Randomized Primal-Dual Methods with Adaptive Step Sizes

Erfan Yazdandoost Hamedani
The University of Arizona

Afrooz Jalilzadeh
The University of Arizona

Necdet S. Aybat
The Pennsylvania State University

Abstract

In this paper we propose a class of randomized primal-dual methods incorporating line search to contend with large-scale saddle point (SP) problems defined by a convex-concave function $\mathcal{L}(\mathbf{x}, y) \triangleq \sum_{i=1}^{M} f_i(x_i) + \Phi(\mathbf{x}, y) - h(y)$. We analyze the convergence rate of the proposed method under mere convexity and strong convexity assumptions of $\mathcal{L}$ in $\mathbf{x}$-variable. In particular, assuming $\nabla_y \Phi(\cdot, \cdot)$ is Lipschitz and $\nabla_{\mathbf{x}} \Phi(\cdot, y)$ is coordinate-wise Lipschitz for any fixed y, the ergodic sequence generated by the algorithm achieves the $\mathcal{O}(M/k)$ convergence rate in the expected primal-dual gap. Furthermore, assuming that $\mathcal{L}(\cdot, y)$ is strongly convex for any y, and that $\Phi(\mathbf{x}, \cdot)$ is affine for any $\mathbf{x}$, the scheme enjoys a faster rate of $\mathcal{O}(M/k^2)$ in terms of primal solution suboptimality. We implemented the proposed algorithmic framework to solve kernel matrix learning problem, and tested it against other state-of-the-art first-order methods.

1 Introduction

Let $(\mathcal{X}_i, \|\cdot\|_{\mathcal{X}_i})$ for $i \in \mathcal{M} \triangleq \{1, 2, \ldots, M\}$ and $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be finite dimensional, normed vector spaces such that $\mathcal{X}_i = \mathbb{R}^{m_i}$ for $i \in \mathcal{M}$. Let $\mathbf{x} = [x_i]_{i \in \mathcal{M}} \in \Pi_{i \in \mathcal{M}} \mathcal{X}_i \triangleq \mathcal{X} = \mathbb{R}^m$ where $m \triangleq \sum_{i \in \mathcal{M}} m_i$. In this paper, we study the following saddle point (SP) problem:

$$(P): \quad \min_{\mathbf{x} \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{L}(\mathbf{x}, y), \qquad \mathcal{L}(\mathbf{x}, y) \triangleq \sum_{i \in \mathcal{M}} f_i(x_i) + \Phi(\mathbf{x}, y) - h(y), \tag{1}$$

where $h : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ and $f_i : \mathcal{X}_i \to \mathbb{R} \cup \{+\infty\}$ for all $i \in \mathcal{M}$ are (possibly nonsmooth) closed μ_i-convex functions with respect to $\|\cdot\|_{\mathcal{X}_i}$ for some $\mu_i \geq 0$, and the coupling function $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex in $\mathbf{x}$ and concave in y and, it satisfies certain differentiability assumptions – see Assumption 1.

Our study is motivated by large-scale problems with a *coordinate-friendly* structure Peng et al. (2016), i.e., for any $i \in \mathcal{M}$, the amount of work to compute the partial-gradient $\nabla_{x_i} \Phi(\mathbf{x}, y)$ is $m_i/m \approx 1/M$ fraction of the work required for the full-gradient $\nabla_{\mathbf{x}} \Phi(\mathbf{x}, y)$ computation. Our objective is to design an efficient first-order *randomized* block-coordinate primal-dual method to compute a saddle point of the structured convex-concave function $\mathcal{L}$ in (1), and to investigate its convergence properties under *mere* and *strong convexity* settings. Typically, the first-order methods rely on the knowledge of global Lipschitz constants to select an appropriate step-size with a convergence guarantee. In practical settings, such constants may not be readily available or it can be difficult to compute them; hence, one may need to consider line-search methods. Since *exact* line-search methods are often difficult to implement, one practical avenue is to adopt backtracking to estimate the *local* Lipschitz constants, which usually leads to larger steps. Hence, in this paper, we propose a randomized block-coordinate primal-dual algorithm with *backtracking* to efficiently solve the SP problem in (1).

Before we discuss important applications, it should be emphasized that (1) covers regularized convex optimization problems with nonlinear constraints as a special case[1], i.e.,

$$\min_{\mathbf{x}} \rho(\mathbf{x}) \triangleq \varphi(\mathbf{x}) + \sum_{i \in \mathcal{M}} f_i(x_i) \ \text{ s.t. } \ g(\mathbf{x}) \in -\mathcal{K}, \tag{2}$$

where $\mathcal{K} \subseteq \mathcal{Y}^*$ is a closed convex cone in the dual space $\mathcal{Y}^*$; $f_i : \mathcal{X}_i \to \mathbb{R} \cup \{+\infty\}$ is a convex (possibly nonsmooth) regularizer function for $i \in \mathcal{M}$; $\varphi : \mathcal{X} \to \mathbb{R}$ is a smooth convex function having a coordinate-wise Lipschitz continuous gradient; and $g : \mathcal{X} \to \mathcal{Y}^*$ is a smooth $\mathcal{K}$-convex, coordinate-wise Lipschitz function with a coordinate-wise Lipschitz Jacobian. This problem can be written as a special case of (1) by setting $\Phi(\mathbf{x}, y) = \varphi(\mathbf{x}) + \langle g(\mathbf{x}), y \rangle$ and

[1] We can show that Assumption 1 is satisfied for this example through arguing that the dual iterate sequence of the proposed method is almost surely bounded. Thus, one does not need the boundedness of dual domain to argue for the existence of global constants $\{L_{x_i x_i}\}_{i \in \mathcal{M}}$, instead bounded dual sequence is sufficient for our proof arguments.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

$h(y) = \mathbb{I}_{\mathcal{K}^*}(y)$, where $\mathcal{K}^* \subseteq \mathcal{Y}$ denotes the dual cone of $\mathcal{K}$ and $\mathbb{I}_{\mathcal{K}^*}(\cdot)$ is the indicator function of $\mathcal{K}^*$. An advantage of such formulation lies in utilizing the corresponding dual variables in order to boost the primal convergence through appropriately controlling the constraint violations.

Application. Many interesting problems arising in machine learning (ML), signal and image processing, finance, etc., can be formulated as a special case of (1). Some instances of such problems include: i) *distributionally robust optimization* Namkoong and Duchi (2016); ii) *kernel matrix learning* Lanckriet et al. (2004); Gönen and Alpaydın (2011); iii) *distance metric learning* Xing et al. (2003); (iv) training *ellipsoidal machines* Shivaswamy and Jebara (2007); (v) *two-player zero-sum game* with nonlinear payoff (Boyd and Vandenberghe, 2004; Chen et al., 2017). In the following, we will briefly discuss some of these problem instances and their formulations as a special case of (1).

Distributionally robust optimization (DRO): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where $\Omega = \{\zeta_1, \ldots, \zeta_m\}$, $\ell : X \times \Omega \to \mathbb{R}$ is a convex loss function over a convex bounded set $X \subset \mathcal{X}$, and we define $\ell_i(u) \triangleq \ell(u; \zeta_i)$. The aim of DRO is to optimize the worst case performance under uncertainty and to compute solutions with some confidence level Namkoong and Duchi (2016). This class of problems can be formulated as

$$\min_{u \in X} \max_{\mathbf{p} \in \mathcal{P}} \ \mathbb{E}_{\zeta \sim \mathbb{P}}[\ell(u; \zeta)] = \sum_{i=1}^{m} p_i \ell_i(u), \tag{3}$$

where $\mathcal{P}$ represents an uncertainty set over the probability distributions. For instance, $\mathcal{P} = \{\mathbf{y} \in \Delta_m : V(\mathbf{p}, \frac{1}{m}\mathbf{1}_m) \leq \rho\}$ is an uncertainty set considered in Namkoong and Duchi (2016), where Δ_m is an m-dimensional probability simplex, and $V(Q, P)$ denotes a divergence measure for probability measures Q and P. Assuming $V(\mathbf{p}, \frac{1}{m}\mathbf{1}_m) = \sum_{i=1}^{m} V_i(p_i, \frac{1}{m})$ and introducing variables $\lambda \in \mathbb{R}_+$ and $\eta \in \mathbb{R}$, we can dualize the divergence constraint in (3) to obtain the following equivalent problem:

$$\min_{\substack{u \in X \\ \lambda \geq 0 \\ \eta \in \mathbb{R}}} \max_{\mathbf{p} \in \mathbb{R}^m_+} \sum_{i=1}^{m} p_i \ell_i(u) - \tfrac{\lambda}{m}\big(V_i(p_i, \tfrac{1}{m}) - \tfrac{\rho}{m}\big) + \eta(p_i - \tfrac{1}{m}).$$

Let $y = [u^\top \ \lambda \ \eta]^\top$ and $\mathbf{x} = \mathbf{p}$, then multiplying the above problem by -1 we can reformulate the problem as (1) by defining $f_i(x_i) = \mathbb{I}_{\mathbb{R}_+}(x_i)$ and $h(y) = \mathbb{I}_{X \times \mathbb{R}_+ \times \mathbb{R}}(y)$ as indicator functions.

Learning a kernel matrix: Suppose we are given a training set consisting of feature vectors $\{\mathbf{a}_i\}_{i=1}^m \subset \mathbb{R}^n$, and the corresponding labels $\{b_i\}_{i=1}^m \subset \{-1, +1\}$. Consider $q \in \mathbb{Z}_+$ different embedding of the data and let $K_i \in \mathbb{S}^m_+$ be the corresponding kernel matrix for $i = 1, \ldots, q$. The objective is to learn a kernel matrix K belonging to a class of kernel matrices $\mathcal{K} \subset \mathbb{S}^m_+$ such that it minimizes the training error of an ℓ_2-norm soft-margin nonlinear SVM over $K \in \mathcal{K}$ – see Lanckriet et al. (2004) for more details. In this setting, it is assumed that the class $\mathcal{K}$ is described as a convex set generated by $\{K_i\}_{i=1}^q$, e.g.,

$$\mathcal{K} \triangleq \big\{ \sum_{i=1}^{q} y_i K_i : \ y_i \geq 0, \ i = 1, \ldots, q \big\} \subset \mathbb{S}^m_+. \tag{4}$$

Then, learning over the class $\mathcal{K}$ in (4) is formulated as

$$\min_{\substack{y \in \mathbb{R}^q_+ : \\ \langle \mathbf{r}, y \rangle = c,}} \ \max_{\substack{\mathbf{x}: \ 0 \leq \mathbf{x} \leq C\mathbf{1}_m, \\ \langle \mathbf{b}, \mathbf{x} \rangle = 0}} 2\mathbf{x}^\top \mathbf{1}_m - \sum_{i=1}^{q} y_i \mathbf{x}^\top H(K_i)\mathbf{x} - \lambda \left\|\mathbf{x}\right\|_2^2, \tag{5}$$

where $c, C > 0$ and $\lambda \geq 0$ are model parameters, $y = [y_i]_{i=1}^q$, $\mathbf{r} = [\text{trace}(K_i)]_{i=1}^q$, $\mathbf{b} = [b_i]_{i=1}^m$ and $H(K_i) \triangleq \mathbf{diag}(\mathbf{b}) K_i \, \mathbf{diag}(\mathbf{b})$. Multiplying the objective function by -1, this problem can be formulated as a special case of (1).

Problems in the aforementioned applications are typically large-scale, and standard primal-dual methods do not scale well with the problem dimension and their iterations are memory expensive; therefore, in terms of the efficiency of work required per-iteration, the advantages of randomized block-coordinate schemes will be evident as problem dimension increases.

1.1 Related Work

Saddle point problems have received a significant attention recently due to their vast applicability and modeling flexibility. Here, we briefly review some recent work that is closely related to ours –see also Table 1 for a detailed comparison.

Bilinear SP: There have been several work proposing efficient algorithms to solve convex-concave SP problems with a *bilinear* coupling function, i.e., $\Phi(x, y) = \langle Ax, y \rangle$ for some linear map $A : \mathcal{X} \to \mathcal{Y}^*$, e.g., Chambolle and Pock (2011); Chen et al. (2014); Chambolle and Pock (2016); He and Monteiro (2016); Li and Yan (2021); Alacaoglu et al. (2022b,a). Chambolle and Pock (2016) considered an SP problem with a composite structure and a convergence rate of $\mathcal{O}(1/K)$ for convex-concave and $\mathcal{O}(1/K^2)$ for strongly convex-affine settings are shown. Later Malitsky and Pock (2018) proposed a primal-dual method with linesearch with the same rate results as in Chambolle and Pock (2016).

Non-bilinear SP: There has been a vast body of research, e.g., Juditsky et al. (2011); He et al. (2015); Kolossoski and Monteiro (2017); Malitsky (2018); Malitsky and Tam (2020); Hamedani and Aybat (2021); Zhang et al. (2021, 2022), studying SP problems with non-bilinear coupling functions. Indeed, non-bilinear SP problems can be viewed as a special case of Variational Inequality (VI) problems. In an important work by Nemirovski (2004), a prox-type extra-gradient based method (known as `Mirror-prox`) is proposed. Assuming that the monotone operator is L-Lipschitz continuous and the constraint set is compact, it is shown that the ergodic iterate sequence converges with $\mathcal{O}(L/K)$ rate –also see He et al. (2015) for extension of `Mirror-prox` to SP problems with a composite structure. Later `Mirror-prox` has been extended to exploit

Table 1: Comparison of different methods in Merely Convex (MC) and Strongly Convex (SC) settings. In convergence rates, k denotes the iteration counter. The work-per-iteration for Juditsky et al. (2011); Malitsky (2018); Hamedani and Aybat (2021) is $\mathcal{O}(M)$ while it is $\mathcal{O}(1)$ for the others.

	Properties			Iteration complexity	
Paper	**Non-bilinear Φ**	**Random blocks**	**Line search**	**C-C**	**SC-C**
Chambolle et al. (2017)	✗	✓	✗	$\mathcal{O}(M/k)$	$\mathcal{O}(M/k^2)$
Xu (2021)	✗	✓	✓	$\mathcal{O}(M/(M+k))$	–
Dang and Lan (2014)	✗	✓	✗	$\mathcal{O}(M/k)$	$\mathcal{O}(M/k^2)$
Tran-Dinh and Liu (2020)	✗	✓	✗	$\mathcal{O}(M/k)$	$\mathcal{O}(M^2/k^2)$
Alacaoglu et al. (2022b)	✗	✓	✗	$\mathcal{O}(M/k)$	–
Alacaoglu et al. (2022a)	✗	✓	✗	$\mathcal{O}(M/k)$	–
Juditsky et al. (2011)	✓	✗	✗	$\mathcal{O}(1/k)$	$\mathcal{O}(1/k^2)$
Malitsky (2018)	✓	✗	✓	$\mathcal{O}(1/k)$	–
Hamedani and Aybat (2021)	✓	✗	✓	$\mathcal{O}(1/k)$	$\mathcal{O}(1/k^2)$
This paper	✓	✓	✓	$\mathcal{O}(M/k)$	$\mathcal{O}(M/k^2)$

SP problems for strongly convex-affine setting; in particular, a *multi-stage* method that repeatedly calls `Mirror-prox` is proposed by Juditsky et al. (2011), and $\mathcal{O}(1/K^2)$ rate is shown for the strongly convex-affine setting when $\mathcal{Y}$ is a *compact* set. Later, Malitsky (2018) also considered a monotone VI problem involving a non-smooth function with an easy-to-compute proximal map. The author proposed a proximal extrapolated gradient method, `PEGM`, with an ergodic convergence rate of $\mathcal{O}(1/K)$. The proposed method enjoys a backtracking scheme to estimate the local Lipschitz constants of the monotone map –for a backtracking line-search method tailored to SP problems, see Hamedani and Aybat (2021) and for a more general setting of monotone inclusion problems, see Malitsky and Tam (2020).

Block coordinate: As we indicated earlier, none of these methods mentioned above exploits the block-coordinate structure of (1). However, in a large-scale setting, the computation of full-gradient and/or prox operator might be prohibitively expensive; hence, presenting a strong motivation for using the partial-gradient and/or separable structure of the problem at each iteration of the algorithm. Therefore, the computation may be broken into smaller pieces; thereby, inducing tractability per iteration, at the cost of possibly slower convergence in terms of overall iteration complexity. There has been a vast body of work on randomized block-coordinate descent schemes for primal optimization problems by Nesterov (2012); Luo and Tseng (1992); Xu and Yin (2013); Richtárik and Takáč (2014); Jalilzadeh et al. (2018); but, there are far fewer studies on randomization of block coordinates for SP algorithms. For some papers motivated by the regularized empirical risk minimization (ERM) of linear predictors arising in ML, see Zhu and Storkey (2015); Yu et al. (2015); Zhang and Xiao (2017); Chambolle et al. (2017).

Furthermore, there are some related recent work by Zhong and Kwok (2014); Gao et al. (2016) on block-coordinate ADMM-type algorithms to solve convex optimization problem with linear constraints. Assuming coordinate-wise Lipschitz differentiability of g and q, $\mathcal{O}(1/(1+\gamma k))$ convergence rate is shown under mere convexity assumption, where $\gamma = \frac{M'}{M} = \frac{N'}{N}$ and M' (N') denotes the number of x_i (y_i) coordinates updated at each iteration.

Majority of the previous work on block-coordinate algorithms for SP problems require a bilinear coupling term in the problem formulation (Dang and Lan, 2014; Fercoq and Bianchi, 2015; Valkonen, 2016). However, to the best of our knowledge, none of the existing methods can handle the framework discussed in this paper – the closest one to ours is Xu (2021) which can exploit the block-coordinate structure and does not assume the knowledge of Lipschitz constants via employing line-search; though, it is for constrained optimization problems, not for more general SP problems considered in this paper.

Xu (2021), which is closely related to our work, considered a convex minimization problem with functional constraints, $\min\{f(\mathbf{x}) + g(\mathbf{x}) \,:\, A\mathbf{x} = b,\ G(\mathbf{x}) \leq 0\}$, where g and component functions of G are Lipschitz differentiable convex functions, and f is a proper closed convex function (possibly nonsmooth). When the function $f(\mathbf{x})$ has a separable structure, $\sum_{i\in\mathcal{M}} f_i(x_i)$, a randomized block-coordinate linearized augmented Lagrangian method, `BLALM`, with a convergence rate of $\mathcal{O}(1/(1+\frac{k}{M}))$ is proposed. Note `BLALM` cannot deal with (1) when Φ is not linear in y.

Finally, in a recent paper by Tran-Dinh and Liu (2020), a randomized block-coordinate primal-dual algorithm for solving convex composite optimization problem with linear constraints is proposed. It is shown that the algorithm achieves a last-iterate convergence of $\mathcal{O}(M/k)$ and $\mathcal{O}(M^2/k^2)$ for convex and strongly convex objective functions.

Notation and Definitions. Throughout the paper $\|\cdot\|$ denotes the Euclidean norm, i.e., $\|\cdot\|_2$. Define $\bar{\mu} \triangleq \max_{i\in\mathcal{M}} \mu_i$ and $\underline{\mu} \triangleq \min_{i\in\mathcal{M}} \mu_i$. Let $F : \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$ such that $F(\mathbf{x}) \triangleq \sup_{y\in\mathcal{Y}} \mathcal{L}(\mathbf{x}, y)$ for $\mathbf{x} \in \mathcal{X}$. Under strongly convex-concave setting, i.e., $\underline{\mu} > 0$, we assume that

$L_{yy} = 0$ which implies that $\Phi(\mathbf{x}, \cdot)$ is affine. In this scenario one can assume that $\Phi(\mathbf{x}, y) = \langle g(\mathbf{x}), y\rangle$ for some continuously differentiable vector-valued function $g : \mathcal{X} \to \mathcal{Y}^*$. Therefore, the problem in (1) can be equivalently represented as $\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x})$, where

$$F(\mathbf{x}) = f(\mathbf{x}) + h^*(g(\mathbf{x})), \quad \forall \mathbf{x} \in \mathcal{X}, \tag{6}$$

and h^* denotes the conjugate function of h.

1.2 Our Contributions

In this paper we studied large-scale SP problems with a general structure: the coupling function is *neither* bilinear *nor* separable. To efficiently handle large-scale SP problems, we propose a randomized block-coordinate primal-dual algorithm with backtracking. The proposed algorithm uses momentum acceleration and is equipped with Bregman distance functions that can generalize previous methods such as Chambolle et al. (2017). These type of schemes are the method of choice for the SP problems with a *coordinate-friendly* structure so that the computational tasks performed on each block coordinate at each iteration are significantly cheaper compared to full-gradient computations.

Let $\mathcal{G} : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ denote the primal-dual gap function defined as follows:

$$\mathcal{G}(\bar{\mathbf{x}}, \bar{y}) \triangleq \sup_{(\mathbf{x},y)\in\mathcal{X}\times\mathcal{Y}} \{\mathcal{L}(\bar{\mathbf{x}}, y) - \mathcal{L}(\mathbf{x}, \bar{y})\}, \tag{7}$$

where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Whenever a saddle point of (1) exists, under Lipschitz continuity of $\nabla_y\Phi(\cdot,\cdot)$ and coordinate-wise Lipschitz continuity of $\nabla_{\mathbf{x}}\Phi(\cdot, y)$ for any fixed y, we prove that the iterate sequence converges to a saddle point $(\mathbf{x}^*, y^*)$ in a.s. sense, and we also show convergence rate guarantee in terms of the expected gap $\mathbf{E}[\mathcal{G}(\bar{\mathbf{x}}^k, \bar{y}^k)]$ when $\underline{\mu} = 0$, and in the solution error using both $\mathbf{E}[\|\mathbf{x}^k - \mathbf{x}^*\|^2]$ and $\mathbf{E}[F(\bar{\mathbf{x}}^K) - F(\mathbf{x}^*)]$ when $\underline{\mu} > 0$, where $\{(\bar{\mathbf{x}}^k, \bar{y}^k)\}_k$ is an ergodic average sequence of the iterates.

Main Result 1. *Suppose global Lipschitz constants are available. Under convex-concave setting, for any $\epsilon > 0$, $\mathbf{z}_\epsilon = (\mathbf{x}_\epsilon, y_\epsilon)$ such that $\mathbf{E}[\mathcal{G}(\mathbf{z}_\epsilon)] \le \epsilon$ can be computed within $\mathcal{O}(M/\epsilon)$ primal-dual oracle calls. Moreover, when Φ is strongly convex in $\mathbf{x}$ and linear in y, an ϵ-optimal primal solution $\mathbf{x}_\epsilon$, i.e., $\mathbf{E}[\|\mathbf{x}_\epsilon - \mathbf{x}^*\|^2] \le \epsilon$ and $\mathbf{E}[F(\mathbf{x}_\epsilon) - F(\mathbf{x}^*)] \le \epsilon$, can be obtained within $\mathcal{O}(\sqrt{M}/\sqrt{\epsilon})$ primal-dual oracle calls. Each call to primal and dual oracles require evaluating $\nabla_{x_i}\Phi$ for some $i \in \mathcal{M}$ and $\nabla_y\Phi$, respectively. See Theorem 2 for details.*

To the best of our knowledge, our proposed method is the only randomized block-coordinate primal-dual algorithm that can handle general SP problems as in (1), and our rate results achieve the lower complexity bounds (Chen et al., 2014) for our setting; hence, they are unimprovable, i.e., *optimal*.

Another contribution that immensely increases the algorithmic applicability is the novel backtracking linesearch scheme adopted within the proposed randomized block-coordinate primal-dual method. The step-size selection for each block is closely related to the coordinatewise Lipschitz constant of the partial gradient corresponding to that block; however, in practice, these constants are largely unknown – one needs to know a constant for each block and the number of blocks could be very large. Moreover, even if they can be estimated correctly – which is not the case in many settings, these estimates lead to very conservative step-size selections due to their being global constants. Our technique not only alleviates the burden of estimating largely unknown constants; but, also make the convergence much faster in practice as it corresponds to using local Lipschitz constants which leads to larger step-sizes while retaining the theoretical convergence rate guarantees of primal-dual methods using the global constants.

Main Result 2. *Suppose that the global Lipschitz constants are not available. The iterates generated by our method with backtracking line search converges to a saddle point point almost surely with the same oracle complexity as in Main Result 1 for both convex and strongly convex settings up to $\mathcal{O}(1)$ constants. See Theorem 1 for details.*

2 Preliminaries

In this section, we state the main definitions and assumptions that we need for our convergence analysis.

Definition 1. *Let $f(\mathbf{x}) \triangleq \sum_{i\in\mathcal{M}} f_i(x_i)$ and $\mathcal{M} = \{1, \dots, M\}$, and define $U_i \in \mathbb{R}^{m\times m_i}$ for $i \in \mathcal{M}$ such that $\mathbf{I}_m = [U_1, \dots, U_M]$, where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. Let $\varphi_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R}$ be a differentiable function on an open set containing $\mathbf{dom}\, h$. Suppose $\varphi_{\mathcal{Y}}$ is 1-strongly convex with respect to $\|\cdot\|_{\mathcal{Y}}$. Let $\mathbf{D}_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a Bregman distance function corresponding to $\varphi_{\mathcal{Y}}$, i.e., $\mathbf{D}_{\mathcal{Y}}(y, \bar{y}) \triangleq \varphi_{\mathcal{Y}}(y) - \varphi_{\mathcal{Y}}(\bar{y}) - \langle\nabla\varphi_{\mathcal{Y}}(\bar{y}), y - \bar{y}\rangle$. The dual space of $\mathcal{Y}$ is denoted by $\mathcal{Y}^*$, and $\|\cdot\|_{\mathcal{Y}^*} : \mathcal{Y}^* \to \mathbb{R}$ such that $\|y'\|_{\mathcal{Y}^*} \triangleq \max\{\langle y', y\rangle : \|y\|_{\mathcal{Y}} \le 1\}$ denotes the dual norm. Similarly, for each $i \in \mathcal{M}$, given an arbitrary norm $\|\cdot\|_{\mathcal{X}_i}$ on $\mathcal{X}_i$, define $\|\cdot\|_{\mathcal{X}_i^*} : \mathcal{X}_i^* \to \mathbb{R}$ and $\mathbf{D}_{\mathcal{X}_i} : \mathcal{X}_i \times \mathcal{X}_i \to \mathbb{R}$ for some $\varphi_{\mathcal{X}_i}$ that is differentiable and 1-strongly convex with respect to $\|\cdot\|_{\mathcal{X}_i}$.*

Definition 2. *Given a diagonal matrix $\mathbf{C} = \mathbf{diag}([c_i]_{i\in\mathcal{M}})$ for some $c_i \ge 0$ for $i \in \mathcal{M}$, define $\|\cdot\|_{\mathbf{C}} : \mathcal{X} \to \mathbb{R}$ such that $\|\mathbf{x}\|_{\mathbf{C}} \triangleq \sqrt{\sum_{i\in\mathcal{M}} c_i\|x_i\|^2_{\mathcal{X}_i}}$; furthermore, $\mathbf{D}^{\mathbf{C}}_{\mathcal{X}}(\mathbf{x}, \bar{\mathbf{x}}) \triangleq \sum_{i\in\mathcal{M}} c_i\mathbf{D}_{\mathcal{X}_i}(x_i, \bar{x}_i)$ for all $\mathbf{x}, \bar{\mathbf{x}} \in \mathcal{X}$.*

Next, we state our assumptions on f, h and Φ.

