A stochastic variance-reduced accelerated primal-dual method for finite-sum saddle-point problems

Erfan Yazdandoost Hamedani1 · Afrooz Jalilzadeh1

Received: 1 June 2021 / Accepted: 3 March 2023 / Published online: 10 April 2023 © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

In this paper, we propose a variance-reduced primal-dual algorithm with Bregman distance functions for solving convex-concave saddle-point problems with finite-sum structure and nonbilinear coupling function. This type of problem typically arises in machine learning and game theory. Based on some standard assumptions, the algorithm is proved to converge with oracle complexities of $O\left(\sqrt{\frac{n}{\epsilon}}\right)$ and $O\left(\frac{n}{\sqrt{\epsilon}} + \frac{1}{\epsilon^{1.5}}\right)$ using constant and non-constant parameters, respectively where $n$ is the number of function components. Compared with existing methods, our framework yields a significant improvement over the number of required primal-dual gradient samples to achieve $\epsilon$-accuracy of the primal-dual gap. We also present numerical experiments to showcase the superior performance of our method compared with state-of-the-art methods.

Keywords Primal-dual · Saddle-point · First-order method · Variance reduction

1 Introduction

Let $(\mathcal{X}, \|\cdot\|_\mathcal{X})$ and $(\mathcal{Y}, \|\cdot\|_\mathcal{Y})$ be finite-dimensional normed vector spaces, with dual spaces $(\mathcal{X}^*, \|\cdot\|_{\mathcal{X}^*})$ and $(\mathcal{Y}^*, \|\cdot\|_{\mathcal{Y}^*})$, respectively, and $\mathcal{N} \triangleq \{1, \ldots, n\}$. We consider the following convex-concave saddle-point (SP) problem
\[
\min_{x \in X} \max_{y \in Y} \mathcal{L}(x, y) \triangleq f(x) + \Phi(x, y) - h(y),
\]  

(1)

where \( X \subseteq \mathcal{X} \) and \( Y \subseteq \mathcal{Y} \) are nonempty, closed, and convex sets; \( f : \mathcal{X} \to \mathbb{R} \triangleq \mathbb{R} \cup \{+\infty\} \) and \( h : \mathcal{Y} \to \mathbb{R} \) are convex, closed, and proper functions (possibly non-smooth); moreover, \( \Phi(x, y) \triangleq \frac{1}{n} \sum_{i \in \mathcal{N}} \Phi_i(x, y) \) is a convex-concave function, i.e., 
\( \Phi(\cdot, y) \) is convex for any \( y \in \mathcal{Y} \) and \( \Phi(x, \cdot) \) is concave for any \( x \in \mathcal{X} \), and \( \Phi_i \) satisfies certain differentiability assumptions for any \( i \in \mathcal{N} \) – see Assumption 1.

We are interested in designing an efficient algorithm for solving (1) which emerges in machine learning and data science problems when \( n \) is large. There has been a lot of efforts to solve large-scale optimization problems efficiently using different approaches, e.g., variance reduction and block-coordinate schemes. Specifically, when the objective has a finite-sum structure, variance reduction schemes close the oracle complexity gap between the deterministic and stochastic settings by providing an unbiased estimator of gradients reducing the variance of the error of gradient estimator, e.g., SAG [28], SVRG [16], SAGA [7].

On the other hand, with the emerging complexities arising in different areas, SP problems are becoming more popular, and various methods have been introduced to solve such problems. Unlike minimization problems, the number of methods solving large-scale SP problems with finite-sum structure is limited, most of which only consider the strongly convex-strongly concave setting where the convergence result is obtained in terms of the distance from the unique solution. In fact, the main challenge in convex-concave setting is to provide a controllable bound for the variance of gradient estimation. In a convex minimization setup, this can be achieved by utilizing convexity and smoothness of the objective function, however, in SP problems due to a nonbilinear structure of the objective function such an approach is not applicable.

To this end, our goal in this paper is to introduce an SVRG-type variance reduction technique for primal-dual algorithms with Bregman distance for solving convex-concave SP problem (1).

### 1.1 Notations

Let \( \mathbb{S}_+^n \) (\( \mathbb{S}_{++}^n \)) be the set of \( n \times n \) symmetric positive (semi-) definite matrices, \( \mathbf{I}_n \) denotes the \( n \times n \) identity matrix, and \( \mathbf{1}_n \) denotes an \( n \)-dimensional vector of ones. \( \mathbb{E}[\cdot] \) denotes the expectation operation and \( \tilde{O}(\cdot) \) denotes \( O(\cdot) \) up to a logarithmic factor. \( \|\cdot\|_2 \) denotes the Euclidean norm.

### 1.2 Applications

There is a wide range of real-life problems arising in machine learning, image processing, game theory, etc. such that they can be formulated as a special case of (1). We briefly introduce some of the interesting examples below.

**I. Distributionally robust optimization (DRO):** Let \( (\Omega, \mathcal{F}, \mathbb{P}) \) be a probability space where \( \Omega = \{\zeta_1, \ldots, \zeta_n\} \), \( \ell : X \times \Omega \to \mathbb{R} \) is a convex loss function, and we define \( \ell_i(u) \triangleq \ell(u; \zeta_i) \). DRO studies worse case performance under uncertainty to find
solutions with some specific confidence level [23]. This problem can be formulated as follows:

\[
\min_{u \in \mathcal{X}} \max_{y \in \mathcal{P}} \mathbb{E}_{\zeta \sim \mathcal{P}}[\ell(u; \zeta)] = \sum_{i=1}^{n} y_i \ell_i(u),
\]

(2)

where \( \mathcal{P} \) represents the uncertainty set. For instance, \( \mathcal{P} = \{ y \in \Delta_n : V(y, \frac{1}{n} 1_n) \leq \rho \} \) is an uncertainty set considered in different papers such as [23], where \( \Delta_n \) is an n-dimensional simplex set, and \( V(Q, P) \) denotes the divergence measure between two sets of probability measures \( Q \) and \( P \). Assuming \( V(y, \frac{1}{n} 1_n) = \sum_{i=1}^{n} V_i(y_i, \frac{1}{n}) \) and using a variable \( \lambda \in \mathbb{R}_+ \) we can relax the divergence constraint in (2) to obtain the following equivalent problem:

\[
\min_{u \in \mathcal{X}} \max_{y \in \Delta_n, \lambda \geq 0} \sum_{i=1}^{n} \left( y_i \ell_i(u) - \frac{\lambda}{n} \left( V_i(y_i, \frac{1}{n}) - \frac{\rho}{n} \right) \right),
\]

(3)

Let \( x = [u^\top \lambda]^\top \), (3) is a special case of (1) by defining \( \Phi_i(x, y) \triangleq n y_i \ell_i(u) - \lambda \left( \frac{1}{2} V_i(y_i, \frac{1}{n}) - \frac{\rho}{n} \right) \), \( f(x) = I_{\mathcal{X} \times \mathbb{R}_+}(u, \lambda) \), and \( h(y) = I_{\Delta_n}(y) \).

**II. Learning a kernel matrix:** Suppose we are given a set of labeled data points consisting of feature vectors \([a_i]_{i=1}^n \subset \mathbb{R}^m\), and the corresponding labels \([b_i]_{i=1}^n \subset \{-1, +1\}\). Consider \( N \) different embedding of the data and let \( K_i \in \mathbb{S}_+^n \) be the corresponding kernel matrix. The objective is to learn a kernel matrix \( K \) belonging to a class of kernel matrices which is a convex set generated by \([K_i]_{i=1}^M\), i.e. \( K \in \mathcal{K} \triangleq \{ \sum_{i=1}^{M} y_i K_i : y_i \geq 0, i = 1, \ldots, M \} \), such that it minimizes the training error of an \( \ell_2 \)-norm soft-margin nonlinear SVM as a function of \( K \) – see [20] for more details. Then one needs to solve the following problem:

\[
\min_{y \in \mathbb{R}_+^M} \max_{x : \langle \mathbf{r}, y \rangle \leq c, \langle \mathbf{b}, x \rangle = 0} 2x^\top 1_n - \sum_{i=1}^{M} y_i x^\top H(K_i)x - \lambda \|x\|^2_2,
\]

(4)

where \( c, C \geq 0 \) and \( \lambda \geq 0 \) are model parameters, \( y = [y_i]_{i=1}^M \), \( \mathbf{r} = [\text{trace}(K_i)]_{i=1}^M \), \( \mathbf{b} = [b_i]_{i=1}^M \) and \( H(K_i) \triangleq \text{diag}(\mathbf{b}) K_i \text{diag}(\mathbf{b}) \). Clearly (4) has a finite sum objective and is a special case of (1).

**III. Two-player zero-sum game with a nonlinear payoff:** This problem arising in game theory, considers computing the equilibrium of a convex-concave two-player game of the following form:

\[
\begin{aligned}
\min_{\mathbf{q} = [q_i]_{i=1}^N} \quad & \max_{\mathbf{p} = [p_i]_{i=1}^N} \quad \sum_{i=1}^{N} \log \left( 1 + \frac{\beta_i p_i}{\sigma_i + q_i} \right) + \sum_{j=1}^{M} \log(1 + \exp(\langle u_j, \mathbf{q} \rangle)) \\
\langle 1_N, \mathbf{q} \rangle = Q & \quad \langle 1_N, \mathbf{p} \rangle = P \\
- \sum_{j=1}^{M} \log(1 + \exp(\langle v_j, \mathbf{p} \rangle)).
\end{aligned}
\]

(5)
The problem in (5) includes some interesting special cases such as the water filling problem arising in information theory (see [3, 5]) when \( u_j = v_j = 0 \), for all \( j \in \{1, \ldots, M\} \). In particular, consider \( N \) given Gaussian communication channels each having signal power \( p_i \) and noise power \( q_i \), for \( i \in \{1, \ldots, N\} \). From Shannon-Hartley equation, the maximum capacity of channel \( i \) is proportional to \( \log(1 + \beta_i p_i \sigma_i + q_i) \) where \( \sigma_i \) is the receiver noise and \( \beta_i > 0 \) is a constant. The goal is to maximize the total capacity given total power \( P > 0 \) while an adversary aims to reduce the total capacity given total noise power \( Q > 0 \). Therefore, we aim to allocate the signal power such that with the worst allocation of noise power, we obtain the largest total capacity for the system.

1.3 Related work

SP problems have become increasingly popular in recent years due to their applicability for solving a wider range of problems. There have been several studies on deterministic first-order primal-dual algorithms for solving (1) when \( \Phi(x, y) \) is bilinear, i.e., \( \Phi(x, y) = \langle Ax, y \rangle \), such as [4, 6, 9, 13, 30], and few others have considered a more general non-bilinear setting [12, 18, 19, 24, 27] in which an optimal rate of \( \mathcal{O}(1/\epsilon) \) has been shown for the convex-concave setting. However, when considering problem (1), this rate is directly affected by the number of function components; hence, the oracle complexity (number of primal-dual sample gradients) is \( \mathcal{O}(n/\epsilon^2) \) which requires a high computational effort for large-scale problems, i.e., \( n \) is large.

Many stochastic primal-dual algorithms have been introduced in different studies with the aim of addressing more general problems having an expectation in the objective function, and achieving lower per iteration complexity. However, this comes at the cost of dropping the oracle complexity to \( \mathcal{O}(1/\epsilon^2) \) for convex-concave setting – see [17, 25, 33]. After introducing variance reduction techniques for minimization optimization problems, different attempts have been made to adopt such techniques in primal-dual algorithms for solving (1) or its special case when \( \Phi \) is linear in \( y \); however, most of these studies only focused on strongly-convex strongly-concave setting such as [8, 21, 27, 32], and few others [14, 31] consider a more general setting which we will briefly describe next.

In [14], a randomized primal-dual smoothing technique has been proposed for solving (1) by inexactly solving a sequence of subproblems when the \( \Phi(x, y) \) has a finite-sum structure. Assumming that each \( \Phi_i(x, y) \) is smooth and convex-concave, and \( f \) is strongly convex primal and dual oracle complexities of \( \tilde{\mathcal{O}}((n + \sqrt{n\kappa})/\sqrt{\epsilon}) \) and \( \tilde{\mathcal{O}}(\sqrt{n}/\epsilon) \), where \( \kappa \) denotes the condition number, have been shown, respectively. In [31] the problem of \( \min_{x \in X} \max_{y \in Y} f(x) + y^T g(x) - h(x) \) has been considered. This problem can be also equivalently written as \( \min_{x \in X} P(x) \) where \( P(x) = h^*(g(x)) + f(x) \) and \( h^*(\cdot) \) denotes the convex conjugate of function \( h(\cdot) \). They proposed a restarted stochastic primal-dual algorithm (RSPD) in which noisy partial primal and dual (sub)gradients are used. The algorithm is restarted periodically and in the outer loop \( \arg\max_{y \in \text{dom} h^*} \mathcal{L}(x, y) \) for some given \( x \) is required to be computed efficiently; however, this operation might be computationally expensive for a more general problem (1). Assuming that the partial (sub)gradients are bounded, \( h^* \) follows
the Hölder condition with constants \((L, \nu)\), \(g(\cdot)\) is Lipschitz continuous, and \(P(\cdot)\) satisfies lower error bound with parameter \(\theta \geq 0\), i.e., \(\text{dist}(x, X^*) \leq c(P(x) - P(x^*))^{\theta}\) for some \(c > 0\) where \(X^*\) denotes the optimal set, they demonstrated an oracle complexity of \(O(\frac{1}{e^{\theta(1-\nu)}})\). Note that if \(P(\cdot)\) does not obey the lower error bound condition, i.e., \(\theta = 0\), then their oracle complexity is \(O(\frac{1}{e^{2}})\). In our recent study [15] we consider problem (1) where \(x\) and \(y\) are assumed to have \(M\) and \(N\) blocks, respectively and the problem has a coordinate-friendly structure. A doubly stochastic block-coordinate primal-dual algorithm has been proposed for deterministic and stochastic settings in which at each iteration only one block of \(x\) and \(y\) are updated. Assuming that the partial gradients of \(\Phi\) are bounded, oracle complexities of \(O(MN/e)\) for deterministic and \(\tilde{O}((MN)^2/e^2)\) for stochastic settings have been achieved.

