RANDOMIZED FIRST-ORDER METHODS FOR SADDLE POINT OPTIMIZATION

CONG D. DANG † AND GUANGHUI LAN ‡

Abstract. In this paper, we present novel randomized algorithms for solving saddle point problems whose dual feasible region is a direct product of many convex sets. Our algorithms can achieve $O(1/N)$ rate of convergence by solving only one dual subproblem at each iteration. Our algorithms can also achieve $O(1/N^2)$ rate of convergence if a strongly convex assumption on the dual problem is made. When applied to linearly constrained problems, they need to solve only one randomly selected subproblem per iteration instead of solving all as in the Alternating Direction Method of Multipliers.

Keywords. Stochastic Optimization, Block Coordinate Descent, Nonsmooth Optimization, Saddle Point Optimization, Alternating Direction Method of Multipliers

1. Introduction. In this paper, we consider the saddle point problem

$$
\min_{x \in X} \left\{ G(x) + \max_{y \in Y_i} \langle x, \sum_{i=1}^p A_i^T y_i \rangle - J(y) \right\},
$$

where $X$ and $Y$ respectively are given closed convex sets in $\mathbb{R}^m$ and $\mathbb{R}^n$, $A_i \in \mathbb{R}^{n_i \times m}$, $i = 1, 2, ..., p$ and $\sum_{i=1}^p n_i = n$, $G(x)$ is a general smooth convex function, $J(y)$ is a reality simple, proper, convex. In addition, we assume that $Y$ has a block structure, i.e.,

$$
Y = Y_1 \times Y_2 \times \cdots \times Y_p,
$$

and $J(y)$ is coordinate-wise separable, i.e.

$$
J(y) = \sum_{i=1}^p J_i(y_i),
$$

in which $J_i : Y_i \to \mathbb{R}$, $i = 1, ..., p$, are reality simple, proper, convex.

These types of saddle point problems found many applications in constrained optimization and data analysis, especially in image processing and machine learning. Convex optimization problems with linear constraints can be easily reformulated to a saddle point problem by introducing a Lagrange multiplier to account for the inequalities constraints. In many data analysis applications, while $G(x)$ is a data fidelity term, the dual problem in (1.1) plays as a certain regularization, e.g., total variation [21], group lasso [10, 9]. Our work also is motivated by solving the optimization problem whose the objective is a summation of many convex functions which was studied in [9]. These problems in the form of (1.1) can be solved effectively by using primal-dual algorithms [8, 2, 5, 23] and Alternating Direction Method of Multipliers (ADMM) methods [1] [12, 6, 19].

This paper is also motivated by the currently active research on block coordinate descent methods (see, e.g., [17, 11, 18, 20, 1, 13, 3]) for solving problems with a separable feasible set. In comparison with regular first-order methods, each iteration of these BCD methods updates only one block of variables. Although simple, these methods are found to be effective in solving huge-scale problems with $n$ as big as $10^8 - 10^{12}$, and hence are very useful for dealing with high-dimensional problems, especially those from large-scale data analysis applications. We refer to [20] for an excellent review on the earlier developments of BCD methods.

In this paper, we propose a novel algorithm, namely Randomized First-order Methods for Saddle Point Optimization, to solve this problem in the form of (1.1). The main idea is to incorporate a block decomposition of dual space into the primal-dual algorithm in [2]. At each iteration, our algorithm requires to solve
only one subproblem in dual space instead of \( p \) subproblems as in primal-dual and ADMM algorithms. We show that our algorithm can also achieve a \( 1/N \) rate of convergence without any strongly convex assumption and achieve \( 1/N^2 \) rate of convergence with strongly convex assumption on the \( J(y) \), where \( N \) is the number of iterations. Furthermore, we demonstrate that our algorithm can deal with the situation when either \( X \) or \( Y \) is bounded, as long as a saddle point of problem (1.1) exists, by using a perturbation-based termination criterion that similar to the one employed by Monteiro and Svaiter [15]. We have found that, in a concurrent work by Zhang and Xiao [22], the problem and algorithm seem to be similar, however our work is quite different and completely independent in the sense of both analysis and termination criterion. In particular, Zhang and Xiao focus on a specific problem, regularized empirical risk minimization (ERM) of linear predictors, which requires strongly convex assumption on both \( G \) and \( J \) in analysis, while our paper consider a more general problem without any strongly convex assumption and with strongly convex assumption on \( J \) only. In addition, Zhang and Xiao use the distance to the unique saddle point and function value as termination criteria while we use duality gap based characteristic to access the quality of a feasible solution.

This paper is organized as follows. We present the Randomized First-order Methods for Saddle Point Optimization in Section 2 and discuss its convergence properties without strongly convex assumption in Section 3. In Section 4, we discuss its convergence properties with a strongly convex assumption on the dual problem. Finally some brief concluding remarks are given in Section 5.

2. Randomized First-order Methods for Saddle Point Optimization. In this section we will introduction some notations to rewrite (1.1) and describe the algorithm formally. Denote \( A = [A_1; A_2; \cdots ; A_p] \), and

\[
U_i = \begin{bmatrix}
0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\
0 & \ddots & 0 & \ddots & \vdots & \vdots & \vdots & 0 \\
\vdots & 0 & 0 & 0 & \ddots & \vdots & \vdots & \vdots \\
\vdots & \vdots & 0 & I_{n_i} & 0 & \ddots & \vdots & 0 \\
\vdots & \vdots & \vdots & 0 & 0 & 0 & \ddots & 0 \\
0 & \vdots & \vdots & \vdots & 0 & \ddots & 0 & \ddots \\
0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0
\end{bmatrix},
\]

where \( I_{n_i} \) is identity matrix size \( n_i \times n_i \). We also denote \( \bar{U}_i \) is a matrix such that \( U_i + \bar{U}_i = I_n \), where \( I_n \) is identity matrix size \( n \times n \). Then, our problem can be rewritten as

\[
\min_{x \in X} \left\{ G(x) + \max_{y \in Y} \langle Ax, y \rangle - J(y) \right\}
\]

(2.1)

Using above notation, our algorithm framework can be described as following
Algorithm 1 Randomized First-order Methods for Saddle Point Optimization

Let \( z^t = (x^t, y^t) \in X \times Y \) and stepsizes \( \{\gamma_t\}_{t \geq 1}, \{\theta_t\}_{t \geq 1}, \{q_t\}_{t \geq 1}, \{\tau_t\}_{t \geq 1}, \{\eta_t\}_{t \geq 1} \).

for \( t = 1, \ldots, N \) do
  1. Generate a random variable \( i_t \) uniformly from \( \{1, 2, \ldots, p\} \).
  2. Update \( y^{t+1} \) and \( x^{t+1} \) by

\[
y^{t+1} = \text{argmin}_{y \in Y} \gamma_t (-U_{i_t} A x^t, y) + \gamma_t J_{i_t}(y_t) + \frac{\gamma_t}{2\tau_t} \| y - y^t \|_2^2,
\]
\[
x^{t+1} = \text{argmin}_{x \in X} q_t G(x) + q_t \langle x, A^T y^{t+1} \rangle + \frac{q_t}{2\eta_t} \| x - x^t \|_2^2.
\]

3. Set \( \tilde{x}^{t+1} = \theta_t (x^{t+1} - x^t) + x^{t+1} \).

end for

Output: Set

\[
\tilde{z}^N = \sum_{t=1}^{N-1} q_t \tilde{z}^{t+1}.
\]

First we will consider the case that \( X \) and \( Y \) are bounded. We assume that there exist a constant \( M > 0 \) such that \( \| x \| \leq M, \forall x \in X \). We also assume that there exists \( \Omega_x, \Omega_y > 0 \) such that

\[
\sup_{x_1, x_2 \in X} \frac{1}{2} \| x_1 - x_2 \|_2^2 \leq \Omega_x^2 \quad \text{and} \quad \sup_{y_1, y_2 \in Y} \frac{1}{2} \| y_1 - y_2 \|_2^2 \leq \Omega_y^2.
\]

3. Convergence Analysis. We first introduce a notion to characterize the solution of (2.1) that will be used to analyze the convergence properties of Algorithm 1.

Definition 3.1. For any \( z = (x, y) \in X \times Y \) and \( \hat{z} = (\hat{x}, \hat{y}) \in X \times Y \), we define

\[
Q(\hat{z}, z) := [G(\hat{x}) + \langle A \hat{x}, y \rangle - J(y)] - [G(x) + \langle A x, \hat{y} \rangle - J(\hat{y})].
\]

We know that \( \hat{z} = (\hat{x}, \hat{y}) \) is a pair of saddle point of (2.1) if and only if for any \( y \in Y \) and \( x \in X \), we have \( Q(\hat{z}, z) \leq 0 \).

The following Lemma presents two obvious observations that can be easily obtained from optimality condition of subproblems (2.2) and (2.3).

Lemma 3.2. Suppose \( \{x^t\}_{t \geq 1} \) and \( \{y^t\}_{t \geq 1} \) are generated by Algorithm 1. Then

a) For all \( y \in Y \), we have

\[
\gamma_t \left[ (-U_{i_t} A x^t, y^{t+1} - y) + J_{i_t}(y_t) - J_{i_t}(y_t) \right] + \frac{\gamma_t}{2\tau_t} \| y^t - y^{t+1} \|_2^2 + \frac{\gamma_t}{2\tau_t} \| y - y^{t+1} \|_2^2 \leq \frac{\gamma_t}{2\tau_t} \| y - y^t \|_2^2.
\]

b) For all \( x \in X \), we have

\[
q_t \left[ G(x^{t+1}) - G(x) + \langle x^{t+1} - x, A^T y^{t+1} \rangle \right] + \frac{q_t}{2\eta_t} \| x^t - x^{t+1} \|_2^2 + \frac{q_t}{2\eta_t} \| x - x^{t+1} \|_2^2 \leq \frac{q_t}{2\eta_t} \| x - x^t \|_2^2.
\]

Proof. The proof follows optimality condition of (2.2) and (2.3) similar to Lemma 2 of [7].

The following Proposition establishes an important recursion.