Assumption 1. *Suppose $f_i : \mathcal{X}_i \to \mathbb{R} \cup \{+\infty\}$ is a closed convex function and its convexity modulus w.r.t. $\|\cdot\|_{\mathcal{X}_i}$ is $\mu_i \ge 0$ for all $i \in \mathcal{M}$ and $h : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ is a closed convex function. Moreover, suppose that $\{f_i\}_{i\in\mathcal{M}}$, h and*

$\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *satisfy the following assumptions:*
(i) *for any fixed* $y \in \mathbf{dom}\, h$, $\Phi(\cdot, y)$ *is convex on* $\mathbf{dom}\, f$, *and coordinate-wise Lipschitz differentiable on an open set containing* $\mathbf{dom}\, f$, *and for all* $i \in \mathcal{M}$, *there exists* $L_{x_i x_i} \geq 0$ *such that for any* $\bar{\mathbf{x}} \in \mathbf{dom}\, f$ *and* $v \in \mathcal{X}_i$ *satisfying* $\bar{\mathbf{x}} + U_i v \in \mathbf{dom}\, f$, *and* $\bar{y} \in \mathbf{dom}\, h$,

$$\|\nabla_{x_i}\Phi(\bar{\mathbf{x}} + U_i v, \bar{y}) - \nabla_{x_i}\Phi(\bar{\mathbf{x}}, \bar{y})\|_{\mathcal{X}_i^*} \leq L_{x_i x_i} \|v\|_{\mathcal{X}_i}; \tag{8}$$

(ii) *for any fixed* $\bar{\mathbf{x}} \in \mathbf{dom}\, f$, $\Phi(\bar{\mathbf{x}}, \cdot)$ *is concave on* $\mathbf{dom}\, h$, *and differentiable on an open set containing* $\mathbf{dom}\, h$, *and there exists* $L_{yy} \geq 0$ *and* $L_{yx_i} > 0$ *for all* $i \in \mathcal{M}$ *such that for any* $y, \bar{y} \in \mathbf{dom}\, h$, $v \in \mathcal{X}_i$ *and* $i \in \mathcal{M}$ *satisfying* $\bar{\mathbf{x}} + U_i v \in \mathbf{dom}\, f$,

$$\begin{aligned}\|\nabla_y \Phi(\bar{\mathbf{x}} + U_i v, \bar{y}) - \nabla_y \Phi(\bar{\mathbf{x}}, y)\|_{\mathcal{Y}^*} \\ \leq L_{yy}\|y - \bar{y}\|_{\mathcal{Y}} + L_{yx_i}\|v\|_{\mathcal{X}_i};\end{aligned} \tag{9}$$

(iii) *for any* $i \in \mathcal{M}$, $\operatorname{argmin}_{x_i \in \mathcal{X}_i}\{tf_i(x_i) + \langle s, x_i\rangle + \mathbf{D}_{\mathcal{X}_i}(x_i, \bar{x}_i)\}$ *can be computed efficiently for any* $\bar{x}_i \in \mathbf{dom}\, f_i$, $s \in \mathcal{X}_i^*$ *and* $t > 0$. *Similarly,* $\operatorname{argmin}_{y \in \mathcal{Y}}\{th(y) + \langle s, y\rangle + \mathbf{D}_{\mathcal{Y}}(y, \bar{y})\}$ *is easy to compute for any* $\bar{y} \in \mathbf{dom}\, h$, $s \in \mathcal{Y}^*$ *and* $t > 0$.

We also define some diagonal matrices to simplify the notation in the rest of the paper:

$$\begin{aligned}&\mathfrak{M} \triangleq \mathbf{diag}([\mu_i]_{i\in\mathcal{M}}), \quad \mathbf{L_{xx}} \triangleq \mathbf{diag}([L_{x_i x_i}]_{i\in\mathcal{M}}),\\ &\mathbf{L_{yx}} \triangleq \mathbf{diag}([L_{yx_i}]_{i\in\mathcal{M}}).\end{aligned} \tag{10}$$

Moreover, we define the largest coordinate-wise Lipschitz constants as follows:

$$\overline{L}_{xx} \triangleq \max_{i\in\mathcal{M}} L_{x_i x_i}, \quad \overline{L}_{yx} \triangleq \max_{i\in\mathcal{M}} L_{yx_i}. \tag{11}$$

3 Randomized Accelerated Primal-dual Algorithm

In this section, we state our proposed algorithm to solve (1) for some given arbitrary norm $\|\cdot\|_{\mathcal{X}_i}$ on $\mathcal{X}_i$ and some Bregman function $\mathbf{D}_{\mathcal{X}_i}$ as in Definition 1 for all $i \in \mathcal{M}$. We propose a randomized block-coordinate accelerated primal-dual (`RB-APD`) method (see Algorithm 1) consists of a single loop primal-dual steps. After the initialization of parameters, a dual ascent step is taken in the direction of $\nabla_y\Phi$ with a momentum term in terms of $\nabla_y\Phi$ to gain acceleration for general convex-concave problems. This can be viewed as a generalization of the approach proposed by Hamedani and Aybat (2021) with $M = 1$ (it also generalizes the commonly used extrapolation step when the function Φ is bilinear[2]). Then, a primal block-coordinate descent step is taken using $\nabla_{x_i}\Phi$ for a uniformly chosen random block coordinate.

[2]Majority of the existing methods use past iterates to gain momentum, e.g., Chambolle and Pock (2011, 2016); Chambolle et al. (2017); Alacaoglu et al. (2022b,a) use the momentum term $(1+\theta^k)\mathbf{x}^k - \theta^k\mathbf{x}^{k-1}$. This iteration can be recovered by our method when Φ is bilinear.

Algorithm 1 Randomized Block-coordinate Accelerated Primal-Dual (`RB-APD`) Algorithm

Input: $(\mathbf{x}_0, y_0) \in \mathbf{dom}\, f \times \mathbf{dom}\, h$, $\{\mu_i\}_{i\in\mathcal{M}} \subseteq \mathbb{R}_+$, $\gamma^0 > 0$, $\bar{\tau} \in \left(0, \frac{1}{\bar{\mu}(M-1)}\right)$, where $\bar{\mu} \triangleq \max_{i\in\mathcal{M}}\mu_i$
$(\mathbf{x}_{-1}, y_{-1}) \leftarrow (\mathbf{x}_0, y_0)$, $\quad \underline{\mu} \leftarrow \min_{i\in\mathcal{M}}\mu_i$
$\tilde{\tau}^0 \leftarrow \bar{\tau}$, $\quad \sigma^{-1} \leftarrow \gamma^0\bar{\tau}$
for $k \geq 0$ **do**
  $\sigma^k \leftarrow \gamma^k\tilde{\tau}^k$, $\quad \theta^k \leftarrow \frac{\sigma^{k-1}}{\sigma^k}$
  $q^k \leftarrow M(\nabla_y\Phi(\mathbf{x}^k, y^k) - \nabla_y\Phi(\mathbf{x}^{k-1}, y^{k-1}))$
  $s^k \leftarrow \nabla_y\Phi(\mathbf{x}^k, y^k) + \theta^k q^k$
  $y^{k+1} \leftarrow \operatorname{argmin}_{y\in\mathcal{Y}} h(y) - \langle s^k, y\rangle + \frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y, y^k)$
  Choose $i_k \in \mathcal{M}$ uniformly at random
  $\tau_{i_k}^k \leftarrow \left(\frac{1}{M}(\mu_{i_k} + \frac{1}{\tilde{\tau}^k}) - \mu_{i_k}\right)^{-1}$
  $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k$
  $x_{i_k}^{k+1} \leftarrow \operatorname{argmin}_{x\in\mathcal{X}_{i_k}} f_{i_k}(x) + \left\langle \nabla_{x_{i_k}}\Phi(\mathbf{x}^k, y^{k+1}), x\right\rangle + \frac{1}{\tau_{i_k}^k}\mathbf{D}_{\mathcal{X}_{i_k}}(x, x_{i_k}^k)$
  $\gamma^{k+1} \leftarrow \gamma^k(1 + \underline{\mu}\tilde{\tau}^k)$
  $\tilde{\tau}^{k+1} \leftarrow \tilde{\tau}^k\sqrt{\frac{\gamma^k}{\gamma^{k+1}}}$, $\quad k \leftarrow k+1$
end for

In many practical settings, typically finding the Lipschitz constants to select an appropriate step-size can be difficult. Therefore, we propose a novel backtracking linesearch that can be combined with `RB-APD` (called `RB-APD-B`) to adaptively select primal and dual step-sizes without the knowledge of Lipschitz constants. To this end, we will define a test function C_*^k that implicitly estimates local Lipschitz constants in order to accept or reject the tested primal-dual step-sizes at each bactracking iteration. In fact, at iteration $k \geq 0$ such a test function can be calculated using only the information related to coordinate i_k, i.e., checking the test function does not involve computing $\nabla_{x_i}\Phi$ for $i \in \mathcal{M} \setminus \{i_k\}$. Starting from an initial arbitrary step-size, at each iteration we reduce the step-sizes by a factor of η until C_*^k falls below a certain threshold. The details of our method is displayed in Algorithm 2. Next, we formally define our test function.

Definition 3. *For any* $k \geq 0$, *given* $\tilde{\tau}^k, \sigma^k, \theta^k > 0$, *define* $\mathbf{T}^k \triangleq \mathbf{diag}\left(\left[\frac{1}{\tau_i^k}\right]_{i\in\mathcal{M}}\right)$ *where* $\tau_i^k \triangleq \left(\frac{1}{M}(\mu_i + \frac{1}{\tilde{\tau}^k}) - \mu_i\right)^{-1}$ *for* $i \in \mathcal{M}$. *We define the test function for the backtracking line-search as follows:*

$$\begin{aligned}C_*^k \triangleq & M\Big(\Phi(\mathbf{x}^{k+1}, y^{k+1}) - \Phi(\mathbf{x}^k, y^{k+1})\\ & - \left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y^{k+1}),\, \mathbf{x}^{k+1} - \mathbf{x}^k\right\rangle\Big)\\ & + \frac{M\sigma^k}{2c_\alpha}\left\|\nabla_y\Phi(\mathbf{x}^{k+1}, y^{k+1}) - \nabla_y\Phi(\mathbf{x}^k, y^{k+1})\right\|_{\mathcal{Y}^*}^2\\ & + \frac{M\sigma^k}{2c_\beta}\left\|\nabla_y\Phi(\mathbf{x}^k, y^{k+1}) - \nabla_y\Phi(\mathbf{x}^k, y^k)\right\|_{\mathcal{Y}^*}^2\\ & - M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}^{k+1}, \mathbf{x}^k)\\ & - \Big(\frac{1 - M(c_\alpha + c_\beta)}{\sigma^k}\Big)\mathbf{D}_{\mathcal{Y}}(y^{k+1}, y^k).\end{aligned} \tag{12}$$

Remark 3.1. *One can also consider an alternative test function involving only partial gradients:*

$$\begin{aligned}\widetilde{C}_*^k \triangleq & M\left\langle \nabla_{x_{i_k}}\Phi(\mathbf{x}^{k+1},y^{k+1}) - \nabla_{x_{i_k}}\Phi(\mathbf{x}^k,y^{k+1}), x_{i_k}^{k+1}-x_{i_k}^k\right\rangle \\ & + \frac{M\sigma^k}{2c_\alpha}\left\|\nabla_y\Phi(\mathbf{x}^{k+1},y^{k+1}) - \nabla_y\Phi(\mathbf{x}^k,y^{k+1})\right\|_{\mathcal{Y}^*}^2 \\ & + \frac{M\sigma^k}{2c_\beta}\left\|\nabla_y\Phi(\mathbf{x}^k,y^{k+1}) - \nabla_y\Phi(\mathbf{x}^k,y^k)\right\|_{\mathcal{Y}^*}^2 \\ & - \frac{M}{\tau_{i_k}^k}\mathbf{D}_{\mathcal{X}_i}(x_{i_k}^{k+1},x_{i_k}^k) \\ & - \Big(\frac{1-M(c_\alpha+c_\beta)}{\sigma^k}\Big)\mathbf{D}_{\mathcal{Y}}(y^{k+1},y^k).\end{aligned}$$

Note that convexity of $\Phi(\cdot,y)$ implies that $\widetilde{C}_*^k$ upper bounds C_*^k. It is worth highlighting that both C_*^k and $\widetilde{C}_*^k$ only use the partial gradient information of $\nabla_{x_{i_k}}\Phi$ and step-size $\tau_{i_k}^k$ related to the randomly picked coordinate $i_k \in \mathcal{M}$, and does not involve computing $\nabla_{x_i}\Phi$ for $i \in \mathcal{M}\setminus\{i_k\}$.

Algorithm 2 Randomized Block-coordinate Accelerated Primal-Dual algorithm with Backtracking (RB−APD−B)

Input: $(\mathbf{x}_0,y_0)\in \mathbf{dom}\, f\times \mathbf{dom}\, h$, $\{\mu_i\}_{i\in\mathcal{M}}\subseteq\mathbb{R}_+$, $c_\alpha,c_\beta,\delta\geq 0$, $\eta\in(0,1)$, $\gamma_0>0$, $\bar\tau\in\Big(0,\frac{1}{\bar\mu(M-1)}\Big)$, where $\bar\mu\triangleq\max_{i\in\mathcal{M}}\mu_i$
$(\mathbf{x}_{-1},y_{-1})\leftarrow(\mathbf{x}_0,y_0)$, $\quad \underline{\mu}\leftarrow\min_{i\in\mathcal{M}}\mu_i$
$\tilde\tau^0\leftarrow\bar\tau$, $\quad\sigma^{-1}\leftarrow\gamma^0\bar\tau$,
for $k\geq 0$ **do**
  Choose $i_k\in\mathcal{M}$ uniformly at random
  loop
    $\sigma^k\leftarrow\gamma^k\tilde\tau^k$, $\quad\theta^k\leftarrow\frac{\sigma^{k-1}}{\sigma^k}$
    $\alpha^{k+1}\leftarrow c_\alpha/\sigma^k$, $\quad\beta^{k+1}\leftarrow c_\beta/\sigma^k$
    $q^k\leftarrow M(\nabla_y\Phi(\mathbf{x}^k,y^k)-\nabla_y\Phi(\mathbf{x}^{k-1},y^{k-1}))$
    $s^k\leftarrow\nabla_y\Phi(\mathbf{x}^k,y^k)+\theta^k q^k$
    $y^{k+1}\leftarrow\operatorname{argmin}_{y\in\mathcal{Y}} h(y)-\langle s^k,y\rangle+\frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y,y^k)$
    $\tau_{i_k}^k\leftarrow\big(\frac{1}{M}(\mu_{i_k}+\frac{1}{\tilde\tau^k})-\mu_{i_k}\big)^{-1}$
    $\mathbf{x}_{i_k}^{k+1}\leftarrow\mathbf{x}^k$
    $x_{i_k}^{k+1}\leftarrow\operatorname{argmin}_{x\in\mathcal{X}_{i_k}} f_{i_k}(x)$ $+\langle\nabla_{x_{i_k}}\Phi(\mathbf{x}^k,y^{k+1}),x\rangle+\frac{1}{\tau_{i_k}^k}\mathbf{D}_{\mathcal{X}_{i_k}}(x,x_{i_k}^k)$
    if $C_*^k\leq-\delta\Big[\frac{M}{\tau_{i_k}^k}\mathbf{D}_{\mathcal{X}_i}(x_{i_k}^{k+1},x_{i_k}^k)+\frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y^{k+1},y^k)\Big]$ **then**
      go to line 21
    else
      $\tilde\tau^k\leftarrow\eta\tilde\tau^k$
    end if
  end loop
  $\gamma^{k+1}\leftarrow\gamma^k(1+\underline{\mu}\tilde\tau^k)$
  $\tilde\tau^{k+1}\leftarrow\tilde\tau^k\sqrt{\frac{\gamma^k}{\gamma^{k+1}}}$, $\quad k\leftarrow k+1$
end for

4 Convergence Analysis

In this section, we discuss the convergence properties of RB−APD−B and RB−APD algorithms in Theorems 1 and 2, respectively, which are the main results of this paper. All related proofs are provided in the appendix. In the rest, $\mathbf{E}[\cdot]$ denotes the expectation operation.

Assumption 2. *For the case $\underline{\mu}>0$, we assume that the Bregman distance generating function $\varphi_{\mathcal{X}_i}(x_i)=\frac{1}{2}\|x_i\|_{\mathcal{X}_i}^2$ for $\|x_i\|_{\mathcal{X}_i}=\sqrt{\langle x_i,x_i\rangle}$ for all $i\in\mathcal{M}$, which leads to $\mathbf{D}_{\mathcal{X}}(\mathbf{x},\mathbf{x}')=\frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|^2$. On the other hand, for the case $\underline{\mu}=0$, a more general distance generating function $\varphi_{\mathcal{X}_i}(x_i)$ can be chosen as defined in Definition 1 assuming that it has a $L_{\varphi_{\mathcal{X}}}$-Lipschitz continuous gradient for all $i\in\mathcal{M}$, after setting $\mathfrak{M}=\mathbf{0}_{M\times M}$ without loss of generality, i.e., treating $\mu_i=0$ for all $i\in\mathcal{M}$. For the case $\underline{\mu}=0$, one can still work with the original $\mathfrak{M}$ without setting it to $\mathbf{0}_{M\times M}$ if one uses quadratic Bregman as in the case $\underline{\mu}>0$ case, i.e., setting $\mathbf{D}_{\mathcal{X}}(\mathbf{x},\mathbf{x}')=\frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|^2$.*

In RB−APD, stated in Algorithm 1, we considered a particular step size sequence. However, RB−APD can be shown to work for a larger class of step sizes. We next describe such step-size conditions to ensure the convergence guarantee for Algorithm 1. Indeed, we provide *(i)* a set of conditions that provide upper bounds on $\tau_{i_k}^k$ and σ^k depending on the Lipschitz constants, for any $k\geq 0$ (see (13a) and (13b)); *(ii)* a set of recursive inequalities that connect primal and dual step-sizes (see (13c) and (13d)).

Assumption 3. *There exists $\{[\tau_i^k]_{i\in\mathcal{M}},\sigma^k,\theta^k\}_{k\geq 0}$ such that $\theta^0=1$, and*

$$\frac{1-\delta}{\tau_{i_k}^k}\geq L_{x_{i_k}x_{i_k}}+\frac{L_{yx_{i_k}}^2}{\alpha^{k+1}}, \tag{13a}$$

$$\frac{1-\delta}{\sigma^k}\geq M\theta^k(\alpha^k+\beta^k)+\frac{ML_{yy}^2}{\beta^{k+1}}, \tag{13b}$$

$$t^k(\mathbf{T}^k+\mathfrak{M})\succeq t^{k+1}\big(\mathbf{T}^{k+1}+(1-\frac{1}{M})\mathfrak{M}\big), \tag{13c}$$

$$\frac{t^k}{\sigma^k}\geq\frac{t^{k+1}}{\sigma^{k+1}}, \quad t^{k+1}\theta^{k+1}=t^k, \tag{13d}$$

where $\mathbf{T}^k\triangleq[\frac{1}{\tau_i^k}\mathbf{I}_{m_i}]_{i\in\mathcal{M}}$, for some positive $\{t^k,\alpha^k\}_{k\geq 0}$ such that $t^0=1$, nonnegative $\{\beta^k\}_{k\geq 0}$ and $\delta\in[0,1)$.

Remark 4.1. *Stepsize update rule in Algorithm 1 implies that for all $k\geq 0$,*

$$\theta^{k+1}=\frac{1}{\sqrt{1+\underline{\mu}\tilde\tau^k}}, \quad \tilde\tau^{k+1}=\theta^{k+1}\tilde\tau^k, \quad \sigma^{k+1}=\sigma^k/\theta^{k+1}.$$

Suppose the parameters are initialized such that $\theta^0=1$ and $[\tau_i^0]_{i\in\mathcal{M}},\sigma^0>0$ such that

$$\Big((1-\delta)\mathbf{T}^0-\mathbf{L}_{\mathbf{xx}}\Big)\frac{1}{\sigma^0}\succeq\frac{1}{c_\alpha}\mathbf{L}_{y\mathbf{x}}^2, \tag{14a}$$

$$1-\big(\delta+M(c_\alpha+c_\beta)\big)\geq\frac{ML_{yy}^2}{c_\beta}(\sigma^0)^2, \tag{14b}$$

for any $\delta, c_\alpha,\ c_\beta \in \mathbb{R}_+$ such that $M(c_\alpha + c_\beta) + \delta \le 1$ satisfying $c_\alpha,\ c_\beta > 0$ when $L_{yy} > 0$, and $c_\alpha > 0$, $c_\beta = 0$ when $L_{yy} = 0$. Then $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}_{k\ge 0}$ satisfies Assumption 3 by selecting $t^k = \sigma^k/\sigma^0$.

As it is apparent from the remark above, for the convex-concave setting, i.e., $\underline{\mu} = 0$, with $\mathfrak{M} = \mathbf{0}$, a constant step size sequence can be selected for the RB−APD, i.e., $\tilde{\tau}^k = \tilde{\tau}^0$, $\sigma^k = \sigma^0$, and $\theta^k = 1$ for all $k \ge 0$ such that $(\tilde{\tau}^0, \sigma^0)$ satisfy (14). We show that this choice implies $\mathcal{O}(1/K)$ rate for the expected gap function; hence, $\mathcal{O}(1/K)$ rate for the primal suboptimality. On the other hand, in strongly convex setting, i.e., $\mathfrak{M} \succ 0$, an accelerated convergence rate of $\mathcal{O}(1/K^2)$ for the primal suboptimality can be obtained by decreasing the primal step-size τ_i^k and increasing the dual step-size σ^k.

Next, we consider the scenario where the Lipschitz constants are not available. We show that RB−APD−B stated in Algorithm 2 is well-defined, i.e., the condition in Line 15 holds in finite number of backtracking steps, and the generated step-size sequence $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$ satisfy the conditions in (13c) and (13d)).

Lemma 1. *Under Assumptions 1 and 2, consider RB−APD−B displayed in Algorithm 2 for any given $\delta \in [0,1)$ and $c_\alpha, c_\beta \ge 0$ such that $M(c_\alpha + c_\beta) + \delta \le 1$. When $L_{yy} > 0$, set $c_\alpha,\ c_\beta > 0$; otherwise, when $L_{yy} = 0$, set $c_\alpha > 0$ and $c_\beta = 0$. The RB−APD−B iterate and step-size sequences, i.e., $\{\mathbf{x}^k, y^k\}_{k\ge 0}$ and $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}_{k\ge 0}$, are well-defined; more precisely, for any $k \ge 0$,*

$$C_*^k \le -\delta\Big[M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}^{k+1}, \mathbf{x}^k) + \frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y^{k+1}, y^k)\Big], \quad (15)$$

holds after finite number of backtracking iterations, where C_*^k is defined in (12) using $\{\sigma^k, \mathbf{T}^k\}_k$ and c_α, c_β as above. Furthermore, $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}_{k\ge 0}$ satisfy (13c) and (13d) for $\{t^k\}_{k\ge 0}$ such that $t^k = \sigma^k/\sigma^0$ for $k \ge 0$.