In contrast to the existing studies mentioned above, we aim to obtain an improved oracle complexity under weaker assumptions for problem (1) where \(\Phi\) is convex-concave and is neither linear in \(x\) nor in \(y\).

1.4 Contribution

We study SP problem (1) with a finite-sum structure where the coupling function \(\Phi\) is not linear in \(x\) nor in \(y\). We develop a stochastic variance-reduced accelerated primal-dual algorithm (SVR-APD) with Bregman distance which is a novel SVRG-type primal-dual algorithm, for solving this problem. Our idea is to consider a new momentum which is a convex combination of the current iterate point and the average of past iterates. This idea combined with an acceleration in terms of partial gradients of \(\Phi\) leads to convergence guarantees in terms of the standard gap function \(E[\sup_{(x, y) \in X \times Y} \mathcal{L}(x^{(k)}, y) - \mathcal{L}(x, y^{(k)})]\) where \(x^{(k)}\) and \(y^{(k)}\) denote the ergodic average of the iterates. More precisely, we demonstrate the oracle complexities of \(O(\sqrt{n}/e)\) and \(O(\frac{n}{\sqrt{e}} + \frac{1}{e^{1.5}})\) using constant and non-constant parameters, respectively.

Comparing to deterministic methods, our oracle complexity of \(O(\sqrt{n}/e)\) shows a clear improvement in the order of \(\sqrt{n}\) magnitude and it has a lower complexity in comparison with \(O(1/e^2)\) for stochastic methods when \(e \leq O(\frac{1}{\sqrt{n}})\) – see Table 1. Moreover, our second result with non-constant parameters leads to an oracle complexity of \(O\left(\frac{n}{\sqrt{e}} + \frac{1}{e^{1.5}}\right)\), which is lower than the stochastic schemes and for \(e \geq O(\frac{1}{n^{1.5}})\) is lower than the deterministic counterpart. Finally, comparing our results, selecting
constant parameters leads to a better complexity for a medium to high accuracy of $\epsilon \leq O\left(\frac{1}{n}\right)$.

Furthermore, we were able to incorporate Bregman-distance functions in the proximal step of an SVRG-type method for convex-concave setting for the first time. The Bregman distance function generalizes the Euclidean distance by providing a significant flexibility. Indeed, it facilitates the computation of the proximal mapping. For example, when the constraint set is a simplex set a closed form solution for the projection can be computed using the entropy-distance function rather than projecting onto the set in the Euclidean space.

1.5 Organization of the paper

In the next section, we precisely state our assumptions, describe the proposed SVRG-type algorithm, and present the oracle complexities of our method under different choices of parameters which are the main results of this paper. Subsequently, in Sect. 4, we provide a convergence analysis proving the main results. Later, in Sect. 5, we apply our SVR-APD method to solve a DRO problem and compare it with competitive methods.

2 Proposed method

For the optimization problem of $\min_{x \in X} f(x) + \frac{1}{n} \sum_{i=1}^{n} g_i(x)$, variants of SVRG method have been developed in [1] when the objective function is merely convex. The main idea is to keep a full gradient at $\tilde{x}^{k-1}$ in the outer loop and use it to provide an estimate of the full gradient, $\xi_t$, such that a tight upper bound on $\|\xi_t - \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\tilde{x}^{k-1})\|$ can be obtained – see [1, Lemma A.2]. However, such an upper bound cannot be obtained for general saddle-point problems. In this section, we propose the Stochastic Variance-reduced Accelerated Primal-dual (SVR-APD) algorithm displayed in Algorithm 1. Our novel idea to resolve this issue is to consider a combination of an iterate with an average of the last loop (see lines 9 and 14) which will help us with controlling the bounds for the error of gradient estimates – see Lemma 2. Moreover, such a technique combined with the average of iterates evaluated at distance generating functions (see line 19) enable us to use the Bregman distance functions in the proposed method which provides a significant flexibility and facilitates the computation of the proximal mapping.

Next, we provide the definitions related to the Bregman distance function and the gap function.

**Definition 1** Let $\psi_X : X \to \mathbb{R}$ and $\psi_Y : Y \to \mathbb{R}$ be continuously differentiable functions on $\text{int(dom } f)$ and $\text{int(dom } h)$, respectively. Moreover, $\psi_X$ and $\psi_Y$ are 1-strongly convex with respect to $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively. We define the Bregman distance function corresponding to the distance generating function $\psi_X$ as $D_X(x, \bar{x}) \triangleq \psi_X(x) - \psi_X(\bar{x}) - \langle \nabla \psi_X(\bar{x}), x - \bar{x} \rangle$, for all $x \in X$ and $\bar{x} \in X^o \triangleq X \cap \text{int(dom } f)$. Similarly we define $D_Y(y, \bar{y}) \triangleq \psi_Y(y) - \psi_Y(\bar{y}) - \langle \nabla \psi_Y(\bar{y}), y - \bar{y} \rangle$, for all $y \in Y$ and $\bar{y} \in Y^o \triangleq Y \cap \text{int(dom } h)$. Moreover, we define the Bregman diameters of $X$ and
Y under $\psi_{\mathcal{X}}$ and $\psi_{\mathcal{Y}}$ as $B_X$ and $B_Y$, respectively, i.e., $B_X \triangleq \sup_{x \in X, \bar{x} \in X^o} D_X (x, \bar{x})$ and $B_Y \triangleq \sup_{y \in Y, \bar{y} \in Y^o} D_Y (y, \bar{y})$.

**Definition 2** For a given pair $(\bar{x}, \bar{y}) \in \mathcal{X} \times \mathcal{Y}$ and a convex compact set $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$, we define the gap function $G_{\mathcal{Z}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ as $G_{\mathcal{Z}} (\bar{x}, \bar{y}) \triangleq \sup_{(x, y) \in \mathcal{Z}} \{ L (\bar{x}, y) - L (x, \bar{y}) \}$.

Moreover, each iteration of the proposed method involves a carefully designed sample gradient approximations to control the corresponding variance. Therefore, we define the conditional expectation and error of estimating sample gradients as follows.

**Definition 3** We denote expectation and conditional expectation with respect to $\mathcal{F}_t^k$ by $\mathbb{E} [\cdot]$ and $\mathbb{E} [\cdot | \mathcal{F}_t^k]$, respectively, such that $\mathcal{F}_t^k \triangleq \sigma (\Psi_t^k)$ where $\sigma (\cdot)$ denotes $\sigma$-algebra and $\Psi_t^k \triangleq \{ j_0^k, i_0^k, \ldots, j_{t-1}^k, i_{t-1}^k \}$. Similarly we define $\mathcal{H}_t^k \triangleq \sigma (\Phi_t^k)$ where $\Phi_t^k \triangleq \{ j_0^k, i_0^k, \ldots, j_{t-1}^k, i_{t-1}^k \}$ and let $\mathbb{E} [\cdot] \triangleq \mathbb{E} [\cdot | \mathcal{F}_t^k]$ and $\mathbb{E} [\cdot | \mathcal{H}_t^k]$.

**Definition 4** Let $\phi_t^k \triangleq \nabla_x \Phi_{j_t^k} (x_{t-1}^k, y_{t-1}^k) - \nabla_y \Phi_{j_t^k} (x_{t-1}^k, y_{t-1}^k)$, $\bar{\phi}_t^k \triangleq \nabla_y \Phi (x_t^k, y_t^k) - \nabla_y \Phi (x_{t-1}^k, y_{t-1}^k)$, $\delta_t^x \triangleq \xi_t - \nabla_x \Phi (x_t^k, y_t^k)$, and $\delta_t^y \triangleq \xi_t - \nabla_x \Phi (x_t^k, y_t^k)$.

Now we consider the following standard assumptions required to achieve the rate result in this paper.

**Assumption 1** Let $D_X$ and $D_Y$ be some Bregman distance functions as in Definition 1; $f$ and $h$ are closed convex functions; and $\Phi$ is continuously differentiable such that

**(i)** for any $x \in \text{dom} \ f \subseteq \mathcal{X}$, $\Phi (\cdot, y)$ is convex; for any $i \in \mathcal{N}$ $\Phi_i (\cdot, y)$ is differentiable; there exist $L_{xx} \geq 0$ and $L_{xy} > 0$ such that for all $x, \bar{x} \in \text{dom} \ f$ and $y, \bar{y} \in \text{dom} \ h$, one has

$$\| \nabla_x \Phi_i (x, y) - \nabla_x \Phi_i (\bar{x}, \bar{y}) \|_{\mathcal{X}^*} \leq L_{xx} \| x - \bar{x} \|_{\mathcal{X}} + L_{xy} \| y - \bar{y} \|_{\mathcal{Y}}, \quad \forall i \in \mathcal{N}, \quad (6)$$

**(ii)** for any $x \in \text{dom} \ f \subseteq \mathcal{X}$, $\Phi (x, \cdot)$ is concave; for any $i \in \mathcal{N}$ $\Phi_i (x, \cdot)$ is differentiable; there exist $L_{yx} > 0$ and $L_{yy} \geq 0$ such that for all $x, \bar{x} \in \text{dom} \ f$ and $y, \bar{y} \in \text{dom} \ h$, one has

$$\| \nabla_y \Phi_i (x, y) - \nabla_y \Phi_i (\bar{x}, \bar{y}) \|_{\mathcal{Y}^*} \leq L_{yy} \| y - \bar{y} \|_{\mathcal{Y}} + L_{yx} \| x - \bar{x} \|_{\mathcal{X}}, \quad \forall i \in \mathcal{N}. \quad (7)$$

Note that (6) and convexity of $\Phi (\cdot, y)$ imply that for any $y \in \text{dom} \ h$ and $x, \bar{x} \in \text{dom} \ f$,

$$0 \leq \Phi (x, y) - \Phi (\bar{x}, y) - \langle \nabla_x \Phi (\bar{x}, y), x - \bar{x} \rangle \leq \frac{L_{xx}}{2} \| x - \bar{x} \|_{\mathcal{X}}^2. \quad (8)$$

**Remark 1** It is worth highlighting that we only assume that $\Phi (\cdot, \cdot)$ is a convex-concave function while each component function $\Phi_i (\cdot, \cdot)$ may not be. Moreover, in the analysis of our method, we use the Lipschitz constants in Assumption 1 to obtain suitable bounds for a one-step progress of iterates as well as providing a bound on the variance of errors of gradients.
Algorithm 1 SVR-APD

1: Initialization: \( \hat{x}^0 \in \mathcal{X}, \hat{y}^0 \in \mathcal{Y} \)
2: \((x_0^0, y_0^0) \leftarrow (\hat{x}^0, \hat{y}^0)\)
3: \((x_0^0, y_0^0) \leftarrow (\nabla \psi_X(x_0^0), \nabla \psi_Y(y_0^0))\)
4: for \( k = 1, \ldots, K \) do
5: \( G_y \leftarrow \nabla_y \Phi(\hat{x}^{k-1}, \hat{y}^{k-1}) \)
6: \( G_x \leftarrow \nabla_x \Phi(\hat{x}^{k-1}, \hat{y}^{k-1}) \)
7: for \( t = 0, \ldots, T^k - 1 \) do
8: Pick \( j_k \in \mathcal{N} \) uniformly at random
9: \( x_t^k \leftarrow \nabla \psi_X((1 - \gamma_t^k) \nabla \psi_Y(x_t^k) + \gamma_t^k \hat{y}^{k-1}) \)
10: \( \xi_t \leftarrow \nabla_x \Phi_{j_k}(x_t^k, y_t^k) - \nabla_y \Phi_{j_k}(\hat{x}^{k-1}, \hat{y}^{k-1}) + G_y \)
11: \( y_t^k \leftarrow \nabla \psi_Y(x_t^k, y_t^k) - \nabla \psi_Y(x_t^k - \gamma_t^k \hat{y}^{k-1} \gamma_t^k - 1) \)
12: \( y_{t+1}^k \leftarrow \arg\min_{y \in \mathcal{Y}} \{ h(y) - (\xi_t + q_t^k y) + \frac{1}{\sigma_t} D_Y(y, \hat{y}^k) \} \)
13: Pick \( i_k \in \mathcal{N} \) uniformly at random
14: \( x_t^k \leftarrow \nabla \psi_X((1 - \gamma_t^k) \nabla \psi_X(x_t^k) + \gamma_t^k \hat{y}^{k-1}) \)
15: \( \xi_t \leftarrow \nabla_x \Phi_{i_k}(x_t^k, y_t^k) - \nabla_y \Phi_{i_k}(\hat{x}^{k-1}, \hat{y}^{k-1}) + G_x \)
16: \( x_{t+1}^k \leftarrow \arg\min_{x \in \mathcal{X}} \{ f(x) + (\xi_t, x) + \frac{1}{\tau_t} D_X(x, \hat{x}^k) \} \)
17: end for
18: \((\hat{x}^k, \hat{y}^k) \leftarrow \frac{1}{T^k} \sum_{t=0}^{T^k - 1} (x_t^k, y_t^k) \)
19: \((\hat{x}^k, \hat{y}^k) \leftarrow \frac{1}{T^k} \sum_{t=0}^{T^k - 1} (\nabla \psi_X(x_{t+1}^k), \nabla \psi_Y(y_{t+1}^k)) \)
20: \((x_0^k, y_0^k, x_1^k, \ldots, x_{T^k}^k, y_{T^k}^k) \leftarrow (x_0^k, y_0^k, x_{T^k}^k, y_{T^k}^k) \)
21: end for

Assumption 2 The Bregman diameters \( B_X \) and \( B_Y \) are bounded.