Proposition 3.3. For all \( t = 1, \ldots, N \), we have

\[
Q(z^t, z^{t+1}) \leq \sum_{s=1}^{t} \text{Lemma 3.2 \( a \text{ or } b \)}.
\]
Proposition 3.3. Suppose \( \{y_t^i\}_{t \geq 1} \) and \( \{x^t\}_{t \geq 1} \) are generated by the above algorithm. Then for any \( y \in Y \) and \( x \in X \), we have

\[
q_t \left[ G(x^{t+1}) - G(x) + \langle Ax^{t+1}, y \rangle - \langle Ax, y^{t+1} \rangle \right] + \gamma_t \left[ J(y^{t+1}) - \frac{p - 1}{p} J(y_t^i) - \frac{1}{p} J(y) \right]
\]

\[
\leq \frac{q_t}{2 \tau} \|x - x^t\|^2 - \frac{q_t}{2 \tau} \|x - x^{t+1}\|^2 + \frac{q_t}{2 \tau} \|x^{t+1} - x^t\|^2 + \frac{\gamma_t}{2} \|x^{t+1} - x^t\|^2 + \frac{\gamma_t}{2} \|x^t - x^{t+1}\|^2
\]

\[
+ \left( \frac{\theta_{t-1} - \gamma_t}{\frac{p - 1}{p} \gamma_t} \right) A x^t - \frac{\theta_{t-1} - \gamma_t}{\frac{p - 1}{p} \gamma_t} Ax^{t+1}, y^t - y \right) + \delta_t + \mu_t,
\]

where \( \delta_t \) are defined as

\[
\delta_t = E_t \left[ \gamma_t U_i (A^{t+1}, y^{t+1}) - \frac{1}{\gamma_t} J(U_i (x_t^{t+1}, y^{t+1})) \right] + \gamma_t E_t \left[ J_i (y^{t+1} - y_t^i) - J_t (y^{t+1} - y_t^i) \right] - \gamma_t E_t \left[ J_i (y^{t+1} - y_t^i) - J_t (y^{t+1} - y_t^i) \right].
\]

and

\[
\mu_t = \gamma_t J(y^{t+1} - y_t^i) E_t \left[ J(y^{t+1}) + \gamma_t E_t \left[ J_i (y^{t+1} - y_t^i) - J_t (y^{t+1} - y_t^i) \right] - \gamma_t E_t \left[ J_i (y^{t+1} - y_t^i) - J_t (y^{t+1} - y_t^i) \right].
\]

Proof. We have

\[
\langle x^{t+1} - x, A^T y^{t+1} \rangle = \langle Ax^{t+1}, y \rangle - \langle Ax, y^{t+1} \rangle - \langle Ax^{t+1}, y \rangle + \langle Ax^{t+1}, y^{t+1} \rangle
\]

Combining the above relation with observation (3.3), we obtain

\[
q_t \left[ G(x^{t+1}) - G(x) + \langle Ax^{t+1}, y \rangle - \langle Ax, y^{t+1} \rangle + \langle Ax^{t+1}, y^{t+1} \rangle \right]
\]

\[
\leq \frac{q_t}{2 \tau} \|x - x^t\|^2 - \frac{q_t}{2 \tau} \|x - x^{t+1}\|^2 + \frac{q_t}{2 \tau} \|x^{t+1} - x^t\|^2 - \frac{q_t}{2 \tau} \|x - x^{t+1}\|^2 - \frac{q_t}{2 \tau} \|y - y^{t+1}\|^2.
\]

(3.7)

Adding up (3.2) and (3.7), we have

\[
q_t \left[ G(x^{t+1}) - G(x) + \langle Ax^{t+1}, y \rangle - \langle Ax, y^{t+1} \rangle \right] + \left[ q_t Ax^{t+1} - \gamma_t U_i A (x^{t+1}, y^{t+1} - y) + \gamma_t J_i (y^{t+1} - y_t^i) - J_t (y^{t+1} - y_t^i) \right]
\]

\[
\leq \frac{q_t}{2 \tau} \|x - x^t\|^2 - \frac{q_t}{2 \tau} \|y - y^{t+1}\|^2 - \frac{q_t}{2 \tau} \|x - x^{t+1}\|^2 - \frac{q_t}{2 \tau} \|y - y^{t+1}\|^2.
\]

(3.8)

Using the definition 2.1 of \( \bar{x}^t \), we have

\[
\langle q_t Ax^{t+1} - \gamma_t U_i A \bar{x}^t, y^{t+1} - y \rangle
\]

\[
= \langle q_t Ax^{t+1} - \gamma_t U_i A (\theta_{t-1} (x^{t-1} + x) + x^t), y^{t+1} - y \rangle
\]

\[
= \langle q_t Ax^{t+1} - \gamma_t U_i A \bar{x}^t, y^{t+1} - y \rangle - \langle \theta_{t-1} \gamma_t U_i A (x^{t-1} + x), y^{t-1} - y \rangle
\]

\[
- \langle \theta_{t-1} \gamma_t U_i A (x^{t-1} + x), y^{t-1} - y \rangle
\]

Subtracting and then adding the right hand side of the above inequality to \( \langle \gamma_t U_i A \bar{x}^t, y^{t+1} - y \rangle \), we obtain

\[
\langle q_t Ax^{t+1} - \gamma_t U_i A \bar{x}^t, y^{t+1} - y \rangle
\]

\[
= \langle q_t Ax^{t+1} - \gamma_t U_i A \bar{x}^t, y^{t+1} - y \rangle - \langle \gamma_t U_i \theta_{t-1} A (x^{t-1} + x), y^{t-1} - y \rangle
\]

\[
- \langle \gamma_t U_i \theta_{t-1} A (x^{t-1} + x), y^{t-1} - y \rangle
\]

(3.9)

We consider

\[
E_t \left[ J(y^{t+1}) - J_t (y^{t+1}) \right] - \sum_{i \neq i} J_i (y_t^i) - J_t (y^{t+1})
\]

\[
= E_t \left[ J(y^{t+1}) - \frac{p - 1}{p} J(y_t^i) - \frac{1}{p} J(y) \right]
\]

\[
= J(y^{t+1}) - \frac{p - 1}{p} J(y_t^i) - \frac{1}{p} J(y) + E_t \left[ J(y^{t+1}) \right] - J(y^{t+1})
\]
that implies
\[
\gamma_t [J_i(y_t) - J_i(y)] = \gamma_t \left[ J(y^{t+1}) - \frac{p-1}{p} J(y^t) - \frac{1}{p} J(y) \right] - \mu_t. \tag{3.10}
\]
Noting that
\[
\mathbb{E}_{u_t} [\langle \theta_{t-1} \gamma_t U_t A(x_t - x^{t-1}), y_t - y \rangle - \langle \gamma_t \tilde{U}_t A x_t, y_t - y \rangle] = \left( \frac{1}{p} \theta_{t-1} \gamma_t A(x_t - x^{t-1}), y_t - y \right) - \frac{p-1}{p} \gamma_t (A x_t, y_t - y),
\]
\[
= \left( \frac{1}{p} \theta_{t-1} \gamma_t - \frac{p-1}{p} \gamma_t \right) A x_t - \frac{1}{p} \theta_{t-1} \gamma_t A x^{t-1}, y_t - y \right),
\]
then from relations (3.9), (3.10), (3.11) and using definitions of \( \delta_t \) and \( \mu_t \), we have
\[
q_t \left[ G(x^{t+1}) - G(x) + \langle A x^{t+1}, y - (A x, y) \rangle \right] + \gamma_t \left[ J(y^{t+1}) - \frac{p-1}{p} J(y^t) - \frac{1}{p} J(y) \right]
\leq \frac{2q_t}{\gamma_t} \| x - x^t \|^2_2 + \frac{2q_t}{\gamma_t} \| x^t - x^{t+1} \|^2_2 + \frac{2q_t}{\gamma_t} \| y - y^t \|^2_2 + \frac{2q_t}{\gamma_t} \| y^t - y^{t+1} \|^2_2
\leq \frac{\gamma t \tau^2_2 L^2}{2} \| x^t - x^{t+1} \|^2_2 + \frac{2q_t}{\gamma_t} \| y^t - y^{t+1} \|^2_2 \tag{3.11}
\]
where \( L_i = \| U_i A \|_2 \) and \( L = \max_{i=1,2,...,p} L_i \). This inequality and (3.11) imply (3.4).

By using Cauchy-Schwartz inequality, we have
\[
\langle \theta_{t-1} \gamma_t U_t A(x_t - x^{t-1}), y^{t+1} - y^t \rangle \leq \frac{\gamma t \tau^2_2 L^2}{2} \| x^t - x^{t+1} \|^2_2 + \frac{2q_t}{\gamma_t} \| y^t - y^{t+1} \|^2_2 \leq \frac{\gamma t \tau^2_2 L^2}{2} \| x^t - x^{t+1} \|^2_2 + \frac{2q_t}{\gamma_t} \| y^t - y^{t+1} \|^2_2
\]
where \( L_i = \| U_i A \|_2 \) and \( L = \max_{i=1,2,...,p} L_i \). This inequality and (3.11) imply (3.4).

**Theorem 3.4.** Suppose \( \{ y^t \}_{t \geq 1} \) and \( \{ x^t \}_{t \geq 1} \) are generated by the above algorithm. We assume that starting point \( z^1 \) is chosen such that \( x^1 = x^0 \) and \( y^1 = \text{argmin}_{y \in Y} J(y) \). We also assume that these parameters are chosen such that,
\[
\gamma_t = \frac{1}{p} \theta_t \gamma_{t+1}, \quad \gamma_1 = \gamma_{N-1}, \quad q_t = \frac{\theta_t \gamma_{t+1} - \frac{p-1}{p} \gamma_{t+1}}{q_{t+1}}, \quad q_{N-1} = \gamma_{N-1}, \quad \frac{q_i}{q_i - \tau_{i-1}} \leq \frac{2q_t}{\gamma_t}, \quad i = 1, ..., N - 1,
\]
\[
\frac{q_i}{q_i - \tau_{i-1}} \geq \frac{\tau_{i+1} \gamma_{i+1} \theta^2_t L^2}{2 \tau_{i-1} \| A \|^2_2}, \quad i = 1, ..., N - 2,
\]
\[
\frac{q_{N-1}}{q_{N-1} - \tau_{N-2}} \geq \frac{\gamma_{N-1} \tau_{N-2} \| A \|^2_2}{2 \tau_{N-2}} \tag{3.12}
\]

Then, for any \( N \geq 1 \), we have
\[
\mathbb{E} [Q(\hat{z}^N, z)] \leq \left( \sum_{t=1}^{N-1} q_t \right)^{-1} \left[ \frac{q_{N-1} \gamma_{N-1} \Omega^2_n}{\gamma_{N-1} \tau_{N-2} \| A \|^2_2} + \frac{2q_t \gamma_{N-1} \Omega^2_n}{\gamma_{N-1} \tau_{N-2} \| A \|^2_2} + 2 \| A \| M \Omega_Y \right], \forall z \in \mathbb{Z}. \tag{3.13}
\]

where \( \hat{z}^N \) is defined in (2.26) and the expectation is taken corresponding to \( i_{[N]} = (i_1, ..., i_{N-1}) \).