We are now ready to state the convergence rate of RB−APD−B, stated in Algorithm 2.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Let $\delta \in [0,1)$, $c_\alpha > 0$ and $c_\beta \ge 0$ are chosen as stated below. For any given $(\mathbf{x}_0, y_0) \in \mathbf{dom}\, f \times \mathbf{dom}\, h$, $\gamma^0 > 0$ and $\bar{\tau} \in \big(0, \frac{1}{\bar{\mu}(M-1)}\big)$, RB−APD−B, stated in Algorithm 2, is well-defined, i.e., the number of inner iterations is finite and bounded by $1+\log_{1/\eta}(\frac{\bar{\tau}}{\Psi})$ uniformly for $k \ge 0$ for some $\Psi > 0$[3]. Let $\{\mathbf{x}^k, y^k\}_{k\ge 0}$ denote the iterate sequence generated by RB−APD−B. For $K \ge 1$, let $T^K \triangleq \sum_{k=0}^{K-1} t^k$, $\bar{\mathbf{x}}^K = \frac{M}{T^K + M - 1}\Big(\sum_{k=0}^{K-2}\big(t^k - (1-\frac{1}{M})t^{k+1}\big)\mathbf{x}^{k+1} + t^{K-1}\mathbf{x}^K\Big)$ and $\bar{y}^K = \frac{1}{T^K}\sum_{k=0}^{K-1} t^k y^{k+1}$ for $\{t^k\}_{k\ge 0}$ such that $t^k = \sigma^k/\sigma^0$ for $k \ge 0$.*

(Part I.) *Suppose $\underline{\mu} = 0$ and $\mathbf{dom}\, f \times \mathbf{dom}\, h$ is compact. Assume $M(c_\alpha + c_\beta) + \delta \le 1$ holds for some $c_\alpha, c_\beta > 0$ if $L_{yy} > 0$; and $c_\beta = 0$, and $Mc_\alpha + \delta \le 1$ for some $c_\alpha > 0$ otherwise. If a saddle point for (1) exists and $\delta > 0$, then $\{(\mathbf{x}^k, y^k)\}_{k\ge 0}$ converges to a saddle point almost surely. Moreover, for all $K \ge 1$, the following bound holds:*

$$\mathbf{E}\left[\mathcal{G}(\bar{\mathbf{x}}^K, \bar{y}^K)\right] \le \frac{1}{T^K}\left(\bar{B} + \sup_{\mathbf{x}\in\mathbf{dom}\, f} B_1(\mathbf{x}) + \sup_{y\in\mathbf{dom}\, h} B_2(y)\right) \quad (16)$$

$$B_1(\mathbf{x}) \triangleq M\mathbf{D}_{\mathcal{X}}^{(1+\frac{1}{M})\mathbf{T}^0+\mathfrak{M}}(\mathbf{x}, \mathbf{x}^0),$$

$$B_2(y) \triangleq \Big(\tfrac{1}{\sigma^0} + \theta^0(M-1)L_{yy}\Big)\mathbf{D}_{\mathcal{Y}}(y, y^0),$$

for some constant[4] $\bar{B} \in \mathbb{R}_+$, and $T^K = \Omega(K)$, implying $\mathcal{O}(1/K)$ sublinear rate for $\mathbf{E}\left[\mathcal{G}(\bar{\mathbf{x}}^K, \bar{y}^K)\right]$.

(Part II.) *Suppose $\underline{\mu} > 0$ and $L_{yy} = 0$. Assume $Mc_\alpha + \delta \in (0,1]$ and $c_\beta = 0$. If a saddle point $(\mathbf{x}^*, y^*)$ for (1) exists, then $\{(\mathbf{x}^k, y^k)\}_{k\ge 0}$ converges to $\mathbf{x}^*$ and $\{y^k\}$ has a limit point almost surely. Moreover, if $\delta > 0$, then any limit point of $\{\mathbf{x}^k, y^k\}$ is a saddle point almost surely, and*

$$\mathbf{E}\left[\tfrac{\gamma^K}{2}\left\|\mathbf{x}^* - \mathbf{x}^K\right\|_{\mathcal{X}}^2 + (1 - Mc_\alpha)\mathbf{D}_{\mathcal{Y}}(y^*, y^K)\right] \le \tfrac{\gamma^0}{2}\left\|\mathbf{x}^* - \mathbf{x}^0\right\|_{\mathcal{X}}^2 + \mathbf{D}_{\mathcal{Y}}(y^*, y^0) + \sigma^0(M-1)\Big(\mathcal{L}(\mathbf{x}^0, y^*) - \mathcal{L}(\mathbf{x}^*, y^*)\Big), \quad (17)$$

$$\mathbf{E}\left[F(\bar{\mathbf{x}}^K) - F(\mathbf{x}^*)\right] \le \frac{1}{T^K}\Big(\tfrac{\gamma^0}{2\sigma^0}\left\|\mathbf{x}^* - \mathbf{x}^0\right\|_{\mathcal{X}}^2 + \tfrac{1}{\sigma^0}\sup_{y\in\mathbf{dom}\, h}\mathbf{D}_{\mathcal{Y}}(y, y^0) + (M-1)\big(F(\mathbf{x}^0) - F(\mathbf{x}^*)\big)\Big) \quad (18)$$

for all $K \ge 1$. Furthermore, both $\gamma^K = \Omega(K^2)$ and $T^K = \Omega(K^2)$; hence, $\mathbf{E}\big[\left\|\mathbf{x}^K - \mathbf{x}^*\right\|_{\mathcal{X}}^2\big] = \mathcal{O}(1/K^2)$ and $0 \le \mathbf{E}\left[F(\bar{\mathbf{x}}^K) - F(\mathbf{x}^*)\right] \le \mathcal{O}(1/K^2)$.

Note that the term $\mathcal{L}(\mathbf{x}^0, y^*) - \mathcal{L}(\mathbf{x}^*, y^*)$ in (17) can be bounded above by $F(\mathbf{x}^0) - F(\mathbf{x}^*)$.

Now we are ready to state convergence rate of Algorithm 1.

Theorem 2. *Suppose Assumptions 1 and 2 hold, and given arbitrary $(\mathbf{x}_0, y_0) \in \mathbf{dom}\, f \times \mathbf{dom}\, h$, $\gamma^0 > 0$ and $\bar{\tau} \in \big(0, \frac{1}{\bar{\mu}(M-1)}\big)$ such that (14) holds for some $c_\alpha, c_\beta \ge 0$ and $\delta \in [0,1)$ as described in Theorem 1, let $\{\mathbf{x}^k, y^k\}_{k\ge 0}$ be the iterate sequence generated by RB−APD. Suppose $\bar{\mathbf{z}}^K = (\bar{\mathbf{x}}^K, \bar{y}^K)$ and T^K are defined for $K \ge 1$ as in Theorem 1.*

(Part I.) *Suppose $\underline{\mu} = 0$ and $\mathbf{dom}\, f \times \mathbf{dom}\, h$ is compact. The stepsize rule in Algorithm 1 implies $\tilde{\tau}^k = \tilde{\tau}^0$, $\sigma^k = \sigma^0$ and $\theta^k = 1$ for all $k \ge 0$; hence, $t^k = 1$ for $k \ge 0$, implying $T^K = K$. If a saddle point for (1) exists and $\delta > 0$, then $\{(\mathbf{x}^k, y^k)\}_{k\ge 0}$ almost surely converges to a saddle point $(\mathbf{x}^*, y^*)$. Moreover, (16) holds for all $K \ge 1$, which implies $\mathbf{E}[\mathcal{G}(\bar{\mathbf{z}}^K)] \le \mathcal{O}(1/K)$.*

[3] Ψ is a function of the problem and algorithm parameters, and its exact form is provided in the appendix.

[4] See appendix for its dependence on algorithm and problem parameters.

(Part II.) *Suppose* $\underline{\mu} > 0$ *and* $L_{yy} = 0$. *If a saddle point for* (1) *exists, then* $\{\mathbf{x}^k\}_{k\geq 0}$ *converges to* $\mathbf{x}^*$ *and* $\{y^k\}$ *has a limit point almost surely.*[5] *Moreover, if* $\delta > 0$, *then any limit point* $(\mathbf{x}^*, y^*)$ *is a saddle point almost surely satisfying* (17) *and* (18) *for all* $K \geq 1$. *Furthermore, as in Theorem 1, both* $\gamma^K = \Omega(K^2)$ *and* $T^K = \Omega(K^2)$; *hence,* $\mathbf{E}\big[\,\|\mathbf{x}^K - \mathbf{x}^*\|_{\mathcal{X}}^2\,\big] = \mathcal{O}(1/K^2)$ *and* $0 \leq \mathbf{E}\left[F(\bar{\mathbf{x}}^K) - F(\mathbf{x}^*)\right] \leq \mathcal{O}(1/K^2)$.

5 Numerical Experiment

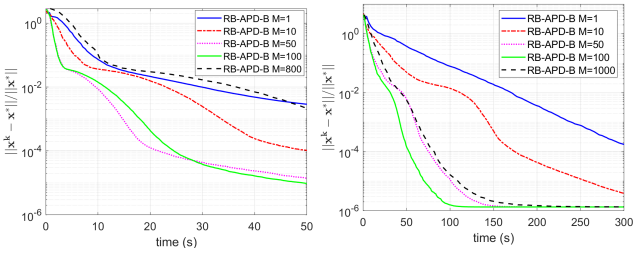
In this section, we implement our scheme on a kernel matrix learning problem described in section 1, and benchmark with `Mirror-prox` by He et al. (2015), proximal extra gradient method (`PEGM`) by Malitsky and Pock (2018), and accelerated primal-dual (`APDB`) by Hamedani and Aybat (2021). The experiments are performed on a machine running 64-bit Windows 10 with Intel i7-8650U @2.11GHz and 16GB RAM.

5.1 Learning a Kernel Matrix

We consider the formulation in (5) for the kernel class $\mathcal{K}$ in (4), and we use a similar setup as in (Lanckriet et al., 2004). In particular, we consider three kernel functions ($q = 3$); polynomial kernel function $\phi_1(\mathbf{a}, \bar{\mathbf{a}}) = (1 + \mathbf{a}^\top \bar{\mathbf{a}})^2$, Gaussian kernel function $\phi_2(\mathbf{a}, \bar{\mathbf{a}}) = \exp(-0.5(\mathbf{a} - \bar{\mathbf{a}})^\top(\mathbf{a} - \bar{\mathbf{a}})/0.1)$, and linear kernel function $\phi_3(\mathbf{a}, \bar{\mathbf{a}}) = \mathbf{a}^\top \bar{\mathbf{a}}$ to compute $K_1, K_2, K_3 \in \mathbb{S}_+^m$, respectively, where m denotes the number data points. We set $\lambda = 1$, $c = \sum_{\ell=1}^3 r_\ell$, where $r_\ell = \text{trace}(K_\ell)$ for $\ell = 1, 2, 3$. The kernel matrices are normalized as in (Lanckriet et al., 2004); thus, $\mathbf{diag}(K_\ell) = \mathbf{1}_m$ and $r_\ell = m$ for each ℓ. We consider two different datasets: a subset of `svmguide1` ($m = 3000$, $n = 4$) and a subset of `MNIST` to classify digits of 4 and 9 ($m = 7500$, $n = 784$) from (Chang and Lin, 2011), where n denotes the number of features. We use 80% of data points as the training set and the rest as the test set.

The algorithms are compared in terms of relative error for the solution ($\|\mathbf{x}^k - \mathbf{x}^*\|_2/\|\mathbf{x}^*\|_2$), where $(\mathbf{x}^*, y^*)$ denotes a saddle point for the considered problem. The purpose of this experiment is to benchmark our method against other methods in terms of convergence behavior observed in practice. To this end, the optimal primal solution $\mathbf{x}^*$ to the problem in (5) is computed calling the commercial optimization solver MOSEK through CVX (Grant et al., 2008).

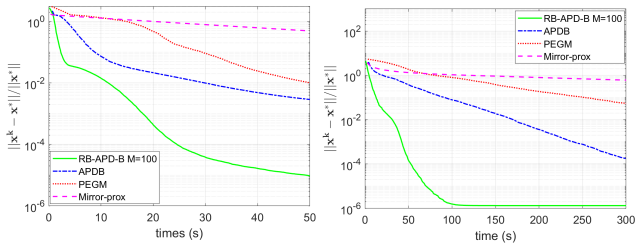
Effect of number of blocks. In Figure 1, we compare the performance of `RB-APD-B` for different numbers of primal blocks, $M \in \{1, 10, 50, 100, 800\}$ for `svmguide1` and $M \in \{1, 10, 50, 100, 1000\}$ for `MNIST` datasets. `RB-APD-B` with $M = 100$ primal blocks has the best performance compared to other block partition strategies. The main reason is that partitioning minimization variable into blocks reduces per iteration complexity and allows more iterations to be performed in a given time interval. That said, employing an excessive number of blocks, e.g., $M = 1000$ for `MNIST` dataset, degrades the overall performance measured by the total number of primal and dual oracle calls, compared to adopting a moderate number of blocks, e.g., $M = 100$ for `MNIST`. Thus, one may conclude that there is a trade-off between the number of blocks and convergence behavior in terms of oracle complexity.



(a) `svmguide1` dataset (b) `MNIST` dataset

Figure 1: Performance of `RB-APD-B` for various number of primal blocks.

Comparison with other methods. The comparison of our method against others is demonstrated in Figure 2. In this experiment, we select the step-size of (He et al., 2015) constant while other methods enjoy a backtracking linesearch. All the methods use the same initial step-size of $\tau = 10^{-2}$ and parameter $\eta = 0.7$. We observe that our algorithm `RB-APD-B` is competitive against the state-of-the-art methods we compare. The reason is that our method potentially benefits from low per-iteration complexity due to using a block-coordinate approach as well as larger block-specific step-sizes due to exploring local coordinate-wise Lipschitz constant via backtracking method. `Mirror-prox` by He et al. (2015) uses constant step-size depending on the global Lipschitz parameter and employs full gradient $\nabla_{\mathbf{x}}\Phi$ to update $\mathbf{x}$ while both `APDB` by Hamedani and Aybat (2021) and `PEGM` by Malitsky and Pock (2018) employ backtracking line search, but similarly use full gradient $\nabla_{\mathbf{x}}\Phi$ to update $\mathbf{x}$.



(a) `svmguide1` dataset (b) `MNIST` dataset

Figure 2: Comparing the performance of `RB-APD-B` with other methods for different datasets.

[5] Since $\underline{\mu} > 0$, $\mathbf{x}^*$ denotes the unique solution to $\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x})$.

5.2 Quadratic constrained quadratic programming (QCQP)

In this subsection, we compare our method against APDB by Hamedani and Aybat (2021) and BLALM by Xu (2021) on randomly generated QCQP problems with various dimensions. In fact, we consider the following QCQP problem

$$\min_{\mathbf{x}\in X} \; f(\mathbf{x}) \triangleq \frac{1}{2}\mathbf{x}^\top A_0\mathbf{x} + b_0^\top \mathbf{x}$$
$$\text{s.t.} \quad g(\mathbf{x}) \triangleq \frac{1}{2}\mathbf{x}^\top A_1\mathbf{x} + b_1^\top \mathbf{x} - c_1 \leq 0,$$

where $X = [-1,1]^m$, $\{A_j\}_{j=0}^1 \subset \mathbb{R}^{m\times m}$ are positive semidefinite matrices generated randomly with a block diagonal structure, $\{b_j\}_{j=0}^1 \subset \mathbb{R}^m$ are generated randomly with elements drawn from standard Gaussian distribution, and $c_1 \in \mathbb{R}$ is generated randomly with elements drawn from uniform distribution over $[0,1]$.

The goal of this experiment is to examine the effect of the dimension of the primal-variable (m) on the runtime of the proposed method and compare it with existing algorithms. To this end, we fix the termination criteria as $\max\{|f(\mathbf{x}_k) - f(\mathbf{x}^*)|, [g(\mathbf{x}_k)]_+\}/m \leq \epsilon$ and increase parameter m to compare the running time of algorithms. In particular, we let $\epsilon = 10^{-6}$ and $m = 10^3 i$ for $i \in \{1,3,5,7,9\}$. The plot is shown in Figure 3 and we observe that RB−APD−B outperforms the other two methods and their gap increases as m increases.

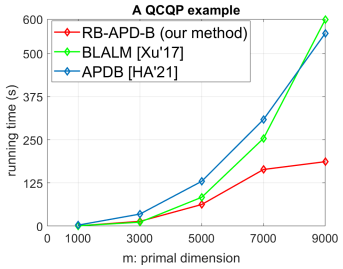


Figure 3: Comparing the effect of m on the running time.

6 Concluding Remarks

The step-sizes for a typical first-order method are selected based on the global Lipschitz constant of the gradient map to guarantee convergence. However, these constants may not be readily available in practice; and even if they are known, they pottentially lead to conservative step-sizes. We developed a novel randomized block coordinate primal-dual method equipped with backtracking line-search to solve large-scale SP problems. The method can contend with non-bilinear, non-separable coupling functions $\Phi(\mathbf{x}, y)$ possibly with multiple primal blocks. At each iteration, a primal block is randomly selected and updated, following a dual update with a momentum term involving $\nabla_y \Phi$.

We showed that for convex-concave setting, the proposed method achieves $\mathcal{O}(M/k)$ convergence rate in the expected primal-dual gap. Furthermore, assuming $\Phi(\mathbf{x}, y)$ is strongly convex in $\mathbf{x}$ and affine in y, our method enjoys a faster rate of $\mathcal{O}(M/k^2)$ in terms of $\mathbf{E}[\|\mathbf{x}_k - \mathbf{x}^*\|^2]$ and $\mathbf{E}[F(\mathbf{x}^k) - F(\mathbf{x}^*)]$, where $F(\cdot)$ denotes the primal function. To the best of our knowledge, our proposed method is the only randomized block-coordinate primal-dual algorithm that can handle the general SP problems as in (1) and our rate results are optimal for this class. Furthermore, our method avoids step-size selection issues related to the use of global Lipschitz constants through employing backtracking line-search scheme we developed. Indeed, the proposed line-search scheme tailored for RB−APD help us alleviate the burden of knowing the global Lipschitz constants by estimating them locally.

References

A. Alacaoglu, V. Cevher, and S. J. Wright. On the complexity of a practical primal-dual coordinate method. *arXiv preprint arXiv:2201.07684*, 2022a.

A. Alacaoglu, O. Fercoq, and V. Cevher. On the convergence of stochastic primal-dual hybrid gradient. *SIAM Journal on Optimization*, 32(2):1288–1318, 2022b.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging application. *arXiv preprint arXiv:1706.04957*, 2017.

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.

C. Dang and G. Lan. Randomized first-order methods for saddle point optimization. *arXiv preprint arXiv:1409.8625*, 2014.

O. Fercoq and P. Bianchi. A coordinate descent primal-dual algorithm with large step size and possibly non separable functions. *arXiv preprint arXiv:1508.04625*, 2015.

X. Gao, Y. Xu, and S. Zhang. Randomized primal-dual proximal block coordinate updates. *arXiv preprint arXiv:1605.05969*, 2016.

M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *Journal of machine learning research*, 12(Jul): 2211–2268, 2011.

M. Grant, S. Boyd, and Y. Ye. Cvx: Matlab software for disciplined convex programming, 2008.

E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

N. He, A. Juditsky, and A. Nemirovski. Mirror prox algorithm for multi-term composite minimization and semiseparable problems. *Computational Optimization and Applications*, 61(2):275–319, 2015.

Y. He and R. D. Monteiro. An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems. *SIAM Journal on Optimization*, 26(1): 29–56, 2016.

A. Jalilzadeh, U. V. Shanbhag, J. H. Blanchet, and P. W. Glynn. Optimal smoothed variable sample-size accelerated proximal methods for structured nonsmooth stochastic convex programs. *arXiv preprint arXiv:1803.00718*, 2018.

A. Juditsky, A. Nemirovski, et al. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, pages 149–183, 2011.

O. Kolossoski and R. Monteiro. An accelerated non-euclidean hybrid proximal extragradient-type algorithm for convex–concave saddle-point problems. *Optimization Methods and Software*, 32(6):1244–1272, 2017.

G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine learning research*, 5 (Jan):27–72, 2004.

Z. Li and M. Yan. New convergence analysis of a primal-dual algorithm with large stepsizes. *Advances in Computational Mathematics*, 47(1):1–20, 2021.

Z.-Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

Y. Malitsky. Proximal extrapolated gradient methods for variational inequalities. *Optimization Methods and Software*, 33(1):140–164, 2018.

Y. Malitsky and T. Pock. A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization*, 28(1): 411–432, 2018.

Y. Malitsky and M. K. Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, pages 2208–2216, 2016.

A. Nemirovski. Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Z. Peng, T. Wu, Y. Xu, M. Yan, and W. Yin. Coordinate friendly structures, algorithms and applications. *arXiv preprint arXiv:1601.00863*, 2016.

P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144 (1-2):1–38, 2014.

H. Robbins and D. Siegmund. *Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971)*, chapter A convergence theorem for non negative almost supermartingales and some applications, pages 233 – 257. New York: Academic Press, 1971.

R. T. Rockafellar. *Convex analysis*. Princeton university press, 2015.

P. K. Shivaswamy and T. Jebara. Ellipsoidal machines. In *Artificial Intelligence and Statistics*, pages 484–491, 2007.

Q. Tran-Dinh and D. Liu. A new randomized primal-dual algorithm for convex optimization with optimal last iterate rates. *arXiv preprint arXiv:2003.01322*, 2020.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Available at `http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf`, 2008.

T. Valkonen. Block-proximal methods with spatially adapted acceleration. *arXiv preprint arXiv:1609.07373*, 2016.

E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528, 2003.

Y. Xu. First-order methods for constrained convex programming based on linearized augmented lagrangian function. *Informs Journal on Optimization*, 3(1):89–117, 2021.

Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications

to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

A. W. Yu, Q. Lin, and T. Yang. Doubly stochastic primal-dual coordinate method for regularized empirical risk minimization with factorized data. *CoRR, abs/1508.03390*, 2015.

X. Zhang, N. S. Aybat, and M. Gürbüzbalaban. Robust accelerated primal-dual methods for computing saddle points. *arXiv preprint arXiv:2111.12743*, 2021.

X. Zhang, N. S. Aybat, and M. Gurbuzbalaban. Sapd+: An accelerated stochastic method for nonconvex-concave minimax problems. *arXiv preprint arXiv:2205.15084, accepted to NeurIPS 2022*, 2022.

Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

W. Zhong and J. Kwok. Fast stochastic alternating direction method of multipliers. In *International Conference on Machine Learning*, pages 46–54, 2014.

Z. Zhu and A. J. Storkey. Adaptive stochastic primal-dual coordinate descent for separable saddle point problems. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 645–658. Springer, 2015.

A SUPPORTING LEMMAS AND DEFINITIONS

Notation: In the rest, $\mathbf{E}[\cdot]$ denotes the expectation operation and the conditional expectation is denoted by $\mathbf{E}^k[\cdot] \triangleq \mathbf{E}[\cdot \mid \mathcal{F}_k]$, where $\mathcal{F}_k \triangleq \sigma\big(\{i_0, \ldots, i_{k-1}\}\big)$ is the σ-algebra generated by $\{i_0, \ldots, i_{k-1}\}$ for $k \geq 1$. Moreover, given a diagonal matrix $\mathcal{A} = \mathbf{diag}([a_i]_{i\in\mathcal{M}})$ for some $\{a_i\}_{i\in\mathcal{M}} \subset \mathbb{R}_{++}$, for any $\mathbf{x} \in \mathcal{X}$ and $\bar{\mathbf{x}} \in \mathbf{dom}\, f$, we define

$$\mathbf{D}_{\mathcal{X}}^{\mathcal{A}}(\mathbf{x}, \bar{\mathbf{x}}) \triangleq \sum_{i\in\mathcal{M}} a_i \mathbf{D}_{\mathcal{X}_i}(x_i, \bar{x}_i).$$

Moreover, for any $\boldsymbol{\delta} \in \mathcal{X}^*$, we define $\|\boldsymbol{\delta}\|_{*,\mathcal{A}} \triangleq \sqrt{\sum_{i\in\mathcal{M}} a_i \|\delta_i\|_{\mathcal{X}_i^*}^2}$.

Note that for any $y \in \mathbf{dom}\, h$, $\bar{\mathbf{x}} \in \mathbf{dom}\, f$ and $i \in \mathcal{M}$, (8) implies that

$$0 \leq \Phi(\bar{\mathbf{x}} + U_i v, y) - \Phi(\bar{\mathbf{x}}, y) - \langle \nabla_{x_i}\Phi(\bar{\mathbf{x}}, y),\ v\rangle \leq \tfrac{1}{2} L_{x_i x_i} \|v\|_{\mathcal{X}_i}^2, \tag{19}$$

for all $v \in \mathcal{X}_i$ such that $\bar{\mathbf{x}} + U_i v \in \mathbf{dom}\, f$. Similarly (9) and concavity of $\Phi(\mathbf{x}, \cdot)$ imply that for any $\mathbf{x} \in \mathbf{dom}\, f$, the following inequality holds for all $y, \bar{y} \in \mathbf{dom}\, h$:

$$0 \geq \Phi(\mathbf{x}, y) - \Phi(\mathbf{x}, \bar{y}) - \langle \nabla_y \Phi(\mathbf{x}, \bar{y}),\ y - \bar{y}\rangle \geq -\tfrac{1}{2} L_{yy} \|y - \bar{y}\|_{\mathcal{Y}}^2. \tag{20}$$

Next, we define a test function $C^k(\mathbf{x}, \mathbf{y})$ for the iteration $k \geq 0$ to accept or reject the given point $(\mathbf{x}, y)$.

Definition 4. *For any $k \geq 0$, given $\tilde{\tau}^k, \sigma^k, \theta^k > 0$, $\mathbf{T}^k = \mathbf{diag}\left(\left[\frac{1}{\tau_i^k}\right]_{i\in\mathcal{M}}\right)$ such that $\tau_i^k \triangleq \left(\frac{1}{M}(\mu_i + \frac{1}{\tilde{\tau}^k}) - \mu_i\right)^{-1}$ for $i \in \mathcal{M}$. We define*

$$\begin{aligned} C^k(\mathbf{x}, y) \triangleq & M\Big(\Phi(\mathbf{x}, y) - \Phi(\mathbf{x}^k, y) - \left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y),\ \mathbf{x} - \mathbf{x}^k\right\rangle\Big) + \frac{M}{2\alpha^{k+1}} \left\|\nabla_y \Phi(\mathbf{x}, y) - \nabla_y \Phi(\mathbf{x}^k, y)\right\|_{\mathcal{Y}^*}^2 \\ & + \frac{M}{2\beta^{k+1}} \left\|\nabla_y \Phi(\mathbf{x}^k, y) - \nabla_y \Phi(\mathbf{x}^k, y^k)\right\|_{\mathcal{Y}^*}^2 - M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \mathbf{x}^k) - \Big(\frac{1}{\sigma^k} - \theta^k M(\alpha^k + \beta^k)\Big)\mathbf{D}_{\mathcal{Y}}(y, y^k) \end{aligned} \tag{21}$$

for some positive sequence $\{\alpha^k, \beta^k\}$.

