Linear Convergence of the Primal-Dual Gradient Method for Convex-Concave Saddle Point Problems without Strong Convexity

Simon S. Du
Carnegie Mellon University
ssdu@cs.cmu.edu

Wei Hu
Princeton University
huwei@cs.princeton.edu

Abstract

We consider the convex-concave saddle point problem $\min_x \max_y f(x)+y^\top Ax-g(y)$ where f is smooth and convex and g is smooth and strongly convex. We prove that if the coupling matrix A has full column rank, the *vanilla* primal-dual gradient method can achieve linear convergence *even if f is not strongly convex.* Our result generalizes previous work which either requires f and g to be quadratic functions or requires proximal mappings for both f and g. We adopt a novel analysis technique that in each iteration uses a "ghost" update as a reference, and show that the iterates in the primal-dual gradient method converge to this "ghost" sequence. Using the same technique we further give an analysis for the primal-dual stochastic variance reduced gradient method for convex-concave saddle point problems with a finite-sum structure.

1 Introduction

We revisit the convex-concave saddle point problems of the form

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} L(x,y) = f(x) + y^\top Ax - g(y), \qquad (1)$$

where both f and g are convex functions and $A \in \mathbb{R}^{d_2\times d_1}$ is a coupling matrix. This formulation has a wide range of applications, including supervised learning [Zhang and Lin, 2015], unsupervised learning [Xu et al., 2005, Bach et al., 2008], reinforcement learning [Du et al., 2017], robust optimization [Ben-Tal et al., 2009], PID control [Hast et al., 2013], etc. See Section 1.2 for some concrete examples.

Algorithm 1 Primal-Dual Gradient Method

```
Inputs: initial points $x_0 \in \mathbb{R}^{d_1}, y_0 \in \mathbb{R}^{d_2}$, step sizes $\eta_1, \eta_2 > 0$
for $t = 0, 1, \ldots$ do
    $x_{t+1} = x_t - \eta_1 \nabla_x L(x_t, y_t)$
          $= x_t - \eta_1 \left(\nabla f(x_t) + A^\top y_t\right)$
    $y_{t+1} = y_t + \eta_2 \nabla_y L(x_t, y_t)$
          $= y_t + \eta_2 \left(Ax_t - \nabla g(y_t)\right)$
end for
```

When the problem dimension is large, the most widely used and sometimes the only scalable methods to solve Problem (1) are first-order methods. Arguably the simplest first-order algorithm is the *primal-dual gradient method* (Algorithm 1), a natural generalization of the gradient descent algorithm, which simultaneously performs gradient descent on the primal variable x and gradient ascent on the dual variable y.

There has been extensive research on analyzing the convergence rate of Algorithm 1 and its variants. It is known that if both f and g are strongly convex and admit efficient proximal mappings, then the *proximal* primal-dual gradient method converges to the optimal solution at a linear rate [Bauschke and Combettes, 2011, Palaniappan and Bach, 2016, Chen and Rockafellar, 1997], i.e., it only requires $O\left(\log\frac{1}{\epsilon}\right)$ iterations to obtain a solution that is ϵ-close to the optimum.

In many applications, however, we only have strong convexity in g but *no* strong convexity in f. This motivates the following question:

Does the primal-dual gradient method converge linearly to the optimal solution if f is not strongly convex?

Intuitively, a linear convergence rate is plausible. Consider the corresponding primal problem of (1):

$$\min_{x\in\mathbb{R}^{d_1}} P(x) = g^*(Ax) + f(x), \qquad (2)$$

where g^* is the conjugate function of g. Because g is smooth and strongly convex, as long as A has full col-

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

umn rank, Problem (2) has a smooth and strongly convex objective and thus vanilla gradient descent achieves linear convergence. Therefore, one might expect a linearly convergent first-order algorithm for Problem (1) as well. However, whether the vanilla primal-dual gradient method (Algorithm 1) has linear convergence turns out to be a nontrivial question.

Two recent results verified this conceptual experiment with additional assumptions: Du et al. [2017] required both f and g to be quadratic functions, and Wang and Xiao [2017] required both f and g to have efficient proximal mappings and uses a *proximal* primal-dual gradient method. In this paper, we give an affirmative answer to this question with minimal assumptions. Our main contributions are summarized below.