Proof. Denote \( z^{[N]} = \{ (x^t, y^t) \}_{t=1}^{N-1} \) and
\[
B_N(z, z^{[N]}) = \sum_{t=1}^{N-1} \left[ \frac{q_t}{2q_t} \| x - x^t \|^2_2 + \frac{q_t}{2q_t} \| x - x^{t+1} \|^2_2 + \sum_{i=1}^{N-1} \left[ \frac{\gamma_t}{2\tau_t} \| y - y^t \|^2_2 + \frac{\gamma_t}{2\tau_t} \| y - y^{t+1} \|^2_2 \right] \right] \tag{3.14}
\]
Taking summation from $t = 1$ to $N - 1$ both sides of (3.4) and using the stepsize choice (3.12), we obtain

$$
\sum_{t=1}^{N-1} q_t \left[ G(x^{t+1}) - G(x) + \langle Ax^{t+1}, y \rangle - \langle Ax, y^{t+1} \rangle \right] + \sum_{t=1}^{N-1} \left[ J(y^{t+1}) - \frac{p-1}{p} J(y^t) - \frac{1}{p} J(y) \right] 
\leq B_N(z, z^{[N]}) - \sum_{t=1}^{N-2} \left( \frac{q_t}{2\eta_t} - \frac{\tau_t + 1}{2\eta_{t+1}} \right) \| x^{t+1} - x^t \|_2^2 + \frac{1}{2} \frac{\gamma_t}{\eta_t} \| x^t - x^{t-1} \|_2^2 - \frac{\tau_t}{2\eta_{t+1}} \| x^t - x^{t-1} \|_2^2 - \frac{1}{2} \frac{\gamma_t}{\eta_t} \| x^t - x^{t-1} \|_2^2 - \frac{\tau_t}{2\eta_{t+1}} \| x^t - x^{t-1} \|_2^2 - \langle q_{N-1} Ax_N - \gamma_{N-1} Ax_{N-1}, y_N - y \rangle
\leq B_N(z, z^{[N]}) - \sum_{t=1}^{N-2} \left( \frac{q_t}{2\eta_t} - \frac{\tau_t + 1}{2\eta_{t+1}} \right) \| x^{t+1} - x^t \|_2^2 + \frac{1}{2} \frac{\gamma_t}{\eta_t} \| x^t - x^{t-1} \|_2^2 - \frac{\tau_t}{2\eta_{t+1}} \| x^t - x^{t-1} \|_2^2 - \frac{1}{2} \frac{\gamma_t}{\eta_t} \| x^t - x^{t-1} \|_2^2 - \frac{\tau_t}{2\eta_{t+1}} \| x^t - x^{t-1} \|_2^2 - \langle q_{N-1} Ax_N - \gamma_{N-1} Ax_{N-1}, y_N - y \rangle
\leq -\frac{p-1}{p} \gamma_t (A x_t, y^t - y) + \sum_{t=1}^{N-1} (\delta_t + \mu_t).
$$

where the second inequality follows (3.12) and the fact that $x^1 = x^0$.

We consider

$$
\sum_{t=1}^{N-1} q_t \left[ J(y^{t+1}) - \frac{p-1}{p} J(y^t) - \frac{1}{p} J(y) \right] 
= \sum_{t=1}^{N-1} \left( \gamma_t - \frac{p-1}{p} \gamma_{t+1} \right) \left[ J(y^{t+1}) - J(y) \right] + \gamma_{N-1} J(y_N) - \gamma_1 J(y_1) - \gamma_{N-1} \frac{1}{p} J(y) 
\leq \sum_{t=1}^{N-2} \left( \gamma_t - \frac{p-1}{p} \gamma_{t+1} \right) \left[ J(y^{t+1}) - J(y) \right] + \gamma_{N-1} \left[ J(y_N) - J(y) \right],
$$

where the last inequality follow from the fact that $\gamma_t = \gamma_{N-1}$ and $y_t = \text{argmin}_{y \in Y} J(y)$. Then from (3.15), the definition (3.11) and the fact that $q_{N-1} = \gamma_{N-1}$, we have

$$
\sum_{t=1}^{N-1} q_t Q(z^{t+1}, z) \leq B_N(z, z^{[N]}) - \frac{\sum_{t=1}^{N-1} \left( \gamma_t - \frac{p-1}{p} \gamma_{t+1} \right) \left[ J(y^{t+1}) - J(y) \right] + \gamma_{N-1} \left[ J(y_N) - J(y) \right]}{2\eta_{t+1}} \| x^N - x^{N-1} \|_2^2 - \gamma_{N-1} \langle Ax_N - Ax_{N-1}, y_N - y \rangle - \frac{p-1}{p} \gamma_1 (A x_1, y_1 - y) + \sum_{t=1}^{N-1} (\delta_t + \mu_t).
$$

By definition of $\hat{x}^N$, we obtain

$$
\left( \sum_{t=1}^{N-1} q_t \right) Q(z^{[N]}, z) \leq B_N(z, z^{[N]}) - \frac{\sum_{t=1}^{N-1} \left( \gamma_t - \frac{p-1}{p} \gamma_{t+1} \right) \left[ J(y^{t+1}) - J(y) \right] + \gamma_{N-1} \left[ J(y_N) - J(y) \right]}{2\eta_{t+1}} \| x^N - x^{N-1} \|_2^2 - \gamma_{N-1} \langle Ax_N - Ax_{N-1}, y_N - y \rangle - \frac{p-1}{p} \gamma_1 (A x_1, y_1 - y) + \sum_{t=1}^{N-1} (\delta_t + \mu_t).
$$

By using Cauchy-Swartz inequality, we have

$$
-\gamma_{N-1} \langle Ax_N - Ax_{N-1}, y_N - y \rangle \leq \frac{\gamma_{N-1} \gamma_{N-1}}{2\eta_{N-1}} \| A \|_2^2 \| x^N - x^{N-1} \|_2^2 + \frac{\gamma_{N-1} \gamma_{N-1}}{2\eta_{N-1}} \| y^N - y \|_2^2.
$$

These three above inequalities imply that

$$
\left( \sum_{t=1}^{N-1} q_t \right) Q(z^{[N]}, z) \leq B_N(z, z^{[N]}) + \frac{\gamma_{N-1}}{2\eta_{N-1}} \| y^N - y \|_2^2 - \frac{\sum_{t=1}^{N-1} \left( \gamma_t - \frac{p}{p} \gamma_{t+1} - \frac{\tau_t + 1}{\eta_{t+1}} \| A \|_2 \right) \| \| x^N - x^{N-1} \|_2^2 + \frac{p-1}{p} \| A \| \| x^1 \|_2 \| y^1 - y \|_2^2}
\leq \frac{q_{N-1} \Omega_Y^2}{2\eta_{N-1}} \sum_{t=1}^{N-2} \left( \frac{q_t}{2\eta_t} - \frac{\tau_t + 1}{2\eta_{t+1}} \right) \| y^t - y^{t+1} \|_2^2 - \sum_{t=1}^{N-1} \left( \frac{q_t}{2\eta_t} - \frac{\tau_t + 1}{2\eta_{t+1}} \right) \| y^t - y^{t+1} \|_2^2
\leq \frac{q_{N-1} \Omega_X^2}{2\eta_{N-1}} \sum_{t=1}^{N-2} \left( \frac{q_t}{2\eta_t} - \frac{\tau_t + 1}{2\eta_{t+1}} \right) \Omega_X^2 + \frac{p-1}{p} \| A \| \| x^1 \|_2 \| y^1 - y \|_2 + \sum_{t=1}^{N-1} (\delta_t + \mu_t).
$$

By the stepsize choice (3.12), we have $\frac{q_t}{\gamma_t} \leq \frac{q_{t+1}}{\gamma_{t+1}} \leq \frac{q_{t+1}}{\eta_{t+1}}$, then by definition (3.14), we conclude that

$$
B_N(z, z^{[N]}) = \sum_{t=1}^{N-2} \left( \frac{q_t}{2\eta_t} - \frac{\tau_t + 1}{2\eta_{t+1}} \right) \| x^{t+1} - x^t \|_2^2 - \sum_{t=1}^{N-2} \left( \frac{q_t}{2\eta_t} - \frac{\tau_t + 1}{2\eta_{t+1}} \right) \| x^{t+1} - x^t \|_2^2 - \frac{\tau_t}{2\eta_{t+1}} \| x^t - x^{t-1} \|_2^2 - \sum_{t=1}^{N-2} \left( \frac{q_t}{2\eta_t} - \frac{\tau_t + 1}{2\eta_{t+1}} \right) \| y^t - y^{t+1} \|_2^2
\leq \frac{q_{N-1} \Omega_X^2}{2\eta_{N-1}} \sum_{t=1}^{N-2} \left( \frac{q_t}{2\eta_t} - \frac{\tau_t + 1}{2\eta_{t+1}} \right) \Omega_X^2 + \frac{p-1}{p} \| A \| \| x^1 \|_2 \| y^1 - y \|_2 + \sum_{t=1}^{N-1} (\delta_t + \mu_t).
$$

Using the above two relations, noting that $\frac{q_{N-1}}{2\eta_{N-1}} \geq \frac{\gamma_{N-1} \gamma_{N-1}}{2\eta_{N-1} \Omega_X^2}$, we obtain

$$
\left( \sum_{t=1}^{N-1} q_t \right) Q(z^{[N]}, z) \leq \frac{q_{N-1}}{2\eta_{N-1}} \Omega_X^2 + \frac{\gamma_{N-1}}{2\eta_{N-1}} \Omega_Y^2 + \frac{p-1}{p} \| A \| \| x^1 \|_2 \| y^1 - y \|_2 + \sum_{t=1}^{N-1} (\delta_t + \mu_t).
$$

(3.16)
Taking expectation w.r.t $i, t = 1, 2, ..., N - 1$, noting that $E_i[\delta_t + \mu_t] = 0$ and $\frac{1}{p} \leq 1$, we obtain

$$E_{\{i, \tau\}}[Q(\hat{z}^N, z)] \leq \left( \sum_{t=1}^{N-1} q_t \right)^{-1} \left[ \frac{q_{N-1} \Omega_X^2}{2\eta_{N-1}} + \frac{\gamma_{N-1} \Omega_Y^2}{2\tau_{N-1}} + \|A\|M\Omega \right].$$

In the following corollary, we describe a specialized convergence result of the above algorithm after properly choosing parameters $\eta_t, \tau_t$.