Remark 2 It is easy to verify that Assumption 2 is satisfied if \( X \) and \( Y \) are bounded sets. Note that Assumptions 1 and 2 imply that problem (1) has an SP solution, hence, it is well-defined. Moreover, in the convergence analysis of the proposed algorithm we need Assumption 2 to bound the distance from the initial iterate points, i.e., \( \sup_{x \in X} \{ D_X(x, x_0^0) \} \leq B_X \) and \( \sup_{y \in Y} \{ D_Y(y, y_0^0) \} \leq B_Y \).

Remark 3 Let \( (\mathcal{U}, \|\cdot\|) \) be a finite-dimensional normed real vector space with dual space \( \mathcal{U}^* \). In the analysis of our method, we use the following fact: there exists \( C_{\mathcal{U}} > 0 \) such that for any vector-valued random variable \( W : \Omega \rightarrow \mathcal{U}, \mathbb{E}[\| W - \mathbb{E}[W]\|_{\mathcal{U}^*}] \leq C_{\mathcal{U}} \mathbb{E}[\| W \|_{\mathcal{U}^*}] \).

Note that this is a property of the vector space \( \mathcal{U} \) which is true for any finite-dimensional vector space. In more details, suppose \( \mathcal{U} \) is a finite-dimensional real vector space equipped with the Euclidean norm denoted by \( \|\cdot\|_2 \). Then for any vector-valued random variable \( W, \mathbb{E}[\| W - \mathbb{E}[W]\|_2^2] = \mathbb{E}[\| W \|_2^2] - \mathbb{E}[\| W \|_2]^2 \leq \|\mathbb{E}[W]\|_2^2; \) hence, \( C_{\mathcal{U}} = 1 \). Moreover, in a finite-dimensional vector space all the norms are equivalent, i.e., for an arbitrary norm \( \|\cdot\|_\alpha \) there exist \( c, C > 0 \) such that \( c \|\cdot\|_2 \leq \|\cdot\|_\alpha \leq C \|\cdot\|_2 \). Therefore, using the equivalency between an arbitrary norm \( \|\cdot\|_\mathcal{U} \) and Euclidean norm one can conclude that \( C_{\mathcal{U}} > 0 \) exists.

Next, we state our assumption on the algorithm’s parameters \( \{\tau^k, \sigma^k, \gamma_X^k, \gamma_Y^k\}_k \). Indeed, these conditions provide upper bounds on the maximum allowable step-size choices in terms of the global Lipschitz constants introduced in Assumption 1.
Assumption 3 (Step-size conditions) There exist $\alpha, \beta > 0$, and $\{\eta^k\}_{k \geq 1} \subset \mathbb{R}_+$, such that for any $k \geq 1$, the step-sizes $\{\tau^k\}_{k \geq 1}$ and $\{\sigma^k\}_{k \geq 1}$, and the momentum parameters $\{\gamma_x^k, \gamma_y^k\}_{k \geq 1} \subset (0, 1]$ satisfy

\[
(6C_X L_{xx}^2 + 8C_Y L_{xy}^2) \eta^k \leq \frac{\gamma_x^k}{\tau^k}, \quad (6C_X L_{xy}^2 + 8C_Y L_{yy}^2) \eta^k \leq \frac{\gamma_y^k}{\sigma^k},
\]

\[
\frac{L_{xy}^2}{\alpha} + 2C_Y L_{xy}^2 \eta^k \leq M_x^k, \quad \frac{L_{yy}^2}{\beta} + 2C_Y L_{yy}^2 \eta^k \leq M_y^k,
\]

where $M_x^k \triangleq \frac{1-\gamma_x^k}{\tau^k} - L_{xx} - (6C_X L_{xx}^2 + 8C_Y L_{xy}^2) \eta^k - \frac{1}{\eta^k}$, and $M_y^k \triangleq \frac{1-\gamma_y^k}{\sigma^k} - (\alpha + \beta) - 8C_Y L_{yy}^2 \eta^k - \frac{1}{\eta^k}$.

Remark 4 Let $L_x \triangleq \sqrt{6C_X L_{xx}^2 + 10C_Y L_{xy}^2}$ and $L_y \triangleq \sqrt{6C_X L_{xy}^2 + 10C_Y L_{yy}^2}$. The following two choices of the algorithm parameters and the design parameter $\eta^k$ satisfy the step-size conditions in Assumption 3.

I) Constant: For $k \geq 1$, $T^k = \tilde{T}$, $\tau^k = \tau$, $\sigma^k = \sigma$, $\gamma_x^k = \gamma_x$, $\gamma_y^k = \gamma_y$, and $\eta^k = \eta$, such that

\[
\tilde{T} = Tn, \quad \gamma_x = \frac{\tilde{\gamma}_x}{n}, \quad \gamma_y = \frac{\tilde{\gamma}_y}{n}, \quad \alpha = L_{yx}, \quad \beta = L_{yy}
\]

\[
\tau = \min \left\{ \frac{1}{L_x \sqrt{n}}, \frac{1-\gamma_x/n}{L_{xx} + L_{xy} + 2L_x} \right\}, \quad \sigma = \min \left\{ \frac{1}{L_y \sqrt{n}}, \frac{1-\gamma_y/n}{L_{yy} + L_{xy} + L_y} \right\},
\]

\[
\eta = \min \left\{ \frac{1}{L_x \sqrt{n}}, \frac{1}{L_y \sqrt{n}}, \frac{b_x + \sqrt{b_x^2 - 4L_x^2}}{2L_x^2}, \frac{b_y + \sqrt{b_y^2 - 4L_y^2}}{2L_y^2} \right\},
\]

for some $T > 0$, $\tilde{\gamma}_x, \tilde{\gamma}_y \in (0, 1)$, where $b_x \triangleq \frac{1-1/n}{\tau} - (L_{xx} + L_{xy})$ and $b_y \triangleq \frac{1-1/n}{\sigma} - (L_{yy} + L_{xy})$.

II) Non-constant: For any $k \geq 1$,

\[
T^k = T(k + 1)^2, \quad \gamma_x^k = \frac{\tilde{\gamma}_x}{k^2}, \quad \gamma_y^k = \frac{\tilde{\gamma}_y}{k^2}, \quad \alpha = L_{yx}, \quad \beta = L_{yy}
\]

\[
\tau^k = \min \left\{ \frac{1}{L_x k}, \frac{1-\gamma_x^k}{L_{xx} + L_{xy} + 2L_x} \right\}, \quad \sigma^k = \min \left\{ \frac{1}{L_y k}, \frac{1-\gamma_y^k}{L_{yy} + L_{xy} + L_y} \right\},
\]

\[
\eta^k = \min \left\{ \frac{\tilde{\gamma}_x}{L_x k}, \frac{\tilde{\gamma}_y}{L_y k}, \frac{b_x + \sqrt{(b_x^2 - 4L_x^2)}}{2L_x^2}, \frac{b_y + \sqrt{(b_y^2 - 4L_y^2)}}{2L_y^2} \right\},
\]

for some $T > 0$, $\tilde{\gamma}_x, \tilde{\gamma}_y \in (0, 1)$, where $b_x^k \triangleq \frac{1-\gamma_x^k}{\tau^k} - (L_{xx} + L_{xy})$ and $b_y^k \triangleq \frac{1-\gamma_y^k}{\sigma^k} - (L_{yy} + L_{xy})$.

It is worth emphasizing that $\eta^k$ is only a design parameter and need not be computed for running the algorithm.

In the analysis of our method, the constants $C_X$, $C_Y$ are appeared in the step-sizes due to converting arbitrary norms into Euclidean norm and vice versa—see Remark 3.
Such an approach is used to obtain a suitable upper bound on the variance of the error of gradient estimates, i.e., $\delta^2$ and $\delta^3$ (see Lemma 2). The magnitude of these constants can be in the order of the dimension of the underlying vector space. For example, in the case of $m$-dimensional vector spaces equipped with $p$-norms these constants are in the order of $m^{1-\frac{1}{\max\{p,q\}}}$ such that $\frac{1}{p} + \frac{1}{q} = 1$. However, in some interesting practical scenarios these constants are small. For instance, if the objective function is separable in one of the variables, e.g., $\mathcal{Y} \subseteq \mathbb{R}^n$ and $\Phi(x, y) = \frac{1}{n} \sum_{i \in \mathcal{N}} \phi_i(x, y_i)$, $\|\cdot\|_X = \|\cdot\|_2$ and $\|\cdot\|_Y = \|\cdot\|_1$ then we observe that $C_X = 1$ and from the fact that $\|\cdot\|_{\infty} \leq \|\cdot\|_2$ and $\nabla_{y_i} \Phi \in \mathbb{R}$ we conclude that $C_Y = 1$.

## 3 Convergence analysis

In the following theorem, we state the main result of this paper by providing a bound on the gap function, i.e., $E = \sup_{x \in \mathcal{Z}} (\mathcal{L}(\tilde{x}, y) - \mathcal{L}(x, \tilde{y}))$, stating the oracle complexities using constant and non-constant step-sizes in the follow-up corollaries.

**Theorem 1** Let $\{x^k, y^k\}_{t,k}$ be the sequence generated by SVR-APD displayed in Algorithm 1 initialized from arbitrary vectors $\tilde{x}^0 \in \mathcal{X}$ and $\tilde{y}^0 \in \mathcal{Y}$. Suppose Assumptions 1 and 2 hold, and the step-size sequence $\{\tau^k, \sigma^k\}_{k \geq 1}$ and the momentum parameter sequence $\{\gamma_x^k, \gamma_y^k\}_{k \geq 1}$ satisfy Assumption 3 for some $\alpha, \beta > 0$, and $\{\eta^k\}_{k \geq 1} \subset \mathbb{R}_+$. Moreover, let $\tilde{z} \triangleq (x, y) \in \mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$, then the following holds:

**I** If the step-sizes and momentum parameters are constant, i.e., $\tau^k = \tau$, $\sigma^k = \sigma$, $\gamma_x^k = \gamma_x$, $\gamma_y^k = \gamma_y$, $T^k = \bar{T}$, $M_X^k \triangleq M_X$, and $M_Y^k \triangleq M_Y$, then for $K \geq 1$,

$$
\mathbb{E}[G_Z(\tilde{x}^{(K)}, \tilde{y}^{(K)})] \leq \frac{1}{K \bar{T}} \left[ \left( \frac{1}{\eta^2} + \frac{Y_x}{\tau} \bar{T} + \frac{1 - Y_x}{\tau} + M_X \right) B_X + \left( \frac{2}{\eta^2} + \frac{Y_y}{\sigma} \bar{T} + \frac{1 - Y_y}{\sigma} + M_Y \right) B_Y \right].
$$

(12)

where $(\tilde{x}^{(K)}, \tilde{y}^{(K)}) \triangleq \frac{1}{K} \sum_{k=1}^{K} (\tilde{x}^k, \tilde{y}^k)$, and $M_X^k, M_Y^k$ are defined in Assumption 3.

**II** If the step-sizes and momentum parameters are non-constant, then for $K \geq 1$,

$$
\mathbb{E}[G_Z(\tilde{x}^{(K)}, \tilde{y}^{(K)})] \leq \frac{1}{S^K} \left[ \left( \frac{1}{\eta^{K}} + \frac{\gamma_x^{K}}{\tau^{K}} T^K + \frac{1 - \gamma_x^{K}}{\tau^{K}} + M_X^K \right) B_X + \left( \frac{2}{\eta^{K}} + \frac{\gamma_y^{K}}{\sigma^{K}} T^K + \frac{1 - \gamma_y^{K}}{\sigma^{K}} + M_Y^K \right) B_Y \right].
$$

(13)

where $(\tilde{x}^{(K)}, \tilde{y}^{(K)}) \triangleq \frac{1}{S^K} \sum_{k=1}^{K} (T^k \tilde{x}^k, T^k \tilde{y}^k)$, $S^K \triangleq \sum_{k=1}^{K} T^k$.

**Corollary 1** Under the premises of Theorem 1 part I, if the step-sizes and parameters of Algorithm 1 are selected as in (10), then $\mathbb{E}[\sup_{z \in \mathcal{Z}} (\mathcal{L}(\tilde{x}^{(K)}, y) - \mathcal{L}(x, \tilde{y}^{(K)}))] \leq O(\frac{1}{K^{\sqrt{n}}})$. Moreover, the oracle complexity is $O(\frac{1}{\sqrt{n}})$. 