Indeed, one can easily verify that $C_*^k = C^k(\mathbf{x}^{k+1}, y^{k+1})$ for $\{\alpha^k, \beta^k\}$ sequence defined as in `RB-APD-B` algorithm.

Lemma 2. *Let $\mathcal{X}$ be a finite dimensional normed vector space with norm $\|.\|_{\mathcal{X}}$, $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ be a closed convex function with convexity modulus $\mu \geq 0$ with respect to $\|.\|_{\mathcal{X}}$, and $\mathbf{D} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a Bregman distance function corresponding to a strictly convex function $\phi : \mathcal{X} \to \mathbb{R}$ that is differentiable on an open set containing $\mathbf{dom}\, f$. Given $\bar{x} \in \mathbf{dom}\, f$ and $t > 0$, let*

$$x^+ = \operatorname*{argmin}_{x\in\mathcal{X}} f(x) + t\mathbf{D}(x, \bar{x}). \tag{22}$$

Then for all $x \in \mathcal{X}$, the following inequality holds:

$$f(x) + t\mathbf{D}(x, \bar{x}) \geq f(x^+) + t\mathbf{D}(x^+, \bar{x}) + t\mathbf{D}(x, x^+) + \frac{\mu}{2}\left\|x - x^+\right\|_{\mathcal{X}}^2. \tag{23}$$

Proof. This result is a trivial extension of Property 1 in Tseng (2008). The first-order optimality condition for (22) implies that $0 \in \partial f(x^+) + t\nabla_x \mathbf{D}(x^+, \bar{x})$ – where $\nabla_x \mathbf{D}$ denotes the partial gradient with respect to the first argument. Note that for any $x \in \mathbf{dom}\, f$, we have $\nabla_x \mathbf{D}(x, \bar{x}) = \nabla\phi(x) - \nabla\phi(\bar{x})$. Hence, $t(\nabla\phi(\bar{x}) - \nabla\phi(x^+)) \in \partial f(x^+)$. Using the convexity inequality for f, we get

$$f(x) \geq f(x^+) + t\left\langle \nabla\phi(\bar{x}) - \nabla\phi(x^+),\ x - x^+\right\rangle + \frac{\mu}{2}\left\|x - x^+\right\|_{\mathcal{X}}^2.$$

The result in (23) immediately follows from this inequality. □

Lemma 3. *Robbins and Siegmund (1971) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and for each $k \geq 0$ suppose a^k and b^k are finite nonnegative $\mathcal{F}^k$-measurable random variables where $\{\mathcal{F}^k\}_{k\geq 0}$ is a sequence sub-σ-algebras of $\mathcal{F}$ such that $\mathcal{F}^k \subset \mathcal{F}^{k+1}$ for $k \geq 0$. If $\mathbb{E}[a^{k+1}|\mathcal{F}^k] \leq a^k - b^k$, then then $a = \lim_{k\to\infty} a^k$ exists almost surely, and $\sum_{k=0}^{\infty} b^k < \infty$.*

Lemma 4. *Given a diagonal matrix $\mathcal{A} = \mathbf{diag}([a_i]_{i\in\mathcal{M}})$ for some $\{a_i\}_{i\in\mathcal{M}} \subset \mathbb{R}_{++}$, and an arbitrary sequence $\{\boldsymbol{\delta}^k\}_{k\geq 0} \subset \mathcal{X}_*$, let $\{\mathbf{v}^k\}_{k\geq 0} \subset \mathcal{X}$ be such that $\mathbf{v}^{k+1} \triangleq \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} \{-\left\langle \boldsymbol{\delta}^k, \mathbf{x}\right\rangle + \mathbf{D}^{\mathcal{A}}_{\mathcal{X}}(\mathbf{x}, \mathbf{v}^k)\}$. Then for all $k \geq 0$ and $\mathbf{x} \in \mathcal{X}$,*

$$\left\langle \boldsymbol{\delta}^k, \mathbf{x} - \mathbf{v}^k\right\rangle \leq \mathbf{D}^{\mathcal{A}}_{\mathcal{X}}(\mathbf{x}, \mathbf{v}^k) - \mathbf{D}^{\mathcal{A}}_{\mathcal{X}}(\mathbf{x}, \mathbf{v}^{k+1}) + \frac{1}{2}\left\|\boldsymbol{\delta}^k\right\|^2_{*,\mathcal{A}^{-1}}.$$

Proof. Since $\mathbf{v}^{k+1}$ computation is separable in $i \in \mathcal{M}$, one can apply Lemma 2 for each coordinate to obtain a bound for $\left\langle \boldsymbol{\delta}^k, \mathbf{x} - \mathbf{v}^{k+1}\right\rangle$. Then we have that

$$\begin{aligned}\left\langle \boldsymbol{\delta}^k, \mathbf{x} - \mathbf{v}^k\right\rangle &= \left\langle \boldsymbol{\delta}^k, \mathbf{x} - \mathbf{v}^{k+1}\right\rangle + \left\langle \boldsymbol{\delta}^k, \mathbf{v}^{k+1} - \mathbf{v}^k\right\rangle\\ &\leq \mathbf{D}^{\mathcal{A}}_{\mathcal{X}}(\mathbf{x}, \mathbf{v}^k) - \mathbf{D}^{\mathcal{A}}_{\mathcal{X}}(\mathbf{x}, \mathbf{v}^{k+1}) - \mathbf{D}^{\mathcal{A}}_{\mathcal{X}}(\mathbf{v}^{k+1}, \mathbf{v}^k) + \left\langle \boldsymbol{\delta}^k, \mathbf{v}^{k+1} - \mathbf{v}^k\right\rangle\\ &\leq \mathbf{D}^{\mathcal{A}}_{\mathcal{X}}(\mathbf{x}, \mathbf{v}^k) - \mathbf{D}^{\mathcal{A}}_{\mathcal{X}}(\mathbf{x}, \mathbf{v}^{k+1}) + \frac{1}{2}\left\|\boldsymbol{\delta}^k\right\|^2_{*,\mathcal{A}^{-1}},\end{aligned}$$

where in the last inequity we used $\left\langle \delta_i^k, v_i^{k+1} - v_i^k\right\rangle \leq \frac{a_i}{2}\left\|v_i^{k+1} - v_i^k\right\|^2_{\mathcal{X}_i} + \frac{1}{2a_i}\left\|\delta_i^k\right\|^2_{\mathcal{X}_i^*}$ for $i \in \mathcal{M}$ together with $\mathbf{D}^{\mathcal{A}}_{\mathcal{X}}(\mathbf{v}^{k+1} - \mathbf{v}^k) \geq \frac{1}{2}\left\|\mathbf{v}^{k+1} - \mathbf{v}^k\right\|^2_{\mathcal{A}}$. □

B Proof of Convergence

Before proving the asymptotic convergence of the iterate sequence and related rate results, we restate the theorems with more details here for completeness. We divide the proof of theorems in three parts: (i) In section B.1, we analyze the proposed backtracking method to derive some key inequalities; (ii) in section B.2, we show asymptotic convergence of the iterate sequence generated by `RB-APD` and `RB-APD-B`; (iii) finally, in section B.3 we establish the convergence rate of the proposed methods.

Remark B.1. *Note that when the problem is strongly convex, we assume that $L_{yy} = 0$ which means that $\Phi(\mathbf{x}, \cdot)$ is affine for any $\mathbf{x} \in \mathcal{X}$. In this scenario, let $\mathbf{x}^*$ be the unique solution to $\min\{F(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ for F defined in (6), it is desirable to provide convergence guarantees for computing an ϵ-optimal point $\mathbf{x}_\epsilon$. In this context, ϵ-optimality can be defined either in function values, i.e., $\mathbf{E}[F(\mathbf{x}_\epsilon) - F(\mathbf{x}^*)] \leq \epsilon$, or in the solution space, i.e., $\mathbf{E}[\|\mathbf{x}_\epsilon - \mathbf{x}^*\|^2_{\mathcal{X}}] \leq \epsilon$.*

We begin by restating the results for `RB-APD-B`, stated in Algorithm 2, under merely convex and strongly convex settings separately.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Let $\delta \in [0, 1)$, $c_\alpha > 0$ and $c_\beta \geq 0$ are chosen as stated below, and define $\underline{\mu} = \min_{i\in\mathcal{M}} \mu_i$, $\bar{\mu} \triangleq \max_{i\in\mathcal{M}} \mu_i$, $\overline{L}_{xx} = \max_{i\in\mathcal{M}} L_{x_ix_i}$, $\overline{L}_{yx} = \max_{i\in\mathcal{M}} L_{yx_i}$, and*

$$\Psi_1 \triangleq \frac{c_\alpha \bar{b}}{2\gamma_0 \overline{L}^2_{yx}}\zeta, \quad \Psi_2 \triangleq \frac{\sqrt{c_\beta(1 - (M(c_\alpha + c_\beta) + \delta))}}{\gamma_0\sqrt{M}L_{yy}}, \tag{24}$$

$$\zeta \triangleq -1 + \sqrt{1 + \frac{4(1-\delta)\gamma_0}{Mc_\alpha}\frac{\overline{L}^2_{yx}}{\bar{b}^2}}, \quad \bar{b} \triangleq \overline{L}_{xx} + \frac{(1-\delta)(M-1)\bar{\mu}}{M}. \tag{25}$$

For any given $(\mathbf{x}_0, y_0) \in \mathbf{dom}\, f \times \mathbf{dom}\, h$ and arbitrary $\gamma^0 > 0$, $\bar{\tau} \in \left(0, \frac{1}{\bar{\mu}(M-1)}\right)$, `RB-APD-B`, stated in Algorithm 2, is well-defined, i.e., the number of backtracking iterations is finite and bounded by $1 + \log_{1/\eta}(\frac{\bar{\tau}}{\Psi})$ uniformly for $k \geq 0$ for some $\Psi > 0$. Let $\{\mathbf{x}^k, y^k\}_{k\geq 0}$ denote the iterate sequence generated by `RB-APD-B`. For all $K \geq 1$, let $T^K \triangleq \sum_{k=0}^{K-1} t^k$ and

$$\bar{\mathbf{x}}^K = \frac{1}{T^K + M - 1}\Big(\sum_{k=0}^{K-2}(Mt^k - (M-1)t^{k+1})\mathbf{x}^{k+1} + Mt^{K-1}\mathbf{x}^K\Big), \quad \bar{y}^K = \frac{1}{T^K}\sum_{k=0}^{K-1} t^k y^{k+1}$$

for $\{t^k\}_{k\geq 0}$ such that $t^k = \sigma^k/\sigma^0$ for $k \geq 0$.
(Part I.) *Suppose $\underline{\mu} = 0$ and $\mathbf{dom}\, f \times \mathbf{dom}\, h$ is compact. Assume $M(c_\alpha + c_\beta) + \delta \leq 1$ holds for some $c_\alpha, c_\beta > 0$ if $L_{yy} > 0$; and $c_\beta = 0$, and $Mc_\alpha + \delta \leq 1$ for some $c_\alpha > 0$ otherwise. For this setting, $\Psi = \Psi_1$ if $L_{yy} = 0$ and $\Psi = \min\{\Psi_1, \Psi_2\}$ if*

$L_{yy} > 0$; moreover, if a saddle point for (1) *exists and $\delta > 0$ is chosen, then $\{(\mathbf{x}^k, y^k)\}_{k\geq 0}$ converges to a saddle point almost surely. Finally, for all $K \geq 1$, the following bounds holds:*

$$\mathbf{E}\left[\mathcal{G}(\bar{\mathbf{x}}^K, \bar{y}^K)\right] \leq \frac{1}{T^K}\left(\bar{B}_1 + \bar{B}_2 + \sup_{\mathbf{x}\in\mathbf{dom}\, f} B_1(\mathbf{x}) + \sup_{y\in\mathbf{dom}\, h} B_2(y)\right) \tag{26a}$$

$$\bar{B}_1 \triangleq \frac{(M-1)L_{\varphi_{\mathcal{X}}}^2}{2}\,\mathbf{E}\Big[\sum_{k=0}^{+\infty} t^k \left\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\right\|^2_{\mathbf{T}^k+\mathfrak{M}}\Big] < +\infty, \tag{26b}$$

$$\bar{B}_2 \triangleq (M-1)\sup\{\mathcal{L}(\mathbf{x}^0, y) - \mathcal{L}(\mathbf{x}, y) :\ (\mathbf{x}, y) \in \mathbf{dom}\, f \times \mathbf{dom}\, h\} < +\infty, \tag{26c}$$

$$B_1(\mathbf{x}) \triangleq M\mathbf{D}_{\mathcal{X}}^{(1+\frac{1}{M})\mathbf{T}^0+\mathfrak{M}}(\mathbf{x}, \mathbf{x}^0), \tag{26d}$$

$$B_2(y) \triangleq \Big(\tfrac{1}{\sigma^0} + \theta^0(M-1)L_{yy}\Big)\mathbf{D}_{\mathcal{Y}}(y, y^0), \tag{26e}$$

and $T^K = \Omega(K)$, implying $\mathcal{O}(1/K)$ sublinear rate for $\mathbf{E}\left[\mathcal{G}(\bar{\mathbf{x}}^K, \bar{y}^K)\right]$.

(Part II.) *Suppose $\underline{\mu} > 0$ and $L_{yy} = 0$; hence, Φ has the following form: $\Phi(\mathbf{x}, y) = \langle g(\mathbf{x}), y\rangle$ for some $g : \mathcal{X} \to \mathcal{Y}^*$ such that Φ is convex in $\mathbf{x}$ on $\mathbf{dom}\, f$ for any $y \in \mathbf{dom}\, h$. Let $F(\mathbf{x}) = \operatorname{argmax}_{y\in\mathcal{Y}} \mathcal{L}(\mathbf{x}, y)$; thus, $F(\mathbf{x}) = f(\mathbf{x}) + h^*(g(\mathbf{x}))$ for $\mathbf{x} \in \mathcal{X}$.*

Assume $Mc_\alpha + \delta \in (0, 1]$ and $c_\beta = 0$. For this setting, $\Psi = \Psi_1$. If a saddle point $(\mathbf{x}^*, y^*)$ for (1) *exists, then $\{(\mathbf{x}^k, y^k)\}_{k\geq 0}$ converges to $\mathbf{x}^*$ and $\{y^k\}$ has a limit point almost surely. Moreover, if $\delta > 0$, then any limit point $(\mathbf{x}^*, y^*)$ is a saddle point almost surely satisfying*

$$\begin{aligned}\mathbf{E}\left[\tfrac{\gamma^K}{2}\left\|\mathbf{x}^* - \mathbf{x}^K\right\|^2_{\mathcal{X}} + (1-Mc_\alpha)\mathbf{D}_{\mathcal{Y}}(y^*, y^K)\right] \leq& \tfrac{\gamma^0}{2}\left\|\mathbf{x}^* - \mathbf{x}^0\right\|^2_{\mathcal{X}} + \mathbf{D}_{\mathcal{Y}}(y^*, y^0)\\ &+ \sigma^0(M-1)\Big(\mathcal{L}(\mathbf{x}^0, y^*) - \mathcal{L}(\mathbf{x}^*, y^*)\Big),\end{aligned} \tag{27}$$

$$\mathbf{E}\Big[F(\bar{\mathbf{x}}^K) - F(\mathbf{x}^*)\Big] \leq \frac{1}{T^K}\Big(\tfrac{\gamma^0}{2\sigma^0}\left\|\mathbf{x}^* - \mathbf{x}^0\right\|^2_{\mathcal{X}} + \tfrac{1}{\sigma^0}\sup_{y\in\mathbf{dom}\, h}\mathbf{D}_{\mathcal{Y}}(y, y^0) + (M-1)\big(F(\mathbf{x}^0) - F(\mathbf{x}^*)\big)\Big), \tag{28}$$

for all $K \geq 1$. Furthermore, both $\gamma^K = \Omega(K^2)$ and $T^K = \Omega(K^2)$; hence, $\mathbf{E}\big[\left\|\mathbf{x}^K - \mathbf{x}^*\right\|^2_{\mathcal{X}}\big] = \mathcal{O}(1/K^2)$ and $0 \leq \mathbf{E}\left[F(\bar{\mathbf{x}}^K) - F(\mathbf{x}^*)\right] \leq \mathcal{O}(1/K^2)$.

Next, we state the convergence rate result for `RB-APD`, displayed in Algorithm 1, in detail.

Theorem 2. *Suppose Assumptions 1 and 2 hold, and let $\underline{\mu} = \min_{i\in\mathcal{M}}\mu_i$, $\bar{\mu} \triangleq \max_{i\in\mathcal{M}}\mu_i$. Given some $\gamma^0 > 0$ and $\bar{\tau} \in \left(0, \frac{1}{\bar{\mu}(M-1)}\right)$ such that* (14) *holds for some $c_\alpha, c_\beta \geq 0$ and $\delta \in [0, 1)$ as described in Theorem 1, let $\{\mathbf{x}^k, y^k\}_{k\geq 0}$ be the iterate sequence generated by `RB-APD`. Suppose $(\bar{\mathbf{x}}^K, \bar{y}^K)$ and T^K are defined for $K \geq 1$ as in Theorem 1.*

(Part I.) *Suppose $\underline{\mu} = 0$ and $\mathbf{dom}\, f \times \mathbf{dom}\, h$ is compact. The stepsize rule in Algorithm 1 implies that $\tilde{\tau}^k = \tilde{\tau}^0$, $\sigma^k = \sigma^0$ and $\theta^k = 1$ for all $k \geq 0$; hence, $t^k = 1$ for $k \geq 0$, implying $T^K = K$. Moreover,* (26) *holds for all $K \geq 1$. Finally, if a saddle point for* (1) *exists and $\delta > 0$, then $\{(\mathbf{x}^k, y^k)\}_{k\geq 0}$ almost surely converges to a saddle point $(\mathbf{x}^*, y^*)$.*

(Part II.) *Suppose $\underline{\mu} > 0$ and $L_{yy} = 0$. If a saddle point for* (1) *exists, then $\{\mathbf{x}^k\}_{k\geq 0}$ converges to $\mathbf{x}^*$ and $\{y^k\}$ has a limit point almost surely,[6] where $\mathbf{x}^*$ denotes the unique solution to $\min\{F(\mathbf{x}) :\ \mathbf{x} \in \mathcal{X}\}$. Moreover, if $\delta > 0$, then any limit point $(\mathbf{x}^*, y^*)$ is a saddle point almost surely satisfying* (27) *and* (28) *for all $K \geq 1$. Furthermore, as in Theorem 1, both $\gamma^K = \Omega(K^2)$ and $T^K = \Omega(K^2)$; hence, $\mathbf{E}\big[\left\|\mathbf{x}^K - \mathbf{x}^*\right\|^2_{\mathcal{X}}\big] = \mathcal{O}(1/K^2)$ and $0 \leq \mathbf{E}\left[F(\bar{\mathbf{x}}^K) - F(\mathbf{x}^*)\right] \leq \mathcal{O}(1/K^2)$.*

To prove the results of Theorems 1 and 2, we first provide a one-step analysis in Lemma 5 to provide a bound on the progress of iterates in terms of the coupling function $\mathcal{L}$. This is the main building block of our convergence analysis.

[6] Since $\underline{\mu} > 0$, $\mathbf{x}^*$ must be the unique $\mathbf{x}$-coordinate of any saddle point.

Lemma 5. *Suppose Assumption 1 holds, let $\{\mathbf{x}^k, y^k\}_{k\geq 0}$ be generated by the following recursion:*

$$q^k \leftarrow M(\nabla_y \Phi(\mathbf{x}^k, y^k) - \nabla_y \Phi(\mathbf{x}^{k-1}, y^{k-1})) \tag{29a}$$

$$s^k \leftarrow \nabla_y \Phi(\mathbf{x}^k, y^k) + \theta^k q^k \tag{29b}$$

$$y^{k+1} \leftarrow \operatorname*{argmin}_y h(y) - \langle s^k, y\rangle + \frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y, y^k) \tag{29c}$$

$$\tilde{x}_i^{k+1} \leftarrow \operatorname*{argmin}_x f_i(x) + \langle \nabla_{x_i}\Phi(\mathbf{x}^k, y^{k+1}), x\rangle + \frac{1}{\tau_i^k}\mathbf{D}_{\mathcal{X}_i}(x, x_i^k), \quad \forall\, i \in \mathcal{M} \tag{29d}$$

$$\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k$$

Choose $i_k \in \mathcal{M}$ uniformly at random

$$x_{i_k}^{k+1} \leftarrow \tilde{x}_{i_k}^{k+1},$$

for some positive parameters $\{\tau_i^k\}_{i\in\mathcal{M}}$, σ^k and θ^k for $k \geq 0$.[7] *Then for any $(\mathbf{x}, y) \in \mathbf{dom}\, f \times \mathbf{dom}\, h$,*

$$\begin{aligned}\mathcal{L}(\mathbf{x}^{k+1}, y) - \mathcal{L}(\mathbf{x}, y^{k+1}) &\leq Q^k(\mathbf{z}) - R^{k+1}(\mathbf{z}) + (M-1)(\theta^k H^k(\mathbf{z}) - H^{k+1}(\mathbf{z})),\\ &+ (M-1)\Big((1-\theta^k)\big(\mathcal{L}(\mathbf{x}^k, y) - \mathcal{L}(\mathbf{x}, y)\big) + (1-\theta^k)_+ L_{yy}\mathbf{D}_{\mathcal{Y}}(y, y^k)\Big) + C^k(\mathbf{x}^{k+1}, y^{k+1}) + \mathcal{E}^k(\mathbf{x}),\end{aligned} \tag{30}$$

holds for all $k \geq 0$ for any $\{\alpha^k\} \subset \mathbb{R}_{++}$, and for any $\{\beta^k\} \subset \mathbb{R}_{++}$ when $L_{yy} > 0$ (and $\beta^k = 0$ for all $k \geq 0$ when $L_{yy} = 0$, defining $0^2/0 = 0$), where $(\cdot)_+ = \max\{\cdot, 0\}$, $Q^k(\cdot)$, $R^{k+1}(\cdot)$, $H^k(\cdot)$, and $\mathcal{E}^k(\mathbf{x})$ are defined as follows:

$$\begin{aligned}Q^k(\mathbf{z}) &\triangleq M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \mathbf{x}^k) + \frac{M-1}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|_{\mathfrak{M}}^2 + \frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y, y^k) + \theta^k\left\langle r^k, y^k - y\right\rangle\\ &+ \frac{M\theta^k}{2\alpha^k}\left\|\nabla_y\Phi(\mathbf{x}^k, y^k) - \nabla_y\Phi(\mathbf{x}^{k-1}, y^k)\right\|_{\mathcal{Y}^*}^2 + \frac{M\theta^k}{2\beta^k}\left\|\nabla_y\Phi(\mathbf{x}^{k-1}, y^k) - \nabla_y\Phi(\mathbf{x}^{k-1}, y^{k-1})\right\|_{\mathcal{Y}^*}^2,\end{aligned} \tag{31a}$$

$$\begin{aligned}R^{k+1}(\mathbf{z}) &\triangleq M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \mathbf{x}^{k+1}) + \frac{M}{2}\left\|\mathbf{x} - \mathbf{x}^{k+1}\right\|_{\mathfrak{M}}^2 + \frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y, y^{k+1}) + \left\langle r^{k+1}, y^{k+1} - y\right\rangle\\ &+ \frac{M}{2\alpha^{k+1}}\left\|\nabla_y\Phi(\mathbf{x}^{k+1}, y^{k+1}) - \nabla_y\Phi(\mathbf{x}^k, y^{k+1})\right\|_{\mathcal{Y}^*}^2 + \frac{M}{2\beta^{k+1}}\left\|\nabla_y\Phi(\mathbf{x}^k, y^{k+1}) - \nabla_y\Phi(\mathbf{x}^k, y^k)\right\|_{\mathcal{Y}^*}^2,\end{aligned} \tag{31b}$$

$$H^k(\mathbf{z}) \triangleq f(\mathbf{x}^k) - f(\mathbf{x}) + \Phi(\mathbf{x}^k, y^k) - \Phi(\mathbf{x}, y), \tag{31c}$$

$$\begin{aligned}\mathcal{E}^k(\mathbf{x}) &\triangleq Mf(\mathbf{x}^{k+1}) - f(\tilde{\mathbf{x}}^{k+1}) - (M-1)f(\mathbf{x}^k) - \left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y^{k+1}), \tilde{\mathbf{x}}^{k+1} - M\mathbf{x}^{k+1} + (M-1)\mathbf{x}^k\right\rangle\\ &+ M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \mathbf{x}^{k+1}) - (M-1)\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \mathbf{x}^k) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \tilde{\mathbf{x}}^{k+1}) + M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}^{k+1}, \mathbf{x}^k)\\ &- \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k) + \frac{M}{2}\left\|\mathbf{x} - \mathbf{x}^{k+1}\right\|_{\mathfrak{M}}^2 - \frac{1}{2}\left\|\mathbf{x} - \tilde{\mathbf{x}}^{k+1}\right\|_{\mathfrak{M}}^2 - \frac{M-1}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|_{\mathfrak{M}}^2,\end{aligned} \tag{31d}$$

where $C^k(\cdot,\cdot)$ defined in (21) *and $r^k \triangleq \nabla_y\Phi(\mathbf{x}^k, y^k) - M\nabla_y\Phi(\mathbf{x}^{k-1}, y^{k-1})$.*

Proof. We define an auxiliary sequence $\{\tilde{\mathbf{x}}^k\}_{k\geq 1} \subseteq \mathcal{X}$ such that $\tilde{x}_i^{k+1}$ is defined as in (29d) for all $k \geq 0$. The auxiliary sequence $\{\tilde{\mathbf{x}}^k\}_{k\geq 1}$ is never actually computed in the implementation of `RB-APD` or `RB-APD-B`, and it is defined for analyzing the convergence behavior of $\{\mathbf{x}^k\}_{k\geq 1} \subseteq \mathcal{X}$. For $k \geq 0$, we apply Lemma 2 for both $\tilde{x}_i$-subproblem in (29d) and the y-subproblem in (29c) (see Line 9 of `RB-APD` and also Line 11 of `RB-APD-B`), we obtain two inequalities holding for any $y \in \mathcal{Y}$ and $\mathbf{x} \in \mathcal{X}$:

$$h(y^{k+1}) - \langle s^k,\ y^{k+1} - y\rangle \leq h(y) + \frac{1}{\sigma^k}\Big[\mathbf{D}_{\mathcal{Y}}(y, y^k) - \mathbf{D}_{\mathcal{Y}}(y, y^{k+1}) - \mathbf{D}_{\mathcal{Y}}(y^{k+1}, y^k)\Big], \tag{32}$$

$$\begin{aligned}&f_i(\tilde{x}_i^{k+1}) + \langle \nabla_{x_i}\Phi(\mathbf{x}^k, y^{k+1}), \tilde{x}_i^{k+1} - x_i\rangle + \frac{\mu_i}{2}\|x_i - \tilde{x}_i^{k+1}\|_{\mathcal{X}_i}^2\\ &\quad\leq f_i(x) + \frac{1}{\tau_i^k}\Big[\mathbf{D}_{\mathcal{X}_i}(x_i, x_i^k) - \mathbf{D}_{\mathcal{X}_i}(x_i, \tilde{x}_i^{k+1}) - \mathbf{D}_{\mathcal{X}_i}(\tilde{x}_i^{k+1}, x_i^k)\Big], \quad \forall i \in \mathcal{M},\end{aligned} \tag{33}$$

where q^k and s^k are defined in (29a) and (29b), respectively. Note that by invoking (19), we may bound the inner product in (33) as follows:

$$\begin{aligned}\left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y^{k+1}), \tilde{\mathbf{x}}^{k+1} - \mathbf{x}\right\rangle &= \left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y^{k+1}), \mathbf{x}^k - \mathbf{x}\right\rangle + \left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y^{k+1}), \tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\right\rangle\\ &\geq \Phi(\mathbf{x}^k, y^{k+1}) - \Phi(\mathbf{x}, y^{k+1}) + \left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y^{k+1}), \tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\right\rangle.\end{aligned} \tag{34}$$

[7] `RB-APD` and `RB-APD-B`, stated in Algorithm 1 and Algorithm 2, respectively, both satisfy this recursion for some primal-dual stepsize sequences.