**Corollary 3.5.** Suppose $\{y^t\}_{t \geq 1}$ and $\{x^t\}_{t \geq 1}$ are generated by the above algorithm. We assume that starting point $z^1$ is chosen such that $x^1 = x^0$ and $y^1 = \text{argmin}_{y \in Y} J(y)$. We also assume that these parameters are chosen such that (3.12) is satisfied and

$$\gamma_t = 1, \ t = 1, 2, ..., N - 1,$$

and

$$\tau_t = \tau = \frac{\Omega_X}{pL\Omega_Y}, \eta_t = \eta = \frac{\Omega_Y}{pL\Omega_X}.$$  

Then for any $N \geq 1$ we have

$$E_{\{i, \tau\}}[Q(\hat{z}^N, z)] \leq \frac{p}{N^2 + p - 2} \left[ pL\Omega_X\Omega_Y + 2\|A\|M\Omega \right], \forall z \in Z.$$  

**Proof.** It is easy to see that from (3.17), we have

$$\theta_t = p, \ t = 1, 2, ..., N - 1,$$

and

$$q_t = \frac{1}{p}, \ t = 1, 2, ..., N - 2 \text{ and } q_{N-1} = 1,$$

that implies

$$\sum_{t=1}^{N-1} q_t = \frac{N + p - 2}{p}.$$  

Moreover, from the stepsizes choice (3.18) we obtain

$$\frac{q_{N-1} \Omega_X^2}{2\eta_{N-1}} = \frac{pL\Omega_X\Omega_Y}{2}$$  

and

$$\frac{\gamma_{N-1} \Omega_Y^2}{2\tau_{N-1}} = \frac{pL\Omega_X\Omega_Y}{2}.$$  

Plugging to (3.13) we obtain (3.19).

In the following corollary, we introduce a stronger termination criterion than $Q$ function that has been used in Theorem 3.4. In particular, to evaluate the quality of a feasible point $\hat{z}$, we use a extended duality gap function $g$ defined as follow

$$g_\sigma(\hat{z}) = \max_{z \in Z} Q_\sigma(\hat{z}, z),$$

where

$$Q_\sigma(\hat{z}, z) = Q(\hat{z}, z) + \sigma(y),$$  

(3.20)
in which \( \sigma(y) \) is a small perturbation, i.e., \( \mathbf{E}_{i_t}[\sigma(y)] = 0 \). This termination criterion is an extension of duality gap which was suggested in [19].

**Corollary 3.6.** Suppose \( \{y^t\}_{t \geq 1} \) and \( \{x^t\}_{t \geq 1} \) are generated by the above algorithm. We assume that starting point \( z^1 \) is chosen such that \( x^1 = x^0 \) and \( y^1 = \arg\min_{y \in \mathcal{Y}} J(y) \). We also assume that these parameters are chosen such that (3.12) is satisfied. Then, for any starting point \( z \), we denote

\[
\sigma = (\sum_{t=1}^{N-1} q_t) \mathbf{E}_{i_t}[g_\sigma(z^N)] \leq \frac{1}{2\eta N} \Omega_X^2 + \frac{1}{2\tau N} \Omega_Y^2 + 2\|A\| M \Omega_Y.
\]

(3.21)

where \( \mathbf{E}_{i_t}[\sigma] = 0 \) and \( \Omega_X, \Omega_Y \) are defined in (2.4).

**Proof.** The proof is almost similar to the proof of Theorem 3.4. The main idea is to break down the perturbation terms \( \delta_t \) and \( \mu_t \) into two pieces, one depends on \( y \) and one is independent from \( y \). In particularly, we denote

\[
\delta_{1t} = \mathbf{E}_{i_t} \left[ (\gamma_t \theta_{t-1} U_{i_t} A(x^t - x^{t-1}), y^t) - (\gamma_t \tilde{U}_{i_t} A x^t, y^t) \right] - \left[ (\gamma_t \theta_{t-1} U_{i_t} A(x^t - x^{t-1}), y^t) - (\gamma_t \tilde{U}_{i_t} A x^t, y^t) \right].
\]

(3.22)

\[
\delta_{2t} = \mathbf{E}_{i_t} \left[ (\gamma_t \theta_{t-1} U_{i_t} A(x^t - x^{t-1}), y) - (\gamma_t \tilde{U}_{i_t} A x^t, y) \right] - \left[ (\gamma_t \theta_{t-1} U_{i_t} A(x^t - x^{t-1}), y) - (\gamma_t \tilde{U}_{i_t} A x^t, y) \right].
\]

(3.23)

and

\[
\mu_{1t} = \gamma_t J(y^{t+1}) - \gamma_t \mathbf{E}_{i_t}[J(y^{t+1})] + \gamma_t \mathbf{E}_{i_t}[J_i(y^{t+1})] - \gamma_t [J_i(y^{t+1})] = \gamma_t [J_i(y^{t+1})].
\]

(3.24)

\[
\mu_{2t} = \gamma_t \mathbf{E}_{i_t}[J_i(y^{t+1})] - \gamma_t [J_i(y^{t+1})].
\]

(3.25)

It is easy to see that \( \delta_t = \delta_{1t} - \delta_{2t} \) and \( \mu_t = \mu_{1t} - \mu_{2t} \). Using exactly the same analysis and Theorem 3.4 except putting the perturbations \( \delta_{2t}, \mu_{2t} \) to the left hand side of (3.16), we have

\[
\left( \sum_{t=1}^{N-1} q_t \right) Q(\hat{z}^N, z) + \sum_{t=1}^{N-1} (\delta_{2t} + \mu_{2t}) \leq \frac{q_{N-1}}{2\eta N-1} \Omega_X^2 + \frac{\gamma_{N-1}}{2\tau N-1} \Omega_Y^2 + \frac{p-1}{p} \|A\| \|x^1\| y^1 - y^2 + \sum_{t=1}^{N-1} (\delta_{1t} + \mu_{1t}).
\]

Let denote \( \sigma = (\sum_{t=1}^{N-1} q_t) \mathbf{E}_{i_t} [\delta_{1t} + \mu_{1t}] \), then from the above inequality we obtain

\[
\left( \sum_{t=1}^{N-1} q_t \right) Q_\sigma(\hat{z}^N, z) \leq \frac{q_{N-1}}{2\eta N-1} \Omega_X^2 + \frac{\gamma_{N-1}}{2\tau N-1} \Omega_Y^2 + \frac{p-1}{p} \|A\| \|x^1\| y^1 - y^2 + \sum_{t=1}^{N-1} (\delta_{1t} + \mu_{1t}).
\]

Maximizing both sides w.r.t. \( z = (x, y) \) then taking expectation w.r.t. \( i_t, t = 1, 2, ..., N-1 \), noting that \( \mathbf{E}_{i_t}[\delta_{1t} + \mu_{1t}] = 0 \) and \( \frac{p}{p-1} \leq 1 \), we obtain

\[
\mathbf{E}_{i_t}[\max_{z \in \mathcal{Z}} Q_\sigma(\hat{z}^N, z)] \leq \left( \sum_{t=1}^{N-1} q_t \right) \mathbf{E}_{i_t} [\delta_{1t} + \mu_{1t}].
\]

In the remaining of this section, we describe the convergence properties for the algorithm applied to Saddle Point Problem with unbounded feasible set \( \mathcal{Z} \). To assess the quality of a feasible solution \( \hat{z} \), we use a perturbation-based criterion recently employed by Monteiro and Svaiter and applied to SPP \[15,16,14\]. One advantage of this termination criterion is that its definition does not depend on the boundedness of the domain. In particular, we define \( \hat{Q} \) as an extension of \( Q \) function such that

\[
\hat{Q}(\hat{z}, z, v) = Q(\hat{z}, z) + \langle v, \hat{z} - z \rangle,
\]

8
and the extended duality gap \( g_\sigma(\hat{z}, v) \) is an extension of \( g_\sigma(\hat{z}) \) such that
\[
\hat{g}_\sigma(\hat{z}, z, v) = \max_{z \in Z} Q_\sigma(\hat{z}, z) + \langle v, \hat{z} - z \rangle,
\]
where \( Q_\sigma(\hat{z}, z) \) is defined in \((3.20)\).

In the following lemma, using the above extended definitions, we will show that for every \( t \leq N \), the distance from \( z^t \) to the saddle point \( \hat{z} \) is bounded.