\(\square\) Springer
Corollary 2 Under the premises of Theorem 1 part II, if the step-sizes and parameters of Algorithm 1 are selected as in (11), then \( \mathbb{E} \left[ \sup_{z \in \mathcal{Z}} \left( \mathcal{L}(\tilde{x}(K), y) - \mathcal{L}(x, \tilde{y}(K)) \right) \right] \leq O(\frac{1}{K^2}). \) Moreover, the oracle complexity is \( O\left( \frac{n}{\sqrt{\epsilon}} + \frac{1}{\epsilon^{1.5}} \right) \).

Remark 5 Note that the best-known available rate for deterministic and stochastic first-order primal-dual methods to solve problem 1 are \( O(\frac{n}{\epsilon}) \) and \( O\left( \frac{n}{\sqrt{\epsilon}} + \frac{1}{\epsilon^{1.5}} \right) \), respectively. Corollaries 1 and 2 show that the proposed SVR-APD method improves these oracle complexity to \( O\left( \frac{n}{\sqrt{\epsilon}} \right) \) and \( O\left( \frac{n}{\sqrt{\epsilon}} + \frac{1}{\epsilon^{1.5}} \right) \) for constant and nonconstant step-sizes, respectively. To the best of our knowledge, this is the first result that provides such an improvement for convex-concave SP problems using Bregman distance functions. It is also worth mentioning that the memory requirement of SVR-APD is \( O(m + d) \) where \( m \triangleq \dim(\mathcal{X}) \) and \( d \triangleq \dim(\mathcal{Y}) \) which is the same as first-order primal-dual methods such as [11, 18, 25, 33].

4 Convergence proof

To show the convergence results, we use the following definition to facilitate the notation.

Definition 5 Let \( D_x^k(x) \triangleq \frac{1}{T_k} \sum_{t=1}^{T_k-1} D_x(x, x_{t-1}^k) \) and similarly \( D_y^k(y) \triangleq \frac{1}{T_k} \sum_{t=1}^{T_k-1} D_y(y, y_{t-1}^k) \), for \( k \geq 1 \).

Let \( \{T^k\}_{k \geq 1} \subset \mathbb{Z}_+ \) be a non-decreasing sequence. We define auxiliary sequences \( \{ u_t^k \}_{t,k}, \{ v_t^k \}_{t,k}, \) and \( \{ w_t^k \}_{t,k} \) which are helpful for the analysis of the algorithm and will never be actually computed in the running of SVR-APD. In particular, for any \( k \geq 1 \) and \( t \in \{0, \ldots, T^k - 1\} \) we define,

\[
\begin{align*}
    u_{t+1}^k & \leftarrow \argmin_{x \in \mathcal{X}} \left\{ -\langle \delta_t^k, x \rangle + \frac{1}{\eta_t} D_x(x, u_t^k) \right\}, \quad (14a) \\
    v_{t+1}^k & \leftarrow \argmin_{y \in \mathcal{Y}} \left\{ \langle \delta_t^k, y \rangle + \frac{1}{\eta_t} D_y(y, v_t^k) \right\}, \quad (14b) \\
    w_{t+1}^k & \leftarrow \argmin_{y \in \mathcal{Y}} \left\{ \langle q_t^k - \tilde{q_t}^k, y \rangle + \frac{1}{\eta_t} D_y(y, w_t^k) \right\}, \quad (14c)
\end{align*}
\]

for any \( \eta^k > 0 \), such that \( (u_0^0, v_0^0, w_0^0) \triangleq (\tilde{x}^0, \tilde{y}^0, \tilde{y}^0) \) and \( (u_{T_k}^k, v_{T_k}^k, w_{T_k}^k) \triangleq (u_0^{k+1}, v_0^{k+1}, w_0^{k+1}) \).

Lemma 1 Let \( \{ x_t^k, y_t^k \}_{t,k} \) be the sequence generated by SVR-APD displayed in Algorithm 1 initialized from arbitrary vectors \( \tilde{x}^0 \in \mathcal{X} \) and \( \tilde{y}^0 \in \mathcal{Y} \). Let \( \{ u_t^k, v_t^k, w_t^k \}_{t,k} \) be the auxiliary sequence defined in (14a)-(14c). Suppose Assumption 1 holds and the \( \delta_t^k, \delta_t^k, q_t^k, \) and \( \tilde{q}_t^k \) are defined in Definition 4. For any \( x \in \mathcal{X}, y \in \mathcal{Y}, \) and \( \eta^k > 0 \) the following results hold for \( k \geq 1 \) and \( t \geq 0, \)
\begin{equation}
\left\{ \delta_t^x, x - x_{t+1}^k \right\} \leq \frac{1}{\eta^k} D_X(x, u_t^k) - D_X(x, u_{t+1}^k) + \eta^k \| \delta_t^x \|_{\chi^s} + \left\{ \delta_t^x, u_t^k - x_t^k \right\} + \frac{1}{\eta^k} D_X(x_{t+1}^k, x_t^k), \quad (15a)
\end{equation}

\begin{equation}
\left\{ \delta_t^y, y_{t+1}^k - y \right\} \leq \frac{1}{\eta^k} D_Y(y, v_t^k) - D_Y(y, v_{t+1}^k) + \eta^k \| \delta_t^y \|_{\gamma^s} + \left\{ \delta_t^y, y_t^k - v_t^k \right\} + \frac{1}{\eta^k} D_Y(y_{t+1}^k, y_t^k), \quad (15b)
\end{equation}

\begin{equation}
\left\{ q_t^k - \bar{q}_t^k, y_t^k - y \right\} \leq \frac{1}{\eta^k} D_Y(y, w_t^k) - D_Y(y, w_{t+1}^k) + \frac{\eta^k}{2} \| q_t^k - \bar{q}_t^k \|_{\gamma^t} + \left\{ q_t^k - \bar{q}_t^k, y_t^k - w_t^k \right\}. \quad (15c)
\end{equation}

**Proof** We start by proving (15a). We split the inner product in the left hand-side into \( \left\{ \delta_t^x, x - x_{t+1}^k \right\} \) and \( \left\{ \delta_t^x, x_{t}^k - x_{t+1}^k \right\} \) and provide an upper bound for each term. Using Lemma 5-(c) in the appendix for (14a) with \( s = \delta_t^x, t = 1/\eta^k \), and \( f \equiv 0 \), we conclude that for any \( x \in X \),

\begin{equation}
\left\{ \delta_t^x, x - x_{t}^k \right\} = \left\{ \delta_t^x, x - u_t^k \right\} + \left\{ \delta_t^x, u_t^k - x_t^k \right\}
\end{equation}

\begin{equation}
\leq \frac{1}{\eta^k} D_X(x, u_t^k) - D_X(x, u_{t+1}^k) + \frac{\eta^k}{2} \| \delta_t^x \|_{\chi^s} + \left\{ \delta_t^x, u_t^k - x_t^k \right\}. \quad (16)
\end{equation}

Moreover, using Young’s inequality and strong convexity of the Bregman distance function we have

\begin{equation}
\left\{ \delta_t^x, x_{t}^k - x_{t+1}^k \right\} \leq \frac{\eta^k}{2} \| \delta_t^x \|_{\chi^s} + \frac{1}{\eta^k} D_X(x_{t+1}^k, x_t^k) \quad (17)
\end{equation}

Adding (17) to (16) gives (15a). Similarly, using (14b) and (14c) one can obtain the results in (15b) and (15c), respectively. \( \square \)

In the following lemma, we derive upper bounds for the error of estimating gradients.

**Lemma 2** Under the premises of Lemma 1, for any \( k \geq 1 \) and \( t \geq 0 \), \( E_{i_t^k} [\delta_t^x] = E_{j_t^k} [\delta_t^y] = E_{j_t^k} [q_t^k - \bar{q}_t^k] = 0 \), the following hold for some \( C_{\chi}, C_{\gamma} > 0 \).

\begin{equation}
E_{i_t^k} \left[ \| \delta_t^x \|_{\chi^s}^2 \right] \leq 6 C_{\chi} E_{i_t^k} \left[ L_{xx}^2 D_X(x_{t+1}^k, x_t^k) + L_{xx}^2 D_Y^k(y_{t+1}^k) + L_{xy}^2 D_Y^k(y_{t+1}^k) \right], \quad (18a)
\end{equation}

\begin{equation}
E_{j_t^k} \left[ \| \delta_t^y \|_{\gamma^t}^2 \right] \leq 8 C_{\gamma} E_{j_t^k} \left[ L_{yx}^2 D_X(x_{t+1}^k, x_t^k) + L_{yx}^2 D_Y(y_{t+1}^k, y_t^k) + L_{yy}^2 D_Y^k(y_{t+1}^k) + L_{yx}^2 D_X^k(x_{t+1}^k) \right]. \quad (18b)
\end{equation}
A stochastic variance-reduced accelerated primal-dual method... 665

Proof The unbiasedness of the stochastic noises \( \delta_t^x \), \( \delta_t^y \), and \( q_t^k - \tilde{q}_t^k \) clearly holds due to the uniform sampling of the sum-function components. Next, from the definition of \( \delta_t^x \) in Definition 4 we have that

\[
\mathbb{E}_{j_t^k} \left[ \left\| \delta_t^x \right\|_{\chi^*}^2 \right] = \mathbb{E}_{j_t^k} \left[ \left\| (\nabla_x \Phi_{j_t^k}(x_t^k, y_{t+1}^k) - \nabla_x \Phi_{j_t^k}(\tilde{x}^{k-1}, \tilde{y}^{k-1})) \right\|_{\chi^*}^2 \right] \\
\quad - \left( \nabla_x \Phi(x_t^k, y_{t+1}^k) - \nabla_x \Phi(\tilde{x}^{k-1}, \tilde{y}^{k-1}) \right) \right\|_{\chi^*}^2 \right] \\
\leq C_{\chi} \mathbb{E}_{j_t^k} \left[ \left\| (\nabla_x \Phi_{j_t^k}(x_t^k, y_{t+1}^k) - \nabla_x \Phi_{j_t^k}(\tilde{x}^{k-1}, \tilde{y}^{k-1})) \right\|_{\chi^*}^2 \right],
\]

where the inequality is concluded from Remark 3 for \( W = \nabla_x \Phi_{j_t^k}(x_t^k, y_{t+1}^k) - \nabla_x \Phi_{j_t^k}(\tilde{x}^{k-1}, \tilde{y}^{k-1}) \). Next, using the triangle inequality, the fact that for any \( \{a_i\}_{i=1}^m \subset \mathbb{R}, (\sum_{i=1}^m a_i)^2 \leq m \sum_{i=1}^m a_i^2 \), and Lipschitz continuity of \( \nabla_x \Phi \) in Assumption 1 for \( i \in \mathcal{N} \) we conclude that

\[
\mathbb{E}_{j_t^k} \left[ \left\| \nabla_x \Phi_{j_t^k}(x_t^k, y_{t+1}^k) - \nabla_x \Phi_{j_t^k}(\tilde{x}^{k-1}, \tilde{y}^{k-1}) \right\|_{\chi^*}^2 \right] \\
\leq 3 \mathbb{E}_{j_t^k} \left[ \left\| \nabla_x \Phi_{j_t^k}(x_t^k, y_{t+1}^k) - \nabla_x \Phi_{j_t^k}(x_{t+1}^k, y_{t+1}^k) \right\|_{\chi^*}^2 \right] \\
+ \left\| \nabla_x \Phi_{j_t^k}(x_{t+1}^k, y_{t+1}^k) - \nabla_x \Phi_{j_t^k}(x_{t+1}^k, \tilde{y}^{k-1}) \right\|_{\chi^*}^2 \right] \\
+ \left\| \nabla_x \Phi_{j_t^k}(x_{t+1}^k, \tilde{y}^{k-1}) - \nabla_x \Phi_{j_t^k} (\tilde{x}^{k-1}, \tilde{y}^{k-1}) \right\|_{\chi^*}^2 \right] \\
\leq 3 \mathbb{E}_{j_t^k} \left[ L_{xx}^2 \left\| x_t^k - x_{t+1}^k \right\|_{\chi}^2 + L_{xy}^2 \left\| x_{t+1}^k - x^{k-1} \right\|_{\chi}^2 + L_{xy}^2 \left\| y_{t+1}^k - y^{k-1} \right\|_{\chi}^2 \right] \\
\quad + \frac{1}{t+1} \sum_{\ell=1}^{t+1} \left( L_{xx}^2 \left\| x_{\ell+1}^k - x_{\ell}^{k-1} \right\|_{\chi}^2 \\
\quad + L_{xy}^2 \left\| y_{t+1}^k - y_{\ell}^{k-1} \right\|_{\chi}^2 \right). \tag{20}
\]

Recall that \( \tilde{x}^{k-1} \) and \( \tilde{y}^{k-1} \) are defined as the average of inner iterates in Algorithm 1. Once again using the triangle inequality we observe that

\[
L_{xx}^2 \left\| x_{t+1}^k - \tilde{x}^{k-1} \right\|_{\chi}^2 + L_{xy}^2 \left\| y_{t+1}^k - \tilde{y}^{k-1} \right\|_{\chi}^2 \leq \frac{1}{(t+1)^2} \left( \sum_{\ell=1}^{t+1} L_{xx} \left\| x_{\ell+1}^k - x_{\ell}^{k-1} \right\|_{\chi}^2 \right) + \left( \sum_{\ell=1}^{t+1} L_{xy} \right. \left\| y_{t+1}^k - y_{\ell}^{k-1} \right\|_{\chi}^2 \right)^2 \right).
\[
\leq \frac{1}{\ell^{k-1}} \sum_{\ell=1}^{T-1} \left( L_{xx}^2 \| x_{t+1}^k - x_{\ell}^{k-1} \|^2_{\mathcal{X}} + L_{xy}^2 \| y_{t+1}^k - y_{\ell}^{k-1} \|^2_{\mathcal{Y}} \right) \\
\leq L_{xx}^2 D_X(x_{t+1}^k) + L_{yy}^2 D_Y(y_{t+1}^k),
\]

where in the last inequality we used the strong convexity of the Bregman distance functions, and Definition of 5.

