max}^t$. Hence at least one of the two selected away-atom coefficients becomes zero after the update. Every clipped away iteration therefore removes at least one atom from the two active sets. In contrast, a joint Frank–Wolfe iteration can add at most one atom to each active set and thus increases their total size by at most two.

Proof of Lemma 4.3. We now count the three types of iterations. Let

$$\mathcal{I}_{\mathrm{nc}}(T) := \{0 \le t < T : \widehat{\gamma}_t \le \gamma_{\max}^t\},$$

and let $\mathcal{I}_{\mathrm{cFW}}(T)$ and $\mathcal{I}_{\mathrm{cA}}(T)$ denote the sets of clipped Frank–Wolfe and clipped away iterations, respectively. Define

$$N_{\mathrm{nc}}(T) := |\mathcal{I}_{\mathrm{nc}}(T)|, \qquad N_{\mathrm{cFW}}(T) := |\mathcal{I}_{\mathrm{cFW}}(T)|, \qquad N_{\mathrm{cA}}(T) := |\mathcal{I}_{\mathrm{cA}}(T)|.$$

These sets form a partition of the first T iterations, so

$$T = N_{\mathrm{nc}}(T) + N_{\mathrm{cFW}}(T) + N_{\mathrm{cA}}(T). \tag{19}$$

Let $m_t := |S_{\boldsymbol{x}}^t| + |S_{\boldsymbol{y}}^t|$ be the total active-set size, and let $N_{\mathrm{FW}}(T)$ denote the number of iterations among the first T iterations at which the joint Frank–Wolfe direction is selected. Since each joint Frank–Wolfe iteration increases m_t by at most two and each clipped away iteration decreases m_t by at least one, while any additional atom removals can only decrease m_t, we have

$$m_T \le m_0 + 2N_{\mathrm{FW}}(T) - N_{\mathrm{cA}}(T).$$

Both active sets remain nonempty, so $m_T \ge 2$. Therefore,

$$N_{\mathrm{cA}}(T) \le 2N_{\mathrm{FW}}(T) + m_0 - 2. \tag{20}$$

Every joint Frank–Wolfe iteration is either non-clipped or a clipped Frank–Wolfe iteration. Hence

$$N_{\mathrm{FW}}(T) \le N_{\mathrm{nc}}(T) + N_{\mathrm{cFW}}(T).$$

Substituting this bound into equation 20 yields

$$N_{\mathrm{cA}}(T) \le 2\big(N_{\mathrm{nc}}(T) + N_{\mathrm{cFW}}(T)\big) + m_0 - 2. \tag{21}$$

Combining equation 21 with equation 19, we obtain

$$N_{\mathrm{nc}}(T) + N_{\mathrm{cFW}}(T) \ge \frac{T - m_0 + 2}{3},$$

which finishes the proof.

□

H.3 Proof of Theorem 3.1

We now combine Proposition 4.1, Proposition 4.2, the clipped-iteration properties above, and the active-set counting bound to establish the rates in Theorem 3.1.

Proof of Theorem 3.1. For later use, define

$$\Delta_\Phi := \Phi^0 - \underline{F}.$$

The quantity Δ_Φ is bounded independently of the horizon T. Throughout the proof, the parameters satisfy the conditions of Proposition 4.1; additional choices required in each case are specified separately. All constants hidden in $\mathcal{O}(\cdot)$ are independent of T.

Case 1: KŁ exponent $\theta \in (1/2, 1)$. Set $\delta = 0$. For each $t \in \mathcal{I}_{\mathrm{nc}}(T)$, define

$$\xi_{\mathrm{Pos}}^t := \max\left\{\frac{1}{32\bar{C}}(g^t)^2,\ \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2,\ \frac{r_{\boldsymbol{y}}}{4\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2\right\},$$
$$\xi_{\mathrm{Neg}}^t := 6r_{\boldsymbol{x}}\beta_{\boldsymbol{x}}\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2.$$

Thus, ξ_{Pos}^t captures the largest of the three positive terms in equation 13, whereas ξ_{Neg}^t is the primal–dual coupling term that must be controlled. We distinguish two mutually exclusive possibilities.

(i) **When $\xi^t_{\text{Pos}}/2 \geq \xi^t_{\text{Neg}}$ for every $t \in \mathcal{I}_{\text{nc}}(T)$:** Under this condition, since ξ^t_{Pos} is no larger than the sum of the three positive terms in equation 13, every non-clipped iteration satisfies

$$\Phi^t - \Phi^{t+1} \geq \frac{1}{64\bar{C}}(g^t)^2 + \frac{r_{\boldsymbol{x}}}{8\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{r_{\boldsymbol{y}}}{8\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\|^2, \qquad t \in \mathcal{I}_{\text{nc}}(T). \tag{22}$$

Summing equation 22 over the non-clipped iterations, together with equation 17 and the non-increase of Φ on the clipped away iterations, gives

$$\sum_{t\in\mathcal{I}_{\text{nc}}(T)}\left[\frac{1}{64\bar{C}}(g^t)^2 + \frac{r_{\boldsymbol{x}}}{8\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{r_{\boldsymbol{y}}}{8\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\|^2\right] + \frac{\bar{C}}{2}N_{\text{cFW}}(T) \leq \Delta_\Phi. \tag{23}$$

In particular,

$$N_{\text{cFW}}(T) \leq \frac{2\Delta_\Phi}{\bar{C}} = \mathcal{O}(1). \tag{24}$$

Combining this with Lemma 4.3, we obtain, for all sufficiently large T,

$$N_{\text{nc}}(T) \geq \frac{T - m_0 + 2}{3} - \frac{2\Delta_\Phi}{\bar{C}} \geq \frac{T}{4}. \tag{25}$$

It follows from equation 23 that there exists some $t \in \mathcal{I}_{\text{nc}}(T)$ satisfying

$$\mathcal{R}_t = g^t + \frac{1}{\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\| + \frac{1}{\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\| \leq \mathcal{O}\left(\frac{1}{\sqrt{\beta_{\boldsymbol{x}}T}}\right),$$

where $\beta_{\boldsymbol{y}} > 0$ is fixed independently of T. Since $\delta = 0$, Lemma E.1 gives an $\mathcal{O}((\beta_{\boldsymbol{x}}T)^{-1/2})$-FWSP.