1.1 Our Contributions

Linear Convergence of the Primal-Dual Gradient Method. We show that as long as f and g are smooth, f is convex, g is strongly convex and the coupling matrix A has full column rank, Algorithm 1 converges to the optimal solution at a linear rate. See Section 3 for a precise statement of our result. This result significantly generalizes previous ones which rely on stronger assumptions. Note that all the assumptions are necessary for linear convergence: without any of them, the primal problem (2) requires at least $\text{poly}(\frac{1}{\epsilon})$ iterations to obtain an ϵ-close solution [Nesterov, 2013], so there is no hope of linear convergence for Problem (1).

New Analysis Technique. To analyze the convergence of an optimization algorithm, a common way is to construct a potential function (also called Lyapunov function in the literature) which decreases after each iteration. For example, for the primal problem (2), a natural potential function is $\|x_t - x^*\|$, the distance between the current iterate and the optimal solution. However, for the primal-dual gradient method, it is difficult to show similar potential functions like $\|x_t - x^*\| + \|y_t - y^*\|$ decrease because the two sequences, $\{x_t\}_{t=0}^{\infty}$ and $\{y_t\}_{t=0}^{\infty}$, are related to each other.

In this paper, we develop a novel method for analyzing the convergence rate of the primal-dual gradient method. The key idea is to consider a "ghost" sequence. For example, in our setting, the "ghost" sequence comes from a gradient descent step for Problem (2). Then we relate the sequence generated by Algorithm 1 to this "ghost" sequence and show they are close in a certain way. See Section 3 for details. We believe this technique is applicable to other problems where we need to analyze multiple sequences.

Extension to Primal-Dual Stochastic Variance Reduced Gradient Method. Many optimization problems in machine learning have a finite-sum structure, and randomized algorithms have been proposed to exploit this structure and to speed up the convergence. There has been extensive research in recent years on developing more efficient stochastic algorithms in such setting [Le Roux et al., 2012, Johnson and Zhang, 2013, Defazio et al., 2014, Xiao and Zhang, 2014, Shalev-Shwartz and Zhang, 2013, Richtárik and Takáč, 2014, Lin et al., 2015, Zhang and Lin, 2015, Allen-Zhu, 2017]. Among them, the stochastic variance reduced gradient (SVRG) algorithm [Johnson and Zhang, 2013] is a popular one with computational complexity $O\left((n+\kappa)d\log\frac{1}{\epsilon}\right)$ for smooth and strongly convex objectives, where n is the number of component functions, d is the dimension of the variable, and κ is a condition number that only depends on problem-dependent parameters like smoothness and strong convexity but not n. Variants of SVRG for saddle point problems have been recently studied by Palaniappan and Bach [2016], Wang and Xiao [2017], Du et al. [2017] and can achieve similar $O\left((n+\kappa)d\log\frac{1}{\epsilon}\right)$ running time.[1] However, these results all require additional assumptions. In this paper, we use our analysis technique developed for Algorithm 1 to show that the primal-dual SVRG method also admits $O\left((n+\kappa)d\log\frac{1}{\epsilon}\right)$ type computational complexity.

1.2 Motivating Examples

In this subsection we list some machine learning applications that naturally lead to convex-concave saddle point problems.

Reinforcement Learning. For policy evaluation task in reinforcement learning, we have data $\{(s_t, r_t, s_{t+1})\}_{t=1}^{n}$ generated by a policy π where s_t is the state at the t-th time step, r_t is the reward and s_{t+1} is the state at the $(t+1)$-th step. We also have a discount factor $0 < \gamma < 1$ and a feature function $\phi(\cdot)$ which maps a state to a feature vector. Our goal is to learn a linear value function $V^{\pi}(s) \approx x^{\top}\phi(s)$ which represents the long term expected reward starting from state s using the policy π. A common way to estimate x is to minimize the empirical mean squared projected Bellman error (MSPBE):

$$\min_{x} (Ax - b)^{\top} C^{-1} (Ax - b), \tag{3}$$

where $A = \sum_{t=1}^{n} \phi(s_t)\left(\phi(s_t) - \gamma\phi(s_{t+1})\right)^{\top}$, $b = \sum_{t=1}^{n} r_t\phi(s_t)$ and $C = \sum_{t=1}^{n} \phi(s_t)\phi(s_t)^{\top}$. Note that directly using gradient descent to solve problem (3) is expensive because we need to invert a matrix C.