**Lemma 3.7.** Suppose that \( z^t = (x^t, y^t), t = 1, 2, ..., N \) are generated by the Algorithm. We assume that starting point \( z^1 \) is chosen such that \( x^1 = x^0 = 0 \) and \( y^1 = \arg\min_{y \in Y} J(y) \). We also assume that these stepsizes satisfy \((3.22)\) and
\[
\frac{\tau_{i-1}}{\gamma_{i-1}} = \frac{2}{\gamma_i}, \quad i = 1, ..., N - 1,
\]
\[
\frac{\eta_{i-1}}{\gamma_{i-1}} = \frac{2}{\eta_i}, \quad i = 1, ..., N - 1,
\]
\[
\frac{\gamma_{i-1}}{\gamma_{i-1}} = \frac{2}{\gamma_i}, \quad i = 1, ..., N - 2.
\]
(3.26)

Then,

- For any \( t \leq N - 1 \), we have
\[
E_{[i]} \left[ ||\hat{x} - x^i||_2^2 \right] \leq 2D^2
\]
(3.27)
and
\[
E_{[i]} \left[ ||\hat{x} - x^N||_2^2 \right] \leq D^2
\]
(3.28)

and
\[
E_{[i]} \left[ ||\hat{y} - y^i||_2^2 \right] \leq \frac{p(2\gamma_{i-1} - q_{i-1})\tau_{i-1}}{(p - 1)\gamma_{i-1}\eta_{i-1}} D^2,
\]
(3.29)
\[
E_{[i]} \left[ ||\hat{y} - y^N||_2^2 \right] \leq \frac{\tau_{N-1}q_{N-1}}{\gamma_{N-1}\eta_{N-1}} D^2
\]
(3.30)

where
\[
D := \sqrt{||x^* - x^1||_2^2 + \frac{\gamma_1\eta_1}{\tau_1q_1} ||y^* - y^1||_2^2},
\]
(3.31)
in which \( z^* = (x^*, y^*) \) is a saddle point of \((2.1)\).

- We have
\[
E_{[i]} \left[ Q(\hat{z}^N, z, v_N) \right] \leq \left( \sum_{t=1}^{N-1} \eta_t \right)^{-1} \left[ \frac{q_{N-1}}{2\eta_{N-1}} E_{[i]} \left[ ||\hat{x}^N - x^1||_2^2 \right] + \frac{\gamma_{N-1}}{2\tau_{N-1}} E_{[i]} \left[ ||\hat{y}^N - y^1||_2^2 \right] \right]
\]
(3.32)

where \( v_N \) is defined as
\[
v_N = \left( \sum_{t=1}^{N-1} q_t \right)^{-1} \left( \frac{q_{N-1}}{\eta_{N-1}} (x^N - x^1), \gamma_{N-1} (Ax^N - Ax^{N-1}) + \frac{\gamma_{N-1}}{\tau_{N-1}} (y^N - y^1) \right).
\]
(3.33)

**Proof.** First, we prove part a). From \((3.15)\), using the stepsize choice \((3.26)\), the fact that \( x^1 = x^0 = 0 \), and replacing \( z \) by \( z^* \), we obtain
\[
\sum_{i=1}^{N-1} q_t Q(z^{t+1}, z^*) \leq \frac{q_1}{\eta_1} ||x^1 - x^*||_2^2 - \frac{q_{N-1}}{\eta_{N-1}} ||x^N - x^*||_2^2 + \frac{\gamma_1}{\tau_1} ||y^1 - y^*||_2^2 - \frac{\gamma_{N-1}}{\tau_{N-1}} ||y^N - y^*||_2^2 + \sum_{i=1}^{N-1} (\delta_t + \mu_t).
\]
(3.33)
Taking expectation w.r.t. \([i_N]\), noting that \(Q(z^{i+1}, z^*) \geq 0, \forall t \geq 1\) and \(E_i[\delta_t + \mu_i] = 0\), we obtain (3.28) and (3.30). For any \(t \leq N - 1\), using similar analysis to (3.15) and the fact that \(x^t = x^0 = 0\), we have

\[
\sum_{i=1}^{t-1} \eta_i \left[ G(x^{i+1}) - G(x) + \langle Ax^{i+1}, y \rangle - \langle Ax, y^{i+1} \rangle \right] + \sum_{i=1}^{t-1} \gamma_i \left[ J(y^{i+1}) - \frac{p-1}{p} J(y^t) - \frac{1}{p} F(y) \right]
\]

\[
\leq B(z, z[0]) - \frac{\eta_{t-1}}{2\eta_{t-1}} \| x^t - x^{t-1} \|_2^2 - \langle q_{t-1} Ax^t - \gamma_{t-1} Ax^{t-1}, y^t - y \rangle + \sum_{i=1}^{t-1} (\delta_i + \mu_i).
\]

We have

\[
-\langle q_{t-1} Ax^t - \gamma_{t-1} Ax^{t-1}, y^t - y \rangle = -\gamma_{t-1} \langle Ax^t - Ax^{t-1}, y^t - y \rangle + (\langle \gamma_{t-1} - q_{t-1} \rangle Ax^t, y^t - y) = -\gamma_{t-1} \langle Ax^t - Ax^{t-1}, y^t - y \rangle + (\langle \gamma_{t-1} - q_{t-1} \rangle \rangle Ax, y^t - \langle Ax^t, y^t \rangle - \langle Ax^t, y^t \rangle - \langle Ax^t, y^t \rangle.
\]

By using Cauchy-Schwartz inequality, we have

\[
-\gamma_{t-1} \langle Ax^t - Ax^{t-1}, y^t - y \rangle \leq \gamma_{t-1} \| A \| \| x^t - x^{t-1} \|_2 \| y^t - y \|_2 
\]

\[
\leq \gamma_{t-1} \gamma_{t-1} \| A \|^2 \| x^t - x^{t-1} \|_2^2 + \gamma_{t-1} \| y^t - y \|_2^2 + \sum_{i=1}^{t-1} (\delta_i + \mu_i),
\]

By Lemma 3.2 we also have

\[
\langle Ax^t - Ax, y^t \rangle \leq \frac{1}{2\eta_{t-1}} \| x^t - x^{t-1} \|_2^2 - \| x^t - x^{t-1} \|_2^2 - \| x^t - x^{t-1} \|_2^2 - \| G(x^t) - G(x) \|
\]

Combining three above relations with (3.31) and replacing \(z\) by \(z^*\), we obtain

\[
\sum_{i=1}^{t-1} \eta_i \Gamma(x^{i+1}, z^*) \leq \frac{\eta_i}{2\eta_i} \| x^t - x^{t-1} \|_2^2 - \| x^t - x^{t-1} \|_2^2 - \| x^t - x^{t-1} \|_2^2 + \sum_{i=1}^{t-1} (\delta_i + \mu_i),
\]

\[
\leq \left( \frac{\eta_{t-1}}{2\eta_{t-1}} + \frac{\gamma_{t-1} - q_{t-1}}{2\eta_{t-1}} \right) E_{[i]} \| x^t - x^{t-1} \|_2^2 + \frac{\gamma_{t-1} - q_{t-1}}{2\eta_{t-1}} \| y^t - y^{t-1} \|_2^2 + \sum_{i=1}^{t-1} (\delta_i + \mu_i),
\]

where \(\tilde{\eta}_i = \eta_i, i = 1, ..., t - 2\) and \(\tilde{\eta}_{t-1} = \gamma_{t-1}\).

By definition of gap function, we know that \(Q(z^{i+1}, z^*) \geq 0 \forall i \geq 1\). Taking expectation w.r.t. \([i_t]\), noting that \(E_i[\delta_i + \mu_i] = 0\), \(2\eta_{t-1} = \frac{\eta_{t-1}}{\eta_{t-1}}, \frac{\eta_{t-1}}{\eta_{t-1}} = \frac{\eta_{t-1}}{\eta_{t-1}}\) and \(\frac{\gamma_{t-1} - q_{t-1}}{2\eta_{t-1}} + \frac{\leq \gamma_{t-1} - q_{t-1}}{2\eta_{t-1}} - \frac{\eta_{t-1} - \eta_{t-1}}{2\eta_{t-1}} \| A \|^2 \geq 0\) we have

\[
\leq \left( \frac{\eta_{t-1}}{2\eta_{t-1}} + \frac{\gamma_{t-1} - q_{t-1}}{2\eta_{t-1}} \right) E_{[i]} \| x^t - x^{t-1} \|_2^2 + \frac{\gamma_{t-1} - q_{t-1}}{2\eta_{t-1}} \| y^t - y^{t-1} \|_2^2 + \sum_{i=1}^{t-1} \| y^t - y^{t-1} \|_2^2,
\]

Dividing both sides of the above inequality by \(\frac{\eta_{t-1}}{2\eta_{t-1}}\), we have

\[
\| x^t - x^{t-1} \|_2^2 + \frac{\gamma_{t-1} - q_{t-1}}{\eta_{t-1} q_{t-1}} \| y^t - y^{t-1} \|_2^2 + \sum_{i=1}^{t-1} \| y^t - y^{t-1} \|_2^2,
\]

\[
(1 + \frac{\gamma_{t-1} - q_{t-1}}{q_{t-1}}) E_{[i]} \| x^t - x^{t-1} \|_2^2 \leq \| x^t - x^{t-1} \|_2^2 + \frac{\gamma_{t-1} - q_{t-1}}{\eta_{t-1} q_{t-1}} \| y^t - y^{t-1} \|_2^2 + \frac{\gamma_{t-1} - q_{t-1}}{q_{t-1}} E_{[i-1]} \| x^t - x^{t-1} \|_2^2.
\]

Since \(1 - \frac{1}{p} \geq 0\), then (3.35) implies

\[
(1 + \frac{\gamma_{t-1} - q_{t-1}}{q_{t-1}}) E_{[i]} \| x^t - x^{t-1} \|_2^2 \leq \| x^t - x^{t-1} \|_2^2 + \frac{\gamma_{t-1} - q_{t-1}}{\eta_{t-1} q_{t-1}} \| y^t - y^{t-1} \|_2^2 + \frac{\gamma_{t-1} - q_{t-1}}{q_{t-1}} E_{[i-1]} \| x^t - x^{t-1} \|_2^2.
\]