(ii) **There exists some $t \in \mathcal{I}_{\text{nc}}(T)$ such that $\xi^t_{\text{Pos}}/2 < \xi^t_{\text{Neg}}$:** In this case, the negative terms dominate, there exists some $t \in \mathcal{I}_{\text{nc}}(T)$ such that

$$\frac{r_{\boldsymbol{y}}}{8\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\|^2 \leq 6r_{\boldsymbol{x}}\beta_{\boldsymbol{x}}\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}^t_+(\boldsymbol{z}^{t+1}))\|^2.$$

Applying Proposition 4.2(ii) yields

$$\|\boldsymbol{v}^t - \boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\| \leq \mathcal{O}(\beta_{\boldsymbol{x}}^{\theta/(2\theta-1)}),$$

and $\xi^t_{\text{Pos}}/2 \leq \xi^t_{\text{Neg}}$ implies

$$g^t \leq \mathcal{O}(\beta_{\boldsymbol{x}}^{\theta/(2\theta-1)}), \qquad \frac{1}{\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\| \leq \mathcal{O}(\beta_{\boldsymbol{x}}^{1/(4\theta-2)}).$$

Therefore, there exists some $t \in \mathcal{I}_{\text{nc}}(T)$ such that $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}\left(\beta_{\boldsymbol{x}}^{1/(4\theta-2)}\right)$-FWSP.

Choosing $\beta_{\boldsymbol{x}} = \Theta\left(T^{-(2\theta-1)/(2\theta)}\right)$, then irrespective of which of the two possibilities above occurs, there exists some $t \in \mathcal{I}_{\text{nc}}(T)$ such that the point $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}\left(T^{-1/(4\theta)}\right)$-FWSP of Problem P.

Case 2: KŁ exponent $\theta \in (0, 1/2]$. Again set $\delta = 0$. In this setting, boundedness of the dual feasible region allows us to convert the KŁ error bound into a quadratic one, so that the negative coupling term can be absorbed uniformly. By Proposition 4.2, we have

$$\begin{aligned}\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}^t_+(\boldsymbol{z}^{t+1}))\|^2 &\leq \omega_1\|\boldsymbol{v}^t - \boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\|^{1/\theta} \\ &\leq \omega_1 \operatorname{diam}(\mathcal{Y})^{1/\theta-2}\|\boldsymbol{v}^t - \boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})\|^2, \end{aligned} \tag{26}$$

where the second inequality follows from the boundedness of $\boldsymbol{v}^t$ and $\boldsymbol{v}^t_+(\boldsymbol{z}^{t+1})$. Fix admissible $\beta_{\boldsymbol{y}} > 0$ and choose $\beta_{\boldsymbol{x}} > 0$ independently of T such that

$$6r_{\boldsymbol{x}}\beta_{\boldsymbol{x}}\omega_1 \operatorname{diam}(\mathcal{Y})^{1/\theta-2} \leq \frac{r_{\boldsymbol{y}}}{8\beta_{\boldsymbol{y}}}, \tag{27}$$

in addition to the parameter conditions in Proposition 4.1. Substituting equation 26 into equation 13 yields the sufficient descent

$$\Phi^t - \Phi^{t+1} \geq \frac{1}{32\bar{C}}(g^t)^2 + \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{r_{\boldsymbol{y}}}{8\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2, \qquad t \in \mathcal{I}_{\mathrm{nc}}(T). \quad (28)$$

Summing equation 28, together with equation 17 and the non-increase on clipped away iterations, gives

$$\sum_{t\in\mathcal{I}_{\mathrm{nc}}(T)} \left[\frac{1}{32\bar{C}}(g^t)^2 + \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{r_{\boldsymbol{y}}}{8\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2\right] + \frac{\bar{C}}{2}N_{\mathrm{cFW}}(T) \leq \Delta_\Phi. \quad (29)$$

Hence, $N_{\mathrm{cFW}}(T) = \mathcal{O}(1)$. Together with Lemma 4.3, this implies $N_{\mathrm{nc}}(T) = \Theta(T)$. In particular, for all sufficiently large T, $N_{\mathrm{nc}}(T) \geq T/4$. Consequently, there exists $t \in \mathcal{I}_{\mathrm{nc}}(T)$ such that

$$\mathcal{R}_t = g^t + \frac{1}{\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\| + \frac{1}{\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\| \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

By the stationarity-transfer result, $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}(T^{-1/2})$-FWSP of Problem P.

We finally consider the concave-dual setting. Unlike the KŁ regime, where we set $\delta = 0$, here the quadratic perturbation turns the dual objective into a strongly concave function. This yields a quadratic proximal error bound and allows the coupling term in equation 13 to be absorbed by choosing $\beta_{\boldsymbol{x}}$ proportional to δ.

Case 3: General concave setting. Suppose now that $f(\boldsymbol{x}, \cdot)$ is concave for every $\boldsymbol{x}$. Because

$$f_\delta(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{y}) - \frac{\delta}{2}\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^2,$$

the perturbed objective $f_\delta(\boldsymbol{x}, \cdot)$ is δ-strongly concave. Consequently, the corresponding dual subproblem satisfies a quadratic error bound. In particular, the proximal error estimate in Proposition 4.2 specializes to

$$\|\boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}(\boldsymbol{z}^{t+1})) - \boldsymbol{x}(\boldsymbol{z}^{t+1}, \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1}))\|^2 \leq \omega_0\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2, \quad (30)$$

where there exist constants $\bar{\delta} > 0$ and $C_{\mathrm{sc}} > 0$, independent of δ and T, such that $\omega_0 \leq \frac{C_{\mathrm{sc}}}{\delta}$ and $0 < \delta \leq \bar{\delta}$. Substituting equation 30 into equation 13, we obtain

$$\Phi^t - \Phi^{t+1} \geq \frac{1}{32\bar{C}}(g^t)^2 + \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \left(\frac{r_{\boldsymbol{y}}}{4\beta_{\boldsymbol{y}}} - \frac{6r_{\boldsymbol{x}}C_{\mathrm{sc}}\beta_{\boldsymbol{x}}}{\delta}\right)\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2.$$