Next, we define two auxiliary sequences for $k \geq 0$:

$$A^{k+1} \triangleq \frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y, y^k) - \frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y, y^{k+1}) - \frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y^{k+1}, y^k), \tag{35a}$$

$$B^{k+1} \triangleq \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \mathbf{x}^k) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \tilde{\mathbf{x}}^{k+1}) - \tfrac{1}{2}\left\|\mathbf{x} - \tilde{\mathbf{x}}^{k+1}\right\|_{\mathfrak{M}}^2 - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k). \tag{35b}$$

Summing (33) over $i \in \mathcal{M}$, and using (34) and (35b) leads to

$$f(\tilde{\mathbf{x}}^{k+1}) \leq f(\mathbf{x}) + \Phi(\mathbf{x}, y^{k+1}) - \Phi(\mathbf{x}^k, y^{k+1}) + B^{k+1} - \left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y^{k+1}), \tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k \right\rangle. \tag{36}$$

For $k \geq 0$, let

$$\Lambda_x^k \triangleq \Phi(\mathbf{x}^{k+1}, y^{k+1}) - \Phi(\mathbf{x}^k, y^{k+1}) - \left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y^{k+1}), \mathbf{x}^{k+1} - \mathbf{x}^k \right\rangle,$$

$$E^k \triangleq f(\mathbf{x}^{k+1}) - \frac{1}{M}f(\tilde{\mathbf{x}}^{k+1}) - (1 - \frac{1}{M})f(\mathbf{x}^k) - \frac{1}{M}\left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y^{k+1}), \tilde{\mathbf{x}}^{k+1} - M\mathbf{x}^{k+1} + (M-1)\mathbf{x}^k \right\rangle,$$

then dividing both sides of (36) by M and rearranging the terms lead to

$$\begin{aligned} &f(\mathbf{x}^{k+1}) + \Phi(\mathbf{x}^{k+1}, y^{k+1}) - f(\mathbf{x}) - \frac{1}{M}\Phi(\mathbf{x}, y^{k+1}) \leq \\ &\left(1 - \frac{1}{M}\right)\left(f(\mathbf{x}^k) - f(\mathbf{x}) + \Phi(\mathbf{x}^k, y^{k+1})\right) + \frac{1}{M}B^{k+1} + \Lambda_x^k + E^k. \end{aligned} \tag{37}$$

Next, multiplying (32) by $\frac{1}{M}$ and adding to (37), then adding $\Phi(\mathbf{x}^{k+1}, y)/M$ to both sides, and rearranging the terms we obtain:

$$\begin{aligned} &\frac{1}{M}\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}) + \Phi(\mathbf{x}^{k+1}, y)\right) + \frac{1}{M}\left(h(y^{k+1}) - h(y) - \Phi(\mathbf{x}, y^{k+1})\right) \\ &\leq \frac{1}{M}\left(\Phi(\mathbf{x}^{k+1}, y) - \Phi(\mathbf{x}^{k+1}, y^{k+1})\right) + \left(1 - \frac{1}{M}\right)\left(f(\mathbf{x}^k) - f(\mathbf{x}) + \Phi(\mathbf{x}^k, y^{k+1})\right) \\ &\quad - \left(1 - \frac{1}{M}\right)\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}) + \Phi(\mathbf{x}^{k+1}, y^{k+1})\right) + \frac{1}{M}B^{k+1} + \frac{1}{M}A^{k+1} \\ &\quad + \frac{1}{M}\left\langle s^k, y^{k+1} - y \right\rangle + \Lambda_x^k + E^k \\ &\leq \frac{1}{M}\left\langle \nabla_y\Phi(\mathbf{x}^{k+1}, y^{k+1}), y - y^{k+1} \right\rangle + \underbrace{\left(1 - \frac{1}{M}\right)\left(f(\mathbf{x}^k) - f(\mathbf{x}) + \Phi(\mathbf{x}^k, y^{k+1}) - \Phi(\mathbf{x}, y)\right)}_{(*)} \\ &\quad - \left(1 - \frac{1}{M}\right)\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}) + \Phi(\mathbf{x}^{k+1}, y^{k+1}) - \Phi(\mathbf{x}, y)\right) + \frac{1}{M}B^{k+1} + \frac{1}{M}A^{k+1} \\ &\quad + \frac{1}{M}\left\langle s^k, y^{k+1} - y \right\rangle + \Lambda_x^k + E^k, \end{aligned} \tag{38}$$

where in the last inequality, we use the concavity of $\Phi(\mathbf{x}^{k+1}, \cdot)$. Note that for any real number $a \in \mathbb{R}$, from (20) we have that for any $\mathbf{x} \in \mathcal{X}$, $\bar{y}, y \in \mathcal{Y}$

$$a\Phi(\mathbf{x}, \bar{y}) \leq a\Phi(\mathbf{x}, y) + a\left\langle \nabla_y\Phi(\mathbf{x}, \bar{y}), \bar{y} - y \right\rangle + \max\{a, 0\} \cdot \frac{L_{yy}}{2}\left\|y - \bar{y}\right\|_{\mathcal{Y}}^2. \tag{39}$$

Next, we provide an upper bound for $(*)$ in (38) as follows:

$$\begin{aligned} (*) &\leq \left(1 - \frac{1}{M}\right)\Big(f(\mathbf{x}^k) - f(\mathbf{x}) + \Phi(\mathbf{x}^k, y^k) - \Phi(\mathbf{x}, y) + \left\langle \nabla_y\Phi(\mathbf{x}^k, y^k),\ y^{k+1} - y^k \right\rangle\Big) \\ &\leq \left(1 - \frac{1}{M}\right)\theta^k H^k(\mathbf{z}) + \left(1 - \frac{1}{M}\right)(1 - \theta^k)\Big(f(\mathbf{x}^k) - f(\mathbf{x}) + \Phi(\mathbf{x}^k, y) - \Phi(\mathbf{x}, y)\Big) \\ &\quad + \left(1 - \frac{1}{M}\right)\frac{L_{yy}(1 - \theta^k)_+}{2}\left\|y - y^k\right\|_{\mathcal{Y}}^2 + \left(1 - \frac{1}{M}\right)\theta^k\left\langle \nabla_y\Phi(\mathbf{x}^k, y^k), y - y^k \right\rangle \\ &\quad + \left(1 - \frac{1}{M}\right)\left\langle \nabla_y\Phi(\mathbf{x}^k, y^k), y^{k+1} - y \right\rangle, \end{aligned} \tag{40}$$

where $(a)_+ = \max\{a, 0\}$ and $H^k(\mathbf{z}) = f(\mathbf{x}^k) - f(\mathbf{x}) + \Phi(\mathbf{x}^k, y^k) - \Phi(\mathbf{x}, y)$; in the first inequality above we used the concavity of $\Phi(\mathbf{x}^k, \cdot)$, and in the second inequality, we split the resulting bound into $\theta^k \geq 0$ and $1 - \theta^k$ fractions, and used (39) for $a = 1 - \theta^k$, $\mathbf{x} = \mathbf{x}^k$ and $\bar{y} = y^k$.

Now before combining (40) with (38), we first simplify the summation of all inner product terms coming from both inequalities. Recall that $s^k = \nabla_y \Phi(\mathbf{x}^k, y^k) + \theta^k q^k$ where $q^k = M(\nabla_y \Phi(\mathbf{x}^k, y^k) - \nabla_y \Phi(\mathbf{x}^{k-1}, y^{k-1}))$ and we define $r^k \triangleq q^k - (M-1)\nabla_y \Phi(\mathbf{x}^k, y^k)$ for all $k \geq 0$. Using these definitions, we can rearrange the sum of all inner products in (40) and (38) as follows:

$$\begin{aligned}
&\frac{1}{M}\left\langle \nabla_y \Phi(\mathbf{x}^{k+1}, y^{k+1}), y - y^{k+1}\right\rangle + \Big(1 - \frac{1}{M}\Big)\theta^k \left\langle \nabla_y \Phi(\mathbf{x}^k, y^k), y - y^k\right\rangle \\
&\quad + \Big(1 - \frac{1}{M}\Big)\left\langle \nabla_y \Phi(\mathbf{x}^k, y^k), y^{k+1} - y\right\rangle + \frac{1}{M}\left\langle s^k, y^{k+1} - y\right\rangle \\
&= \left\langle \frac{1}{M}\nabla_y \Phi(\mathbf{x}^{k+1}, y^{k+1}) - \nabla_y \Phi(\mathbf{x}^k, y^k), y - y^{k+1}\right\rangle \\
&\quad + (1 - \frac{1}{M})\theta^k \left\langle \nabla_y \Phi(\mathbf{x}^k, y^k), y - y^k\right\rangle + \frac{\theta^k}{M}\left\langle q^k, y^k - y\right\rangle + \frac{\theta^k}{M}\left\langle q^k, y^{k+1} - y^k\right\rangle \\
&= -\frac{1}{M}\left\langle r^{k+1}, y^{k+1} - y\right\rangle + \frac{\theta^k}{M}\left\langle r^k, y^k - y\right\rangle + \frac{\theta^k}{M}\left\langle q^k, y^{k+1} - y^k\right\rangle. \qquad (41)
\end{aligned}$$

Hence, using (40) and (41) within (38), and the definition of $\mathcal{L}(\cdot, \cdot)$ we obtain the following inequality:

$$\begin{aligned}
&\frac{1}{M}\left(\mathcal{L}(\mathbf{x}^{k+1}, y) - \mathcal{L}(\mathbf{x}, y^{k+1})\right) \leq \\
&\quad \frac{1}{M}\Big(B^{k+1} + A^{k+1} - \left\langle r^{k+1}, y^{k+1} - y\right\rangle + \theta^k \left\langle r^k, y^k - y\right\rangle + \theta^k \underbrace{\left\langle q^k, y^{k+1} - y^k\right\rangle}_{(**)}\Big) \\
&\quad + \Big(1 - \frac{1}{M}\Big)\Big((1 - \theta^k)\big(\mathcal{L}(\mathbf{x}^k, y) - \mathcal{L}(\mathbf{x}, y)\big) + (1 - \theta^k)_+ L_{yy}\mathbf{D}_{\mathcal{Y}}(y, y^k)\Big) \\
&\quad + \Big(1 - \frac{1}{M}\Big)\Big(\theta^k H^k(\mathbf{z}) - H^{k+1}(\mathbf{z})\Big) + \Lambda_x^k + E^k, \qquad (42)
\end{aligned}$$

where we also used the fact that $\mathbf{D}_{\mathcal{Y}}(y, \bar{y}) \geq \frac{1}{2}\|y - \bar{y}\|^2$.

In order to provide a bound for term $(**)$, we provide a general bound for $\left\langle q^k, y - y^k\right\rangle$ for any $y \in \mathcal{Y}$ as follows. Let $p_x^k \triangleq \nabla_y \Phi(\mathbf{x}^k, y^k) - \nabla_y \Phi(\mathbf{x}^{k-1}, y^k)$ and $p_y^k \triangleq \nabla_y \Phi(\mathbf{x}^{k-1}, y^k) - \nabla_y \Phi(\mathbf{x}^{k-1}, y^{k-1})$. Such definitions immediately imply that $q^k = M(p_x^k + p_y^k)$, for all $k \geq 0$. Hence, using Young's inequality twice, once for $\left\langle p_x^k, y - y^k\right\rangle$ and once for $\left\langle p_y^k, y - y^k\right\rangle$ and the fact that $\mathbf{D}_{\mathcal{Y}}(y, \bar{y}) \geq \frac{1}{2}\|y - \bar{y}\|_{\mathcal{Y}}^2$, for any $y, \bar{y} \in \mathcal{Y}$, we obtain that for all $k \geq 0$,

$$\left|\left\langle q^k, y - y^k\right\rangle\right| \leq M(\alpha^k + \beta^k)\mathbf{D}_{\mathcal{Y}}(y, y^k) + \frac{M}{2\alpha^k}\left\|p_x^k\right\|_{\mathcal{Y}^*}^2 + \frac{M}{2\beta^k}\left\|p_y^k\right\|_{\mathcal{Y}^*}^2, \qquad (43)$$

for any $\alpha^k, \beta^k > 0$. Moreover, if $L_{yy} = 0$, then $\left\|p_y^k\right\|_{\mathcal{Y}^*} = 0$; hence, $\left|\left\langle q^k, y - y^k\right\rangle\right| \leq \alpha^k \mathbf{D}_{\mathcal{Y}}(y, y^k) + \frac{1}{2\alpha^k}\left\|p_x^k\right\|_{\mathcal{Y}^*}^2$, for any $\alpha^k > 0$. Therefore, first using (43) within (42) for some $\alpha^k, \beta^k > 0$ (with the exception of $\beta^k = 0$ for when $L_{yy} = 0$), then multiplying both sides of the resulting inequality by M; and finally, adding and subtracting $M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \mathbf{x}^{k+1}) - (M-1)\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \mathbf{x}^k) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}, \tilde{\mathbf{x}}^{k+1}) + M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}^{k+1}, \mathbf{x}^k) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k)$ and $\frac{M}{2}\left\|\mathbf{x} - \mathbf{x}^{k+1}\right\|_{\mathfrak{M}}^2 - \frac{1}{2}\left\|\mathbf{x} - \tilde{\mathbf{x}}^{k+1}\right\|_{\mathfrak{M}}^2 - \frac{M-1}{2}\left\|\mathbf{x} - \mathbf{x}^k\right\|_{\mathfrak{M}}^2$ to the right-hand side, and rearranging the terms yield the desired result in (30). □

B.1 Backtracking Step-size Analysis

Lemma 6. *Suppose the sequence* $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}_{k\geq 0}$ *satisfy* (13a) *and* (13b) *for some positive* $\{\alpha^k\}_{k\geq 0}$*, nonnegative* $\{\beta^k\}_{k\geq 0}$*, and* $\delta \in [0, 1)$*. Let* $\{\mathbf{x}^k, y^k\}$ *be generated according to the recursion in* (29) *using the given parameter sequence* $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$*. Then* $\{\mathbf{x}^k, y^k\}$ *and* $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$ *satisfy* (15) *with the same* $\{\alpha^k, \beta^k\}$ *and* δ*.*

Proof. The proof is similar to Lemma 3.4. in Hamedani and Aybat (2021). □

Lemma 7. *Given arbitrary $\{\tilde{\tau}^k\}_{k\geq 0} \subset \mathbb{R}_{++}$, and $\bar{\tau}, \gamma^0 > 0$, let $\sigma^{-1} = \gamma^0\bar{\tau}$, and for $k \geq 0$, let $\sigma^k = \gamma^k\tilde{\tau}^k$, $\theta^k = \sigma^{k-1}/\sigma^k$, and $\gamma^{k+1} = \gamma^k(1 + \underline{\mu}\tilde{\tau}^k)$. Moreover, for $i \in \mathcal{M}$ and $k \geq 0$, let $\tau_i^k = \left(\frac{1}{M}(\mu_i + \frac{1}{\tilde{\tau}^k}) - \mu_i\right)^{-1}$. Then, $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$ satisfies* (13c) *and* (13d) *with $t^k = \sigma^k/\sigma^0$.*

Proof. Since $t^k = \sigma^k/\sigma^0$ for $k \geq 0$, we get

$$\frac{t^k}{\sigma^k} = \frac{1}{\sigma^0} = \frac{t^{k+1}}{\sigma^{k+1}}, \qquad t^{k+1}\theta^k = \frac{\sigma^{k+1}}{\sigma^0}\frac{\sigma^k}{\sigma^{k+1}} = \frac{\sigma^k}{\sigma^0} = t^k;$$

thus, the first condition in (13d) holds with equality and the second condition is also satisfied for the given choice of parameters.

Furthermore, since $t^k = \sigma^k/\sigma^0$ and $\mu_i + \frac{1}{\tau_i^k} = \frac{1}{M}(\mu_i + \frac{1}{\tilde{\tau}^k})$ for all $i \in \mathcal{M}$, using these choice of parameters (13c) can be equivalently written as follows:

$$\frac{1}{M}\Big(\frac{1}{\tilde{\tau}^k} + \mu_i\Big) \geq \frac{\sigma^{k+1}}{\sigma^k}\Big(\frac{1}{M}\Big(\frac{1}{\tilde{\tau}^{k+1}} + \mu_i\Big) - \frac{\mu_i}{M}\Big), \quad \forall\, i \in \mathcal{M}, \tag{44}$$

which is equivalent to $\frac{1}{\tilde{\tau}^k} + \mu_i \geq \frac{\sigma^{k+1}}{\sigma^k}\frac{1}{\tilde{\tau}^{k+1}} = \frac{\gamma^{k+1}\tilde{\tau}^{k+1}}{\gamma^k\tilde{\tau}^k}\frac{1}{\tilde{\tau}^{k+1}} = \frac{1}{\tilde{\tau}^k}(1+\underline{\mu}\tilde{\tau}^k) = \frac{1}{\tilde{\tau}^k} + \underline{\mu}$, which trivially holds for all $i \in \mathcal{M}$ as $\mu_i \geq \underline{\mu}$ for $i \in \mathcal{M}$, where the first equality above follows from $\sigma^k = \gamma^k\tilde{\tau}^k$, the second equality uses $\gamma^{k+1} = \gamma^k(1+\underline{\mu}\tilde{\tau}^k)$. This completes the proof.

□

Lemma 8. *Consider $\{\tilde{\tau}^k\}_{k\geq 0}$ generated by* `RB-APD-B` *displayed in Algorithm 2 for some $\delta \in [0, 1)$ and $c_\alpha, c_\beta \geq 0$ such that $M(c_\alpha + c_\beta) + \delta \leq 1$. When $L_{yy} > 0$, set $c_\alpha,\ c_\beta > 0$; otherwise, when $L_{yy} = 0$, set $c_\alpha > 0$ and $c_\beta = 0$. There exists a positive sequence $\{\hat{\tau}^k\}_{k\geq 0}$ such that $\tilde{\tau}^k \geq \eta\hat{\tau}^k$ for all $k \geq 0$. Furthermore, when $L_{yy} > 0$ and $\underline{\mu} = 0$, $\hat{\tau}^k \geq \min\{\Psi_1, \Psi_2\}$ for $k \geq 0$; when $L_{yy} = 0$ and $\underline{\mu} = 0$, $\hat{\tau}^k \geq \Psi_1$ for $k \geq 0$; and when $L_{yy} = 0$ and $\underline{\mu} > 0$, $\hat{\tau}^k \geq \Psi_1\sqrt{\gamma^0/\gamma^k}$ for $k \geq 0$, where Ψ_1 and Ψ_2 are defined in* (24).[8]

Proof. Let us fix arbitrary $k \geq 0$ and $i_k \in \mathcal{M}$. Lemma 6 implies that if (13a) and (13b) hold then (15) holds as well. First, to show that the backtracking condition is satisfied in a finite number of steps, we will show that there exists $\hat{\tau}^k > 0$ such that (13a) and (13b) are true for all $\tilde{\tau}^k \in (0, \hat{\tau}^k]$. Using $\sigma^k = \gamma^k\tilde{\tau}^k$, $\theta^k = \sigma^{k-1}/\sigma^k$, $\overline{L}_{xx} = \max_{i\in\mathcal{M}}\{L_{x_ix_i}\}$, and $\overline{L}_{yx} = \max_{i\in\mathcal{M}}\{L_{yx_i}\}$ inequalities in (13a) and (13b) hold if

$$0 \geq -(1-\delta) + \overline{L}_{xx}\tau_{i_k}^k + \frac{\overline{L}_{yx}^2}{c_\alpha}\gamma^k\tilde{\tau}^k\tau_{i_k}^k, \quad 1 - (\delta + M(c_\alpha + c_\beta)) \geq \frac{ML_{yy}^2}{c_\beta}(\gamma^k\tilde{\tau}^k)^2. \tag{45}$$

Recall that $\bar{\mu} = \max_{i\in\mathcal{M}}\mu_i$. Suppose $L_{yy} > 0$, then $\tau_i^k = \left(\frac{1}{M}(\mu_i + \frac{1}{\tilde{\tau}^k}) - \mu_i\right)^{-1}$ for all $i \in \mathcal{M}$ implies that $\tau_{i_k}^k \leq M(1/\tilde{\tau}^k - (M-1)\bar{\mu})^{-1} = M\tilde{\tau}^k/(1-(M-1)\bar{\mu}\tilde{\tau}^k)$ implies that (45) holds for all $\tilde{\tau}^k \in (0, \hat{\tau}^k]$, where

$$\hat{\tau}^k \triangleq \min\left\{\frac{-\bar{b} + \sqrt{\bar{b}^2 + 4(1-\delta)\overline{L}_{yx}^2\gamma^k/(Mc_\alpha)}}{2\overline{L}_{yx}^2\gamma^k/c_\alpha}, \quad \frac{\sqrt{c_\beta(1-(M(c_\alpha+c_\beta)+\delta))}}{\gamma^k\sqrt{M}L_{yy}}\right\}, \tag{46}$$

$$\bar{b} \triangleq \overline{L}_{xx} + \frac{(1-\delta)(M-1)\bar{\mu}}{M}. \tag{47}$$

Note that when $L_{yy} = 0$, the second inequality in (45) always holds due to our choice of $\delta \in [0, 1)$ and $c_\alpha, c_\beta \geq 0$ satisfying $M(c_\alpha + c_\beta) + \delta \leq 1$; hence, $\hat{\tau}_k$ is defined by the first term in (46), i.e., treating $1/0$ in the second term as $+\infty$. Since in each step of backtracking, $\tilde{\tau}^k$ is decreased by a factor of $\eta \in (0, 1)$, when the backtracking terminates, $\tilde{\tau}^k \geq \eta\hat{\tau}^k$. Next, we provide a lower bound on $\hat{\tau}^k$ by considering the following two cases: (Case I) $\underline{\mu} > 0$; and (Case II) $\underline{\mu} = 0$. In particular, we will also use the following useful inequality: for any $a \geq 0$ and $b, c > 0$, we have $\sqrt{a^2 + cb^2} \geq a + \sqrt{\bar{c}}bd$ where $d \triangleq -\frac{a}{b\sqrt{\bar{c}}} + \sqrt{\frac{a^2}{b^2\bar{c}} + 1}$ holds for any $\bar{c} \in (0, c]$.

[8] $\underline{\mu} = 0$ implies $\gamma^k = \gamma^0$ for $k \geq 0$, while $\underline{\mu} > 0$ implies $\gamma^{k+1} > \gamma^k$ for $k \geq 0$.

For (Case I), from the assumption we know that $L_{yy} = 0$; therefore, $\hat{\tau}^k = \frac{-\bar{b}+\sqrt{\bar{b}^2+4(1-\delta)\overline{L}_{yx}^2\gamma^k/(Mc_\alpha)}}{2\overline{L}_{yx}^2\gamma^k/c_\alpha}$. Using the fact that $\gamma_{k+1} \geq \gamma_k \geq \gamma_0 > 0$ for all $k \geq 0$, and the above useful inequality for $a = \bar{b}$, $b = 2\sqrt{\frac{1-\delta}{Mc_\alpha}}\overline{L}_{yx}$, $c = \gamma_k$ and $\bar{c} = \gamma_0$ we conclude that $\hat{\tau}^k \geq \Psi_1\sqrt{\gamma_0/\gamma_k}$.