[1] κ may be different in the primal and the primal-dual settings.

Du et al. [2017] considered the equivalent saddle point formulation:

$$\min_x \max_y L(x,y) = -y^\top Ax - \frac{1}{2}y^\top Cy + b^\top y.$$

The gradient of L can be computed more efficiently than the original formulation (3), and L has a finite-sum structure.

Empirical Risk Minimization. Consider the classical supervised learning problem of learning a linear predictor $x \in \mathbb{R}^d$ given n data points $(a_i, b_i) \in \mathbb{R}^d \times \mathbb{R}$. Denote by $A \in \mathbb{R}^{n\times d}$ the data matrix whose i-th row is $a_i^\top$. Then the empirical risk minimization (ERM) problem amounts to solving

$$\min_{x\in\mathbb{R}^d} \ell(Ax) + f(x),$$

where ℓ is induced by some loss function and f is a regularizer; both f and ℓ are convex functions. Equivalently, we can solve the dual problem $\max_{y\in\mathbb{R}^n}\left\{-\ell^*(y) - f^*(-A^\top y)\right\}$ or the saddle point problem $\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^n}\left\{y^\top Ax - \ell^*(y) + f(x)\right\}$. The saddle point formulation is favorable in many scenarios, e.g., when such formulation admits a finite-sum structure [Zhang and Lin, 2015, Wang and Xiao, 2017], reduces communication complexity in the distributed setting [Xiao et al., 2017] or exploits sparsity structure [Lei et al., 2017].

Robust Optimization. The robust optimization framework [Ben-Tal et al., 2009] aims at minimizing an objective function with uncertain data, which naturally leads to a saddle point problem, often with the following form:

$$\min_x \max_y \mathbb{E}_{\xi\sim P(y)}\left[f(x,\xi)\right], \tag{4}$$

where f is some loss function we want to minimize and the distribution of the data is parametrized by $P(y)$. For certain special cases [Liu et al., 2017], Problem (4) has the bilinear form as in (1).

1.3 Comparison with Previous Results

There have been many attempts to analyze the primal-dual gradient method or its variants. In particular, Chen and Rockafellar [1997], Chambolle and Pock [2011], Palaniappan and Bach [2016] showed that if both f and g are strongly convex and have efficient proximal mappings, then the proximal primal-dual gradient method achieves a linear convergence rate.[2] In fact, even without proximal mappings, as long as both f and g are smooth and strongly convex, Algorithm 1 achieves a linear convergence rate. In Appendix B we give a simple proof of this fact.

Two recent papers show that it is possible to achieve linear convergence even without strong convexity in f. The key is the additional assumption that A has full column rank, which helps "transfer" g's strong convexity to f. Du et al. [2017] considered the case when both f and g are quadratic functions, i.e., when Problem (1) has the following special form:

$$L(x,y) = x^\top Bx + b^\top x + y^\top Ax - y^\top Cy + c^\top y.$$

Note that B does not have to be positive definite (but C has to be), and thus strong convexity is not necessary in the primal variable. Their analysis is based on writing the gradient updates as a linear dynamic system (c.f. Equation (41) in [Du et al., 2017]):

$$\begin{bmatrix} x_{t+1} - x^* \\ \sqrt{\frac{\eta_1}{\eta_2}}\left(y_{t+1} - y^*\right)\end{bmatrix} = (I - G)\begin{bmatrix} x_t - x^* \\ \sqrt{\frac{\eta_1}{\eta_2}}\left(y_t - y^*\right)\end{bmatrix}, \tag{5}$$

where G is some fixed matrix that depends on A, B, C and step sizes. Next, it suffices to bound the spectral norm of G (which can be made strictly less than 1) to show that $\left(x_t - x^*, \sqrt{\frac{\eta_1}{\eta_2}}\left(y_t - y^*\right)\right)$ converges to $(0,0)$ at a linear rate. However, it is difficult to generalize this approach to general saddle point problem (1) since only when f and g are quadratic do we have the linear form (5).

Wang and Xiao [2017] considered the proximal primal-dual gradient method. They construct a potential function (c.f. Page 15 in [Wang and Xiao, 2017]) and show it decreases at a linear rate. However, this potential function heavily relies on the proximal mappings so it is difficult to use this technique to analyze Algorithm 1.

In Table 1, we summarize different assumptions sufficient for linear convergence used in different papers.