By further imposing $\beta_{\boldsymbol{x}} \leq \frac{r_{\boldsymbol{y}}\delta}{48\beta_{\boldsymbol{y}}r_{\boldsymbol{x}}C_{\mathrm{sc}}}$, we have every non-clipped iteration satisfies the sufficient descent inequality

$$\Phi^t - \Phi^{t+1} \geq \frac{1}{32\bar{C}}(g^t)^2 + \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{r_{\boldsymbol{y}}}{8\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2, \qquad t \in \mathcal{I}_{\mathrm{nc}}(T). \quad (31)$$

Therefore, summing over all iteration types gives

$$\sum_{t\in\mathcal{I}_{\mathrm{nc}}(T)} \left[\frac{1}{32\bar{C}}(g^t)^2 + \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{r_{\boldsymbol{y}}}{8\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2\right] + \frac{\bar{C}}{2}N_{\mathrm{cFW}}(T) \leq \Delta_\Phi. \quad (32)$$

It follows that $N_{\mathrm{cFW}}(T) = \mathcal{O}(1), N_{\mathrm{nc}}(T) = \Theta(T)$. Hence there exists some $t \in \mathcal{I}_{\mathrm{nc}}(T)$ satisfying

$$\mathcal{R}_t = g^t + \frac{1}{\beta_{\boldsymbol{x}}}\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\| + \frac{1}{\beta_{\boldsymbol{y}}}\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\| \leq \mathcal{O}\left(\frac{1}{\sqrt{T\beta_{\boldsymbol{x}}}}\right) = \mathcal{O}\left(\frac{1}{\sqrt{\delta T}}\right).$$

where the last equality follows from $\beta_{\boldsymbol{x}} = \Theta(\delta)$.

Applying Lemma E.1, we obtain

$$\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}^t, \boldsymbol{y}^t) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}^t, \boldsymbol{y}^t) \leq \mathcal{O}\left(\frac{1}{\sqrt{\delta T}} + \delta\right). \quad (33)$$

Balancing the last two terms gives $\delta = \Theta(T^{-1/3})$ and $\beta_{\boldsymbol{x}} = \Theta(T^{-1/3})$. Substituting these choices into equation 33, we conclude that

$$\min_{0\le t<T} \left\{\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}^t, \boldsymbol{y}^t) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}^t, \boldsymbol{y}^t)\right\} = \mathcal{O}(T^{-1/3}). \tag{34}$$

Thus, in the general concave setting, the iteration complexity for an ϵ-FWSP is $\mathcal{O}(\epsilon^{-3})$. Combining the three cases completes the proof. □

I Proof of Theorem 3.2

We now exploit the additional primal convexity available in the convex–concave setting. Since C–C problems are a special case of NC–C problems, the one-step Lyapunov estimates, the control of the primal–dual coupling term, and the clipped-iteration analysis developed above remain valid.

Under the C–C assumptions, the shifted Lyapunov function can be controlled by the algorithmic residual defined in equation 6. This relation relies on the convex–concave structure of the regularized problem and is the additional ingredient needed beyond the NC–C analysis. Combining it with the sufficient descent estimate for non-clipped iterations yields a quadratic recursion for the Lyapunov function. We then combine this recursion with the active-set counting argument to obtain the improved stationarity rate.

Throughout this section, $f(\cdot, \boldsymbol{y})$ is convex on $\mathcal{X}$ for every $\boldsymbol{y} \in \mathcal{Y}$, and $f(\boldsymbol{x}, \cdot)$ is concave on $\mathcal{Y}$ for every $\boldsymbol{x} \in \mathcal{X}$. Define the shifted Lyapunov function

$$\widetilde{\Phi}^t := \Phi^t - \underline{F}, \qquad \underline{F} := \min_{\boldsymbol{z}\in\mathbb{R}^n} q_\delta(\boldsymbol{z}).$$

Since $\Phi^t \ge \underline{F}$, we have $\widetilde{\Phi}^t \ge 0$. Moreover, the shift is independent of t, so all one-step estimates for Φ^t hold unchanged for $\widetilde{\Phi}^t$.

I.1 Lyapunov control by the algorithmic residual

The additional primal convexity allows the shifted Lyapunov function to be controlled by the algorithmic residual defined in equation 6.

Lemma I.1 (Lyapunov control by the algorithmic residual)**.** Suppose that the parameters satisfy the conditions of Proposition 4.1. Then there exists a constant $C_{\mathrm{cc}} > 0$, independent of t, T, δ, and the admissible choice of $\beta_{\boldsymbol{x}}$, such that every non-clipped iteration t satisfies

$$0 \le \widetilde{\Phi}^t \le C_{\mathrm{cc}} \mathcal{R}_t.$$

□

Proof. Recall that

$$\widetilde{\Phi}^t := \Phi^t - \underline{F}, \qquad \underline{F} = \min_{\boldsymbol{z}} q_\delta(\boldsymbol{z}).$$

By the definition of Φ, we can decompose

$$\begin{aligned}\widetilde{\Phi}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) &= F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) - 2d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) + 2q_\delta(\boldsymbol{z}) - \underline{F} \\ &= F_\delta(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) - d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) + p_\delta(\boldsymbol{z}, \boldsymbol{v}) - d_\delta(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) \\ &\quad + q_\delta(\boldsymbol{z}) - p_\delta(\boldsymbol{z}, \boldsymbol{v}) + q_\delta(\boldsymbol{z}) - \underline{F}.\end{aligned}$$

Each term on the right-hand side is nonnegative. We bound the four terms separately.