For (Case II), when $\underline{\mu} = 0$, $\gamma^k = \gamma^0$ for $k \geq 0$. Hence, from (46), we have $\hat{\tau}^k = \hat{\tau}^0$ for $k \geq 0$; thus, when $L_{yy} = 0$, we get $\hat{\tau}^0 \geq \Psi_1$, and when $L_{yy} > 0$, we get $\hat{\tau}_0 \geq \min\{\Psi_1, \Psi_2\}$. □

Lemma 9. *Suppose $\underline{\mu} > 0$, and $L_{yy} = 0$. Stepsize sequences generated by both* `RB-APD` *and* `RB-APD-B`*, displayed in Algorithms 1 and 2, respectively, satisfy $\sigma^k = \Omega(k)$, $\tilde{\tau}^k = \Omega(1/\sigma^k)$, and $\tilde{\tau}^k/\sigma^k = \mathcal{O}(1/k^2)$ for $k \geq 0$. Indeed, $\sigma^k \geq \frac{\Gamma^2}{3\underline{\mu}}k$, $\tilde{\tau}^k\sigma^k \geq \Gamma^2/\underline{\mu}^2$ and $(\gamma^k)^{-1} = \tilde{\tau}^k/\sigma^k \leq 9/(\Gamma^2k^2)$ for $k \geq 0$, where $\Gamma = \underline{\mu}\tilde{\tau}_0\sqrt{\gamma_0}$ for* `RB-APD` *and $\Gamma = \underline{\mu}\eta\Psi_1\sqrt{\gamma_0}$ for* `RB-APD-B` *with Ψ_1 as defined in* (24)*. Furthermore, for all $\epsilon > 0$, $\sigma^k \geq \frac{\Gamma^2}{(2+\epsilon)\underline{\mu}}k$ and $\tilde{\tau}^k/\sigma^k \leq (2+\epsilon)^2/(\Gamma^2k^2)$ for $k \geq \lceil\frac{1}{\epsilon}\rceil$.*

Proof. The proof follows directly from Lemma 3.7. in Hamedani and Aybat (2021). □

Next, we prove Lemma 1.

Proof of Lemma 1. Indeed, Lemma 7 implies that $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$ generated by `RB-APD-B` satisfies (13c) and (13d) for $\{t^k\}$ such that $t^k = \sigma^k/\sigma^0$ for $k \geq 0$. Moreover, Lemma 8 shows that for any $k \geq 0$, the backtracking condition in Algorithm 2 holds after a finite number of inner iterations. Thus, the results of Lemma 1 clearly hold.

Before proceeding to the proof of our main results, we would like to remind the reader Assumption 2 stating our assumptions on the Bregman distance generating function $\varphi_{\mathcal{X}_i}(\cdot)$ for $i \in \mathcal{M}$.

B.2 Asymptotic Convergence Analysis

To fix the notation, suppose $\mathbf{F} : \Omega \to \mathbb{R}$ is a random variable, $\mathbf{F}(\omega)$ denotes a particular realization of $\mathbf{F}$ corresponding to $\omega \in \Omega$ where Ω denotes the sample space.

First, recall that $U_i \in \mathbb{R}^{m\times m_i}$ for $i \in \mathcal{M}$ such that $\mathbf{I}_m = [U_1, \ldots, U_M]$, where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix –see Definition 1. Therefore, we can write $\mathbf{x}^{k+1}$ equivalently as follows:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + U_{i_k}U_{i_k}^\top(\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k). \tag{48}$$

Also recall that for all $k \geq 1$, $\mathbf{E}^k[\cdot] = \mathbf{E}[\cdot \mid \mathcal{F}_k]$, where $\mathcal{F}_k = \sigma\left(\{i_0, \ldots, i_{k-1}\}\right)$ is the σ-algebra generated by i.i.d. random variables $\{i_0, \ldots, i_k\}$. Thus,

$$\mathbf{E}^k[\mathbf{x}^{k+1}] = \mathbf{x}^k + \mathbf{E}^k[U_{i_k}U_{i_k}^\top](\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k) = \frac{1}{M}\tilde{\mathbf{x}}^{k+1} + \Big(1 - \frac{1}{M}\Big)\mathbf{x}^k. \tag{49}$$

Furthermore, for any $\psi : \mathcal{X} \to \mathbb{R}$ such that $\psi(\mathbf{x}) = \sum_{i\in\mathcal{M}}\psi_i(x_i)$ for some $\psi_i : \mathcal{X}_i \to \mathbb{R}$, we also have

$$\mathbf{E}^k[\psi(\mathbf{x}^{k+1})] = \frac{1}{M}\sum_{i\in\mathcal{M}}\big(\psi(\mathbf{x}^k) + \psi_i(\tilde{x}^{k+1}) - \psi_i(x_i^k)\big) = \frac{1}{M}\psi(\tilde{\mathbf{x}}^{k+1}) + \Big(1 - \frac{1}{M}\Big)\psi(\mathbf{x}^k). \tag{50}$$

Therefore, for $\mathcal{E}^k(\cdot)$ defined in (31d), we can conclude that $\mathbf{E}^k[\mathcal{E}^k(\mathbf{x})] = 0$ for any fixed $\mathbf{x}$ and $k \geq 0$.

Let $\mathbf{z}^\# = (\mathbf{x}^\#, y^\#)$ be a saddle point of $\mathcal{L}$ in (1), and Bregman distance generating functions are selected according to Assumption 2. Lemma 7 implies that $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$ sequences generated by `RB-APD` and `RB-APD-B`, both satisfy (13c) and (13d) for $t^k = \frac{\sigma^k}{\sigma^0}$ for $k \geq 0$. Thus, we can conclude that for $k \geq 1$,

$$t^{k-1}R^k(\mathbf{z}^\#) \geq t^kQ^k(\mathbf{z}^\#). \tag{51}$$

Finally, note that $C_*^k = C^k(\mathbf{x}^{k+1}, y^{k+1})$ for $\{\alpha^k, \beta^k\}$ sequence defined as $\alpha^k = c_\alpha/\sigma^{k-1}$ and $\beta^k = c_\beta/\sigma^{k-1}$ for all $k \geq 1$ for some $c_\alpha, c_\beta \geq 0$ as stated in Theorems 1 and 2.

Now multiplying inequality (30) by t^k evaluated at $(\mathbf{x}, y) = (\mathbf{x}^\#, y^\#)$ taking conditional expectation, and using the facts that $\mathbf{E}[\mathcal{E}^k(\mathbf{x}^\#) \mid \mathcal{F}_k] = 0$, $(1-\theta^k)_+ L_{yy}{=}0$ [9], and $C_*^k \le -\delta[M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}^{k+1}, \mathbf{x}^k) + \frac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y^{k+1}, y^k)]$ combined with (51) lead to

$$\begin{aligned} &t^k \mathbf{E}^k[\mathcal{L}(\mathbf{x}^{k+1}, y^\#) - \mathcal{L}(\mathbf{x}^\#, y^{k+1})] + \mathbf{E}^k[t^k R^{k+1}(z^\#) + (M-1)t^k H^{k+1}(\mathbf{z}^\#)] \\ &\le t^{k-1} R^k(\mathbf{z}^\#) + (M-1)t^{k-1} H^k(\mathbf{z}^\#) + (M-1)t^k(1-\theta^k)\big(\mathcal{L}(\mathbf{x}^k, y^\#) - \mathcal{L}(\mathbf{x}^\#, y^\#)\big) \\ &\quad - t^k \delta \mathbf{E}^k[M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}^{k+1}, \mathbf{x}^k) + \tfrac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y^{k+1}, y^k)]. \end{aligned} \tag{52}$$

Note when $\underline{\mu} > 0$ from the step-size rules we have that for any $k \ge 1$,

$$\theta^k = \frac{\sigma^{k-1}}{\sigma^k} = \frac{\gamma^{k-1}\tilde{\tau}^{k-1}}{\gamma^k \tilde{\tau}^k} \ge \frac{\gamma^{k-1}}{\gamma^k} = \frac{1}{1+\underline{\mu}\tilde{\tau}^{k-1}} \ge \frac{1}{1+\underline{\mu}\tilde{\tau}^0} \ge \frac{M-1}{M}, \tag{53}$$

$$\implies (M-1)(1-\theta^k)t^k \le t^{k-1}. \tag{54}$$

On the other hand, when $\underline{\mu} = 0$, then $\theta^k \ge 1$ which immediately implies that $(M-1)(1-\theta^k)t^k \le t^{k-1}$. Therefore, we can rewrite (52) as follows by noting that $\mathcal{L}(\mathbf{x}^\#, y^\#) - \mathcal{L}(\mathbf{x}^\#, y^{k+1}) \ge 0$,

$$\mathbf{E}^k[a^{k+1}] = \mathbf{E}[a^{k+1} \mid \mathcal{F}_k] \le a^k - b^k, \tag{55}$$

where $a^k, b^k \in \mathcal{F}_k$ are defined as follows

$$a^k = t^{k-1} R^k(\mathbf{z}^\#) + (M-1)t^{k-1} H^k(\mathbf{z}^\#) + t^{k-1}\big(\mathcal{L}(\mathbf{x}^k, y^\#) - \mathcal{L}(\mathbf{x}^\#, y^\#)\big), \tag{56}$$

$$b^k = t^k \delta \left(\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k) + \tfrac{1}{\sigma^k}\mathbf{D}_{\mathcal{Y}}(y^{k+1}, y^k)\right) \ge 0. \tag{57}$$

Moreover, from concavity of $\Phi(\mathbf{x}, \cdot)$ and the fact that $\mathcal{L}(\mathbf{x}^k, y^\#) - \mathcal{L}(\mathbf{x}^\#, y^\#) \ge 0$ we can provide a lower bound on a^k as follows:

$$\begin{aligned} a^k &\ge t^{k-1}\Big[(M-1)\big(\mathcal{L}(\mathbf{x}^k, y^\#) - \mathcal{L}(\mathbf{x}^\#, y^\#)\big) + M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^{k-1}}(\mathbf{x}, \mathbf{x}^k) + \frac{M}{2}\left\|\mathbf{x}^\# - \mathbf{x}^k\right\|_{\mathfrak{M}}^2 + \frac{1}{\sigma^{k-1}}\mathbf{D}_{\mathcal{Y}}(y^\#, y^k) \\ &\quad + \left\langle q^k, y^k - y^\#\right\rangle + \frac{M}{2\alpha^k}\left\|p_x^k\right\|_{\mathcal{Y}^*}^2 + \frac{M}{2\beta^k}\left\|p_y^k\right\|_{\mathcal{Y}^*}^2\Big], \\ &\ge \frac{M}{2}\left\|\mathbf{x}^\# - \mathbf{x}^k\right\|_{t^{k-1}(\mathbf{T}^{k-1}+\mathfrak{M})}^2 + t^{k-1}\Big(\tfrac{1}{\sigma^{k-1}} - M(\alpha^k + \beta^k)\Big)\mathbf{D}_{\mathcal{Y}}(y^\#, y^k), \\ &\ge \frac{M}{2}t^k\ \left\|\mathbf{x}^\# - \mathbf{x}^k\right\|_{\mathbf{T}^k + (1-\frac{1}{M})\mathfrak{M}}^2 + t^k\Big(\tfrac{1}{\sigma^k} - \theta^k M(\alpha^k + \beta^k)\Big)\mathbf{D}_{\mathcal{Y}}(y^\#, y^k), \end{aligned} \tag{58}$$

where the second inequality follows from (43) and $\mathbf{D}_{\mathcal{X}_i}(x_i, x_i') \ge \frac{1}{2}\left\|x_i - x_i'\right\|_{\mathcal{X}_i}^2$ for $i \in \mathcal{M}$, and the third one follows from (13c) and (13d).

Note that for all $i \in \mathcal{M}$, we have

$$Mt^k\Big(\frac{1}{\tau_i^k} + \big(1 - \frac{1}{M}\big)\mu_i\Big) = \frac{t^k}{\tilde{\tau}^k} = \frac{\sigma^k}{\sigma^0}\frac{1}{\tilde{\tau}^k} = \frac{\gamma^k}{\gamma_0}\frac{1}{\tilde{\tau}^0} \ge \frac{1}{\bar{\tau}}, \quad \forall\, k \ge 0, \tag{59}$$

where we used $t^k = \frac{\sigma^k}{\sigma^0}$, $\sigma^k = \gamma^k\tilde{\tau}^k$, and $\gamma^k \ge \gamma^0$ for all $k \ge 0$, and $\tilde{\tau}^0 \le \bar{\tau}$. Moreover, for all $k \ge 0$, setting $\alpha^{k+1} = c_\alpha/\sigma^k$ and $\beta^{k+1} = c_\beta/\sigma^k$ for $c_\alpha, c_\beta \ge 0$ such that $1 - M(c_\alpha + c_\beta) \ge \delta > 0$, implies

$$t^k\Big(\tfrac{1}{\sigma^k} - \theta^k M(\alpha^k + \beta^k)\Big) = \frac{1}{\sigma^0} - M\frac{\sigma^{k-1}}{\sigma^0}\frac{c_\alpha + c_\beta}{\sigma^{k-1}} \ge \frac{\delta}{\sigma^0} \ge \frac{\delta}{\gamma^0\bar{\tau}}, \quad \forall\, k \ge 0, \tag{60}$$

where we used $t^k = \frac{\sigma^k}{\sigma^0}$ and $\theta^k = \sigma^{k-1}/\sigma^k$ for all $k \ge 0$, and $\sigma^0 = \gamma^0\tilde{\tau}^0 \le \gamma^0\bar{\tau}$. Finally, combining (58) with the lower bounds given in (59) and (60), and using $\mathbf{D}_{\mathcal{Y}}(y, y') \ge \frac{1}{2}\left\|y - y'\right\|_{\mathcal{Y}}^2$ for any $y \in \mathcal{Y}$ and $y' \in \mathbf{dom}\, h$, we get

$$a^k \ge \frac{\delta_x'}{2}\left\|\mathbf{x}^\# - \mathbf{x}^k\right\|_{\mathcal{X}}^2 + \frac{\delta_y'}{2}\left\|y^\# - y^k\right\|_{\mathcal{Y}}^2 \ge 0, \quad \forall\, k \ge 0, \tag{61}$$

[9] When $\underline{\mu} = 0$, for both `RB-APD` and `RB-APD-B`, $\theta^k \ge 0$ for $k \ge 0$; thus, $(1-\theta^k)_+ = 0$, which leads to $(1-\theta^k)_+ L_{yy} = 0$ for $k \ge 0$. On the other hand, for the case $\underline{\mu} > 0$, we assume that $L_{yy} = 0$, i.e., $\Phi(\mathbf{x}, \cdot)$ is affine for every fixed $\mathbf{x}$; hence, we again get $(1-\theta^k)_+ L_{yy} = 0$, even though $\theta^k < 1$ for some $k \ge 0$ is possible for this scenario.

where $\delta'_x = \frac{1}{\bar{\tau}} > 0$ and $\delta'_y = \frac{\delta}{\gamma^0 \bar{\tau}} > 0$. Therefore, invoking Lemma 3, we can conclude that $\lim_{k\to+\infty} a^k \geq 0$ and $\sum_{k=0}^{\infty} b^k \in \mathbb{R}_{++}$ exist in a.s. sense, i.e.,

$$\sum_{k=0}^{+\infty} t^k \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\tilde{\mathbf{x}}^{k+1}, \mathbf{x}^k) < \infty \quad a.s. \tag{62}$$

Since $\{a^k\}$ is an a.s. bounded sequence, (61) implies that $\{\mathbf{z}^k(\omega)\}$ is a bounded sequence for any $\omega \in \Omega$, where $\mathbf{z}^k = (\mathbf{x}^k, y^k)$; hence, it has a convergent subsequence $\mathbf{z}^{k_n}(\omega) \to \mathbf{z}^*(\omega)$ as $n \to \infty$ for some $\mathbf{z}^*(\omega) \in \mathcal{X} \times \mathcal{Y}$ – note that k_n also depends on ω which is omitted to simplify the notation. Define $\mathbf{z}^* = (\mathbf{x}^*, y^*)$ such that $\mathbf{z}^* = [\mathbf{z}^*(\omega)]_{\omega\in\Omega}$.

Next, we argue that $\mathbf{z}^{k_n\pm1} \to \mathbf{z}^*$ almost surely as $n \to \infty$. To this aim, first we analyze $\{t^k\mathbf{T}^k\}$ sequence that appears in the definition of b^k in (57). Note that (59) implies that

$$\frac{t^k}{\tau_i^k} = \frac{1}{M}\frac{\gamma^k}{\gamma^0\tilde{\tau}^0} - t^k\Big(1-\frac{1}{M}\Big)\mu_i = \frac{\gamma^k}{\gamma^0\tilde{\tau}^0}\Big(\frac{1}{M} - \tilde{\tau}^k\Big(1-\frac{1}{M}\Big)\mu_i\Big), \quad \forall\, i \in \mathcal{M},\ \forall\, k \geq 0, \tag{63}$$

where we used $t^k = \frac{\sigma^k}{\sigma^0} = \frac{\gamma^k\tilde{\tau}^k}{\gamma^0\tilde{\tau}^0}$ for $k \geq 0$. Since $\bar{\tau} \geq \tilde{\tau}^k$ for $k \geq 0$, choosing $\mathbb{R}_{++} \ni \bar{\tau} \leq \frac{\bar{\delta}}{\bar{\mu}(M-1)}$ for some $\bar{\delta} \in (0,1)$ and $\bar{\mu} = \max_{i\in\mathcal{M}} \mu_i$ implies that for all $k \geq 0$, we have

$$\frac{t^k}{\tau_i^k} \geq \frac{\gamma^k}{\gamma^0\tilde{\tau}^0}\Big(\frac{1}{M} - \bar{\tau}\Big(1-\frac{1}{M}\Big)\mu_i\Big) \geq \frac{\gamma^k}{\gamma^0}\frac{1-\bar{\delta}}{M\bar{\tau}} \geq \frac{1-\bar{\delta}}{M\bar{\tau}} > 0, \quad \forall\, i \in \mathcal{M}, \tag{64}$$

which follows from $\gamma^k \geq \gamma^0$ for $k \geq 0$; hence, $t^k\mathbf{T}^k \succeq \frac{1}{\bar{\tau}}\frac{1-\bar{\delta}}{M}\mathbf{I}$ for $k \geq 0$. Finally, we also have $\frac{t^k}{\sigma^k} = \frac{1}{\sigma^0} \geq \frac{1}{\gamma^0\bar{\tau}} > 0$ for $k \geq 0$. Now, now combining these two results with $\sum_{k=0}^{\infty} b^k < \infty$ (due to Lemma 3), we can conclude that $b_k \to 0$ implying

$$0 \leq \big\|\mathbf{x}^{k+1} - \mathbf{x}^k\big\|_{\mathcal{X}}^2 \leq \big\|\tilde{\mathbf{x}}^{k+1} - \mathbf{x}^k\big\|_{\mathcal{X}}^2 \to 0, \quad \big\|y^{k+1} - y^k\big\|_{\mathcal{Y}}^2 \to 0, \tag{65}$$

almost surely as $k \to \infty$. Therefore, for any realization $\omega \in \Omega$ and $\zeta > 0$, there exists $N_1(\omega)$ such that for any $n \geq N_1(\omega)$, we have $\max\{\big\|\mathbf{z}^{k_n}(\omega) - \mathbf{z}^{k_n-1}(\omega)\big\|,\ \big\|\mathbf{z}^{k_n}(\omega) - \mathbf{z}^{k_n+1}(\omega)\big\|\} < \frac{\zeta}{2}$. Convergence of $\{\mathbf{z}^{k_n}(\omega)\}$ sequence also implies that there exists $N_2(\omega)$ such that for any $n \geq N_2(\omega)$, $\big\|\mathbf{z}^{k_n}(\omega) - \mathbf{z}^*(\omega)\big\| < \frac{\zeta}{2}$. Thus, for $\omega \in \Omega$, letting $N(\omega) \triangleq \max\{N_1(\omega), N_2(\omega)\}$, we conclude that $\big\|\mathbf{z}^{k_n\pm1}(\omega) - \mathbf{z}^*(\omega)\big\| < \zeta$, i.e., $\mathbf{z}^{k_n\pm1} \to \mathbf{z}^*$ almost surely as $n \to \infty$.

Fix an arbitrary $\omega \in \Omega$ and consider the subsequence $\{k_n\}_{n\geq1}$. For all $n \in \mathbb{Z}_+$, $\mathbf{x}$- and y-updates imply that

$$\frac{1}{\tau_{i_{k_n}}^{k_n}}\Big(\nabla\varphi_{\mathcal{X}_{i_{k_n}}}\big(\mathbf{x}^{k_n}(\omega)\big) - \nabla\varphi_{\mathcal{X}_{i_{k_n}}}\big(\mathbf{x}^{k_n+1}(\omega)\big)\Big) - \nabla_{x_{i_{k_n}}}\Phi(\mathbf{x}^{k_n}(\omega), y^{k_n+1}(\omega)) \in \partial f_{i_{k_n}}\big(x_{i_{k_n}}^{k_n+1}(\omega)\big), \tag{66a}$$

$$\frac{1}{\sigma^{k_n}}\Big(\nabla\varphi_{\mathcal{Y}}\big(y^{k_n}(\omega)\big) - \nabla\varphi_{\mathcal{Y}}\big(y^{k_n+1}(\omega)\big)\Big) + s^{k_n}(\omega) \in \partial h\big(y^{k_n+1}(\omega)\big), \tag{66b}$$

where we define $s^k = \nabla_y\Phi(\mathbf{x}^k, y^k) + \theta^k q^k$ and $q^k = M(\nabla_y\Phi(\mathbf{x}^k, y^k) - \nabla_y\Phi(\mathbf{x}^{k-1}, y^{k-1}))$ for $k \geq 0$.

Note that the sequence of randomly chosen block coordinates in RB-APD or RB-APD-B, i.e., $\{i_{k_n}\}_{n\geq1}$, is a Markov chain containing a single recurrent class. More specifically, the states are represented by $\mathcal{M}$ and starting from state $i \in \mathcal{M}$ the probability of eventually returning to state i is strictly positive for all $i \in \mathcal{M}$. Therefore, for any $i \in \mathcal{M}$, we can select a further subsequence $\mathcal{K}^i \subseteq \{k_n\}_{n\in\mathbb{Z}_+}$ such that $i_\ell = i$ for all $\ell \in \mathcal{K}^i$. Note that $\mathcal{K}^i$ is an infinite subsequence w.p. 1 and $\{\mathcal{K}^i\}_{i\in\mathcal{M}}$ is a partition of $\{k_n\}_{n\in\mathbb{Z}_+}$. For any $i \in \mathcal{M}$, one can conclude from (66b) and (66a) that for all $\ell \in \mathcal{K}^i$,

$$u_i^k \triangleq \frac{1}{\tau_i^\ell}\Big(\nabla\varphi_{\mathcal{X}_i}\big(\mathbf{x}^\ell(\omega)\big) - \nabla\varphi_{\mathcal{X}_i}\big(\mathbf{x}^{\ell+1}(\omega)\big)\Big) - \nabla_{x_i}\Phi(\mathbf{x}^\ell(\omega), y^{\ell+1}(\omega)) \in \partial f_i\big(x_i^{\ell+1}(\omega)\big), \tag{67a}$$

$$v^k \triangleq \frac{1}{\sigma^\ell}\Big(\nabla\varphi_{\mathcal{Y}}\big(y^\ell(\omega)\big) - \nabla\varphi_{\mathcal{Y}}\big(y^{\ell+1}(\omega)\big)\Big) + s^\ell(\omega) \in \partial h\big(y^{\ell+1}(\omega)\big). \tag{67b}$$

Since $\mathcal{K}^i \subseteq \{k_n\}_{n\in\mathbb{Z}_+}$, we have that $\lim_{\ell\in\mathcal{K}^i} \mathbf{z}^\ell(\omega) = \lim_{\ell\in\mathcal{K}^i} \mathbf{z}^{\ell+1}(\omega) = \mathbf{z}^*(\omega)$.

(Part I) of Theorems 2 and 1. Here we consider the case $\underline{\mu} = 0$. We first show that for any $\omega \in \Omega$, $\mathbf{z}^*(\omega)$ is a saddle point of (1) by considering the optimality conditions for the updates of $\mathbf{x}^{k+1}$ and y^{k+1} of the RB-APD and RB-APD-B algorithms. Then, we argue that $\mathbf{z}^*$ is indeed the unique limit point of $\{\mathbf{z}^k\}$, i.e., $\mathbf{z}^k \to \mathbf{z}^*$ as $k \to \infty$.