1.4 Paper Organization

The rest of the paper is organized as follows. We give necessary definitions in Section 2. In Section 3, we present our main result for the primal-dual gradient method and its proof. In Section 4, we extend our analysis to the primal-dual stochastic variance reduced gradient method. In Section 5, we use some preliminary experiments to verify our theory. We conclude in Section 6 and put omitted proofs in the appendix.

2 Preliminaries

Let $\|\cdot\|$ denote the Euclidean (L_2) norm of a vector, and let $\langle\cdot,\cdot\rangle$ denote the standard Euclidean inner product

[2] Chen and Rockafellar [1997], Palaniappan and Bach [2016] considered a more general formulation than Problem (1). Here we specialize in the bi-linear saddle point problem.

Paper	f smooth	f s.c.	g smooth	g s.c.	A full column rank	Other Assumptions
[Chen and Rockafellar, 1997]	\	Yes	\	Yes	No	Prox maps for f and g
[Du et al., 2017]	Yes	No	Yes	Yes	Yes	f and g are quadratic
[Wang and Xiao, 2017]	\	No	\	Yes	Yes	Prox maps for f and g
Folklore	Yes	Yes	Yes	Yes	No	No
This Paper	Yes	No	Yes	Yes	Yes	No

Table 1: Comparisons of assumptions that lead to the linear convergence of primal-dual gradient method for solving Problem (1). When we have proximal mappings for f and g, we do not need their smoothness.

between two vectors. For a matrix $A \in \mathbb{R}^{m\times n}$, let $\sigma_i(A)$ be its i-th largest singular value, and let $\sigma_{\max}(A) := \sigma_1(A)$ and $\sigma_{\min}(A) := \sigma_{\min\{m,n\}}(A)$ be the largest and the smallest singular values of A, respectively. For a function f, we use ∇f to denote its gradient. Denote $[n] := \{1, 2, \ldots, n\}$. Let I_d be the identity matrix in $\mathbb{R}^{d\times d}$.

The smoothness and the strong convexity of a function are defined as follows:

Definition 2.1. *For a differentiable function $\phi : \mathbb{R}^d \to \mathbb{R}$, we say*

- *ϕ is β-smooth if $\|\nabla\phi(u) - \nabla\phi(v)\| \le \beta \|u - v\|$ for all $u, v \in \mathbb{R}^d$;*
- *ϕ is α-strongly convex if $\phi(v) \ge \phi(u) + \langle\nabla\phi(u), v - u\rangle + \frac{\alpha}{2}\|u - v\|^2$ for all $u, v \in \mathbb{R}^d$.*

We also need the definition of conjugate function:

Definition 2.2. *The conjugate of a function $\phi : \mathbb{R}^d \to \mathbb{R}$ is defined as*

$$\phi^*(y) := \sup_{x\in\mathbb{R}^d} \{\langle x, y\rangle - \phi(x)\}, \qquad \forall y \in \mathbb{R}^d.$$

It is well-known that if ϕ is closed and convex, then $\phi^{**} = \phi$. If ϕ is smooth and strongly convex, its conjugate ϕ^* has the following properties:

Fact 2.1. *If $\phi : \mathbb{R}^d \to \mathbb{R}$ is β-smooth and α-strongly convex ($\beta \ge \alpha > 0$), then*

(i) ([Kakade et al., 2009]) $\phi^* : \mathbb{R}^d \to \mathbb{R}$ is $\frac{1}{\alpha}$-smooth and $\frac{1}{\beta}$-strongly convex.

(ii) ([Rockafellar, 1970]) The gradient mappings $\nabla\phi$ and $\nabla\phi^*$ are inverse of each other.

3 Linear Convergence of the Primal-Dual Gradient Method

In this section we show the linear convergence of Algorithm 1 on Problem (1) under the following assumptions:

Assumption 3.1. *f is convex and ρ-smooth ($\rho \ge 0$).*

Assumption 3.2. *g is β-smooth and α-strongly convex ($\beta \ge \alpha > 0$).*

Assumption 3.3. *The matrix $A \in \mathbb{R}^{d_2\times d_1}$ satisfies* $\mathrm{rank}(A) = d_1$.