For the first term, by convexity of $F_\delta(\cdot, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t)$,

$$F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) \le \langle \nabla_{\boldsymbol{x}} F_\delta(\boldsymbol{x}^t, \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{x}^t - \boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t)\rangle \le g^t_{\mathrm{FW},\boldsymbol{x}} \le g^t. \tag{35}$$

We next bound the second term. By concavity of $d_\delta(\cdot, \boldsymbol{z}^t, \boldsymbol{v}^t)$ and Danskin's theorem, we have

$$\begin{aligned}p_\delta(\boldsymbol{z}^t, \boldsymbol{v}^t) - d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) &\le \langle \nabla_{\boldsymbol{y}} d_\delta(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{y}^t\rangle \\ &= \langle \nabla_{\boldsymbol{y}} F_\delta(\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{y}^t\rangle \\ &\le g^t_{\mathrm{FW},\boldsymbol{y}} + L \operatorname{diam}(\mathcal{Y}) \|\boldsymbol{x}^t - \boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t)\| \\ &\le g^t + L \operatorname{diam}(\mathcal{Y}) \sigma_6 g^t_{\mathrm{PFW},\boldsymbol{x}} \le (1 + 2L \operatorname{diam}(\mathcal{Y}) \sigma_6)\, g^t.\end{aligned} \tag{36}$$

For the third term, recall that $p_\delta(\boldsymbol{z}, \cdot)$ is concave and differentiable, with $\nabla_{\boldsymbol{v}} p_\delta(\boldsymbol{z}, \boldsymbol{v}) = r_{\boldsymbol{y}}(\boldsymbol{y}(\boldsymbol{z}, \boldsymbol{v}) - \boldsymbol{v})$. Let $\boldsymbol{v}(\boldsymbol{z}) \in \operatorname{argmax}_{\boldsymbol{v}} p_\delta(\boldsymbol{z}, \boldsymbol{v})$. Since the maximization is unconstrained, first-order optimality gives $\boldsymbol{v}(\boldsymbol{z}) = \boldsymbol{y}(\boldsymbol{z}, \boldsymbol{v}(\boldsymbol{z})) \in \mathcal{Y}$. The algorithm also maintains $\boldsymbol{v}^t \in \mathcal{Y}$. Hence, by concavity,

$$q_\delta(\boldsymbol{z}^t) - p_\delta(\boldsymbol{z}^t, \boldsymbol{v}^t) \le \langle \nabla_{\boldsymbol{v}} p_\delta(\boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{v}(\boldsymbol{z}^t) - \boldsymbol{v}^t \rangle \le r_{\boldsymbol{y}} \operatorname{diam}(\mathcal{Y}) \|\boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{v}^t\|. \tag{37}$$

Using the Lipschitz continuity of $\boldsymbol{y}(\boldsymbol{z}, \boldsymbol{v})$ with respect to $\boldsymbol{z}$, we obtain

$$\begin{aligned} \|\boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{v}^t\| &\le \|\boldsymbol{y}(\boldsymbol{z}^{t+1}, \boldsymbol{v}^t) - \boldsymbol{v}^t\| + \|\boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{y}(\boldsymbol{z}^{t+1}, \boldsymbol{v}^t)\| \\ &\le \frac{\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|}{\beta_{\boldsymbol{y}}} + \sigma_3 \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|. \end{aligned} \tag{38}$$

Combining equation 37 and equation 38 yields

$$q_\delta(\boldsymbol{z}^t) - p_\delta(\boldsymbol{z}^t, \boldsymbol{v}^t) \le \frac{r_{\boldsymbol{y}} \operatorname{diam}(\mathcal{Y}) \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|}{\beta_{\boldsymbol{y}}} + \sigma_3 r_{\boldsymbol{y}} \operatorname{diam}(\mathcal{Y}) \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|. \tag{39}$$

It remains to control $q_\delta(\boldsymbol{z}^t) - \underline{F}$. Let $\boldsymbol{z}(\boldsymbol{v}) \in \arg\min_{\boldsymbol{z} \in \mathbb{R}^n} p_\delta(\boldsymbol{z}, \boldsymbol{v})$. Since $p_\delta(\cdot, \boldsymbol{v})$ is convex and coercive, such a minimizer exists. Moreover, $\nabla_{\boldsymbol{z}} p_\delta(\boldsymbol{z}, \boldsymbol{v}) = r_{\boldsymbol{x}}(\boldsymbol{z} - \boldsymbol{x}(\boldsymbol{z}, \boldsymbol{v}))$, so first-order optimality gives $\boldsymbol{z}(\boldsymbol{v}) = \boldsymbol{x}(\boldsymbol{z}(\boldsymbol{v}), \boldsymbol{v}) \in \mathcal{X}$. Similarly, we have $\boldsymbol{v}(\boldsymbol{z}) = \boldsymbol{y}(\boldsymbol{z}, \boldsymbol{v}(\boldsymbol{z})) \in \mathcal{Y}$. Noting that $\underline{F} = \min_{\boldsymbol{z}} q_\delta(\boldsymbol{z}) \ge \min_{\boldsymbol{z}} p_\delta(\boldsymbol{z}, \boldsymbol{v}) = p_\delta(\boldsymbol{z}(\boldsymbol{v}), \boldsymbol{v})$, we have

$$\begin{aligned} q_\delta(\boldsymbol{z}^t) - \underline{F} &= q_\delta(\boldsymbol{z}^t) - p_\delta(\boldsymbol{z}^t, \boldsymbol{v}^t) + p_\delta(\boldsymbol{z}^t, \boldsymbol{v}^t) - \underline{F} \\ &\le q_\delta(\boldsymbol{z}^t) - p_\delta(\boldsymbol{z}^t, \boldsymbol{v}^t) + p_\delta(\boldsymbol{z}^t, \boldsymbol{v}^t) - p_\delta(\boldsymbol{z}(\boldsymbol{v}^t), \boldsymbol{v}^t). \end{aligned} \tag{40}$$

The first term has already been controlled in equation 39. For the second term, convexity of $p_\delta(\cdot, \boldsymbol{v}^t)$ gives