Next, we argue that $\sup_{k\geq 0}\frac{1}{\tau_i^k}<\infty$ for $i\in\mathcal{M}$ and $\sup_{k\geq 0}\max\{\frac{1}{\sigma^k},\theta^k\}<\infty$. Once we have this result, using the fact that for any $i\in\mathcal{M}$, $\nabla\varphi_{\mathcal{X}_i}$ and $\nabla\varphi_{\mathcal{Y}}$ are continuously differentiable on $\mathbf{dom}\, f_i$ and $\mathbf{dom}\, h$, respectively, it follows from Theorem 24.4 in Rockafellar (2015) that by taking the limit of both sides of (67) we get $\mathbf{0}\in\nabla_{x_i}\Phi\big(\mathbf{x}^*(\omega),y^*(\omega)\big)+\partial f_i(x_i^*(\omega))$, and $\mathbf{0}\in\partial h(y^*(\omega))-\nabla_y\Phi(\mathbf{x}^*(\omega),y^*(\omega))$, which implies that $\mathbf{z}^*(\omega)$ is a saddle point of (1) for any $\omega\in\Omega$. Indeed, since $\underline{\mu}=0$, for $k\geq 0$, $\gamma^k=\gamma^0$; hence, $\sigma^k=\gamma^0\tilde{\tau}^k$. Moreover, from Lemma 8, we have $\eta\Psi\leq\tilde{\tau}^k\leq\bar{\tau}$ for $k\geq 0$, where $\Psi=\Psi_1$ if $L_{yy}=0$ and $\Psi=\min\{\Psi_1,\Psi_2\}$ if $L_{yy}>0$. Note that these bounds on $\tilde{\tau}^k$ hold for both `RB-APD` and `RB-APD-B`. Next, using the τ_i^k update rule, $\sigma^k=\gamma^0\tilde{\tau}^k$ and $\theta^k=\frac{\sigma^{k-1}}{\sigma^k}$, we get the following uniform bounds holding for all $k\geq 0$:

$$0<\frac{1}{\tau_i^0}\leq\frac{1}{\tau_i^k}\leq\frac{1}{M}\frac{1}{\tilde{\tau}^k}-\frac{M-1}{M}\mu_i\leq\frac{1}{\eta\Psi M},\quad\forall\, i\in\mathcal{M},\tag{68}$$

$$\frac{1}{\gamma^0\bar{\tau}}\leq\frac{1}{\sigma^k}=\frac{1}{\gamma^0\tilde{\tau}^k}\leq\frac{1}{\eta\Psi\gamma^0},\quad 0\leq\theta^k=\frac{\sigma^{k-1}}{\sigma^k}=\frac{\tilde{\tau}^{k-1}}{\tilde{\tau}^k}\leq\frac{\bar{\tau}}{\eta\Psi}.\tag{69}$$

Next, we show that $\mathbf{z}^k\to\mathbf{z}^*$ almost surely as $k\to\infty$, and for this result we will use the following bound on $\{t^k\}$:

$$0\leq t^k=\frac{\sigma^k}{\sigma^0}=\frac{\gamma^0\tilde{\tau}^k}{\sigma^0}\leq\frac{\bar{\tau}}{\tilde{\tau}^0}\leq\frac{\bar{\tau}}{\eta\Psi}.\tag{70}$$

Since (52) is true for any saddle point $\mathbf{z}^{\#}$, letting $\mathbf{z}^{\#}=\mathbf{z}^*$ and repeating the same arguments we used for showing (55), we can conclude that

$$\mathbf{E}^k[d^{k+1}]\leq d^k-b^k,\tag{71}$$

where $d^k\triangleq t^{k-1}[(M-1)H^k(\mathbf{z}^*)+R^k(\mathbf{z}^*)+\mathcal{L}(\mathbf{x}^k,y^*)-\mathcal{L}(\mathbf{x}^*,y^*)]$ and b^k is defined in (57). Moreover, similar to (61), we can show that

$$d^k\geq\frac{\delta_x'}{2}\left\|\mathbf{x}^*-\mathbf{x}^k\right\|_{\mathcal{X}}^2+\frac{\delta_y'}{2}\left\|y^*-y^k\right\|_{\mathcal{Y}}^2\geq 0,\quad\forall\, k\geq 0,\tag{72}$$

where δ_x',δ_y' are defined after inequality (61). Next, invoking Lemma 3 again for (71), one can conclude that $d_*\triangleq\lim_{k\to\infty}d^k\geq 0$ exists almost surely. Now, we show that $d^{k+1}\to 0$ as $k\to\infty$. Let $\{\mathbf{z}^{k_n}\}_{n\geq 0}$ be the subsequence we considered earlier such that $\mathbf{z}^{k_n}\to\mathbf{z}^*$ a.s. as $n\to\infty$. Using (65), it is trivial to check that $H^{k_n+1}(\mathbf{z}^*)\to 0$ and $\mathcal{L}(\mathbf{x}^{k_n+1},y^*)-\mathcal{L}(\mathbf{x}^*,y^*)\to 0$ as $k\to\infty$. Thus, $\lim_{n\to\infty}d^{k_n+1}=\lim_{n\to\infty}t^{k_n}R^{k_n+1}(\mathbf{z}^*)$. Consider

$$\begin{aligned}R^{k_n+1}(\mathbf{z}^*)\triangleq&M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^{k_n}}(\mathbf{x}^*,\mathbf{x}^{k_n+1})+\frac{M}{2}\left\|\mathbf{x}^*-\mathbf{x}^{k_n+1}\right\|_{\mathfrak{M}}^2+\frac{1}{\sigma^{k_n}}\mathbf{D}_{\mathcal{Y}}(y^*,y^{k_n+1})\\&+\left\langle\nabla_y\Phi(\mathbf{x}^{k_n+1},y^{k_n+1})-M\nabla_y\Phi(\mathbf{x}^{k_n},y^{k_n}),y^{k_n+1}-y^*\right\rangle\\&+\frac{M}{2c_\alpha}\sigma^{k_n}\left\|\nabla_y\Phi(\mathbf{x}^{k_n+1},y^{k_n+1})-\nabla_y\Phi(\mathbf{x}^{k_n},y^{k_n+1})\right\|_{\mathcal{Y}^*}^2\\&+\frac{M}{2c_\beta}\sigma^{k_n}\left\|\nabla_y\Phi(\mathbf{x}^{k_n},y^{k_n+1})-\nabla_y\Phi(\mathbf{x}^{k_n},y^{k_n})\right\|_{\mathcal{Y}^*}^2,\end{aligned}$$

where we set $\alpha^{k+1}=c_\alpha/\sigma^k$ and $\beta^{k+1}=c_\beta/\sigma^k$ for some $c_\alpha,c_\beta\geq 0$ as described in Lemma 8. Using (68), (69) and (70) together with the fact that $\mathbf{z}^{k_n\pm 1}(\omega)\to\mathbf{z}^*(\omega)$ for any $\omega\in\Omega$, we conclude that $0=t^{k_n}R^{k_n+1}(\mathbf{z}^*(\omega))=\lim_{n\to\infty}d^{k_n+1}(\omega)$ for any $\omega\in\Omega$, where we also used the fact that $\{\mathbf{z}^{k_n}\}$ is a bounded sequence which implies $\{\nabla_y\Phi(\mathbf{x}^{k_n+1},y^{k_n+1})-M\nabla_y\Phi(\mathbf{x}^{k_n},y^{k_n})\}_{n\geq 0}$ is bounded as well due to continuity of $\nabla_y\Phi$. Henceforth, $\lim_{k\to\infty}d^k=0$ almost surely which together with (72) implies that $\mathbf{z}^k\to\mathbf{z}^*$ almost surely.

(Part II) of Theorems 2 and 1. Suppose $\underline{\mu}>0$. Recall that in this case, for all $i\in\mathcal{M}$, we set $\varphi_{\mathcal{X}_i}(x_i)=\frac{1}{2}\left\|x_i\right\|_{\mathcal{X}_i}^2$, where $\left\|x_i\right\|_{\mathcal{X}_i}=\sqrt{\langle x_i,x_i\rangle}$. Indeed, since $\sum_{k=0}^{\infty}b^k<\infty$, $t^k\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\tilde{\mathbf{x}}^{k+1},\mathbf{x}^k)\to 0$ holds. Moreover, (64) shows that for any $i\in\mathcal{M}$ and $k\geq 0$, $\frac{t^k}{\tau_i^k}\geq\frac{\gamma^k}{\gamma_0}\frac{1-\bar{\delta}}{M\bar{\tau}}$; therefore, we have $0\leq\gamma^k\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|_{\mathcal{X}}^2\leq\gamma^k\left\|\tilde{\mathbf{x}}^{k+1}-\mathbf{x}^k\right\|_{\mathcal{X}}^2\to 0$. Moreover, we have that $\gamma^k=\sigma^k/\tilde{\tau}^k\geq(\Gamma/(\underline{\mu}\tilde{\tau}^k))^2\geq(M\Gamma/(\underline{\mu}\tau_i^k))^2$ for all $k\geq 0$, where the first inequality follows from Lemma 9 and the last one uses (68). Therefore, one can easily conclude that $\left\|\mathbf{x}^{k+1}-\mathbf{x}^k\right\|_{\mathcal{X}}/\tau_i^k\to 0$ for any $i\in\mathcal{M}$. Thus, invoking invoking (Rockafellar, 2015, Theorem 24.4), (67a) implies that $-\nabla_{\mathbf{x}}\Phi(\mathbf{x}^*,y^*)\in\partial f(\mathbf{x}^*)$ assuming $\nabla\varphi_{\mathcal{X}_i}$ is Lipschitz. Finally, it follows from (69) and (67b) that $\nabla_y\Phi(\mathbf{x}^*,y^*)\in\partial h(y^*)$, where we used (Rockafellar, 2015, Theorem 24.4) assuming $\nabla\varphi_{\mathcal{Y}}$ is continuous. Therefore, we establish that any limit point of $\{\mathbf{z}^k\}$ is a saddle point of (1). □

B.3 Convergence Rate Analysis

Next we use the one-step result shown in Lemma 5 to derive a useful bound for the ergodic sequence generated by either RB−APD or RB−APD−B, which will help us establish the desired convergence rate results.

Lemma 10. *Suppose Assumptions 1 and 2 hold. Given some $\gamma^0 > 0$ and $\bar{\tau} \in \left(0, \frac{1}{\bar{\mu}(M-1)}\right)$, let $\{\mathbf{x}^k, y^k\}_{k\geq 0}$ be the iterate sequence generated by either RB−APD−B, stated in Algorithm 2, or by RB−APD, stated in Algorithm 1. If RB−APD is used, we assume that (14) holds for some $c_\alpha, c_\beta \geq 0$ and $\delta \in [0,1)$ as described in Theorem 1. Then for any $(\mathbf{x}, y) \in \mathbf{dom}\, f \times \mathbf{dom}\, h$ and $K \geq 1$,*

$$\begin{aligned} T^K\big(\mathcal{L}(\bar{\mathbf{x}}^K, y) - \mathcal{L}(\mathbf{x}, \bar{y}^K)\big) \\ \leq M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^0}(\mathbf{x}, \mathbf{x}^0) + \frac{M-1}{2}\left\|\mathbf{x}-\mathbf{x}^0\right\|_{\mathfrak{M}}^2 + \big(\frac{1}{\sigma^0} + \theta^0(M-1)L_{yy}\big)\mathbf{D}_{\mathcal{Y}}(y, y^0) \\ - t^K\Big(M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^K}(\mathbf{x}, \mathbf{x}^K) + \frac{M-1}{2}\left\|\mathbf{x}-\mathbf{x}^K\right\|_{\mathfrak{M}}^2 + \big(\frac{1}{\sigma^K} - M\theta^K(\alpha^K+\beta^K)\big)\mathbf{D}_{\mathcal{Y}}(y, y^K)\Big) \\ + (M-1)\Big(\mathcal{L}(\mathbf{x}^0, y) - \mathcal{L}(\bar{\mathbf{x}}^K, y) + \sum_{k=0}^{K-1} t^k(1-\theta^k)_+ L_{yy}\mathbf{D}_{\mathcal{Y}}(y, y^k)\Big) + \sum_{k=0}^{K-1} t^k\mathcal{E}^k(\mathbf{x}), \end{aligned} \tag{73}$$

where $T^K = \sum_{k=0}^{K-1} t^k$, $\bar{\mathbf{x}}^K = \frac{1}{T^K+M-1}\Big(\sum_{k=0}^{K-2}(Mt^k-(M-1)t^{k+1})\mathbf{x}^{k+1} + Mt^{K-1}\mathbf{x}^K\Big)$ and $\bar{y}^K = \frac{1}{T^K}\sum_{k=0}^{K-1} t^k y^{k+1}$ for $\{t^k\}_{k\geq 0} \subset \mathbb{R}_{++}$ such that $t^k = \sigma^k/\sigma^0$ for $k \geq 0$.

Proof. By employing Lemma 5, we aim to provide a convergence rate analysis for both convex-concave setting, i.e., $\underline{\mu} = 0$, and strongly convex-concave setting, i.e., $\underline{\mu} = 0$ and $L_{yy} = 0$. Suppose the Bregman distance generating functions are set according to Assumption 2. Lemma 7 implies that $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$ sequences generated by RB−APD and RB−APD−B, both satisfy (13c) and (13d) for $t^k = \frac{\sigma^k}{\sigma^0}$ for $k \geq 0$. Thus, one can easily verify that for any $\mathbf{z} \in \mathbf{dom}\, f \times \mathbf{dom}\, h$, we have $t^{k+1}Q^{k+1}(\mathbf{z}) - t^k R^{k+1}(\mathbf{z}) \leq 0$ for all $k \geq 0$.

Now, multiplying both sides of (30) by $t^k = \sigma^k/\sigma^0 > 0$, summing over $k = 0$ to $K-1$, we obtain

$$\begin{aligned} \sum_{k=0}^{K-1} t^k\left(\mathcal{L}(\mathbf{x}^{k+1}, y) - \mathcal{L}(\mathbf{x}, y^{k+1})\right) \leq Q^0(\mathbf{z}) - t^{K-1}R^K(\mathbf{z}) + (M-1)\big(\theta^0 H^0(\mathbf{z}) - t^{K-1}H^K(\mathbf{z})\big) \\ + \sum_{k=0}^{K-1} t^k(M-1)\Big((1-\theta^k)(\mathcal{L}(\mathbf{x}^k, y) - \mathcal{L}(\mathbf{x}, y)) + (1-\theta^k)_+ L_{yy}\mathbf{D}_{\mathcal{Y}}(y, y^k)\Big) + \sum_{k=0}^{K-1} t^k(C_*^k + \mathcal{E}^k(\mathbf{x})), \end{aligned} \tag{74}$$

where we used $t^0 = 1$ and $t^{k+1}\theta^{k+1} = t^k$ for $k \geq 0$. Due to initialization $\mathbf{x}^{-1} = \mathbf{x}^0$, and $y^{-1} = y^0$, the definitions of $Q^k(\cdot)$ and $H^k(\cdot)$ given in (31a) and (31c), respectively, imply that

$$\begin{aligned} Q^0(\mathbf{z}) &= M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^0}(\mathbf{x}, \mathbf{x}^0) + \frac{M-1}{2}\left\|\mathbf{x}-\mathbf{x}^0\right\|_{\mathfrak{M}}^2 + \frac{1}{\sigma^0}\mathbf{D}_{\mathcal{Y}}(y, y^0) + \theta^0(M-1)\left\langle\nabla_y\Phi(\mathbf{x}^0, y^0),\ y - y^0\right\rangle, \\ H^0(\mathbf{z}) &= f(\mathbf{x}^0) - f(\mathbf{x}) + \Phi(\mathbf{x}^0, y^0) - \Phi(x, y). \end{aligned}$$

Using the bound (20), we have that

$$\begin{aligned} H^0(\mathbf{z}) + \left\langle\nabla_y\Phi(\mathbf{x}^0, y^0), y - y^0\right\rangle &\leq f(\mathbf{x}^0) - f(\mathbf{x}) + \Phi(\mathbf{x}^0, y) - \Phi(\mathbf{x}, y) + \frac{L_{yy}}{2}\left\|y - y^0\right\|_{\mathcal{Y}}^2 \\ &\leq \mathcal{L}(\mathbf{x}^0, y) - \mathcal{L}(\mathbf{x}, y) + L_{yy}\mathbf{D}_{\mathcal{Y}}(y, y^0). \end{aligned} \tag{75}$$

Moreover, using concavity of $\Phi(\mathbf{x}, \cdot)$, for any $\mathbf{x} \in \mathcal{X}$, we can find a lower bound on $H^K(\mathbf{z})$,

$$\begin{aligned} H^K(\mathbf{z}) &\geq f(\mathbf{x}^K) - f(\mathbf{x}) + \Phi(\mathbf{x}^K, y) - \Phi(\mathbf{x}, y) + \left\langle\nabla_y\Phi(\mathbf{x}^K, y^K), y^K - y\right\rangle \\ &= \mathcal{L}(\mathbf{x}^K, y) - \mathcal{L}(\mathbf{x}, y) + \left\langle\nabla_y\Phi(\mathbf{x}^K, y^K), y^K - y\right\rangle. \end{aligned} \tag{76}$$

Within $R^K(\mathbf{z})$, there is $\left\langle r^K, y^K - y\right\rangle$ term, where $r^K = \nabla_y\Phi(\mathbf{x}^K, y^K) - M\nabla_y\Phi(\mathbf{x}^{K-1}, y^{K-1})$, and in order to upper

bound the right-hand-side of (74), we first provide an intermediate inequality:

$$
\begin{aligned}
&(M-1)H^K(\mathbf{z}) + \langle r^K, y^K - y\rangle \\
&\quad \geq (M-1)\Big(\mathcal{L}(\mathbf{x}^K, y) - \mathcal{L}(\mathbf{x}, y)\Big) + \langle q^K, y^K - y\rangle \\
&\quad \geq (M-1)\Big(\mathcal{L}(\mathbf{x}^K, y) - \mathcal{L}(\mathbf{x}, y)\Big) - M(\alpha^K + \beta^K)\mathbf{D}_{\mathcal{Y}}(y, y^K) \\
&\quad - \frac{M}{2\alpha^K}\left\|\nabla_y \Phi(\mathbf{x}^K, y^K) - \nabla_y \Phi(\mathbf{x}^{K-1}, y^K)\right\|_{\mathcal{Y}^*}^2 - \frac{M}{2\beta^K}\left\|\nabla_y \Phi(\mathbf{x}^{K-1}, y^K) - \nabla_y \Phi(\mathbf{x}^{K-1}, y^{K-1})\right\|_{\mathcal{Y}^*}^2,
\end{aligned} \tag{77}
$$

where the first inequality follows from (76) and $q^K = M\big(\nabla_y \Phi(\mathbf{x}^K, y^K) - \nabla_y \Phi(\mathbf{x}^{K-1}, y^{K-1})\big)$ and for the second inequality we used (43) to lower bound $\langle q^K, y - y^K\rangle$. Therefore, (74), (75) and (77) together imply that

$$
\begin{aligned}
&\sum_{k=0}^{K-1} t^k\big(\mathcal{L}(\mathbf{x}^{k+1}, y) - \mathcal{L}(\mathbf{x}, y^{k+1})\big) + \sum_{k=1}^{K-1} t^k(M-1)(\theta^k - 1)\big(\mathcal{L}(\mathbf{x}^k, y) - \mathcal{L}(\mathbf{x}, y)\big) \\
&+ (M-1)t^{K-1}\big(\mathcal{L}(\mathbf{x}^K, y) - \mathcal{L}(\mathbf{x}, y)\big) \leq M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^0}(\mathbf{x}, \mathbf{x}^0) + \frac{M-1}{2}\left\|\mathbf{x} - \mathbf{x}^0\right\|_{\mathfrak{M}}^2 \\
&\quad + \big(\frac{1}{\sigma^0} + \theta^0(M-1)L_{yy}\big)\mathbf{D}_{\mathcal{Y}}(y, y^0) + (M-1)\big(\mathcal{L}(\mathbf{x}^0, y) - \mathcal{L}(\mathbf{x}, y)\big) \\
&\quad - t^K\Big(M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^K}(\mathbf{x}, \mathbf{x}^K) + \frac{M-1}{2}\left\|\mathbf{x} - \mathbf{x}^K\right\|_{\mathfrak{M}}^2 + \big(\frac{1}{\sigma^K} - M\theta^K(\alpha^K + \beta^K)\big)\mathbf{D}_{\mathcal{Y}}(y, y^K)\Big) \\
&\quad + \sum_{k=0}^{K-1} t^k(M-1)(1-\theta^k)_+ L_{yy}\mathbf{D}_{\mathcal{Y}}(y, y^k) + \sum_{k=0}^{K-1} t^k(C_*^k + \mathcal{E}^k(\mathbf{x})),
\end{aligned} \tag{78}
$$

where we used the fact that (13c) and (13d) hold for both `RB-APD` and `RB-APD-B`.

Note that the left hand side of (78) can be lower bounded using Jensen's inequality twice:

$$
\begin{aligned}
&\sum_{k=0}^{K-1} t^k\big(\mathcal{L}(\mathbf{x}^{k+1}, y) - \mathcal{L}(\mathbf{x}, y^{k+1})\big) + \sum_{k=1}^{K-1} t^k(M-1)(\theta^k - 1)\big(\mathcal{L}(\mathbf{x}^k, y) - \mathcal{L}(\mathbf{x}, y)\big) \\
&+ (M-1)t^{K-1}\big(\mathcal{L}(\mathbf{x}^K, y) - \mathcal{L}(\mathbf{x}, y)\big) \geq \\
&\quad (T^K + M - 1)\big(\mathcal{L}(\bar{\mathbf{x}}^K, y) - \mathcal{L}(\mathbf{x}, y)\big) + T^K\big(\mathcal{L}(\mathbf{x}, y) - \mathcal{L}(\mathbf{x}, \bar{y}^K)\big),
\end{aligned} \tag{79}
$$

which follows from convexity of $\mathcal{L}(\cdot, y)$ and $-\mathcal{L}(\mathbf{x}, \cdot)$ for every fixed $(\mathbf{x}, y)$. In the above application of Jensen's inequality on $-\mathcal{L}(\mathbf{x}, \cdot)$, the convex combination coefficients are $\{t^k\}_{k=0}^{K-1}$, which satisfy $t^k = \sigma^k/\sigma^0 \geq 0$ and $T^K = \sum_{k=0}^{K-1} t^k$, while in the application of Jensen's inequality on $\mathcal{L}(\cdot, y)$, the convex combination coefficients are $\{Mt^k - (M-1)t^{k+1}\}_{k=0}^{K-1}$ and $Mt^{K-1} \geq 0$ –note that they sum to $T^K + M - 1$, and $Mt^k - (M-1)t^{k+1} \geq 0$ follows from $\theta^{k+1} \geq \frac{M-1}{M}$ for all $k \geq 0$; indeed, we have already argued that when $\underline{\mu} = 0$, $\theta^k \geq 1$ for $k \geq 0$, and when $\underline{\mu} > 0$, (53) shows that $\theta^{k+1} \geq \frac{M-1}{M}$ for all $k \geq 0$.

Note that for `RB-APD` since the parameter choice satisfy (13a) and (13b) with some $\delta \in [0, 1)$, $\alpha^{k+1} = c_\alpha/\sigma^k$ and $\alpha^{k+1} = c_\beta/\sigma^k$ for $k \geq 0$; therefore, Lemma 6 implies that (15) holds with the same $\{\alpha^k, \beta^k\}$ and δ. Moreover, Lemma 8 implies that (15) always holds for `RB-APD-B`. Thus, we have that $C_*^k \leq 0$ for $k \geq 0$; hence, combining (79) with (78) leads to the desired result. □

As we discussed before, we provide a uniform analysis of the rate results for `RB-APD` and `RB-APD-B`. Now we are ready to provide the rate result for **(Part I)** and **(Part II)** of Theorem 1 and 2.

Indeed, Lemma 1 implies that the step-size sequence $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$ selected in `RB-APD-B` algorithm is well-defined satisfying (13c) and (13d) for $\{t^k\}$ such that $t^k = \sigma^k/\sigma^0$ for $k \geq 0$, and $C_*^k \leq 0$ for $k \geq 0$. Next, we show that $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$ selected in `RB-APD` algorithm satisfies Assumption 3. Indeed, since $\theta^k = \sigma^{k-1}/\sigma^k$, for $\alpha^k = c_\alpha/\sigma^{k-1}$ and $\beta^k = c_\beta/\sigma^{k-1}$, (13a) and (13b) can be written as

$$
\frac{1-\delta}{\tau_{i_k}^k} \geq L_{x_{i_k}x_{i_k}} + \frac{L_{yx_{i_k}}^2}{c_\alpha}\sigma^k, \quad 1 - \delta - M(c_\alpha + c_\beta) \geq \frac{ML_{yy}^2}{c_\beta}(\sigma^k)^2. \tag{80}
$$

Clearly, the initial step-sizes selected as in Remark 4.1 implies that (80) holds for $k = 0$. When $\mu = 0$, i.e., (Part I), we have $\gamma^k = \gamma^0$ and $\theta^k = 1$ for $k \geq 0$; hence, $\tilde{\tau}^k = \tilde{\tau}^0$ and $\sigma^k = \sigma^0$ for $k \geq 0$. Thus, (13a) and (13b) hold for all $k \geq 0$ for $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$ produced by RB−APD. For the case $\mu > 0$, i.e., (Part II), we will use induction to show that (80) holds. Recall that for this case, we assume $L_{yy} = 0$; hence, the second condition in (80) holds for any σ^k as long as $1 \geq \delta + M(c_\alpha + c_\beta)$. Now suppose the first condition in (80) holds for some $k \geq 0$, using $\sigma^{k+1} = \sigma^k\sqrt{\gamma^{k+1}/\gamma^k}$ and $\gamma^{k+1}/\gamma^k \geq 1$, we get

$$\begin{aligned}\frac{1-\delta}{\tau_i^{k+1}} &= \frac{1-\delta}{M}\Big(\frac{1}{\tilde{\tau}^{k+1}} - \mu_i\Big) - (1-\delta)\mu_i = \frac{1-\delta}{M\tilde{\tau}^k}\sqrt{\frac{\gamma^{k+1}}{\gamma^k}} - (1-\delta)\Big(1-\frac{1}{M}\Big)\mu_i,\\ &= \frac{1-\delta}{\tilde{\tau}_i^k}\sqrt{\frac{\gamma^{k+1}}{\gamma^k}} + (1-\delta)\left(\sqrt{\frac{\gamma^{k+1}}{\gamma^k}} - 1\right)\Big(1-\frac{1}{M}\Big)\mu_i,\\ &\geq \Big(L_{x_ix_i} + \frac{L_{yx_i}^2}{c_\alpha}\sigma^k\Big)\sqrt{\frac{\gamma^{k+1}}{\gamma^k}} \geq L_{x_ix_i} + \frac{L_{yx_i}^2}{c_\alpha}\sigma^{k+1}, \quad \forall\, i \in \mathcal{M}.\end{aligned}$$

This completes the induction. Moreover, Lemma 7 implies that $\{[\tau_i^k]_{i\in\mathcal{M}}, \sigma^k, \theta^k\}$ generated by RB−APD satisfies (13c) and (13d) for $\{t^k\}$ such that $t^k = \sigma^k/\sigma^0$ for $k \geq 0$. Thus, Assumption 3 holds for $\{\alpha^k, \beta^k, t^k\}_{k\geq 0}$ as in the algorithm.

Proof of Theorem 1 and 2 (Part I). Suppose the Bregman distance generating functions are set according to Assumption 2, and for the results in (Part I), we assume that $Z \triangleq \mathbf{dom}\, f \times \mathbf{dom}\, h$ is a compact set. In order to show the convergence rate for the expected gap, we will use the result in Lemma 10 by taking the supremum over Z, and then computing the expectation of an appropriate upper bound on the supremum with respect to randomness in coordinate selection.