While the first two assumptions on f and g are standard in convex optimization literature, the third one is important for ensuring linear convergence of Problem (1). Note, for example, that if A is the all-zero matrix, then there is no interaction between x and y, and to solve the convex optimization problem on x we need at least $\Omega\left(\frac{1}{\sqrt{\epsilon}}\right)$ iterations [Nesterov, 2013] instead of $O\left(\log\frac{1}{\epsilon}\right)$.

Denote by $(x^*, y^*) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ the optimal solution to Problem (1). For simplicity, we let $\sigma_{\max} := \sigma_{\max}(A)$ and $\sigma_{\min} := \sigma_{\min}(A)$.

Recall the first-order optimality condition:

$$\begin{cases}\nabla_x L(x^*, y^*) = \nabla f(x^*) + A^\top y^* = 0,\\ \nabla_y L(x^*, y^*) = -\nabla g(y^*) + Ax^* = 0.\end{cases} \tag{6}$$

Theorem 3.1. *In the setting of Algorithm 1, define $a_t := \|x_t - x^*\|$ and $b_t := \|y_t - \nabla g^*(Ax_t)\|$. Let $\lambda := \frac{2\beta\sigma_{\max}\cdot\left(\rho+\frac{\sigma_{\max}^2}{\alpha}\right)}{\alpha\sigma_{\min}^2}$ and $P_t := \lambda a_t + b_t$. If we choose $\eta_1 = \frac{\alpha}{(\alpha+\beta)\left(\frac{\sigma_{\max}^2}{\alpha}+\lambda\sigma_{\max}\right)}$ and $\eta_2 = \frac{2}{\alpha+\beta}$, then we have*

$$P_{t+1} \le \left(1 - C\cdot\frac{\alpha^2\sigma_{\min}^4}{\beta^3\sigma_{\max}^2\cdot\left(\rho+\frac{\sigma_{\max}^2}{\alpha}\right)}\right)P_t$$

for some absolute constant $C > 0$.

In this theorem, we use $P_t = \lambda a_t + b_t$ as the potential function and show that this function shrinks at a geometric rate. Note that from (6) and Fact 2.1 (ii) we have $y^* = (\nabla g)^{-1}(Ax^*) = \nabla g^*(Ax^*)$. Then we have upper bounds $\|x_t - x^*\| = a_t \le \frac{1}{\lambda}P_t$ and $\|y_t - y^*\| \le \|y_t - \nabla g^*(Ax_t)\| + \|\nabla g^*(Ax_t) - y^*\| = b_t + \|\nabla g^*(Ax_t) - \nabla g^*(Ax^*)\| \le b_t + \frac{\sigma_{\max}}{\alpha}a_t \le$

$\max\left\{1, \frac{\sigma_{\max}}{\alpha\lambda}\right\} P_t$, which imply that if P_t is small then (x_t, y_t) will be close to the optimal solution (x^*, y^*). Therefore a direct corollary of Theorem 3.1 is:

Corollary 3.1. *For any $\epsilon > 0$, after $O^*\left(\log \frac{P_0}{\epsilon}\right)$ iterations, we have $\|x_t - x^*\| \le \epsilon$ and $\|y_t - y^*\| \le \epsilon$, where $O^*(\cdot)$ hides polynomial factors in $\beta, 1/\alpha, \sigma_{\max}, 1/\sigma_{\min}$ and ρ.*

We remark that our theorem suggests that step sizes depend on problem parameters which may be unknown. In practice, we may try to use a small amount of data to estimate them first or use the adaptive tuning heuristic introduced in [Wang and Xiao, 2017].

3.1 Proof of Theorem 3.1

Now we present the proof of Theorem 3.1.

First recall the standard linear convergence guarantee of gradient descent on a smooth and strongly convex objective. See Theorem 3.12 in [Bubeck, 2015] for a proof.

Lemma 3.1. *Suppose $\phi : \mathbb{R}^d \to \mathbb{R}$ is γ-smooth and δ-strongly convex, and let $\bar{x} := \operatorname{argmin}_{x \in \mathbb{R}^d} \phi(x)$. For any $0 < \eta \le \frac{2}{\gamma+\delta}$, $x \in \mathbb{R}^d$, letting $\tilde{x} = x - \eta \nabla \phi(x)$, we have*

$$\|\tilde{x} - \bar{x}\| \le (1 - \delta\eta) \|x - \bar{x}\|.$$

Step 1: Bounding the Decrease of $\|x_t - x^*\|$ via a One-Step "Ghost" Algorithm.[3] Our technique is to consider the following one-step "ghost" algorithm for the primal variable, which corresponds to a gradient descent step for the primal problem (2). We define an auxiliary variable $\tilde{x}_{t+1}$: given x_t, let