We denote

\[
a = \frac{\gamma_{t-1} - q_{t-1}}{q_{t-1}}, i = 1, ..., t,
\]
Then the above inequality becomes, for any $t \leq N$,
\[
E_{[t]} \left[ \|x^* - x^t\|^2 \right] \leq \frac{D^2}{\tau t + a} + \frac{q_{t-1}}{\tau t + a} E_{[t-1]} \left[ \|x^* - x^{t-1}\|^2 \right] \\
\leq \frac{D^2}{\tau t + a} + \frac{q_{t-1}}{\tau t + a} \left( \frac{D^2}{\tau t + a} + \frac{q_{t-2}}{\tau t + a} E_{[t-2]} \left[ \|x^* - x^{t-2}\|^2 \right] \right) \\
= \frac{D^2}{\tau t + a} \left( 1 + \frac{q_{t-2}}{\tau t + a} \right) + \frac{q_{t-2}}{\tau t + a} \frac{D^2}{\tau t + a} E_{[t-2]} \left[ \|x^* - x^{t-2}\|^2 \right] \\
\leq \ldots \\
\leq \frac{D^2}{\tau t + a} \left( 1 + \frac{q_{t-2}}{\tau t + a} + \ldots + \frac{q_{t-1}}{\tau t + a} \right) + \frac{q_{t-1}}{\tau t + a} \frac{D^2}{\tau t + a} \|x^* - x^{t-1}\|^2 \\
= \frac{D^2}{\tau t + a} \left( 1 + \frac{q_{t-1}}{\tau t + a} \right) + \frac{q_{t-1}}{\tau t + a} \frac{D^2}{\tau t + a} \|x^* - x^{t-1}\|^2 \\
\leq 2D^2.
\]
Then from (3.35), for any $t \leq N - 1$, we have
\[
\frac{(p - 1) \gamma_{t-1} \eta t - 1}{p t - 1 q_{t-1}} E_{[t]} \left[ \|\hat{y} - y^t\|^2 \right] \leq D^2 + \frac{\gamma_{t-1} - q_{t-1}}{q_{t-1}} 2D^2,
\]
or
\[
E_{[t]} \left[ \|\hat{y} - y^t\|^2 \right] \leq \frac{p(2\gamma_{t-1} - q_{t-1}) t - 1}{(p - 1) \gamma_{t-1} \eta_{t-1}} D^2.
\]
We now prove part b). Noting that,
\[
\|x - x^1\|^2 - \|x - x^N\|^2 = 2(x^N - x^1, x) + \|x^1\|^2 - \|x^N\|^2 \\
= 2(x^N - x^1, x - \hat{x}^N) + 2(x^N - x^1, \hat{x}^N) + \|x^1\|^2 - \|x^N\|^2 \\
= 2(x^N - x^1, x - \hat{x}^N) + \|x^1 - \hat{x}^N\|^2 - \|x^N - \hat{x}^N\|^2.
\]
Applying this relation to (3.34), using the fact that $x^1 = x^0 = 0$, we obtain
\[
\sum_{t=1}^{N-1} \frac{q_{t} Q(z^{t+1}, z) + \gamma_{N-1}(Ax^N - Ax^{N-1}, \hat{y}^N - y)}{2q_{N-1}} \leq \frac{2q_{N-1}}{\gamma_{N-1}} \|\hat{x}^N - x^1\|^2 - \frac{2q_{N-1}}{\gamma_{N-1}} \|\hat{x}^N - x^N\|^2 + \frac{2q_{N-1}}{\gamma_{N-1}} \|x^1 - \hat{x}^N\|^2 - \frac{2q_{N-1}}{\gamma_{N-1}} \|x^N - \hat{x}^N\|^2 \\
- \gamma_{N-1}(Ax^N - Ax^{N-1}, y^N - \hat{y}^N) + \sum_{t=1}^{N-1} (\delta_t + \mu_t).
\]
Because $Q(z^{t+1}, z)$ is linear, then from the definition of gap function, we have
\[
\sum_{t=1}^{N-1} \frac{q_{t} Q(z^N, z) + \langle v_N, \hat{z}^N - z \rangle}{2q_{N-1}} \leq \frac{q_{N-1}}{2q_{N-1}} \|\hat{x}^N - x^1\|^2 - \frac{q_{N-1}}{2q_{N-1}} \|\hat{x}^N - x^N\|^2 + \frac{q_{N-1}}{2q_{N-1}} \|x^1 - \hat{x}^N\|^2 - \frac{q_{N-1}}{2q_{N-1}} \|x^N - \hat{x}^N\|^2 \\
- \gamma_{N-1}(Ax^N - Ax^{N-1}, y^N - \hat{y}^N) + \sum_{t=1}^{N-1} (\delta_t + \mu_t).
\]
By using Cauchy-Schwartz inequality, we have
\[
- \gamma_{N-1}(Ax^N - Ax^{N-1}, y^N - \hat{y}^N) \leq \gamma_{N-1} \|A\| \|x^N - x^{N-1}\|^2 \|y^N - \hat{y}^N\|^2 \\
\leq \frac{\gamma_{N-1} \|A\|^2}{2} \|x^N - x^{N-1}\|^2 + \frac{\gamma_{N-1}}{2} \|y^N - \hat{y}^N\|^2
\]
then apply to (3.36), note that $q_{N-1} \geq \gamma_{N-1} \eta_{N-1} t - 1 \|A\|^2$, we have
\[
Q(z^N, z) + \langle v_N, \hat{z}^N - z \rangle \leq \left( \sum_{t=1}^{N-1} \eta_t \right)^{-1} \left[ \frac{q_{N-1}}{2q_{N-1}} \|\hat{x}^N - x^1\|^2 + \frac{\gamma_{N-1}}{2} \|y^N - \hat{y}^N\|^2 + \sum_{t=1}^{N-1} (\delta_t + \mu_t) \right].
\]
Taking expectation w.r.t $[\eta_N]$, noting that $E_{[\eta]} [\delta_t + \mu_t] = 0$, we obtain
\[
E_{[\eta]} \left[ Q(z^N, z) + \langle v_N, \hat{z}^N - z \rangle \right] \leq \left( \sum_{t=1}^{N-1} \eta_t \right)^{-1} \left[ \frac{q_{N-1}}{2q_{N-1}} E_{[\eta]} \left[ \|\hat{x}^N - x^1\|^2 \right] + \frac{\gamma_{N-1}}{2} E_{[\eta]} \left[ \|y^N - \hat{y}^N\|^2 \right] \right].
\]
In the following result we show that the algorithm can compute a nearly optimal solution with a small perturbation.

**Theorem 3.8.** Suppose that \( z^t = (x^t, y^t), t = 1, 2, ..., N \) are generated by the Algorithm. We assume that starting point \( x^1 = x^0 = 0 \) and \( y^1 = \text{argmin}_{y \in Y} J(y) \). We also assume that these stepsizes satisfy (3.12) and (3.20). Then there exists a perturbation vector \( v_N \) such that

\[
E_{[i_N]} \left[ \bar{Q}(\bar{z}^N, z, v_N) \right] \leq \left( \sum_{t=1}^{N-1} \eta_t \right)^{-1} \left( 3 + \frac{p}{p-1} \frac{2\tau t - 1 - \eta t - 1}{\eta t - 1} \right) D^2, \tag{3.37}
\]

where

\[
E ||v_N|| \leq \frac{KD}{\sum_{t=1}^{N-1} \eta_t}. \tag{3.38}
\]

**Proof.** First, we prove that \( E ||v_N||_2 \) is bounded. Using (3.27), (3.28), (3.29) and (3.30), we have

\[
E \left[ \left\| \frac{q_{N-1}}{q_{N-1}} (x^N - x^1) \right\|_2, \gamma_{N-1} (Ax^N - Ax^{N-1}) + \frac{7N-1}{\eta_{N-1}} (y^N - y^1) \right\|_2 \right]
\leq E \gamma_{N-1} \left[ \|Ax^N - Ax^{N-1}\|_2 + \frac{q_{N-1}}{q_{N-1}} \|x^N - x^1\|_2 + \frac{7N-1}{\eta_{N-1}} \|y^N - y^1\|_2 \right]
\leq E \gamma_{N-1} \left[ \|Ax^N - Ax^{N-1}\|_2 + \frac{2q_{N-1}}{q_{N-1}} \|x^N - x^1\|_2 + \frac{2q_{N-1}}{q_{N-1}} \|y^N - y^1\|_2 \right]
\leq 2\frac{q_{N-1}}{q_{N-1}} D + 2 \sqrt{\frac{2N-1}{\gamma_{N-1}}D} + \gamma_{N-1} \|A\|_2(D + \sqrt{2D}) = KD,
\]

where

\[
K = \frac{2q_{N-1}}{\eta_{N-1}} + 2 \sqrt{\frac{2N-1}{\gamma_{N-1}}} + \gamma_{N-1} \|A\|_2(1 + \sqrt{2}).
\]

Then the above inequality and the definition of \( v_N \) imply (3.38). Second, we prove that \( \frac{q_{N-1}}{2q_{N-1}} E_{[i_N]} \left[ \|x^N - x^1\|_2 \right] + \frac{7N-1}{2q_{N-1}} E_{[i_N]} \left[ \|y^N - y^1\|_2 \right] \) is bounded. Using (3.27), (3.28), (3.29) and (3.30), we have

\[
\frac{q_{N-1}}{q_{N-1}} E_{[i_N]} \left[ \|x^N - x^1\|_2 \right] + \frac{7N-1}{2q_{N-1}} E_{[i_N]} \left[ \|y^N - y^1\|_2 \right]
\leq \frac{q_{N-1}}{q_{N-1}} D^2 + \frac{1}{\sum_{t=1}^{N-1} q_t} \left[ \sum_{t=1}^{N-1} q_t \left( \frac{q_{t-1}}{\eta_{t-1}} \|x^t - x^{t-1}\|_2 + \frac{7N-1}{\eta_{t-1}} \|y^t - y^{t-1}\|_2 \right) \right]
\leq \frac{q_{N-1}}{q_{N-1}} D^2 + \frac{1}{\sum_{t=1}^{N-1} q_t} \left[ \sum_{t=1}^{N-1} q_t \left( \frac{q_{t-1}}{\eta_{t-1}} \|x^t - x^{t-1}\|_2 + \frac{7N-1}{\eta_{t-1}} \|y^t - y^{t-1}\|_2 \right) \right]
\leq \frac{q_{N-1}}{q_{N-1}} D^2 + \frac{1}{\sum_{t=1}^{N-1} q_t} \left[ \sum_{t=1}^{N-1} q_t \left( \frac{q_{t-1}}{\eta_{t-1}} D^2 + \frac{7N-1}{\eta_{t-1}} \right) \right]
\leq \left( \frac{3q_{N-1}}{q_{N-1}} + \frac{p}{p-1} \frac{2N-1}{\eta_{N-1}} \right) D^2.
\]

In the following the corollary, we describe a specialized convergence result of Algorithm [1] after properly choosing parameters.