Finally, combining (20) with (21) and the resulting inequality with (19) lead to (18a).

For proving the inequalities in (18b) and (18c) one can use a similar argument. In particular,

\[
E_{j_t} \left[ \| \delta_t^y \|^2_{\mathcal{Y}^*} \right] \\
= E_{j_t} \left[ \left\| \nabla_y \Phi_{j_t^k} (x_t^k, y_t^k) - \nabla_y \Phi_{j_t^k} (\tilde{x}^{k-1}, \tilde{y}^{k-1}) - (\nabla_y \Phi (x_t^k, y_t^k) - \nabla_y \Phi (\tilde{x}^{k-1}, \tilde{y}^{k-1})) \right\|^2_{\mathcal{Y}^*} \right] \\
\leq C_\gamma E_{j_t} \left[ \left\| \nabla_y \Phi_{j_t^k} (x_t^k, y_t^k) - \nabla_y \Phi_{j_t^k} (\tilde{x}^{k-1}, \tilde{y}^{k-1}) \right\|^2_{\mathcal{Y}^*} \right] \\
\leq 4C_\gamma E_{j_t} \left[ L_{yx}^2 \left\| x_t^k - x_{t+1}^k \right\|^2_{\mathcal{X}} + L_{yy}^2 \left\| y_t^k - y_{t+1}^k \right\|^2_{\mathcal{Y}} \right] \\
+ L_{yx}^2 \left\| x_{t+1}^k - \tilde{x}^{k-1} \right\|^2_{\mathcal{X}} \\
\leq 4C_\gamma E_{j_t} \left[ L_{yx}^2 \left\| x_t^k - x_{t+1}^k \right\|^2_{\mathcal{X}} + L_{yy}^2 \left\| y_t^k - y_{t+1}^k \right\|^2_{\mathcal{Y}} \right] \\
+ \frac{1}{\ell^{k-1}} \sum_{\ell=1}^{T-1} \left( L_{yx}^2 \left\| y_{t+1}^k - y_{\ell}^{k-1} \right\|^2_{\mathcal{Y}} + L_{yx}^2 \left\| x_{t+1}^k - x_{\ell}^{k-1} \right\|^2_{\mathcal{X}} \right).
\]

Moreover,

\[
E_{j_t^k} \left[ \| q_t^k - \tilde{q}_t^k \|^2_{\mathcal{Y}^*} \right] = E_{j_t^k} \left[ \left\| \nabla_y \Phi_{j_t^k} (x_t^k, y_t^k) - \nabla_y \Phi_{j_t^k} (x_{t-1}^k, y_{t-1}^k) \right\|^2_{\mathcal{Y}^*} \right] \\
- (\nabla_y \Phi (x_t^k, y_t^k) - \nabla_y \Phi (x_{t-1}^k, y_{t-1}^k)) \right\|^2_{\mathcal{Y}^*} \right] \\
\leq C_\gamma E_{j_t^k} \left[ \left\| \nabla_y \Phi_{j_t^k} (x_t^k, y_t^k) - \nabla_y \Phi_{j_t^k} (x_{t-1}^k, y_{t-1}^k) \right\|^2_{\mathcal{Y}^*} \right] \\
\leq 2C_\gamma E_{j_t^k} \left[ L_{yx}^2 \left\| x_t^k - x_{t-1}^k \right\|^2_{\mathcal{X}} + L_{yy}^2 \left\| y_t^k - y_{t-1}^k \right\|^2_{\mathcal{Y}} \right].
\]

Lemma 3 Suppose Assumption 1 holds, then the following inequality holds for any \( x \in X, \tilde{x} \in X^*, y, \tilde{y} \in Y, \) and \( \tilde{y} \in Y^* \),

\[
\left\| \nabla_y \Phi (x, y) - \nabla_y \Phi (\tilde{x}, \tilde{y}), y - \tilde{y} \right\| \leq \frac{L_{yx}^2}{\alpha} D_X(x, \tilde{x}) + \frac{L_{yy}^2}{\beta} D_Y(y, \tilde{y}) \\
+ (\alpha + \beta) D_Y(\tilde{y}, y),
\]

\( \square \) Springer
for any $\alpha, \beta > 0$. Moreover, if $L_{yy} = 0$, then for any $\alpha > 0$:

$$
\left\| \nabla_y \Phi(x, y) - \nabla_y \Phi(\bar{x}, \bar{y}), y - \bar{y} \right\| \leq \frac{L_{yx}^2}{\alpha} \text{D}_X(x, \bar{x}) + \alpha \text{D}_X(y, \bar{y}),
$$

(23)

**Proof** By adding and subtracting $\nabla_y \Phi(\bar{x}, y)$ and triangle inequality we obtain

$$
\left\| \nabla_y \Phi(x, y) - \nabla_y \Phi(\bar{x}, \bar{y}), y - \bar{y} \right\|
\leq \left\| \nabla_y \Phi(x, y) - \nabla_y \Phi(\bar{x}, y), y - \bar{y} \right\| + \left\| \nabla_y \Phi(\bar{x}, y) - \nabla_y \Phi(\bar{x}, \bar{y}), y - \bar{y} \right\|.
$$

(24)

Next, using Young’s inequality and Assumption 1 we conclude that

$$
\left\| \nabla_y \Phi(x, y) - \nabla_y \Phi(\bar{x}, \bar{y}), y - \bar{y} \right\|
\leq \frac{L_{yx}^2}{2\alpha} \|x - \bar{x}\|_X^2 + \frac{\alpha}{2} \|y - \bar{y}\|_Y^2 + \frac{L_{yy}^2}{2\beta} \|y - \bar{y}\|_Y^2 + \frac{\beta}{2} \|y - \bar{y}\|_Y^2.
$$

The result in (22) immediately follows using strong convexity of Bregman distance functions. Moreover, if $L_{yy} = 0$, then the second inner product in the right hand side of (24) will be zero and the result can be concluded.

An immediate consequence of Lemma 3 by setting $(x, y) = (x^k_t, y^k_t)$ and $(\bar{x}, \bar{y}) = (x^k_{t-1}, y^k_{t-1})$ is that for any $\bar{y} \in \mathcal{Y}, k \geq 1$, and $t \geq 0$

$$
\left\| \hat{q}^k_t, y^k_t - \bar{y} \right\| \leq \frac{L_{yx}^2}{\alpha} \text{D}_X(x^k_t, x^k_{t-1}) + \frac{L_{yy}^2}{\beta} \text{D}_Y(y^k_t, y^k_{t-1}) + (\alpha + \beta) \text{D}_Y(\bar{y}, y^k_t).
$$

(25)

In the next lemma, we provide a one-step analysis for SVR-APD which is the main building block for showing the rate result stated in Theorem 1.

**Lemma 4** Let $(x^k_t, y^k_t)_{t,k}$ be the sequence generated by SVR-APD displayed in Algorithm 1 initialized from arbitrary vectors $\bar{x}^0 \in \mathcal{X}$ and $\bar{y}^0 \in \mathcal{Y}$. Let $(u^k_t, v^k_t, w^k_t)_{t,k}$ be the auxiliary sequence defined in (14a)-(14c). Suppose Assumption 1 holds and the $\delta^k_t, \bar{q}^k_t$, and $\bar{q}^k_t$ are defined in Definition 4. Then for any $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $\eta^k > 0$ the following inequality holds for $k \geq 1$ and $t \in \{0, \ldots, T^k - 1\}$.

$$
\mathcal{L}(x^k_{t+1}, y) - \mathcal{L}(x, y_{t+1}) \leq

A^k_t - B^k_t + \left(q^k_t, y^k_t - y\right) - \left(q^k_{t+1}, y^k_{t+1} - y\right) + \frac{1}{\eta^k} \left(\text{D}_X(x, u^k_t) - \text{D}_X(x, u^k_{t+1})\right)

+ \frac{1}{\eta^k} \left(\text{D}_Y(y, v^k_t) - \text{D}_Y(y, v^k_{t+1})\right) + \frac{1}{\eta^k} \left(\text{D}_Y(y, w^k_t) - \text{D}_Y(y, w^k_{t+1})\right)

+ \frac{\gamma^k}{\tau^k} \left(\text{D}_X^k(x) - \text{D}_X(x, x^k_{t+1})\right) + \frac{1 - \gamma^k}{\tau^k} \left(\text{D}_X(x, x^k_t) - \text{D}_X(x, x^k_{t+1})\right)

\square
\[ + \left( (6C_X L_{xx}^2 + 8C_Y L_{xy}^2) \eta^k - \frac{\gamma^k}{\tau} \right) D_X^k (x^k_{t+1}) \]
\[ + \left( \frac{L_{xx}}{\alpha} + 2C_Y L_{xy}^2 \eta^k \right) D_X (x^k_t, x^k_{t-1}) - M^k_X D_X (x^k_{t+1}, x^k_t) \]
\[ + \frac{\gamma^k}{\sigma} \left( D_Y^k (y) - D_Y (y, y_t^k) \right) + \frac{1 - \gamma^k}{\sigma} \left( D_Y (y, y_t^k) - D_Y (y, y_{t+1}^k) \right) \]
\[ + \left( (6C_X L_{xx}^2 + 8C_Y L_{xy}^2) \eta^k - \frac{\gamma^k}{\tau} \right) D_Y^k (y^k_{t+1}) \]
\[ + \left( \frac{L_{yy}^2}{\beta} + 2C_Y L_{yy}^2 \eta^k \right) D_Y (y^k_t, y^k_{t-1}) - M^k_Y D_Y (y^k_{t+1}, y^k_t). \quad (26) \]

\[ A^k_t \triangleq \left\{ \delta^k_x, u^k_t - x^k_t \right\} + \left\{ \delta^k_y, y^k_t - v^k_t \right\} + \left\{ q^k_t - \bar{q}^k_t, y^k_t - w^k_t \right\} \]
\[ + \eta^k \left( \| \delta^k_x \|_{\mathcal{A}^*}^2 + \| \delta^k_y \|_{\mathcal{Y}^*}^2 + \frac{1}{2} \left\| q^k_t - \bar{q}^k_t \right\|_{\mathcal{Y}^*}^2 \right). \quad (27) \]

\[ B^k_t \triangleq \eta^k \left( (6C_X L_{xx}^2 + 8C_Y L_{xy}^2) (D_X (x^k_{t+1}, x^k_t) + D_X^k (x^k_{t+1})) \right. \]
\[ + 8C_Y L_{xy}^2 D_Y (y^k_{t+1}, y^k_t) + (6C_X L_{xy}^2 + 8C_Y L_{yy}^2) D_Y^k (y^k_{t+1}) \]
\[ + \left. 2C_Y (L_{xy}^2 D_X (x^k_t, x^k_{t-1}) + L_{yy}^2 D_Y (y^k_t, y^k_{t-1})) \right). \quad (28) \]

where \( M^k_X \) and \( M^k_Y \) are defined in Assumption 3.