$$\begin{aligned} & p_\delta(\boldsymbol{z}^t, \boldsymbol{v}^t) - p_\delta(\boldsymbol{z}(\boldsymbol{v}^t), \boldsymbol{v}^t) \\ &\le \langle \nabla_{\boldsymbol{z}} p_\delta(\boldsymbol{z}^t, \boldsymbol{v}^t), \boldsymbol{z}^t - \boldsymbol{z}(\boldsymbol{v}^t) \rangle \\ &\le r_{\boldsymbol{x}} \operatorname{diam}(\mathcal{X}) \|\boldsymbol{z}^t - \boldsymbol{x}(\boldsymbol{z}^t, \boldsymbol{v}^t)\| \\ &\le r_{\boldsymbol{x}} \operatorname{diam}(\mathcal{X}) \left(\|\boldsymbol{z}^t - \boldsymbol{x}^t\| + \|\boldsymbol{x}^t - \boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t)\| + \|\boldsymbol{x}(\boldsymbol{y}^t, \boldsymbol{z}^t, \boldsymbol{v}^t) - \boldsymbol{x}(\boldsymbol{z}^t, \boldsymbol{v}^t)\| \right) \\ &\le r_{\boldsymbol{x}} \operatorname{diam}(\mathcal{X}) \left(\frac{\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|}{\beta_{\boldsymbol{x}}} + \sigma_6 g^t_{\mathrm{PFW},\boldsymbol{x}} + \sigma_1 \|\boldsymbol{y}^t - \boldsymbol{y}(\boldsymbol{z}^t, \boldsymbol{v}^t)\| \right) \\ &\le r_{\boldsymbol{x}} \operatorname{diam}(\mathcal{X}) \left(\frac{\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|}{\beta_{\boldsymbol{x}}} + \sigma_6 g^t_{\mathrm{PFW},\boldsymbol{x}} + \sigma_1 \left(\sigma_8 g^t_{\mathrm{PFW},\boldsymbol{y}} + \sigma_9 g^t_{\mathrm{PFW},\boldsymbol{x}} \right) \right) \\ &\le \frac{r_{\boldsymbol{x}} \operatorname{diam}(\mathcal{X}) \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|}{\beta_{\boldsymbol{x}}} + 2 r_{\boldsymbol{x}} \operatorname{diam}(\mathcal{X}) \left(\sigma_6 + \sigma_1 \max\{\sigma_8, \sigma_9\} \right) g^t, \end{aligned}$$

where the third inequality uses the triangle inequality; the fourth uses the update of z^{t+1}, Lemma D.3(i), and Lemma D.1(i); the fifth uses Lemma D.3(iii); and the last inequality follows from $g^t_{\mathrm{PFW},x} + g^t_{\mathrm{PFW},y} = g^t_{\mathrm{PFW}} \le 2g^t$. Combining this bound with equation 40, we conclude that

$$\begin{aligned} q_\delta(\boldsymbol{z}^t) - \underline{F} \le 2 r_{\boldsymbol{x}} \operatorname{diam}(\mathcal{X}) \left(\sigma_6 + \sigma_1 \max\{\sigma_8, \sigma_9\} \right) g^t + \frac{r_{\boldsymbol{y}} \operatorname{diam}(\mathcal{Y}) \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|}{\beta_{\boldsymbol{y}}} \\ + \left(r_{\boldsymbol{x}} \operatorname{diam}(\mathcal{X}) + \sigma_3 r_{\boldsymbol{y}} \operatorname{diam}(\mathcal{Y}) \beta_{\boldsymbol{x}} \right) \frac{\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|}{\beta_{\boldsymbol{x}}}. \end{aligned} \tag{41}$$

Putting all pieces together finishes the proof. □

We now combine Lemma I.1 with the one-step descent estimate and the active-set counting argument to prove Theorem 3.2.

Proof of Theorem 3.2. For every non-clipped iteration t, define

$$\mathcal{D}_t \coloneqq \frac{1}{32\bar{C}} (g^t)^2 + \frac{r_{\boldsymbol{x}}}{4\beta_{\boldsymbol{x}}} \|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|^2 + \frac{r_{\boldsymbol{y}}}{8\beta_{\boldsymbol{y}}} \|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|^2.$$

Choose $\beta_{\boldsymbol{x}} = c_{\boldsymbol{x}}\delta$, where $c_{\boldsymbol{x}} > 0$ is sufficiently small so that the coupling term in Proposition 4.1 can be absorbed as in the NC–C analysis. Then every non-clipped iteration satisfies

$$\widetilde{\Phi}^t - \widetilde{\Phi}^{t+1} \geq \mathcal{D}_t. \tag{42}$$

On the other hand, Lemma I.1 gives

$$\widetilde{\Phi}^t \leq C_{\mathrm{cc}} \left(g^t + \frac{\|\boldsymbol{z}^{t+1} - \boldsymbol{z}^t\|}{\beta_{\boldsymbol{x}}} + \frac{\|\boldsymbol{v}^t - \boldsymbol{v}_+^t(\boldsymbol{z}^{t+1})\|}{\beta_{\boldsymbol{y}}} \right).$$

Since $\beta_{\boldsymbol{y}}$ is fixed and $\beta_{\boldsymbol{x}}$ is sufficiently small, the definition of $\mathcal{D}_t$ implies that, for some constant $C_0 > 0$ independent of T,

$$(\widetilde{\Phi}^t)^2 \leq \frac{C_0}{\beta_{\boldsymbol{x}}} \mathcal{D}_t.$$

Combining this inequality with equation 42, we obtain

$$\widetilde{\Phi}^t - \widetilde{\Phi}^{t+1} \geq c_0 \beta_{\boldsymbol{x}} (\widetilde{\Phi}^t)^2 \tag{43}$$

for some constant $c_0 > 0$ independent of T. Let $0 \leq t_1 < t_2 < \cdots < t_N < T$ denote the non-clipped iterations among the first T iterations, where $N = N_{\mathrm{nc}}(T)$. Since every clipped iteration is Lyapunov non-increasing, equation 43 yields

$$\widetilde{\Phi}^{t_{k+1}} \leq \widetilde{\Phi}^{t_k} - c_0 \beta_{\boldsymbol{x}} (\widetilde{\Phi}^{t_k})^2.$$

The standard reciprocal argument therefore gives

$$\widetilde{\Phi}^{t_k} = \mathcal{O}\left(\frac{1}{\beta_{\boldsymbol{x}} k} \right). \tag{44}$$

We next convert this bound into a stationarity bound. Set $m = \lceil \frac{N}{2} \rceil$. By equation 44,

$$\widetilde{\Phi}^{t_m} = \mathcal{O}\left(\frac{1}{\beta_{\boldsymbol{x}} N} \right).$$

Summing equation 42 over the non-clipped iterations $t_m, \ldots, t_N$, and using the monotonicity of the Lyapunov function at the intervening clipped iterations, gives

$$\sum_{k=m}^{N} \mathcal{D}_{t_k} \leq \widetilde{\Phi}^{t_m} - \widetilde{\Phi}^T \leq \widetilde{\Phi}^{t_m} = \mathcal{O}\left(\frac{1}{\beta_{\boldsymbol{x}} N} \right). \tag{45}$$