**Corollary 3.9.** Suppose that \( z^t = (x^t, y^t), t = 1, 2, ..., N \) are generated by the Algorithm. We assume that starting point \( z^1 \) is chosen such that \( x^1 = x^0 = 0 \) and \( y^1 = \text{argmin}_{y \in Y} J(y) \). We also assume that these
By choosing parameters properly, we show that Algorithm 1 can obtain stepsizes satisfying (3.12) and (3.26). Then there exists a perturbation vector of subproblem (2.2) under the strongly convex assumption of \( J \).

\[
\theta_t = p, \quad \gamma_t = 1, \quad q_t = \frac{1}{p}, \quad \tau_t = \frac{1}{Lp^2}, \quad \eta_t = \frac{1}{Lp^2}, \quad \eta_{N-1} = \frac{1}{Lp^2}, \quad i = 1, \ldots, N - 2, \quad \eta_{N-1} = 1, \quad \tau_t \leq 1, \quad \gamma_t = 1, \quad q_t = 1, \quad \forall t = 1, 2, \ldots, N - 2, \]

then we have

\[
E_{[i_N]}[\tilde{Q}(\tilde{z}^N, z, v_N)] \leq \frac{p}{N + p - 2} \left( 3 + \frac{p}{p - 1} \right) Lp^2 D^2, \tag{3.39}
\]

and

\[
E[\|v_N\|^2] \leq \frac{p}{N + p - 2} \left( 2Lp^2 D + 2LpD + (D + \sqrt{2}D) \right), \tag{3.40}
\]

Proof. We have

\[
\sum_{t=1}^{N-1} q_t = \frac{(N + p - 2)}{p}.
\]

Then apply this observation to (3.38), using definition of \( K \), and (3.37) we obtain (3.41) and (3.40) respectively.

Corollary 3.10. Suppose that \( z^t = (x^t, y^t), t = 1, 2, \ldots, N \) are generated by the Algorithm. We assume that starting point \( z^1 \) is chosen such that \( x^1 = x^0 = 0 \) and \( y^1 = \arg\min_{y \in Y} J(y) \). We also assume that these stepsizes satisfy (3.12) and (3.26). Then there exists a perturbation vector \( v_N \) such that

\[
E_{[i_N]}[\tilde{g}_\sigma(\tilde{z}^N, z, v_N)] \leq \left( \sum_{t=1}^{N-1} \eta_t \right)^{-1} \left( 3 + \frac{p}{p - 1} \frac{2\tau_{t-1} - \eta_{t-1}}{\eta_{t-1}} \right) D^2,
\]

where

\[
E[\|v_N\|] \leq \frac{KD}{\sum_{t=1}^{N-1} q_t}.
\]

Proof. This proof is almost similar to the proof of Theorem 3.3 except using the same technique of the proof of Corollary 3.6.

4. Convergence analysis with strongly convex assumption. In this section, we present the convergence properties of Algorithm 1 in strongly convex setting. In particular, we assume that \( J_i(y_i), i = 1, 2, \ldots, p \) are strongly convex functions. WLOG, we also assume that these strong convexity parameters of \( J_i(y_i) \) are 1. By choosing parameters properly, we show that Algorithm 1 can obtain \( 1/N^2 \) rate of convergence.

The following Lemma describes an important observation that can be obtained from optimality condition of subproblem (2.2) under the strongly convex assumption of \( J(y) \).

Lemma 4.1. Suppose \( \{x^t\}_{t \geq 1} \) and \( \{y^t\}_{t \geq 1} \) are generated by Algorithm 1. Then

- a) For all \( y \in Y \), we have

\[
\gamma_t \left[ (-U_t, A\tilde{x}^t, y^{t+1} - y) + J_t(y_t) - J_t(y_t) \right] + \frac{2\eta_t}{\tau_t} \|y^t - y^{t+1}\|^2 + \frac{\|y_t - y^{t+1}\|^2}{2} \leq \frac{\gamma_t}{\tau_t} \|y_t - y^{t+1}\|^2. \tag{4.1}
\]

13
The following Proposition establishes an important recursion.

**Proposition 4.2.** Suppose \( \{y^t\}_{t \geq 1} \) and \( \{x^t\}_{t \geq 1} \) are generated by the above algorithm. Then for any \( y \in Y \) and \( x \in X \), we have

\[
q_t \left[ G(x^{t+1}) - G(x) + (Ax^{t+1}, y) - (Ax^t, y^{t+1}) \right] + \gamma_t \left[ J(y^{t+1}) - \frac{p-1}{p} J(y^t) - \frac{1}{p} J(y) \right]
\]

\[
\leq \frac{q_t}{2\eta_t} \|x - x^t\|^2 - \frac{q_t}{2\eta_t} \|x - x^{t+1}\|^2 - \frac{q_t}{2\eta_t} \|x^{t+1} - x^t\|^2
\]

\[
+ \frac{2q_t}{2\eta_t} \|y - y^t\|^2 - \frac{q_t}{2\eta_t} \|y - y^{t+1}\|^2 - \frac{2q_t}{2\eta_t} \|y^{t+1} - y^t\|^2 - \gamma_t \left( 1 + \frac{1}{2\tau_t} \right) \|y - y^{t+1}\|^2
\]

\[
- (q_t Ax^{t+1} - \gamma_t Ax^t, y^{t+1} - y) - (\theta_t - 1) \gamma_t L \|x^t - x^{t-1}\|_2 \|y^{t+1} - y^t\|_2 + \delta_t + \mu_t,
\]

where \( \delta_t \) and \( \mu_t \) are defined in (4.5) and (5.7).

**Proof.** Using similar analysis of proof of Proposition 3.3, we have

\[
q_t \left[ G(x^{t+1}) - G(x) + (Ax^{t+1}, y) - (Ax^t, y^{t+1}) \right] + \gamma_t \left[ J(y^{t+1}) - \frac{p-1}{p} J(y^t) - \frac{1}{p} J(y) \right]
\]

\[
\leq \frac{q_t}{2\eta_t} \|x - x^t\|^2 - \frac{q_t}{2\eta_t} \|x - x^{t+1}\|^2 - \frac{q_t}{2\eta_t} \|x^{t+1} - x^t\|^2
\]

\[
+ \frac{2q_t}{2\eta_t} \|y - y^t\|^2 - \frac{q_t}{2\eta_t} \|y - y^{t+1}\|^2 - \frac{2q_t}{2\eta_t} \|y^{t+1} - y^t\|^2 - \gamma_t \left( 1 + \frac{1}{2\tau_t} \right) \|y - y^{t+1}\|^2
\]

\[
- (q_t Ax^{t+1} - \gamma_t Ax^t, y^{t+1} - y) - (\theta_t - 1) \gamma_t L \|x^t - x^{t-1}\|_2 \|y^{t+1} - y^t\|_2 + \delta_t + \mu_t,
\]

The result follows the fact that

\[-(\theta_t - 1) \gamma_t L \|x^t - x^{t-1}\|_2 \|y^{t+1} - y^t\|_2 \leq \theta_t - 1 \gamma_t L \|x^t - x^{t-1}\|_2 \|y^{t+1} - y^t\|_2.\]

The convergence properties of the Algorithm II are described in the following theorem.

**Theorem 4.3.** Suppose \( \{y^t\}_{t \geq 1} \) and \( \{x^t\}_{t \geq 1} \) are generated by the above algorithm. We assume that these parameters are chosen such that,

\[
\begin{align*}
\gamma_t &= \frac{1}{2} \theta_t \gamma_{t+1}, \ i = 1, \ldots, N - 1, \\
q_t &= \frac{1}{2} \theta_t \gamma_{t+1} - \frac{p-1}{p} \gamma_{t+1}, \ i = 1, \ldots, N - 2 \\
q_N^{-1} &= \gamma_{N-1}, \\
\gamma_t \left( 1 + \frac{1}{2\tau_t} \right) &\geq \frac{2\tau_t}{2\tau_t + 1}, \ i = 1, \ldots, N - 1, \\
\frac{q_t}{2\eta_t} &\geq \frac{1}{2\eta_t}, \ i = 1, \ldots, N - 1, \\
\frac{q_t}{2\eta_t} \gamma_t &\geq \theta_t - 1 \gamma_t L^2, \ i = 1, \ldots, N - 1, \\
\frac{q_N^{-1}}{2N^{-1}} &\geq \tau_t^{-1} \gamma_{N-1} A_k.
\end{align*}
\]

Then, for any \( N \geq 1 \), we have

\[
E_{\{x_i\} \mid Q(z^N, z)} \leq \left( \sum_{t=1}^{N-1} q_t \right)^{-1} \left[ \frac{q_t}{2\eta_t} \Omega_x^2 + \frac{1}{2\tau_t} \Omega_y^2 + \|A\| \Omega_Y \right].
\]

where and the expectation is taken corresponding to \( i_{\{N\}} = (i_1, \ldots, i_{N-1}) \).