**Proof** Applying Lemma 5-(a), on the update rule of \( y^k_{t+1} \), implies that for any \( y \in Y \),

\[ h(y^k_{t+1}) - h(y) \leq \left\{ \xi_t + q^k_t, x^k_{t+1} - y \right\} + \frac{1}{\sigma^k} \left\{ \nabla \psi \gamma (y^k_{t+1}) - \nabla \psi \gamma (y^k_t), y - y^k_{t+1} \right\} \]
\[ = \left\{ \xi_t + q^k_t, x^k_{t+1} - y \right\} \]
\[ + \frac{1}{\sigma^k} \left\{ \nabla \psi \gamma (y^k_{t+1}) - (1 - \gamma^k) \nabla \psi \gamma (y^k_t) - \gamma^k y^k_{t+1}, y - y^k_{t+1} \right\} \]
\[ = \left\{ \xi_t + q^k_t, y^k_{t+1} - y \right\} + \frac{1 - \gamma^k}{\sigma^k} \left\{ \nabla \psi \gamma (y^k_{t+1}) - \nabla \psi \gamma (y^k_t), y - y^k_{t+1} \right\} \]
\[ + \frac{\gamma^k}{\sigma^k T^k-1} \sum_{\ell=1}^{T^k-1} \left\{ \nabla \psi \gamma (y^k_{t+1}) - \nabla \psi \gamma (y^k_{\ell-1}), y - y^k_{t+1} \right\}, \]

where in the first equality we used Lemma 5-(d) and update rule of \( y^k_t \) in line 7 of Algorithm 1. Using the generalized three-point property of Bregman distance in Lemma 5-(b) twice; for \( \left\{ \nabla \psi \gamma (y^k_{t+1}) - \nabla \psi \gamma (y^k_t), y - y^k_{t+1} \right\} \) and \( \left\{ \nabla \psi \gamma (y^k_{t+1}) - \nabla \psi \gamma (y^k_{\ell-1}), y - y^k_{t+1} \right\} \), the last inequality can be written as

\[ h(y^k_{t+1}) - h(y) \leq \left\{ \xi_t + q^k_t, x^k_{t+1} - y \right\} \]

© Springer
A stochastic variance-reduced accelerated primal-dual method...

\[ + \frac{1 - \gamma^k_Y}{\sigma^k} \left( D_Y(y^k, y^k_T) - D_Y(y^k, y^k_{t+1}) - D_Y(y^k_{t+1}, y^k_T) \right) \]

\[ + \frac{\gamma^k_Y}{\sigma^k T^{k-1}} \sum_{t=1}^{T^{k-1}} \left( D_Y(y^k_T, y^k_{t-1}) - D_Y(y^k, y^k_{t+1}) - D_Y(y^k_{t+1}, y^k_{t-1}) \right). \]

(29)

Using the definition of \( \delta^y_T \), i.e., \( \delta^y_T = \xi_T - \nabla_y \Phi(x^k_T, y^k_T) \), Definition 5, and rearranging the terms we obtain that

\[ h(y^k_{t+1}) - h(y) \leq \left\{ \delta^y_T, y^k_{t+1} - y \right\} + \left\{ \nabla_y \Phi(x^k_T, y^k_T), y^k_{t+1} - y \right\} \]

\[ + \left\{ q^k_t, y^k_T - y \right\} + \left\{ q^k_t, y^k_{t+1} - y^k_T \right\} \]

\[ + \frac{1 - \gamma^k_Y}{\sigma^k} \left( D_Y(y^k_T, y^k_T) - D_Y(y^k, y^k_{t+1}) - D_Y(y^k_{t+1}, y^k_T) \right) \]

\[ + \frac{\gamma^k_Y}{\sigma^k} \left( D^k_Y - D_Y(y^k_{t+1}) - D^k_Y(y^k_{t+1}) \right). \]

(30)

Adding \( \left\{ \nabla_y \Phi(x^k_{t+1}, y^k_{t+1}), y - y^k_{t+1} \right\} \) to both sides of (30) and adding and subtracting \( \left\{ \bar{q}^k_t, y^k_T - y \right\} \) to the right hand side of (30) lead to

\[ h(y^k_{t+1}) - h(y) + \left\{ \nabla_y \Phi(x^k_{t+1}, y^k_{t+1}), y - y^k_{t+1} \right\} \]

\[ \leq \left\{ \delta^y_T, y^k_{t+1} - y \right\} - \left\{ \bar{q}^k_{t+1}, y^k_{t+1} - y \right\} \]

\[ + \left\{ \bar{q}^k_t - \bar{q}^k_T, y^k_T - y \right\} + \left\{ \bar{q}^k_t, y^k_{t+1} - y^k_T \right\} \]

\[ + \frac{1 - \gamma^k_Y}{\sigma^k} \left( D_Y(y^k_T, y^k_T) - D_Y(y^k, y^k_{t+1}) - D_Y(y^k_{t+1}, y^k_T) \right) \]

\[ + \frac{\gamma^k_Y}{\sigma^k} \left( D^k_Y - D_Y(y^k_{t+1}) - D^k_Y(y^k_{t+1}) \right). \]

(31)

Now using (25) we obtain

\[ \left\{ q^k_t, y^k_{t+1} - y^k_T \right\} \leq \frac{L^2_{yx}}{\alpha} D_X(x^k_T, x^k_{t-1}) + \frac{L^2_{yy}}{\beta} D_Y(y^k_T, y^k_{t-1}) + (\alpha + \beta) D_Y(y^k_{t+1}, y^k_T). \]

(32)

for any \( \alpha, \beta > 0 \), and note that if \( L_{yx} = 0 \), then \( \beta = 0 \) and we define \( 0^2/0 = 0 \). Moreover, \( \Phi(x, \cdot) \) is a concave function, for any \( x \in \mathcal{X} \), therefore, we have that

\[ \Phi(x^k_{t+1}, y) - \Phi(x^k_{t+1}, y^k_{t+1}) \leq \left\{ \nabla_y \Phi(x^k_{t+1}, y^k_{t+1}), y - y^k_{t+1} \right\}. \]

(33)
Using (32) within (31) and adding (33) to the result implies that

\[
    h(y_{t+1}^k) - h(y) + \Phi(x_{t+1}^k, y) - \Phi(x_{t+1}^k, y_{t+1}^k)
    \]

\[
    \leq \left( \delta_t^k, y_{t+1}^k - y \right) + \left( \gamma_t^k, y_{t+1}^k - y \right) + \left( \hat{\gamma}_t^k, y_{t+1}^k - y \right) - \left( q_t^k, y_{t+1}^k - y \right)
    \]

\[
    + \frac{1 - y_t^k}{\sigma_k} \left( D_Y(y, y_{t+1}^k) - D_Y(y, y_{t+1}^k) - D_Y(y_{t+1}^k, y_{t+1}^k) \right)
    \]

\[
    \leq \frac{1 - y_t^k}{\sigma_k} \left( D_Y(y, y_{t+1}^k) - D_Y(y, y_{t+1}^k) - D_Y(y_{t+1}^k, y_{t+1}^k) \right)
    \]

\[
    + \frac{y_t^k}{\sigma_k} \left( D_Y(y, y_{t+1}^k) - D_Y(y, y_{t+1}^k) - D_Y(y_{t+1}^k, y_{t+1}^k) \right)
    \]

\[
    + \left( \frac{L^2_{xy}}{\alpha} D_X(x_t^k, x_{t-1}^k) + \frac{L^2_{yy}}{\beta} D_Y(y_t^k, y_{t-1}^k) + (\alpha + \beta) D_Y(y_{t+1}^k, y_{t+1}^k) \right). \tag{34}
    \]

Next, we analyze the update rule of \( x_{t+1}^k \). Applying Lemma 5-(a) on the update rule of \( x_{t+1}^k \), implies that for any \( x \in X \),

\[
    f(x_{t+1}^k) - f(x) \leq \left( \zeta_t, x - x_{t+1}^k \right) + \frac{1}{\tau_k} \left( \nabla \psi_X(x_{t+1}^k) - \nabla \psi_X(x_t^k), x - x_{t+1}^k \right)
    \]

\[
    = \left( \zeta_t, x - x_{t+1}^k \right) + \frac{1 - y_t^k}{\tau_k} \left( \nabla \psi_X(x_{t+1}^k) - \nabla \psi_X(x_t^k), x - x_{t+1}^k \right)
    \]

\[
    + \frac{y_t^k}{\tau_k} \left( \nabla \psi_X(x_{t+1}^k) - r_x^{-1}, x - x_{t+1}^k \right), \tag{35}
    \]

where the equality holds by using Lemma 5-(d) and the update rule of \( \hat{x}_t^k \) in line 13 of Algorithm 1. Using the three-point property of Bregman distance in Lemma 5-(b), the above inequality can be rewritten as

\[
    f(x_{t+1}^k) - f(x) \leq \left( \zeta_t, x - x_{t+1}^k \right) + \frac{1 - y_t^k}{\tau_k} \left( D_X(x, x_t^k) - D_X(x, x_{t+1}^k) - D_X(x_{t+1}^k, x_t^k) \right)
    \]

\[
    + \frac{y_t^k}{\tau_k} \left( D_X(x, x_{t+1}^k) - D_X(x, x_{t+1}^k) - D_X(x_{t+1}^k, x_{t+1}^k) \right)
    \]

\[
    + \frac{y_t^k}{\tau_k} \left( D_X(x, x_{t+1}^k) - D_X(x, x_{t+1}^k) - D_X(x_{t+1}^k, x_{t+1}^k) \right) \tag{36}
    \]

Recall that \( \delta_x^k = \zeta_t - \nabla \Phi(x_t^k, y_{t+1}^k) \). we add and subtract \( \nabla \Phi(x_t^k, y_{t+1}^k), x - x_{t+1}^k \) to the right-hand side of (36), rearranging the terms, and using Definition 5 lead to

\[
    f(x_{t+1}^k) - f(x) \leq \left( \nabla \Phi(x_t^k, y_{t+1}^k), x - x_{t+1}^k \right) + \left( \delta_x^k, x - x_{t+1}^k \right)
    \]

\[
    + \frac{1 - y_t^k}{\tau_k} \left( D_X(x, x_t^k) - D_X(x, x_{t+1}^k) - D_X(x_{t+1}^k, x_t^k) \right)
    \]

\[
    + \frac{y_t^k}{\tau_k} \left( D_X(x, x_{t+1}^k) - D_X(x, x_{t+1}^k) - D_X(x_{t+1}^k, x_{t+1}^k) \right) \tag{37}
    \]
The first inner product in the right hand side of (37) can be bounded as follows
\[
\left\langle \nabla_x \Phi(x^k_t, y^k_{t+1}), x - x^k_{t+1} \right\rangle \\
= \left\langle \nabla_x \Phi(x^k_t, y^k_{t+1}), x - x^k_t \right\rangle + \left\langle \nabla_x \Phi(x^k_t, y^k_{t+1}), x^k_t - x^k_{t+1} \right\rangle \\
\leq \Phi(x, y^k_{t+1}) - \Phi(x^k_t, y^k_{t+1}) + \left\langle \nabla_x \Phi(x^k_t, y^k_{t+1}), x^k_t - x^k_{t+1} \right\rangle \\
\leq \Phi(x, y^k_{t+1}) - \Phi(x^k_t, y^k_{t+1}) + L_{xx} D_X(x^k_{t+1}, x^k_t),
\tag{38}
\]
where the first inequality hold due to convexity of $\Phi(\cdot, y)$, for any $y \in \mathcal{Y}$; in the second inequality (8) and the fact that $D_X(x, \bar{x}) \geq \frac{1}{2} \|x - \bar{x}\|_{\mathcal{X}}^2$, for any $x, \bar{x} \in \mathcal{X}$, are used. Now, we use (38) within (37), then rearranging the terms leads to the following result.
\[
f(x^k_{t+1}) - f(x) - \Phi(x, y^k_{t+1}) + \Phi(x^k_{t+1}, y^k_{t+1}) \\
\leq \left\langle \delta^x_t, x - x^k_{t+1} \right\rangle + L_{xx} D_X(x^k_{t+1}, x^k_t) \\
+ \left( 1 - \gamma^k_t \right) D_X(x, x^k_t) - D_X(x, x^k_{t+1}) - D_X(x^k_{t+1}, x^k_t) \\
+ \frac{\gamma^k_t}{\tau^k} \left( D^k_X(x) - D_X(x, x^k_{t+1}) - D^k_X(x^k_{t+1}) \right).
\tag{39}
\]
Combining the results of (34) and (39) and using the definition of $\mathcal{L}(x, y)$ as well as Definition 5 lead to the following:
\[
\mathcal{L}(x^k_{t+1}, y) - \mathcal{L}(x, y^k_{t+1}) \leq \left\langle \delta^y_{t+1}, y^k_{t+1} - y \right\rangle + \left\langle \delta^x_t, x - x^k_{t+1} \right\rangle + \left\langle q^k_t - \bar{q}^k_t, y^k_{t+1} - y \right\rangle \\
+ \left\langle \tilde{q}^k_t, y^k_{t+1} - y \right\rangle - \left\langle \tilde{q}^k_{t+1}, y^k_{t+1} - y \right\rangle + \frac{\gamma^k_t}{\tau^k} \left( D^k_X(x) - D_X(x, x^k_{t+1}) \right) \\
+ \left( 1 - \gamma^k_x \right) D_X(x, x^k_t) - D_X(x, x^k_{t+1}) - D_X(x^k_{t+1}, x^k_t) \\
- \left( 1 - \gamma^k_x \right) L_{xx} D_X(x^k_{t+1}, x^k_t) + \frac{\gamma^k_y}{\sigma^k} \left( D^k_Y(y) - D_Y(y, y^k_{t+1}) \right) \\
+ \frac{1 - \gamma^k_y}{\sigma^k} \left( D_Y(y, y^k_{t+1}) - D_Y(y, y^k_{t+1}) \right) - \frac{\gamma^k_y}{\sigma^k} D^k_Y(y^k_{t+1}) \\
+ \frac{L^2_{\mathcal{Y}}}{\beta} D_Y(y^k_{t+1}, y^k_{t+1}) - \left( 1 - \frac{\gamma^k_y}{\sigma^k} \right) (\alpha + \beta) D_Y(y^k_{t+1}, y^k_{t+1}).
\tag{40}
\]
Finally, we use Lemma 1 to bound the first three inner products in the right hand side of (40), then we add and subtract $B^k_t$ to the right hand side which conclude the result. 
\[\square\]