Since $N - m + 1 = \Theta(N)$, there exists $k_\star \in \{m, \ldots, N\}$ such that

$$\mathcal{D}_{t_{k_\star}} = \mathcal{O}\left(\frac{1}{\beta_{\boldsymbol{x}} N^2} \right). \tag{46}$$

It follows from the definition of $\mathcal{D}_t$ that

$$g^{t_{k_\star}} + \frac{\|\boldsymbol{z}^{t_{k_\star}+1} - \boldsymbol{z}^{t_{k_\star}}\|}{\beta_{\boldsymbol{x}}} + \frac{\|\boldsymbol{v}^{t_{k_\star}} - \boldsymbol{v}_+^{t_{k_\star}}(\boldsymbol{z}^{t_{k_\star}+1})\|}{\beta_{\boldsymbol{y}}} = \mathcal{O}\left(\frac{1}{\beta_{\boldsymbol{x}} N} \right). \tag{47}$$

It remains to relate N to T. The active-set counting argument gives

$$N_{\mathrm{nc}}(T) + N_{\mathrm{c}FW}(T) = \Omega(T).$$

Moreover, every clipped Frank–Wolfe iteration produces a uniform Lyapunov decrease. Since the total Lyapunov decrease is bounded, $N_{\mathrm{c}FW}(T) = \mathcal{O}(1)$. Therefore,

$$N = N_{\mathrm{nc}}(T) = \Theta(T).$$

Substituting this bound into equation 47 yields

$$g^{t_{k_\star}} + \frac{\|\boldsymbol{z}^{t_{k_\star}+1} - \boldsymbol{z}^{t_{k_\star}}\|}{\beta_{\boldsymbol{x}}} + \frac{\|\boldsymbol{v}^{t_{k_\star}} - \boldsymbol{v}_+^{t_{k_\star}}(\boldsymbol{z}^{t_{k_\star}+1})\|}{\beta_{\boldsymbol{y}}} = \mathcal{O}\left(\frac{1}{\beta_{\boldsymbol{x}} T} \right).$$

Applying Lemma E.1 at the non-clipped iteration $t_{k_\star}$ and using $\beta_{\boldsymbol{x}} = \Theta(\delta)$, we obtain

$$\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}^{t_{k_\star}}, \boldsymbol{y}^{t_{k_\star}}) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}^{t_{k_\star}}, \boldsymbol{y}^{t_{k_\star}}) = \mathcal{O}\left(\frac{1}{\delta T} + \delta \right).$$

Choosing $\delta = \Theta(T^{-1/2})$ balances the two terms and gives

$$\mathcal{G}_{\mathcal{X}}(\boldsymbol{x}^{t_{k_\star}}, \boldsymbol{y}^{t_{k_\star}}) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x}^{t_{k_\star}}, \boldsymbol{y}^{t_{k_\star}}) = \mathcal{O}(T^{-1/2}).$$

The claimed $\mathcal{O}(\epsilon^{-2})$ iteration complexity follows. □

J KŁ–NC AND C–NC PROBLEMS

This section gives the primal-side counterparts of the NC–KŁ and NC–C results in Theorem 3.1. We first state the corresponding structural conditions. For the KŁ–NC setting, we assume that, for every $\boldsymbol{y} \in \mathcal{Y}$, the problem $\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}, \boldsymbol{y})$ has a nonempty solution set and a finite optimal value, and that there exist constants $\tau_{\boldsymbol{x}} > 0$ and $\theta \in (0, 1)$ such that

$$\left(f(\boldsymbol{x}, \boldsymbol{y}) - \min_{\boldsymbol{x}'\in\mathcal{X}} f(\boldsymbol{x}', \boldsymbol{y})\right)^{\theta} \leq \frac{1}{\tau_{\boldsymbol{x}}} \operatorname{dist}(\mathbf{0}, \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}) + \partial \mathbf{1}_{\mathcal{X}}(\boldsymbol{x}))$$

for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$. For the C–NC setting, we assume that $f(\cdot, \boldsymbol{y})$ is convex on $\mathcal{X}$ for every $\boldsymbol{y} \in \mathcal{Y}$.

The primal-side counterpart of equation 1 is

$$F_\delta^{\mathrm{p}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{v}) \coloneqq f(\boldsymbol{x}, \boldsymbol{y}) + \frac{r_{\boldsymbol{x}}}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 - \frac{r_{\boldsymbol{y}}}{2}\|\boldsymbol{y} - \boldsymbol{v}\|^2 + \frac{\delta}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2, \tag{48}$$

where $\bar{\boldsymbol{x}} \in \mathcal{X}$ is fixed. The corresponding algorithm is the symmetric primal-side variant of Algorithm 1, obtained by replacing F_δ with F_δ^{p} and interchanging the roles of the primal and dual problems.

Theorem J.1 (Convergence rates for KŁ–NC and C–NC problems)**.** Let $T > 1$. For each of the following settings, the algorithmic parameters can be chosen so that there exists $t \in \{0, \dots, T-1\}$ for which the corresponding guarantee holds.

(i) (KŁ–NC, $\theta \in (0, 1/2]$) Suppose the primal-side KŁ condition holds with $\theta \in (0, 1/2]$. Let $\delta = 0$ and choose $\beta_{\boldsymbol{y}} > 0$ sufficiently small and independently of T. Then $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}(T^{-1/2})$-FWSP of Problem P.

(ii) (KŁ–NC, $\theta \in (1/2, 1)$) Suppose the primal-side KŁ condition holds with $\theta \in (1/2, 1)$. Let $\delta = 0$ and choose

$$\beta_{\boldsymbol{y}} = \Theta\left(T^{-(2\theta-1)/(2\theta)}\right).$$

Then $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}(T^{-1/(4\theta)})$-FWSP of Problem P.

(iii) (C–NC) Suppose that $f(\cdot, \boldsymbol{y})$ is convex on $\mathcal{X}$ for every $\boldsymbol{y} \in \mathcal{Y}$. Using equation 48, choose $\delta = \Theta(T^{-1/3})$ and $\beta_{\boldsymbol{y}} = \Theta(T^{-1/3})$. Then $(\boldsymbol{x}^t, \boldsymbol{y}^t)$ is an $\mathcal{O}(T^{-1/3})$-FWSP of Problem P.