$$\tilde{x}_{t+1} := x_t - \eta_1 \left(\nabla f(x_t) + \nabla h(x_t)\right). \tag{7}$$

where $h(x) := g^*(Ax)$. Note that $\tilde{x}_{t+1}$ is defined only for the purpose of the proof. Our main idea is to use this "ghost" algorithm as a reference and bound the distance between the primal-dual gradient iterate x_{t+1} and this "ghost" variable $\tilde{x}_{t+1}$. We first prove with this "ghost" algorithm, the distance between the primal variable and the optimum x^* decreases at a geometric rate.

Proposition 3.1. *If $\eta_1 \le \frac{2}{\rho + \sigma_{\max}^2/\alpha + \sigma_{\min}^2/\beta}$, then*

$$\|\tilde{x}_{t+1} - x^*\| \le \left(1 - \frac{\sigma_{\min}^2}{\beta}\eta_1\right) \|x_t - x^*\|.$$

Proof. Since (7) is a gradient descent step for the primal problem (2) whose objective is $P(x) = h(x) + f(x)$ where $h(x) = g^*(Ax)$, it suffices to show that P is smooth and strongly convex in order to apply Lemma 3.1. Note that g^* is $\frac{1}{\alpha}$-smooth and $\frac{1}{\beta}$-strongly convex according to Fact 2.1.

We have $\nabla h(x) = A^\top \nabla g^*(Ax)$. Then for any $x, x' \in \mathbb{R}^d$ we have

$$\begin{aligned}
&\|\nabla P(x) - \nabla P(x')\| \\
\le& \|\nabla f(x) - \nabla f(x')\| + \left\|A^\top \nabla g^*(Ax) - A^\top \nabla g^*(Ax')\right\| \\
\le& \rho \|x - x'\| + \sigma_{\max} \|\nabla g^*(Ax) - \nabla g^*(Ax')\| \\
\le& \rho \|x - x'\| + \frac{\sigma_{\max}}{\alpha} \|Ax - Ax'\| \\
\le& \rho \|x - x'\| + \frac{\sigma_{\max}^2}{\alpha} \|x - x'\| \\
=& (\rho + \sigma_{\max}^2/\alpha) \|x - x'\|,
\end{aligned}$$

where we have used the ρ-smoothness of f, the $\frac{1}{\alpha}$-smoothness of g^*, and the bound on $\sigma_{\max}(A)$. Therefore P is $(\rho + \sigma_{\max}^2/\alpha)$-smooth.

On the other hand, for any $x, x' \in \mathbb{R}^d$ we have

$$\begin{aligned}
&P(x') - P(x) \\
=& f(x') - f(x) + g^*(Ax') - g^*(Ax) \\
\ge& \langle \nabla f(x), x' - x \rangle + \langle \nabla g^*(Ax), Ax' - Ax \rangle \\
&+ \frac{1/\beta}{2} \|Ax' - Ax\|^2 \\
=& \langle \nabla f(x) + A^\top \nabla g^*(Ax), x' - x \rangle + \frac{1}{2\beta} \|Ax' - Ax\|^2 \\
\ge& \langle \nabla P(x), x' - x \rangle + \frac{1}{2\beta} \sigma_{\min}^2 \|x' - x\|^2,
\end{aligned}$$

where we have used the convexity of f, the $\frac{1}{\beta}$-strong convexity of g^*, and that A has full column rank. Therefore P is $\sigma_{\min}^2/\beta$-strongly convex.

With the smoothness and the strong convexity of P, the proof is completed by applying Lemma 3.1. □

Proposition 3.1 suggests that if we use the "ghost" algorithm (7), we have the desired linear convergence property. The following proposition gives an upper bound on $\|x_{t+1} - x^*\|$ by bounding the distance between x_{t+1} and $\tilde{x}_{t+1}$.