Now we are ready to prove the results in Theorem 1 and Corollary 1, 2.
Proof of Theorem 1 Consider the result in Lemma 4, summing the inequality over \( t = 0, \ldots, T^k - 1 \), divide by \( T^k \), and using the step-size conditions (9b) in Assumption 3 one can conclude that for \( k \geq 1 \),

\[
\frac{1}{T^k} \sum_{t=0}^{T^k-1} (L(x^k_{t+1}, y) - L(x, y^k_{t+1})) \\
\leq \frac{1}{T^k} \sum_{t=0}^{T^k-1} (A^k_t - B^k_t) + \frac{1}{T^k} \left( \langle q^k_0, y^k_0 - y \rangle - \langle q^{k+1}_0, y^{k+1}_0 - y \rangle \right) \\
+ \frac{1}{\eta^k T^k} \left( D_X(x, u^k_0) - D_X(x, u^{k+1}_0) \right) + \frac{1}{\eta^k T^k} \left( D_Y(y, v^k_0) - D_Y(y, v^{k+1}_0) \right) \\
+ \frac{1}{\tau^k T^k} \left( D_X(x, x^k_T) - D_X(x, x^{k+1}_T) \right) \\
+ \left( (6C_\chi L^2_{xx} + 8C_\gamma L^2_{yy}) \eta^k - \frac{y^k_0}{\tau^k} \right) \frac{1}{T^k} \sum_{t=0}^{T^k-1} D^k_X(x^k_{t+1}) \\
+ \frac{M^k_x}{T^k} \left( D_X(x^k_0, x^{k+1}_0) - D_X(x^k_{T^k-1}, x^{k+1}_{T^k-1}) \right) \\
+ \frac{\gamma^k_y}{\sigma^k} \left( D^k_Y(y) - D^{k+1}_Y(y) \right) + \frac{1 - \gamma^k_y}{\sigma^k T^k} \left( D_Y(y, y^k_0) - D_Y(y, y^{k+1}_0) \right) \\
+ \left( (6C_\chi L^2_{xy} + 8C_\gamma L^2_{yy}) \eta^k - \frac{y^k_0}{\sigma^k} \right) \frac{1}{T^k} \sum_{t=0}^{T^k-1} D^k_Y(y^k_{t+1}) \\
+ \frac{M^k_y}{T^k} \left( D_Y(y^k_0, y^{k+1}_0) - D_Y(y^{k+1}_T, y^{k+1}_{T^k-1}) \right). \tag{41}
\]

Recall that \((x^{k+1}_0, y^{k+1}_0) = (x^{k+1}_T, y^{k+1}_T)\), \((x^{k+1}_0, y^{k+1}_0) = (x^{k+1}_T, y^{k+1}_T)\) for any \( k \geq 1 \), and \((\bar{x}^k, \bar{y}^k) = \frac{1}{T^k} \sum_{t=0}^{T^k-1} (x^k_{t+1}, y^k_{t+1})\), then using (9a) in Assumption 3 and Jensen’s inequality, i.e., \( \frac{1}{T} \sum_{t=0}^{T-1} \phi(z_t) \geq \phi \left( \frac{1}{T} \sum_{t=0}^{T-1} z_t \right) \) for any \( T > 0 \) and any convex function \( \phi \), for the convex functions \( L(\cdot, y) \) and \( -L(x, \cdot) \), we obtain the following

\[
L(\bar{x}^k, y) - L(x, \bar{y}^k) \\
\leq \frac{1}{T^k} \sum_{t=0}^{T^k-1} (A^k_t - B^k_t) + \frac{1}{T^k} \left( \langle \bar{q}^k_0, y^k_0 - y \rangle - \langle \bar{q}^{k+1}_0, y^{k+1}_0 - y \rangle \right) \\
+ \frac{1}{\eta^k T^k} \left( D_X(x, u^k_0) - D_X(x, u^{k+1}_0) \right)
\]
\[ + \frac{1}{\eta^k T^k} \left( D_Y(y, v_0^k) - D_Y(y, v_{0+1}^k) \right) + \frac{1}{\eta^k T^k} \left( D_Y(y, w_0^k) - D_Y(y, w_{0+1}^k) \right) + \frac{\gamma_x}{\tau^k} \left( D_X^k(x) - D_{X+1}^k(x) \right) + \frac{1 - \gamma_x}{\tau^k T^k} \left( D_X(x, x_0^k) - D_X(x, x_{0+1}^k) \right) + M_x \left( D_X(x_0^k, x_{-1}^k) - D_X(x_{0+1}^k, x_{-1+1}^k) \right) + \frac{\gamma_y}{\sigma^k} \left( D_Y^k(y) - D_{Y+1}^k(y) \right) + \frac{1 - \gamma_y}{\sigma^k T^k} \left( D_Y(y, y_0^k) - D_Y(y, y_{0+1}^k) \right) + M_y \left( D_Y(y_0^k, y_{-1}^k) - D_Y(y_{0+1}^k, y_{-1+1}^k) \right). \] (42)

Now multiplying (42) by \( T^k \), leads to

\[ T^k (\mathcal{L}(\tilde{x}^k, y) - \mathcal{L}(x, \tilde{y}^k)) \leq \sum_{i=0}^{T^k-1} \left( A_i^k - B_i^k \right) + \left( \tilde{q}_0^k, y_0^k - y \right) - \left( \tilde{q}_{0+1}^k, y_{0+1}^k - y \right) + \frac{1}{\eta^k} (Q^k - Q^{k+1}) + \frac{\gamma_x}{\tau^k} \left( D_X^k(x) - D_{X+1}^k(x) \right) + \frac{1 - \gamma_x}{\tau^k T^k} \left( D_X(x, x_0^k) - D_X(x, x_{0+1}^k) \right) + M_x \left( D_X(x_0^k, x_{-1}^k) - D_X(x_{0+1}^k, x_{-1+1}^k) \right) + \frac{\gamma_y}{\sigma^k} \left( D_Y^k(y) - D_{Y+1}^k(y) \right) + \frac{1 - \gamma_y}{\sigma^k T^k} \left( D_Y(y, y_0^k) - D_Y(y, y_{0+1}^k) \right) + M_y \left( D_Y(y_0^k, y_{-1}^k) - D_Y(y_{0+1}^k, y_{-1+1}^k) \right), \] (43)

where \( Q^k \triangleq D_X(x, u_0^k) + D_Y(x, v_0^k) + D_Y(x, w_0^k) \) for any \( k \geq 1 \). Next, we consider two cases depending on selecting the step-sizes. 

**Part I** In this scenario, we consider constant step-sizes and parameters, i.e., \( T^k = \tilde{T}, \tau^k = \tau, \sigma^k = \sigma, \gamma_x^k = \gamma_x, \gamma_y^k = \gamma_y, \eta^k = \eta \). Then, summing (43) over \( k \) from 1 to \( K \), using (47) and Jensen’s inequality we conclude that

\[ \tilde{T} K (\mathcal{L}(\tilde{x}^{(K)}, y) - \mathcal{L}(x, \tilde{y}^{(K)})) \leq \sum_{k=1}^{K} \sum_{i=0}^{\tilde{T}-1} \left( A_i^k - B_i^k \right) - \left( \tilde{q}_{0+1}^k, y_{0+1}^k - y \right) + \frac{1}{\eta} (Q^0 - Q^{K+1}) + \frac{\gamma_x}{\tau} \left( D_X^k(x) - D_{X+1}^k(x) \right) + \frac{1 - \gamma_x}{\tau} \left( D_X(x, x_0^k) - D_X(x, x_{0+1}^k) \right) + M_x \left( D_X(x_0^k, x_{-1}^k) - D_X(x_{0+1}^k, x_{-1+1}^k) \right) + \frac{\gamma_y}{\sigma} \left( D_Y^k(y) - D_{Y+1}^k(y) \right) + \frac{1 - \gamma_y}{\sigma} \left( D_Y(y, y_0^k) - D_Y(y, y_{0+1}^k) \right) + M_y \left( D_Y(y_0^k, y_{-1}^k) - D_Y(y_{0+1}^k, y_{-1+1}^k) \right). \] (44)
Note that the inner product term $\langle q_0^{K+1}, y_0^{K+1} - y \rangle$ in the right hand side of (48) can be lower bounded using (25) as follows

$$\langle q_0^{K+1}, y_0^{K+1} - y \rangle \geq - \frac{L^2_{xx}}{\alpha} D_X(x_0^{K+1}, x_{-1}^{K+1}) - \frac{L^2_{yy}}{\beta} D_Y(y_0^{K+1}, y_{-1}^{K+1}) - (\alpha + \beta)D_Y(y, y_0^{K+1}).$$

(45)

Using (45) within (44), the step-size condition (9b) in Assumption 3, and then dropping the nonpositive terms lead to

$$\bar{T} K (\mathcal{L}(\tilde{x}^{(K)}, y) - \mathcal{L}(x, \tilde{y}^{(K)}))$$

$$\leq \sum_{k=1}^{K} \sum_{t=0}^{T_k-1} (A_t^k - B_t^k) + \frac{1}{\eta} Q^0 + \frac{\gamma_x}{\tau} T D_X^1(x) + \frac{1 - \gamma_x}{\tau} D_X(x, x_1^1) + M_x D_X(x_0^1, x_1^{1-1}) + \frac{\gamma_y}{\sigma} T D_Y^1(y) + \frac{1 - \gamma_y}{\sigma} D_Y(y, y_1^0) + M_y D_Y(y_0^1, y_1^{1-1}).$$

(46)

Finally, taking supremum over $z = (x, y) \in Z = X \times Y$, then taking expectation $E[\cdot]$ and using the fact that $E[A_t^k] \leq B_t^k$ from Lemma 2 leads to the result in (12).

**Part II** Let $\{s^k\}_{k \geq 1} \subset \mathbb{R}_+$ be a sequence and $\{P^k\}_{k \geq 1} \subset \mathbb{R}_+$ be a bounded sequence, i.e., there exists $\Delta > 0$ such that $\sup_{k \geq 1} P^k \leq \Delta$, then

$$\sum_{k=1}^{K} s^k (P^k - P^{k+1}) = s^1 P^1 + \left[ \sum_{k=2}^{K} (s^k - s^{k-1}) \Delta \right] - s^K P^{K+1} \leq s^K (\Delta - P^{K+1}).$$

(47)

Therefore, assuming that the Bregman diameter is bounded, one can use (47) for each term in the right hand side of (43) involving differences of two consecutive Bregman distance functions. Hence, summing (43) over $k$ from 1 to $K$, using (47) and Jensen’s inequality we conclude that for $K \geq 1$

$$S^K (\mathcal{L}(\tilde{x}^{(K)}, y) - \mathcal{L}(x, \tilde{y}^{(K)}))$$

$$\leq \sum_{k=1}^{K} \sum_{t=0}^{T_k-1} (A_t^k - B_t^k) - \left\langle q_0^{K+1}, y_0^{K+1} - y \right\rangle + \frac{1}{\eta K} (B_X + 2B_Y - Q^{K+1})$$

$$+ \frac{\gamma_x}{\tau K} T^K (B_X - D_X^{K+1}(x)) + \frac{1 - \gamma_x}{\tau K} (B_X - D_X(x, x_0^{K+1}))$$

$$+ \frac{\gamma_y}{\sigma K} T^K (B_Y - D_Y^{K+1}(y)) + \frac{1 - \gamma_y}{\sigma K} (B_Y - D_Y(y, y_0^{K+1})).$$
where $S^K = \sum_{k=1}^K T^k$. Similar to the previous case the inner product can be bounded by (45) and using the step-size condition (9b) we obtain,

$$S^K(\mathcal{L}(\tilde{x}^{(K)}, y) - \mathcal{L}(x, \tilde{y}^{(K)}))
\leq \sum_{k=1}^K T^k \sum_{t=0}^{T^k-1} (A^k_t - B^k_t) + \left(\frac{1}{\eta^K} + \frac{\gamma_x^K}{\tau^K} T^K + \frac{1 - \gamma_x^K}{\tau^K} + M_x^K\right)B_X
+ \left(\frac{2}{\eta^K} + \frac{\gamma_y^K}{\sigma^K} T^K + \frac{1 - \gamma_y^K}{\sigma^K} + M_y^K\right)B_Y.$$  

Finally, with a similar argument as in Part I and using Lemma 2 the result in (13) can be concluded. \hfill \qed

**Proof of Corollary 1** Assume that the step-sizes and parameters are selected as in (10). We are interested in finding the magnitude of the constants involved in the right-hand side of (12). In particular, $\tau = O(1/\sqrt{n})$ and $\sigma = O(1/\sqrt{n})$, hence, we can find the magnitude of $b_x$ and $b_y$ defined in Remark 4 part I which is $b_x = O(\sqrt{n})$ and $b_y = O(\sqrt{n})$. This immediately implies that $\eta = O(\sqrt{n})$. Moreover, we conclude that $M_x^K = M_y^K = O(\sqrt{n})$ and $M_y^K = M_y = O(\sqrt{n})$. It is easy to verify that $\max\{\frac{1}{\eta^K}, \frac{\gamma_x^K}{\tau^K}, \frac{1 - \gamma_x^K}{\tau^K}, \frac{\gamma_y^K}{\sigma^K}, \frac{1 - \gamma_y^K}{\sigma^K}, M_x, M_y\} = O(\sqrt{n})$ and $\bar{T} = O(n)$; hence, we conclude that the right side of (12) has the rate of $O(\frac{1}{K^{1/2}})$. Therefore, the total number of gradients to achieve $\epsilon$-gap is $nK + \sum_{k=1}^K T^k = O(nK) = O(\frac{n}{\epsilon}).$ \hfill \qed