□

Proof. Consider the transformed minimax problem

$$\min_{\boldsymbol{y}\in\mathcal{Y}} \max_{\boldsymbol{x}\in\mathcal{X}} \widetilde{f}(\boldsymbol{y}, \boldsymbol{x}), \qquad \widetilde{f}(\boldsymbol{y}, \boldsymbol{x}) := -f(\boldsymbol{x}, \boldsymbol{y}),$$

where $\boldsymbol{y}$ and $\boldsymbol{x}$ play the primal and dual roles, respectively. The smoothness assumptions are preserved under this transformation.

The primal-side KŁ condition for f is exactly the dual-side KŁ condition for $\widetilde{f}$, with the same exponent θ. Indeed,

$$\max_{\boldsymbol{x}'\in\mathcal{X}} \widetilde{f}(\boldsymbol{y}, \boldsymbol{x}') - \widetilde{f}(\boldsymbol{y}, \boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{y}) - \min_{\boldsymbol{x}'\in\mathcal{X}} f(\boldsymbol{x}', \boldsymbol{y}), \quad \text{and}$$
$$-\nabla_{\boldsymbol{x}} \widetilde{f}(\boldsymbol{y}, \boldsymbol{x}) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}).$$

Hence, Theorem 3.1(ii)–(iii), applied to $\widetilde{f}$, gives parts (i)–(ii) after interchanging the primal and dual variables and the corresponding algorithmic parameters.

For the C–NC setting, convexity of $f(\cdot, \boldsymbol{y})$ implies concavity of $\widetilde{f}(\boldsymbol{y}, \cdot)$. Moreover, under the same interchange, the primal perturbation $\frac{\delta}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2$ in equation 48 becomes the dual perturbation $-\frac{\delta}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2$ for the transformed problem. Therefore, Theorem 3.1(i), applied to $\widetilde{f}$, gives $\delta = \Theta(T^{-1/3})$, $\beta_{\boldsymbol{y}} = \Theta(T^{-1/3})$, and the $\mathcal{O}(T^{-1/3})$ stationarity rate.

Finally, the Frank–Wolfe stationarity measures are preserved under the transformation:

$$\widetilde{\mathcal{G}}_{\mathcal{Y}}(\boldsymbol{y},\boldsymbol{x}) = \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x},\boldsymbol{y}), \qquad \widetilde{\mathcal{G}}_{\mathcal{X}}(\boldsymbol{y},\boldsymbol{x}) = \mathcal{G}_{\mathcal{X}}(\boldsymbol{x},\boldsymbol{y}).$$

Thus, the FWSP guarantees obtained for the transformed problem are exactly the stated FWSP guarantees for Problem P. □

K Additional numerical experiment details

Here, we provide the remaining details for the experiments reported in Section 5. In all experiments, performance is measured using the two-sided Frank–Wolfe stationarity gap for the original problem,

$$\mathcal{G}_{\mathcal{X}}(\boldsymbol{x},\boldsymbol{y}) + \mathcal{G}_{\mathcal{Y}}(\boldsymbol{x},\boldsymbol{y}). \tag{49}$$

For each benchmark problem, we generate 10 independent problem instances and run all compared methods with the same problem-specific initialization. The stationarity gap equation 49 is evaluated every 10 iterations. Since equation 49 need not decrease monotonically, for each problem instance we record the best stationarity gap observed up to each reported iteration. Figure 2 then reports, at each reported iteration, the median of these best-so-far gap values across the 10 iterations, with shaded regions indicating the corresponding interquartile range.

K.1 Graphical matching

We consider the graphical matching game used in the experiments of (Gidel et al., 2017). The game models a zero-sum competition between two universities over s students, where each university must assign pairs of students into dorms. Assignments (or matchings) can be modeled with a complete graph on s vertices, which has an edge for every possible roommate pair. Hence, there are $d = \binom{s}{2}$ possible edges. Denote by $\mathcal{P}_s \subseteq \mathbb{R}^d$ the corresponding perfect matching polytope. The LMO over $\mathcal{P}_s$ can be solved in $\mathcal{O}(s^3)$ time using the Blossom algorithm (Edmonds, 1965). Following Gidel et al. (2017), each student i is assigned an underlying value $\mu_i \sim \mathcal{U}[0,1]$, which is shared between the two universities. Each university U observes a noisy version of this value, $b_i^{(U)} \sim \mathcal{N}(\mu_i, 0.1)$. Each student also has a randomly generated preference ordering over the other students. We write $\sigma_i(j)$ for the rank assigned by student i to student j, with a smaller rank representing a higher preference. For two possible roommates j and ℓ of student i, define the signed contribution of student i to the payoff as

$$q_i(j,\ell) = \begin{cases} b_i^{(1)}, & \sigma_i(j) < \sigma_i(\ell), \\ -b_i^{(2)}, & \sigma_i(j) > \sigma_i(\ell), \\ \dfrac{b_i^{(1)} - b_i^{(2)}}{2}, & j = \ell. \end{cases}$$

The payoff matrix $M \in \mathbb{R}^{d\times d}$ is indexed by possible roommate pairs. For two edges $e, f \in E(K_s)$, only students that belong to both edges contribute to the corresponding entry of M. In particular, for each $i \in e \cap f$, let j denote the other endpoint of e and let ℓ denote the other endpoint of f. We define

$$M_{e,f} = \sum_{i \in e\cap f} q_i(j,\ell),$$

when $e \cap f \neq \emptyset$, and $M_{e,f} = 0$ otherwise. If $e = f$, $M_{e,f} = M_{e,e} = q_i(j,j) + q_i(\ell,\ell)$.

Let $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{P}_s$ denote the strategies of the two universities. The resulting graphical matching problem is the bilinear saddle-point problem

$$\min_{\boldsymbol{x}\in\mathcal{P}_s} \max_{\boldsymbol{y}\in\mathcal{P}_s} \langle \boldsymbol{x}, M\boldsymbol{y}\rangle,$$

which is convex in $\boldsymbol{x}$ and concave in $\boldsymbol{y}$. In our experiment we take $s = 32$, so that $d = 496$. The LMO over $\mathcal{P}_s$ reduces to a minimum-weight perfect-matching problem on the complete graph K_s, which we solve using the matching routine in NetworkX based on Edmonds' Blossom algorithm (Hagberg et al., 2008). The smoothness constant L is computed for each instance by estimating the largest singular value of M using a matrix-free iterative singular-value computation.