Proposition 3.2. *If $\eta_1 \le \frac{2}{\rho + \sigma_{\max}^2/\alpha + \sigma_{\min}^2/\beta}$, then*

$$\begin{aligned}
\|x_{t+1} - x^*\| \le& \left(1 - \frac{\sigma_{\min}^2}{\beta}\eta_1\right) \|x_t - x^*\| \\
&+ \sigma_{\max}\eta_1 \|y_t - \nabla g^*(Ax_t)\|.
\end{aligned} \tag{8}$$

Proof. We have $\tilde{x}_{t+1} - x_{t+1} = \eta_1 A^\top (y_t - \nabla g^*(Ax_t))$, which implies

$$\|\tilde{x}_{t+1} - x_{t+1}\| \le \eta_1 \sigma_{\max} \|y_t - \nabla g^*(Ax_t)\|.$$

Then the proposition follows by applying the triangle inequality and Proposition 3.1. □

[3] $\|x_t - x^*\|$ may not decrease as t increases. Here what we mean is to upper bound $\|x_{t+1} - x^*\|$ using $\|x_t - x^*\|$ and an error term.

Step 2: Bounding the Decrease of $\|y_t - \nabla g^*(Ax_t)\|$. One may want to show the decrease of $\|y_t - y^*\|$ similarly using a "ghost" update for the dual variable. However, the objective function in the dual problem $\max_y \left\{-g(y) - f^*(-A^\top y)\right\}$ might be non-smooth, which means we cannot obtain a result similar to Proposition 3.1. Instead, we show that $\|y_t - \nabla g^*(Ax_t)\|$ decreases geometrically up to an error term.

Proposition 3.3. *We have*

$$\|x_{t+1} - x_t\| \le \left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)\eta_1 \|x_t - x^*\| + \sigma_{\max}\eta_1 \|y_t - \nabla g^*(Ax_t)\| .$$

Proof. Using the gradient update formula of the primal variable, we have

$$\begin{aligned} &\frac{1}{\eta_1}\|x_{t+1} - x_t\| = \left\|\nabla f(x_t) + A^\top y_t\right\| \\ &\le \left\|\nabla f(x_t) + A^\top \nabla g^*(Ax_t)\right\| + \left\|A^\top (y_t - \nabla g^*(Ax_t))\right\| \\ &\le \left\|\nabla f(x_t) + A^\top \nabla g^*(Ax_t)\right\| + \sigma_{\max}\|y_t - \nabla g^*(Ax_t)\| . \end{aligned} \tag{9}$$

Recall that the primal objective function $P(x) = f(x) + g^*(Ax)$ is $(\rho + \sigma_{\max}^2/\alpha)$-smooth (see the proof of Proposition 3.1). So we have

$$\begin{aligned} &\left\|\nabla f(x_t) + A^\top \nabla g^*(Ax_t)\right\| = \|\nabla P(x_t)\| \\ &= \|\nabla P(x_t) - \nabla P(x^*)\| \le (\rho + \sigma_{\max}^2/\alpha)\|x_t - x^*\| . \end{aligned}$$

Plugging this back to (9) we obtain the desired result. □

Proposition 3.4. *If $\eta_2 \le \frac{2}{\alpha+\beta}$, then*

$$\begin{aligned} &\|y_{t+1} - \nabla g^*(Ax_{t+1})\| \\ &\le \left(1 - \alpha\eta_2 + \frac{\sigma_{\max}^2}{\alpha}\eta_1\right)\|y_t - \nabla g^*(Ax_t)\| \\ &\quad + \frac{\sigma_{\max}}{\alpha}\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)\eta_1\|x_t - x^*\| . \end{aligned}$$

Proof. For fixed x_t, the update rule $y_{t+1} = y_t - \eta_2(\nabla g(y_t) - Ax_t)$ is a gradient descent step for the objective function $\tilde{g}(y) := g(y) - y^\top Ax_t$ which is also β-smooth and α-strongly convex. By the optimality condition, the minimizer $\tilde{y}^* = \operatorname{argmin}_{y\in\mathbb{R}^d}\tilde{g}(y)$ satisfies $\nabla g(\tilde{y}^*) = Ax_t$, i.e., $\tilde{y}^* = \nabla g^*(Ax_t)$. Then from Lemma 3.1 we know that

$$\|y_{t+1} - \nabla g^*(Ax_t)\| \le (1 - \alpha\eta_2)\|y_t - \nabla g^*(Ax_t)\| . \tag{10}$$