**Proof of Corollary 2** Assume that the step-sizes and parameters are selected as in (11), then from the fact that $T^k = T(k + 1)^2$ we have that $S^K = \sum_{k=1}^K T(k + 1)^2 = T \left(\frac{(K+1)^3}{3} + \frac{(K+1)^2}{2} + \frac{(K+1)}{6} - 1\right) \geq T\frac{(K+1)^3}{3}$. Moreover, $\tau^k = O(\frac{1}{K})$, $\sigma^k = O(\frac{1}{K})$, $\gamma_x^k = O(\frac{1}{K})$, $\gamma_y^k = O(\frac{1}{K})$, and $\eta^k = O(\frac{1}{K})$ which implies that $\max\{\frac{1}{\eta^K}, \frac{\gamma_x^K}{\tau^K} T^K, \frac{1 - \gamma_x^K}{\tau^K} T^K, \frac{\gamma_y^K}{\sigma^K} T^K, \frac{1 - \gamma_y^K}{\sigma^K} T^K, M_x^K, M_y^K\} = O(K)$; hence, we conclude that the right side of (13) has the rate of $O(1/K^2)$. Therefore, the total number of gradients is $nK + \sum_{k=1}^K T^k = nK + S^K = O\left(\frac{n}{\sqrt{\epsilon}} + \frac{1}{\epsilon^{1/2}}\right)$. \hfill \qed

### 5 Numerical experiment

In this section, we implement SVR-APD with constant (SVR-APD-I) and non-constant (SVR-APD-II) step-sizes for solving DRO problem (3) described in Sect. 1.2, and compare them with the state-of-the-art first-order methods designed for solving large-scale convex-concave SP problems including Stochastic Primal-Dual Hybrid Gradient
Table 2 Datasets from LIBSVM 34: https://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/. For MNIST dataset the training task is performed to classify digits of 4 and 9

|          | Mushrooms | Phishing | a7a   | MNIST  | Synthetic |
|----------|-----------|----------|-------|--------|-----------|
| # of samples | 8124      | 11055    | 16100 | 11791  | 5000      |
| # of features | 112       | 64       | 122   | 780    | 1000      |

Fig. 1 Comparison of the methods in terms of the running time for different datasets from left to right: Mushrooms, Phishing, a7a, MNIST, and Synthetic

(SPDHG) [33], Stochastic Mirror Descent (SMD) [25], and Stochastic Mirror-prox (SMP) [17].

Similar to the setup in [23], we consider \( \{a_i\}_{i=1}^n \subset \mathbb{R}^m \) to be a set of features with labels \( \{b_i\}_{i=1}^n \subset \{-1, +1\}^n \). We consider the logistic loss function, i.e., \( \ell_i(x) = \log(1 + \exp(-b_i a_i^\top x)) \), and Chi-square divergence measure \( V(y, \frac{1}{n}1_n) = \frac{1}{2} \| n y - 1_n \|_2^2 \), and we set \( X = [-10, 10]^d \) and \( \rho = 50 \). Our goal is to compare the performance of the methods when \( n \) (number of samples) is large.

Different datasets have been used for the experiments and summary of the information can be found in Table 2. To generate the synthetic data, we choose a true classifier \( x^* \sim U[-1, 1] \) and the feature vectors \( A = [a_i]_{i=1}^n \in \mathbb{R}^{n \times m} \) are generated independent and identically distributed from \( \mathcal{N}(0, 1) \). We set the labels to be \( b_i = \text{sign}(a_i^\top x^*) \) and flip them with probability of 10% similar to [23].

To compute the projection onto the simplex-set constraint, \( y \in \Delta_n \), in problem (3), we choose the entropy Bregman distance generating function \( \psi_Y(y) = \sum_{i=1}^n y_i \log(y_i) \), where \( y = [y_i]_{i=1}^n \in \mathbb{R}^n \)—see [26] for more details. We also choose the step-sizes \( \tau_k \) and \( \sigma_k \) as in (10b) and (11b) for SVR-APD-I and SVR-APD-II, respectively. Note that in this example \( C_X = C_Y = 1 \) — see the discussion after Remark 4. Since the global Lipschitz constants often provide conservative step-sizes, for all the methods we tuned the step-sizes and plotted the best performance.
Table 3  Comparison of the methods in terms of the gap function $\sup_{(x,y) \in Z} [\mathcal{L}(x_K, y) - \mathcal{L}(x, y_K)]$ for different datasets

| Data set  | SVR-APD-I | SVR-APD-II | SPDHG | SMD | SMP |
|-----------|-----------|-----------|-------|-----|-----|
| Mushrooms | 1.1e-04   | 1.3e-04   | 5.3e-02 | 5.2e-02 | 1.4e-02 |
| Phishing  | 1.2e-02   | 1.8e-02   | 5.7e-01 | 5.6e-01 | 6.1e-01 |
| a7a       | 7.6e-04   | 3.7e-03   | 6.0e-01 | 6.2e-01 | 5.5e-01 |
| MNIST     | 4.5e-02   | 5.9e-02   | 5.7e-01 | 5.0e-01 | 7.0e-01 |
| Synthetic | 6.7e-06   | 1.3e-04   | 1.7e-02 | 1.3e+00 | 1.3e+00 |

We plot the results in terms of the difference of Lagrangian functions, i.e., $\mathcal{L}(x^k, y^*) - \mathcal{L}(x^*, y^k)$, versus the running time of the algorithms in Figure 1. Moreover, we computed the gap function for the last iterate of each method presented in Table 3. For all three experiments our methods outperforms the other three by achieving lower gap function and the superiority of our method is more evident as the total number of samples increases. Moreover, consistent with our results, SVR-APD-I with $O(\sqrt{n}/\epsilon)$ oracle complexity eventually outperform SVR-APD-II with $O(1/\epsilon^{1.5})$ for higher accuracy.

Data Availability  The datasets analysed during the current study are available in the LIBSVM 34, https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

Declarations

Conflict of interest  The authors declare no conflict of interest.

Appendix

The following lemma provides some fundamental properties associated with the Bregman distance functions–see [2, 29] for the proofs.

Lemma 5  Let $(\mathcal{U}, \| \cdot \|_{\mathcal{U}})$ be a finite-dimensional normed vector space with the dual space $(\mathcal{U}^*, \| \cdot \|_{\mathcal{U}^*})$, $f : \mathcal{U} \to \mathbb{R} \cup \{ +\infty \}$ be a closed convex function, $U \subset \mathcal{U}$ is a closed convex set, $\psi : \mathcal{U} \to \mathbb{R}$ be a distance generating function which is continuously differentiable on an open set containing $\text{dom} f$ and 1-strongly convex with respect to $\| \cdot \|_{\mathcal{U}}$, and $D_U : \mathcal{U} \times (\mathcal{U} \cap \text{dom} f) \to \mathbb{R}$ be a Bregman distance function associated with $\psi$. Then, the following result holds.

a) Given $\bar{x} \in U \cap \text{dom} f$, $s \in \mathcal{U}$ and $t > 0$, let $x^+ = \arg\min_{x \in U} f(x) - \langle s, x \rangle + t D_U(x, \bar{x})$. Then for all $x \in U$, the following inequality holds:

$$f(x) \geq f(x^+) + \langle s, x - x^+ \rangle + t \langle \nabla \psi(\bar{x}) - \nabla \psi(x^+), x - x^+ \rangle.$$  \hspace{1cm} (50)

b) For all $x \in U$ and $z \in U \cap \text{dom} f$, $\langle \nabla \psi(z) - \nabla \psi(y), x - z \rangle = D_U(x, y) - D_U(x, z) - D_U(z, y)$.  

 Springer
c) Given the update of $x^+$ in (a), for all $x \in U$ the following inequality holds:

$$
 f(x^+) - f(x) + \langle s, x - \bar{x} \rangle \leq t(D_U(x, \bar{x}) - D_U(x, x^+)) + \frac{1}{2t} \|s\|_U^2. \quad (51)
$$

d) Assuming $\psi(\cdot)$ is a closed function, then $\nabla \psi^*(\nabla \psi(x)) = x$, for all $x \in \text{dom} \nabla \psi \subset U$, and $\nabla \psi(\nabla \psi^*(y)) = y$, for all $y \in \text{dom} \nabla \psi^* \subset U^*$.

References

1. Allen-Zhu, Z., Yuan, Y.: Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In: International conference on machine learning, pp. 1080–1089 (2016)
2. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J., Lafferty, J.: Clustering with bregman divergences. J. Mach. Learn. Res. 6, 10 (2005)
3. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press, Cambridge (2004)
4. Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal-dual algorithm. Math. Program. 159(1–2), 253–287 (2016)
5. Chen, Y., Lan, G., Ouyang, Y.: Accelerated schemes for a class of variational inequalities. Math. Program. 165(1), 113–149 (2017)
6. Dang, C., Lan, G.: Randomized first-order methods for saddle point optimization. http://arxiv.org/abs/1409.8625 (2014)
7. Defazio, A., Bach, F., Lacoste-Julien, S.: Saga: a fast incremental gradient method with support for non-strongly convex composite objectives. Adv. Neural Inf. Process. Syst. 27, 1646–1654 (2014)
8. Devraj, A.M., Chen, J.: Stochastic variance reduced primal dual algorithms for empirical composition optimization. Adv. Neural Inf. Process. Syst. 45, 9882–9892 (2019)
9. Du, S.S., Hu, W.: Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. http://arxiv.org/abs/1802.01504 (2018)
10. Hamedani, E.Y., Aybat, N.S.: A primal-dual algorithm for general convex-concave saddle point problems. http://arxiv.org/abs/1803.01401 (2018)
11. Hamedani, E.Y., Aybat, N.S.: A primal-dual algorithm with line search for general convex-concave saddle point problems. SIAM J. Optim. 31(2), 1299–1329 (2021)
12. He, N., Juditsky, A., Nemirovski, A.: Mirror prox algorithm for multi-term composite minimization and semi-separable problems. Comput. Optim. Appl. 61(2), 275–319 (2015)
13. He, Y., Monteiro, R.D.: An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems. SIAM J. Optim. 26(1), 29–56 (2016)
14. Hien, L.T.K., Zhao, R., Haskell, W.B.: An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems. http://arxiv.org/abs/1711.03669 (2017)
15. Jalilzadeh, A., Hamedani, E.Y., Aybat, N.S., Shanbhag, U.V.: A randomized block-coordinate primal-dual method for large-scale saddle point problems. http://arxiv.org/abs/1907.03886 (2019)
16. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. Adv. Neural Inf. Process. Syst 26, 315–323 (2013)
17. Juditsky, A., Nemirovski, A., Tautu, C.: Solving variational inequalities with stochastic mirror-prox algorithm. Stoch. Syst. 1(1), 17–58 (2011)
18. Juditsky, A., Nemirovski, A., et al.: First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. Optim. Mach. Learn. 30, 149–183 (2011)
19. Kolosowski, O., Monteiro, R.D.: An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex-concave saddle-point problems. Optim. Methods Softw. 32(6), 1244–1272 (2017)
20. Lanier, G.R., Cristianini, N., Bartlett, P., Ghouti, L.E., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. J. Mach. Learn. Res. 5(Jan), 27–72 (2004)
21. Lian, X., Wang, M., Liu, J.: Finite-sum composition optimization via variance reduced gradient descent. Artif. Intell. Stat. 45, 1159–1167 (2017)
22. Malitsky, Y.: Proximal extrapolated gradient methods for variational inequalities. Optim. Methods Softw. 33(1), 140–164 (2018)
23. Namkoong, H., Duchi, J.C.: Stochastic gradient methods for distributionally robust optimization with f-divergences. Adv. Neural Inf. Process. Syst. 29, 2208–2216 (2016)
24. Nemirovski, A.: Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipshitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM J. Optim. 15(1), 229–251 (2004)
25. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. 19(4), 1574–1609 (2009)
26. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. 103(1), 127–152 (2005)
27. Palaniappan, B., Bach, F.: Stochastic variance reduction methods for saddle-point problems. Adv. Neural Inf. Process. Syst. 29, 1416–1424 (2016)
28. Roux, N.L., Schmidt, M., Bach, F.R.: A stochastic gradient method with an exponential convergence rate for finite training sets. Adv. Neural Inf. Process. Syst. 25, 2663–2671 (2012)
29. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization (2008). Available at http://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf
30. Wang, J., Xiao, L.: Exploiting strong convexity from data with primal-dual first-order algorithms. http://arxiv.org/abs/1703.02624 (2017)
31. Yan, Y., Xu, Y., Lin, Q., Zhang, L., Yang, T.: Stochastic primal-dual algorithms with faster convergence than $o(1/\sqrt{T})$ for problems without bilinear structure. http://arxiv.org/abs/1904.10112 (2019)
32. Zhang, J., Xiao, L.: A composite randomized incremental gradient method. In: International Conference on Machine Learning, pp. 7454–7462 (2019)
33. Zhao, R.: Accelerated stochastic algorithms for convex-concave saddle-point problems. Math. Op. Res. 47(2), 1443–1473 (2022)
34. Chih-Chung, Chang Chih-Jen, Lin LIBSVM ACM Transactions on Intelligent Systems and Technology 2(3) 1–27 (2011). https://doi.org/10.1145/1961189.1961199

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.