**Proof.** Denote

\[
\tilde{B}_N(z, z^{[N]}) = \sum_{t=1}^{N-1} \left[ \frac{q_t}{2\eta_t} \|x - x^t\|^2 - \frac{q_t}{2\eta_t} \|x - x^{t+1}\|^2 \right] + \sum_{t=1}^{N-1} \left[ \frac{q_t}{2\eta_t} \|y - y^t\|^2 - \gamma_t \left( 1 + \frac{1}{2\tau_t} \right) \|y - y^{t+1}\|^2 \right].
\]
Taking summation from $t = 1$ to $N - 1$ both sides of (4.2) and using the stepsize choice (4.3), we obtain
\[
\sum_{t=1}^{N-1} q_t \left[ G(x_t^{t+1}) - G(x) + \langle Ax_t^{t+1}, y \rangle - \langle Ax, y_t^{t+1} \rangle \right] + \sum_{t=1}^{N-1} \left[ J(y_t^{t+1}) - \frac{p-1}{p} J(y^*) - \frac{1}{p} J(y) \right] 
\leq \hat{B}_N(z, z[N]) - \sum_{t=1}^{N-1} \left( \frac{q_t - \eta_t}{2\tau_t-1} \|x_t - x_t^{t+1}\|_2^2 + \frac{2\tau_t}{2\tau_t-1} \|y_t^{t+1} - y_t\|_2^2 \right) 
- \frac{(N-1)\eta_t}{2\tau_t} \|y_t - y\|_2^2 + \sum_{t=1}^{N-1} \theta_t \gamma_t \|x_t - x_t^{t+1}\|_2^2 + \gamma_t \|y_t - y\|_2^2 
\leq \hat{B}_N(z, z[N]) - \sum_{t=1}^{N-1} \left( \frac{2\gamma_t}{2\eta_t} - \frac{\eta_t}{2\tau_t-1} \|x_t - x_t^{t+1}\|_2^2 + \frac{2\tau_t}{2\tau_t-1} \|y_t - y\|_2^2 \right) 
- \frac{(N-1)\eta_t}{2\tau_t} \|y_t - y\|_2^2 + \sum_{t=1}^{N-1} \left( \delta_t + \mu_t \right) 
\leq \hat{B}_N(z, z[N]) - \frac{(N-1)\eta_t}{2\tau_t} \|x_t - x_t^{t+1}\|_2^2 + \frac{2\tau_t}{2\tau_t-1} \|y_t^{t+1} - y_t\|_2^2 
- \frac{(N-1)\eta_t}{2\tau_t} \|y_t - y\|_2^2 + \sum_{t=1}^{N-1} \left( \delta_t + \mu_t \right) 
\leq \hat{B}_N(z, z[N]) - \frac{(N-1)\eta_t}{2\tau_t} \|x_t - x_t^{t+1}\|_2^2 + \frac{2\tau_t}{2\tau_t-1} \|y_t^{t+1} - y_t\|_2^2 
- \frac{(N-1)\eta_t}{2\tau_t} \|y_t - y\|_2^2 + \sum_{t=1}^{N-1} \left( \delta_t + \mu_t \right) 
\text{where the first inequality follows (4.3) and the fact that } x^1 = x^0, \text{ the second inequality follows the inequality } a^2 + b^2 \geq 2ab \text{ and the third inequality follows the fact that } \frac{a}{\sqrt{2} - 1} \geq \sqrt{2} \gamma_t \frac{L}{\gamma_t}.
\]

Similar to the Theorem 4.3, by stepsize choice (4.3), we have
\[
\left( \sum_{t=1}^{N-1} q_t \right) Q(z^N, z) \leq \hat{B}_N(z, z[N]) + \frac{2\gamma_t}{2\eta_t-1} \|y_t - y\|_2^2 - \frac{(N-1)\eta_t}{2\tau_t} \|x_t - x_t^{t+1}\|_2^2 + \frac{2\tau_t}{2\tau_t-1} \|y_t^{t+1} - y_t\|_2^2 + \sum_{t=1}^{N-1} \left( \delta_t + \mu_t \right).
\]

By the stepsize choice (4.3), we have $1 + \frac{1}{2\tau_t-1} \geq \frac{1}{2\tau_t-1}$ and $\frac{1}{\eta_t} \geq \frac{1}{\eta_t} \geq \frac{1}{\eta_t}$, then by definition (3.14), we conclude that
\[
B_N(z, z[N]) = \frac{q_t}{2\eta_t} \|x_t - x_t^{t+1}\|_2^2 - \sum_{t=1}^{N-2} \left( \frac{q_t}{2\eta_t} - \frac{\eta_t}{2\tau_t-1} \|x_t - x_t^{t+1}\|_2^2 - \frac{2\tau_t}{2\tau_t-1} \|x_t^{t+1} - x_t\|_2^2 \right) + \frac{2\tau_t}{2\tau_t-1} \|y_t^{t+1} - y_t\|_2^2 - \sum_{t=1}^{N-1} \gamma_t \left( 1 + \frac{1}{2\tau_t} \right) \|y_t - y_t^{t+1}\|_2^2 - \gamma_t \left( 1 + \frac{1}{2\tau_t} \right) \|y_t - y\|_2^2 
\leq \frac{q_t}{2\eta_t} \|x_t - x_t^{t+1}\|_2^2 + \frac{2\tau_t}{2\tau_t-1} \|y_t^{t+1} - y_t\|_2^2 - \sum_{t=1}^{N-1} \gamma_t \left( 1 + \frac{1}{2\tau_t} \right) \|y_t - y\|_2^2 
\leq \frac{q_t}{2\eta_t} \Omega_X^2 + \frac{2\tau_t}{2\tau_t-1} \Omega_Y^2 - \sum_{t=1}^{N-1} \gamma_t \left( 1 + \frac{1}{2\tau_t} \right) \|y_t - y\|_2^2.
\]

Using the above two relations, noting that $\frac{q_t - \eta_t}{2\tau_t-1} \geq \frac{\gamma_t \eta_t}{2\tau_t} \|x_t - x_t^{t+1}\|_2^2$ and $\frac{2\tau_t}{2\tau_t-1} \|y_t^{t+1} - y_t\|_2^2 + \sum_{t=1}^{N-1} \left( \delta_t + \mu_t \right)$. We obtain
\[
\left( \sum_{t=1}^{N-1} q_t \right) Q(z^N, z) \leq \frac{q_t}{2\eta_t} \Omega_X^2 + \frac{\gamma_t \eta_t}{2\tau_t} \Omega_Y^2 + \frac{p-1}{p} \|A\| \|x_0 - x\|_2 \|y_0 - y\|_2 + \sum_{t=1}^{N-1} \left( \delta_t + \mu_t \right).
\]

Taking expectation w.r.t $i_t, t = 1, 2, ..., N - 1$, noting that $\mathbb{E}[i_t] = 0$ and $\frac{p-1}{p} \leq 1$, we obtain the result.

Below we provide a specialized convergence result for the Algorithm 1 to solve strongly convex saddle point problems after properly selecting these parameters.

**Corollary 4.4.** Suppose $\{y_t\}_{t=1}^\infty$ and $\{x_t\}_{t=1}^\infty$ are generated by the above algorithm. We assume that these parameters are chosen such that (4.3) is satisfied and
\[
\gamma_t = t + p, \quad t = 1, 2, ..., N - 1, \quad \text{and} \quad \tau_t = \frac{1}{t-1}, \eta_t = \frac{t}{p^4 L^2}.
\]

Then for any $N \geq 1$ we have
\[
\mathbb{E} \left[ Q(z^N, z) \right] \leq \frac{p^3 L^2 \Omega_X^2 + \Omega_Y^2}{2} \left( \frac{p-1}{p} \right) + \sum_{t=1}^{N-1} \left( \delta_t + \mu_t \right).
\]
Proof. It is easy to see that from (4.6), we have
\[
\theta_t = \frac{p(t + p)}{t + p + 1}, \quad t = 1, 2, \ldots, N - 1,
\]
\[
q_t = \frac{t + 1}{p}, \quad t = 1, 2, \ldots, N - 2 \quad \text{and} \quad q_{N-1} = N + p - 1,
\]
that implies
\[
\sum_{t=1}^{N-1} q_t = N + p - 1 + \sum_{t=1}^{N-2} \frac{t + 1}{p} \geq \frac{(N - 2)^2}{2p}.
\]
Moreover, from the stepsizes choice (4.7) we obtain
\[
\frac{q_1}{2\eta_1} = \frac{p^3 L^2}{2}
\]
and
\[
\frac{\gamma_1}{2\tau_1} = 0.
\]
Plugging to (4.8) we obtain (4.9).

5. Conclusions. In this paper, we present a new algorithm, namely Randomized First-order Methods, for Saddle Point Optimization in which we incorporate the block coordinate decomposition technique into solving dual problem. Indeed, each iteration, our algorithm require to solve only one subproblem rather than solving all subproblem as in primal dual algorithms and ADMMs. We also present 1/N rate of convergence for both bounded and unbounded cases and 1/N^2 rate of convergence for strongly convex saddle point problem in which we just need to assume strong convexity on J(y).

REFERENCES

[1] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. Technical report. submitted to SIAM Journal on Optimization.
[2] Antonin Chambolle and Thomas 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.
[3] Cong D Dang and Guanghui Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. arXiv preprint arXiv:1309.2249, 2013.
[4] Jr. Douglas, Jim and Jr. Rachford, H. H. On the numerical solution of heat conduction problems in two and three space variables. Transactions of the American Mathematical Society, 82(2):pp. 421–439, 1956.
[5] Ernie Esser, Xiaoqun Zhang, and Tony F. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM J. Imag. Sci., 3(4):1015–1046, December 2010.
[6] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Computers and Mathematics with Applications, 2(1):17–40, 1976.
[7] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. SIAM Journal on Optimization, 2013. to appear.
[8] H. O. Hartley. Journal of the American Statistical Association, 55(292):pp. 758–760, 1960.
[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference and prediction. Springer, 2 edition, 2008.
[10] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. pages 433–440, 2009.
[11] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: Convergence rates and conditioning. Mathematics of Operations Research, 35:641–654, 2010.
[12] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. SIAM Journal on Numerical Analysis, 16(6):pp. 964–979, 1979.
[13] Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods. Manuscript, 2013.

[14] R. Monteiro and B. Svaiter. Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM Journal on Optimization, 23(1):475–507, 2013.

[15] R.D.C. Monteiro and B.F. Svaiter. Complexity of variants of tsengs modified f-b splitting and korpelevich’s methods for hemi-variational inequalities with applications to saddle-point and convex optimization problems. Manuscript, School of ISyE, Georgia Tech, Atlanta, GA, 30332, USA, June 2010.

[16] R.D.C. Monteiro and B.F. Svaiter. On the complexity of the hybrid proximal projection method for the iterates and the ergodic mean. SIAM Journal on Optimization, 20:2755–2787, 2010.

[17] Y. E. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, February 2010.

[18] Y. E. Nesterov. Subgradient methods for huge-scale optimization problems. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, February 2012.

[19] Yuyuan Ouyang, Yunmei Chen, Guanghui Lan, and Eduardo Pasiliao Jr. An accelerated linearized alternating direction method of multipliers. arXiv preprint arXiv:1401.6607, 2014.

[20] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Mathematical Programming, 2012. to appear.

[21] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. Physica D, 60:259–268, 1992.

[22] Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. Manuscript, September 2014.

[23] M. Zhu and T. Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. UCLA CAM Report, pages 08–34, 2008.