Now, recall the bound (73) established in Lemma 10. Since $\underline{\mu} = 0$, we know that $\{\theta^k\}_{k\geq 0}$ generated by either RB−APD or RB−APD−B both satisfy $\theta^k \geq 1$ for all $k \geq 0$; thus, we have that $(1-\theta^k)_+ L_{yy}\mathbf{D}_{\mathcal{Y}}(y, y^k) = 0$ for all $k \geq 0$. Furthermore, since $\alpha^{k+1} = c_\alpha/\sigma^k$ and $\beta^{k+1} = c_\beta/\sigma^k$ for all $k \geq 0$ for some $c_\alpha, c_\beta \geq 0$ such that $M(c_\alpha + c_\beta) < 1$, we know that $\frac{1}{\sigma^K} - M\theta^K(\alpha^K + \beta^K) \geq 0$. Thus, after dropping the nonpositive terms on the right-hand side of (73), taking supremum of the resulting bound over $\mathbf{z} = (\mathbf{x}, y) \in \mathbf{dom}\, f \times \mathbf{dom}\, h$, we get

$$\mathcal{G}(\bar{\mathbf{z}}^K) \leq \frac{1}{T^K}\Big(\sup_{\mathbf{x}\in\mathbf{dom}\, f} B_1^K(\mathbf{x}) + \sup_{y\in\mathbf{dom}\, h} B_2^K(y)\Big). \tag{81}$$

$$B_1^K(\mathbf{x}) \triangleq M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^0+(1-\frac{1}{M})\mathfrak{M}}(\mathbf{x}, \mathbf{x}^0) + \sum_{k=0}^{K-1} t^k\mathcal{E}^k(\mathbf{x}), \tag{82}$$

$$B_2^K(y) \triangleq (\tfrac{1}{\sigma^0} + \theta^0(M-1)L_{yy})\mathbf{D}_{\mathcal{Y}}(y, y^0) + (M-1)\big(\mathcal{L}(\mathbf{x}^0, y) - \mathcal{L}(\bar{\mathbf{x}}^K, y)\big). \tag{83}$$

Due to compactness of $Z = \mathbf{dom}\, f \times \mathbf{dom}\, h$ and continuity of $\mathcal{L}(\cdot,\cdot)$, there exists $\bar{B}_2 < +\infty$ such that $(M-1)\sup_{y\in\mathbf{dom}\, h}\{\mathcal{L}(\mathbf{x}^0, y) - \mathcal{L}(\bar{\mathbf{x}}^K, y)\} \leq \bar{B}_2$ for all $K \geq 1$. One can easily construct a crude bound: $\bar{B}_2 = (M-1)\sup\{\mathcal{L}(\mathbf{x}^0, y) - \mathcal{L}(\mathbf{x}, y) : (\mathbf{x}, y) \in \mathbf{dom}\, f \times \mathbf{dom}\, h\} < +\infty$. That said, in many practical situations a much tighter bound can be obtained, e.g., in case $\mathcal{L}(\cdot, y)$ is Lipschitz with a uniform constant $L_x > 0$ for all $y \in \mathbf{dom}\, h$, then $\bar{B}_2 = (M-1)L_xD_x$, where $D_x = \sup_{\mathbf{x},\mathbf{x}'\in\mathbf{dom}\, f}\|\mathbf{x} - \mathbf{x}'\|_{\mathcal{X}}$ denotes the diameter of $\mathbf{dom}\, f$. For instance, let f be the indicator function of a compact convex set X, then for every $y \in \mathbf{dom}\, h$, $\mathcal{L}(\cdot, y)$ is indeed Lipschitz with a uniform constant $L_x = \sup\{\|\nabla_{\mathbf{x}}\Phi(\mathbf{x}, y)\|_{\mathcal{X}^*} : \mathbf{x} \in X,\ y \in \mathbf{dom}\, h\} < \infty$ due to continuity of $\nabla_{\mathbf{x}}\Phi$ and compactness of $X \times \mathbf{dom}\, h$. Thus,

$$\exists \bar{B}_2 < +\infty : \quad \sup_{y\in\mathbf{dom}\, h} B_2^K(y) \leq \bar{B}_2 + \Big(\tfrac{1}{\sigma^0} + \theta^0(M-1)L_{yy}\Big)\sup_{y\in\mathbf{dom}\, h}\mathbf{D}_{\mathcal{Y}}(y, y^0). \tag{84}$$

Next, we claim that $\mathbf{E}[\sup_{\mathbf{x}\in\mathbf{dom}\, f} B_1^K(\mathbf{x})]$ can be bounded as follows:

Claim 1: $\exists\ \bar{B}_1 < +\infty$ such that

$$\begin{aligned}
&\mathbf{E}\Big[\sup_{\mathbf{x}\in\mathbf{dom}\, f} B_1^K(\mathbf{x})\Big]\\
&\leq \sup_{\mathbf{x}\in\mathbf{dom}\, f} M\mathbf{D}_{\mathcal{X}}^{(1+\frac{1}{M})\mathbf{T}^0+\mathfrak{M}}(\mathbf{x},\mathbf{x}^0) + \frac{(M-1)L_{\varphi_{\mathcal{X}}}^2}{2}\mathbf{E}\Big[\sum_{k=0}^{K-1} t^k \left\|\tilde{\mathbf{x}}^{k+1}-\mathbf{x}^k\right\|^2_{\mathbf{T}^k+\mathfrak{M}}\Big]\\
&\leq \sup_{\mathbf{x}\in\mathbf{dom}\, f} M\mathbf{D}_{\mathcal{X}}^{(1+\frac{1}{M})\mathbf{T}^0+\mathfrak{M}}(\mathbf{x},\mathbf{x}^0) + \frac{(M-1)L_{\varphi_{\mathcal{X}}}^2}{2}\cdot \bar{B}_1 < +\infty, \quad \forall\ K\geq 1.
\end{aligned}$$

It follows from (49) and (50) that

$$\mathbf{E}^k\big[Mf(\mathbf{x}^{k+1}) - f(\tilde{\mathbf{x}}^{k+1}) - (M-1)f(\mathbf{x}^k)\big] = 0, \tag{85a}$$

$$\mathbf{E}^k\big[\left\langle \nabla_{\mathbf{x}}\Phi(\mathbf{x}^k, y^{k+1}), \tilde{\mathbf{x}}^{k+1} - M\mathbf{x}^{k+1} + (M-1)\mathbf{x}^k\right\rangle\big] = 0, \tag{85b}$$

$$\mathbf{E}^k\big[M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\mathbf{x}^{k+1},\mathbf{x}^k) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k}(\tilde{\mathbf{x}}^{k+1},\mathbf{x}^k)\big] = 0. \tag{85c}$$

For $k\geq 0$, we define

$$\Xi^k(\mathbf{x}) \triangleq M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x},\mathbf{x}^{k+1}) - (M-1)\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x},\mathbf{x}^k) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x},\tilde{\mathbf{x}}^{k+1}); \tag{86}$$

hence, (85) implies that

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\mathbf{dom}\, f} B_1^K(\mathbf{x})\right] = \mathbf{E}\left[\sup_{\mathbf{x}\in\mathbf{dom}\, f}\left\{M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^0+(1-\frac{1}{M})\mathfrak{M}}(\mathbf{x},\mathbf{x}^0) + \sum_{k=0}^{K-1} t^k\Xi^k(\mathbf{x})\right\}\right]. \tag{87}$$

Furthermore, for all $k\geq 0$, we also define

$$\tilde{\Gamma}_1^{k+1} \triangleq (\mathbf{T}^k+\mathfrak{M})(\nabla\varphi_{\mathcal{X}}(\tilde{\mathbf{x}}^{k+1}) - \nabla\varphi_{\mathcal{X}}(\mathbf{x}^k)), \quad \Gamma_1^{k+1} \triangleq (\mathbf{T}^k+\mathfrak{M})(\nabla\varphi_{\mathcal{X}}(\mathbf{x}^{k+1}) - \nabla\varphi_{\mathcal{X}}(\mathbf{x}^k)); \tag{88}$$

hence, from the definition of Bregman distances, we get

$$\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x},\mathbf{x}^k) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x},\tilde{\mathbf{x}}^{k+1}) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\tilde{\mathbf{x}}^{k+1},\mathbf{x}^k) = \left\langle\tilde{\Gamma}_1^{k+1},\ \mathbf{x}-\tilde{\mathbf{x}}^{k+1}\right\rangle, \quad \forall\,\mathbf{x}\in\mathcal{X}, \tag{89}$$

$$\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x},\mathbf{x}^k) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x},\mathbf{x}^{k+1}) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^{k+1},\mathbf{x}^k) = \left\langle\Gamma_1^{k+1},\ \mathbf{x}-\mathbf{x}^{k+1}\right\rangle, \quad \forall\,\mathbf{x}\in\mathcal{X}. \tag{90}$$

Therefore,

$$\begin{aligned}
\Xi^k(\mathbf{x}) &= \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\tilde{\mathbf{x}}^{k+1},\mathbf{x}^k) + \left\langle\tilde{\Gamma}_1^{k+1},\mathbf{x}-\tilde{\mathbf{x}}^{k+1}\right\rangle - M\Big(\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^{k+1},\mathbf{x}^k) + \left\langle\Gamma_1^{k+1},\mathbf{x}-\mathbf{x}^{k+1}\right\rangle\Big)\\
&= M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^k,\mathbf{x}^{k+1}) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^k,\tilde{\mathbf{x}}^{k+1}) + \left\langle\tilde{\Gamma}_1^{k+1} - M\Gamma_1^{k+1},\mathbf{x}-\mathbf{x}^k\right\rangle,
\end{aligned} \tag{91}$$

where we used $\left\langle\tilde{\Gamma}_1^{k+1},\mathbf{x}^k-\tilde{\mathbf{x}}^{k+1}\right\rangle = -\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^k,\tilde{\mathbf{x}}^{k+1}) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\tilde{\mathbf{x}}^{k+1},\mathbf{x}^k)$ and $\left\langle\Gamma_1^{k+1},\mathbf{x}^k-\mathbf{x}^{k+1}\right\rangle = -\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^k,\mathbf{x}^{k+1}) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^{k+1},\mathbf{x}^k)$. Next, we define an auxiliary sequence $\{\mathbf{v}^k\}_{k\geq 0}$ by initializing $\mathbf{v}^0=\mathbf{x}^0\in\mathbf{dom}\, f$, and invoking Lemma 4 with $\boldsymbol{\delta}^k = \tilde{\Gamma}_1^{k+1} - M\Gamma_1^{k+1}$ and $\mathcal{A}=\mathbf{T}^k+\mathfrak{M}$ for all $k\geq 0$. Therefore, (91) implies

$$\begin{aligned}
\Xi^k(\mathbf{x}) \leq\ & M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^k,\mathbf{x}^{k+1}) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^k,\tilde{\mathbf{x}}^{k+1}) + \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x},\mathbf{v}^k) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x},\mathbf{v}^{k+1})\\
&+\left\langle\tilde{\Gamma}_1^{k+1} - M\Gamma_1^{k+1},\mathbf{v}^k-\mathbf{x}^k\right\rangle + \frac{1}{2}\left\|\tilde{\Gamma}^{k+1} - M\Gamma^{k+1}\right\|^2_{*,(\mathbf{T}^k+\mathfrak{M})^{-1}}.
\end{aligned} \tag{92}$$

Thus, (92) immediately implies that

$$\begin{aligned}
\sum_{k=0}^{K-1} t^k\Xi^k(\mathbf{x}) \leq\ & \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^0+\mathfrak{M}}(\mathbf{x},\mathbf{x}^0) + \sum_{k=0}^{K-1} t^k\Big(M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^k,\mathbf{x}^{k+1}) - \mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^k,\tilde{\mathbf{x}}^{k+1})\Big)\\
&+\sum_{k=0}^{K-1} t^k\Big(\left\langle\tilde{\Gamma}_1^{k+1} - M\Gamma_1^{k+1},\mathbf{v}^k-\mathbf{x}^k\right\rangle + \frac{1}{2}\left\|\tilde{\Gamma}^{k+1} - M\Gamma^{k+1}\right\|^2_{*,(\mathbf{T}^k+\mathfrak{M})^{-1}}\Big),
\end{aligned} \tag{93}$$

which follows from $t^k(\mathbf{T}^k+\mathfrak{M}) \succeq t^{k+1}(\mathbf{T}^{k+1}+\mathfrak{M})$ for $k\geq 0$ whenever $\underline{\mu}=0$[10], and in the above inequality, we also used $t^0\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^0+\mathfrak{M}}(\mathbf{x},\mathbf{v}^0)=\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^0+\mathfrak{M}}(\mathbf{x},\mathbf{x}^0)$, and $t^{K-1}\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^{K-1}+\mathfrak{M}}(\mathbf{x},\mathbf{v}^K)\geq 0$ for $\mathbf{x}\in\mathcal{X}$.

Next, similar to (85), one can easily verify that

$$\mathbf{E}^k\big[M\Gamma_1^{k+1}\big]=\tilde{\Gamma}_1^{k+1},\quad \mathbf{E}^k\big[\|M\Gamma_1^{k+1}\|^2_{*,(\mathbf{T}^k+\mathfrak{M})^{-1}}\big]=M\|\tilde{\Gamma}_1^{k+1}\|^2_{*,(\mathbf{T}^k+\mathfrak{M})^{-1}}, \tag{94}$$

which implies that

$$\begin{aligned}\mathbf{E}^k[\|\tilde{\Gamma}_1^{k+1}-M\Gamma_1^{k+1}\|^2_{*,(\mathbf{T}^k+\mathfrak{M})^{-1}}] &= (M-1)\left\|\nabla\varphi_{\mathcal{X}}(\tilde{\mathbf{x}}^{k+1})-\nabla\varphi_{\mathcal{X}}(\mathbf{x}^k)\right\|^2_{*,(\mathbf{T}^k+\mathfrak{M})}\\ &\leq (M-1)L^2_{\varphi_{\mathcal{X}}}\left\|\tilde{\mathbf{x}}^{k+1}-\mathbf{x}^k\right\|^2_{\mathbf{T}^k+\mathfrak{M}},\end{aligned} \tag{95}$$

where in the equality we used (94) and the identity $\mathbf{E}[\|X-\mathbf{E}[X]\|_2^2]=\mathbf{E}[\|X\|_2^2]-\|\mathbf{E}[X]\|_2^2$ holding for any random variable X; and for the inequality above, we used the Lipschitz continuity of $\nabla\varphi_{\mathcal{X}}(\cdot)$. Furthermore, for $k\geq 0$, one also has

$$\mathbf{E}^k[M\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^k,\mathbf{x}^{k+1})-\mathbf{D}_{\mathcal{X}}^{\mathbf{T}^k+\mathfrak{M}}(\mathbf{x}^k,\tilde{\mathbf{x}}^{k+1})]=0, \tag{96a}$$

$$\mathbf{E}^k[\left\langle\tilde{\Gamma}_1^{k+1}-M\Gamma_1^{k+1},\mathbf{v}^k-\mathbf{x}^k\right\rangle]=0, \tag{96b}$$

where the first one follows from the same arguments we used for showing (85c), and the second one follows from (94).

Finally, (87), (93), (95) and (96) together with the tower property of expectation imply that

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\mathbf{dom}\,f}B_1^K(\mathbf{x})\right]\leq\sup_{\mathbf{x}\in\mathbf{dom}\,f}M\mathbf{D}_{\mathcal{X}}^{(1+\frac{1}{M})\mathbf{T}^0+\mathfrak{M}}(\mathbf{x},\mathbf{x}^0)+\frac{(M-1)L^2_{\varphi_{\mathcal{X}}}}{2}\mathbf{E}\left[\sum_{k=0}^{K-1}t^k\left\|\tilde{\mathbf{x}}^{k+1}-\mathbf{x}^k\right\|^2_{\mathbf{T}^k+\mathfrak{M}}\right].$$

Note that (62), (68) and (70) imply that

$$\sum_{k=0}^{K}t^k\left\|\tilde{\mathbf{x}}^{k+1}-\mathbf{x}^k\right\|^2_{\mathbf{T}^k+\mathfrak{M}}\to\sum_{k=0}^{+\infty}t^k\left\|\tilde{\mathbf{x}}^{k+1}-\mathbf{x}^k\right\|^2_{\mathbf{T}^k+\mathfrak{M}}<+\infty\quad a.s.\quad K\to+\infty.$$

Since we assume $\mathbf{dom}\,f$ is compact, Lebesgue's dominated convergence theorem implies that

$$\bar{B}_1\triangleq\frac{(M-1)L^2_{\varphi_{\mathcal{X}}}}{2}\mathbf{E}\left[\sum_{k=0}^{+\infty}t^k\left\|\tilde{\mathbf{x}}^{k+1}-\mathbf{x}^k\right\|^2_{\mathbf{T}^k+\mathfrak{M}}\right]<+\infty. \tag{97}$$

Thus, the uniform bound in (87) implies that

$$\mathbf{E}\left[\sup_{\mathbf{x}\in\mathbf{dom}\,f}B_1^K(\mathbf{x})\right]\leq\bar{B}_1+\sup_{\mathbf{x}\in\mathbf{dom}\,f}M\mathbf{D}_{\mathcal{X}}^{(1+\frac{1}{M})\mathbf{T}^0+\mathfrak{M}}(\mathbf{x},\mathbf{x}^0),\quad\forall\,K\geq 1. \tag{98}$$

This completes the proof of **Claim 1**. Therefore, the result in (26a) can be deduced from (81), (84), and **Claim 1**. Furthermore, since $\sigma^k=\gamma^0\tilde{\tau}^k$ for $k\geq 0$, we conclude that $T^K=\sum_{k=0}^{K-1}\sigma^k/\sigma^0\geq\frac{\eta\Psi}{\tilde{\tau}^0}K$. □

Proof of Theorem 1 and 2 (Part II). Let $(\mathbf{x}^*,y^*)$ be a sadde point of $\mathcal{L}$. In strongly convex-concave setting, i.e., $\underline{\mu}>0$, we assume that $L_{yy}=0$ and $\varphi_{\mathcal{X}}(\cdot)=\frac{1}{2}\|\cdot\|^2$. When $L_{yy}=0$, defining $0^2/0=0$, one-step result in Lemma 5 continues to holds with $\beta^k=0$ for all $k\geq 0$. Thus, consider setting $\alpha^k=c_\alpha/\sigma^{k-1}$ for $k\geq 0$ for some $c_\alpha>0$ and $\delta\in[0,1)$ such that $Mc_\alpha+\delta\leq 1$. Therefore, evaluating the result of Lemma 10 given in (73) at $\mathbf{x}=\mathbf{x}^*$, and substituting $L_{yy}=0$ and $\alpha^k=c_\alpha/\sigma^{k-1}$ for $k\geq 0$, we get

$$\begin{aligned}T^K\big(\mathcal{L}(\bar{\mathbf{x}}^K,y)-\mathcal{L}(\mathbf{x}^*,\bar{y}^K)\big)&+(M-1)\big(\mathcal{L}(\bar{\mathbf{x}}^K,y)-\mathcal{L}(\mathbf{x}^*,y)\big)\\ &\leq\tfrac{M}{2}\left\|\mathbf{x}^*-\mathbf{x}^0\right\|^2_{\mathbf{T}^0+(1-\frac{1}{M})\mathfrak{M}}+\tfrac{1}{\sigma^0}\mathbf{D}_{\mathcal{Y}}(y,y^0)+(M-1)\Big(\mathcal{L}(\mathbf{x}^0,y)-\mathcal{L}(\mathbf{x}^*,y)\Big)\\ &+\sum_{k=0}^{K-1}t^k\mathcal{E}^k(\mathbf{x}^*)-t^K\left(\tfrac{M}{2}\left\|\mathbf{x}^*-\mathbf{x}^K\right\|^2_{\mathbf{T}^K+(1-\frac{1}{M})\mathfrak{M}}+\tfrac{1}{\sigma^K}\big(1-Mc_\alpha\big)\mathbf{D}_{\mathcal{Y}}(y,y^K)\right).\end{aligned} \tag{99}$$

[10]For $i\in\mathcal{M}$ and $k\geq 0$, $\frac{1}{\tau_i^k}+\mu_i=\frac{1}{M}\Big(\frac{1}{\tilde{\tau}^k}+\mu_i\Big)$; hence, $t^k(\frac{1}{\tau_i^k}+\mu_i)\geq t^{k+1}(\frac{1}{\tau_i^{k+1}}+\mu_i)$ is equivalent to $\frac{1}{\tilde{\tau}^k}+\mu_i\geq\frac{t^{k+1}}{t^k}\Big(\frac{1}{\tilde{\tau}^{k+1}}+\mu_i\Big)=\frac{\gamma^{k+1}}{\gamma^k}\Big(\frac{1}{\tilde{\tau}^k}+\mu_i\frac{\tilde{\tau}^{k+1}}{\tilde{\tau}^k}\Big)$, which clearly holds because $\underline{\mu}=0$ implies $\gamma^{k+1}=\gamma^k$ and we also have $\tilde{\tau}^{k+1}\leq\tilde{\tau}^k$, for all $k\geq 0$.

Recall that (85) implies $\mathbf{E}^k[\mathcal{E}(\mathbf{x}^*)] = 0$. Furthermore, since $(\mathbf{x}^*, y^*)$ is a saddle point, we have $\mathcal{L}(\mathbf{x}^*, y^*) \geq \mathcal{L}(\mathbf{x}^*, y)$ and $\mathcal{L}(\mathbf{x}, y^*) \geq \mathcal{L}(\mathbf{x}^*, y^*)$ for any $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$. Therefore, when we substitute $y = y^*$ in (99), the left-hand side is non-negative, and we get the following inequality:

$$\begin{aligned} t^K \mathbf{E}\left[\tfrac{M}{2}\left\|\mathbf{x}^* - \mathbf{x}^K\right\|^2_{\mathbf{T}^K + (1-\frac{1}{M})\mathfrak{M}} + \tfrac{1}{\sigma^K}\big(1 - Mc_\alpha\big)\mathbf{D}_{\mathcal{Y}}(y^*, y^K)\right] \\ \leq \tfrac{M}{2}\left\|\mathbf{x}^* - \mathbf{x}^0\right\|^2_{\mathbf{T}^0 + (1-\frac{1}{M})\mathfrak{M}} + \tfrac{1}{\sigma^0}\mathbf{D}_{\mathcal{Y}}(y^*, y^0) + (M-1)\Big(\mathcal{L}(\mathbf{x}^0, y^*) - \mathcal{L}(\mathbf{x}^*, y^*)\Big). \end{aligned} \tag{100}$$

According to `RB-APD` and `RB-APD-B`, we have $M\mathbf{T}^k + (M-1)\mathfrak{M} = \frac{1}{\tilde{\tau}^k}\mathbf{I}_m$ for $k \geq 0$; hence,

$$t^k(M\mathbf{T}^k + (M-1)\mathfrak{M}) = \frac{\sigma^k}{\sigma^0}\frac{1}{\tilde{\tau}^k}\mathbf{I}_m = \frac{\gamma^k}{\sigma^0}\mathbf{I}_m, \quad \forall\, k \geq 0,$$

which follows from $\sigma^k = \gamma^k \tilde{\tau}^k$ for $k \geq 0$. Thus, (100) leads to the desired result in (27).

Next, we will show the convergence rate in terms of the primal objective function value of the ergodic primal iterate sequence. Let $y^*(\mathbf{x}) = \operatorname{argmax}_{y \in \mathcal{Y}} \mathcal{L}(\mathbf{x}, y)$ be the unique maximizer for any given $\mathbf{x} \in \mathbf{dom}\, f$. After dropping non-positive terms from the right-hand side of (99), substituting $y = y^*(\bar{\mathbf{x}}^K)$ and taking the expectation of both sides, we get

$$\begin{aligned} T^K \mathbf{E}\left[\mathcal{L}(\bar{\mathbf{x}}^K, y^*(\bar{\mathbf{x}}^K)) - \mathcal{L}(\mathbf{x}^*, \bar{y}^K)\right] \\ \leq \tfrac{\gamma^0}{2\sigma^0}\left\|\mathbf{x}^* - \mathbf{x}^0\right\|^2_{\mathcal{X}} + \tfrac{1}{\sigma^0}\sup_{y \in \mathbf{dom}\, h} \mathbf{D}_{\mathcal{Y}}(y, y^0) + (M-1)\mathbf{E}\Big[\mathcal{L}(\mathbf{x}^0, y^*(\bar{\mathbf{x}}^K)) - \mathcal{L}(\bar{\mathbf{x}}^k, y^*(\bar{\mathbf{x}}^K))\Big], \end{aligned} \tag{101}$$

where we used $(M\mathbf{T}^0 + (M-1)\mathfrak{M}) = \frac{\gamma^0}{\sigma^0}\mathbf{I}_m$. Note that $\mathcal{L}(\mathbf{x}^0, y^*(\bar{\mathbf{x}}^K)) - \mathcal{L}(\bar{\mathbf{x}}^k, y^*(\bar{\mathbf{x}}^K)) \leq F(\mathbf{x}^0) - F(\bar{\mathbf{x}}^k) \leq F(\mathbf{x}^0) - F(\mathbf{x}^*)$ and $F(\bar{\mathbf{x}}^K) - F(\mathbf{x}^*) \leq \mathcal{L}(\bar{\mathbf{x}}^K, y^*(\bar{\mathbf{x}}^K)) - \mathcal{L}(\mathbf{x}^*, \bar{y}^K)$ w.p. 1. Therefore, the result in (28) can be concluded immediately.

Finally, $\mathcal{O}(1/K^2)$ rate for both (27) and (28) follows from Lemma 9, which implies that $\gamma^K = \sigma^K/\tilde{\tau}^K = \Omega(K^2)$ and $T^K = \sum_{k=1}^K \sigma^k/\sigma^0 = \Omega(K^2)$. □