All graphical-matching experiments are run for $T = 2000$ iterations and are initialized from the same pair of randomly generated perfect matchings $(\boldsymbol{x}^0, \boldsymbol{y}^0)$ for each problem instance. For the algorithm we propose, we use

$$r_{\boldsymbol{x}} = 0.01L, \qquad r_{\boldsymbol{y}} = 0.001L, \qquad \beta_{\boldsymbol{x}} = T^{-1/2}, \qquad \beta_{\boldsymbol{y}} = 0.05,$$

together with

$$\delta = 0.01L\,T^{-1/2}, \qquad \bar{C} = 0.7L.$$

The perturbation anchor is the fractional perfect matching

$$\bar{\boldsymbol{y}} = \frac{1}{s-1}\mathbf{1}.$$

For `LMO-LMO`, we use

$$\tau_t = \frac{1.5}{t+10}, \qquad \beta_t = \frac{0.003L}{(t+10)^{1/5}}.$$

For `SP-FW`, we use

$$\gamma_t = \frac{2}{t+2},$$

while for `SP-AFW`, we use

$$\gamma_t = \frac{2}{2+n_t},$$

where n_t denotes the number of previous non-drop iterations.

K.2 Robust budget allocation

For the nonconvex-concave setting, we consider the robust budget allocation problem introduced in (Wilder, 2018). The problem models an advertiser allocating a finite budget across a collection of advertising channels in order to maximize the expected profit from the customers reached, when the values of the customers are uncertain.

Let m denote the number of advertising channels, n the number of customers, and B the allocation budget. Following (Wilder, 2018), we encode the integer budget allocation by introducing B identical copies of each advertising channel, with each copy representing one unit of budget. Relaxing the resulting binary selection variables to $[0, 1]$, the feasible set is

$$\mathcal{X} = \left\{\boldsymbol{x} \in [0,1]^{mB} : \langle \mathbf{1}, \boldsymbol{x}\rangle = B\right\}.$$

For each copied advertising channel i and customer j, let p_{ij} denote the probability that one unit of advertising on the corresponding channel reaches customer j. The multilinear extension of the probability that customer j is reached is

$$c_j(\boldsymbol{x}) = 1 - \prod_{i=1}^{mB}(1 - p_{ij}x_i).$$

If customer j is reached, the advertiser assigns a value $w_j > 0$ to that customer, interpreted as the expected profit from reaching them. Uncertainty in these values is modeled by an adversary $\boldsymbol{y} \in [0,1]^n$ that may reduce the value of customer j from w_j to $w_j(1 - y_j)$, where

$$\mathcal{Y} = \{\boldsymbol{y} \in [0,1]^n : \langle \mathbf{1}, \boldsymbol{y}\rangle \le \Gamma\}.$$

Here, $\Gamma \ge 0$ is a chosen uncertainty budget that controls the total amount by which the customer values may be reduced. The robust budget allocation problem takes the form

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

where

$$f(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{\sum_{j=1}^n w_j} \sum_{j=1}^n w_j(1 - y_j)c_j(\boldsymbol{x}).$$

The objective is affine, and hence concave in $\boldsymbol{y}$, while it is nonconvex in $\boldsymbol{x}$.

We generate random instances following (Wilder, 2018). For each channel–customer pair, we include an edge with probability 0.2, independently across pairs, and assign each included edge a reach probability drawn uniformly from $[0, 0.2]$. Customer values are drawn independently from $\mathcal{U}[0.5, 1.5]$. The same reach probabilities are used for all B copies of a channel, and we set $\Gamma = n/2$.

We consider a small problem with $m = 50 = n$, $B = 5, \Gamma = 25$, as well as a larger problem with $m = n = 100, B = 10, \Gamma = 50$.

In our experiments Γ is integral, and the LMOs over $\mathcal{X}$ and $\mathcal{Y}$ are available in closed form. The primal oracle selects the B coordinates corresponding to the smallest entries of $\nabla_{\boldsymbol{x}} F_\delta$ and sets them to one. The dual oracle selects up to Γ coordinates corresponding to the largest positive entries of $\nabla_{\boldsymbol{y}} F_\delta$ and sets them to one. All remaining coordinates are set to zero.

The smoothness constant L is computed for each instance from a global bound on the Hessian norm, obtained using matrix-free power iteration. Both budget allocation experiments are run for $T = 20000$ iterations, and all methods are initialized from the same point $(\boldsymbol{x}^0, \boldsymbol{y}^0)$, where $\boldsymbol{x}^0$ is a randomly selected vertex of $\mathcal{X}$ and $\boldsymbol{y}^0 = \frac{1}{2}\mathbf{1}$.

For the algorithm we propose, we use the following parameters for both problem sizes, with L computed independently for each instance.

$$r_{\boldsymbol{x}} = 0.06L, \qquad r_{\boldsymbol{y}} = 0.001L, \qquad \beta_{\boldsymbol{x}} = 2T^{-1/3}, \qquad \beta_{\boldsymbol{y}} = 0.8,$$

together with

$$\delta = 0.01L\,T^{-1/3}, \qquad \bar{C} = L.$$

The perturbation anchor is $\bar{\boldsymbol{y}} = \boldsymbol{y}^0$. We initialize $S^0_{\boldsymbol{y}} = \{\boldsymbol{b}, \mathbf{1} - \boldsymbol{b}\}$, where $b_i = 1$ for $i \le n/2$ and $b_i = 0$ otherwise, with weight $1/2$ on each atom. For `LMO-LMO`, we use

$$\tau_t = \frac{0.4}{(t+1)^{5/6}}, \qquad \beta_t = \frac{0.01L}{(t+1)^{1/6}}.$$

K.3 Parameter selection and implementation details

For each algorithm, parameters were tuned empirically for each benchmark and then fixed across all 10 reported instances. For our proposed method, $\beta_{\boldsymbol{x}}$ and δ scale as $T^{-1/2}$ in the convex–concave graphical matching experiment and as $T^{-1/3}$ in the nonconvex–concave budget allocation experiments, following the rates suggested by our analysis. Although the experimental choices of r_x and r_y do not satisfy the conservative sufficient bounds used in the theory, these less conservative choices were found to perform well across the reported instances.