Since we want to upper bound $\|y_{t+1} - \nabla g^*(Ax_{t+1})\|$, we need to take into account the difference between x_{t+1} and x_t. We prove an upper bound on $\|x_{t+1} - x_t\|$ in Proposition 3.3. Using Proposition 3.3 and (10), we have

$$\begin{aligned} &\|y_{t+1} - \nabla g^*(Ax_{t+1})\| \\ &\le \|y_{t+1} - \nabla g^*(Ax_t)\| + \|\nabla g^*(Ax_{t+1}) - \nabla g^*(Ax_t)\| \\ &\le \|y_{t+1} - \nabla g^*(Ax_t)\| + \frac{\sigma_{\max}}{\alpha}\|x_{t+1} - x_t\| \\ &\le (1 - \alpha\eta_2)\|y_t - \nabla g^*(Ax_t)\| \\ &\quad + \frac{\sigma_{\max}}{\alpha}\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)\eta_1\|x_t - x^*\| \\ &\quad + \frac{\sigma_{\max}^2}{\alpha}\eta_1\|y_t - \nabla g^*(Ax_t)\| . \end{aligned}$$

□

Note that the upper bound on $\|x_{t+1} - x_t\|$ given in Proposition 3.3 is proportional to η_1, not to η_2. This allows us to choose a relatively small η_1 to ensure that the factor $1 - \alpha\eta_2 + \frac{\sigma_{\max}^2}{\alpha}\eta_1$ in Proposition 3.4 is indeed less than 1, i.e., $\|y_t - \nabla g^*(Ax_t)\|$ is approximately decreasing.

Step 3: Putting Things Together. Now we are ready to finish the proof of Theorem 3.1. From Propositions 3.2 and 3.4 we have

$$a_{t+1} \le \left(1 - \frac{\sigma_{\min}^2}{\beta}\eta_1\right)a_t + \sigma_{\max}\eta_1 b_t, \tag{11}$$

$$\begin{aligned} b_{t+1} \le &\frac{\sigma_{\max}}{\alpha}\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)\eta_1 a_t \\ &+ \left(1 - \alpha\eta_2 + \frac{\sigma_{\max}^2}{\alpha}\eta_1\right)b_t. \end{aligned} \tag{12}$$

To prove the convergence of sequences $\{a_t\}$ and $\{b_t\}$ to 0, we consider a linear combination $P_t = \lambda a_t + b_t$ with a free parameter $\lambda > 0$ to be determined. Combining (11) and (12), with some routine calculations, we can show that our choices of λ, η_1 and η_2 given in Theorem 3.1 can ensure $P_{t+1} \le cP_t$ for some $0 < c < 1$, as desired. We give the remaining details in Appendix A.1.

4 Extension to Primal-Dual SVRG

In this section we consider the case where the saddle point problem (1) admits a finite-sum structure:[4]

$$\min_{x\in\mathbb{R}^{d_1}}\max_{y\in\mathbb{R}^{d_2}} L(x,y) = \frac{1}{n}\sum_{i=1}^{n} L_i(x,y), \tag{13}$$

where $L_i(x,y) := f_i(x) + y^\top A_i x - g_i(y)$. Optimization problems with finite-sum structure are ubiquitous in

[4] For ease of presentation we assume f, g and K can be split into n terms. It is not hard to generalize our analysis to the case where f, g and A can be split into different numbers of terms.

machine learning, because loss functions can often be written as a sum of individual loss terms corresponding to individual observations.

In this section, we make the following assumptions:

Assumption 4.1. *Each f_i is ρ-smooth ($\rho \geq 0$), and $f = \frac{1}{n}\sum_{i=1}^{n} f_i$ is convex.*

Assumption 4.2. *Each g_i is β-smooth, and $g = \frac{1}{n}\sum_{i=1}^{n} g_i$ is α-strongly convex ($\beta \geq \alpha > 0$).*

Assumption 4.3. *Each A_i satisfies $\sigma_{\max}(A_i) \leq M$, and $A = \frac{1}{n}\sum_{i=1}^{n} A_i$ has rank d_1.*

Note that we only require component functions f_i and g_i to be smooth; they are not necessarily convex. However, the overall objective function $L(x,y) = f(x) + y^\top Ax - g(y)$ still has to satisfy Assumptions 3.1-3.3.

Given the finite-sum structure (13), we denote the individual gradient of each L_i as

$$B_i\left(x,y\right) := \begin{bmatrix} \nabla_x L_i(x,y) \\ \nabla_y L_i(x,y) \end{bmatrix} = \begin{bmatrix} \nabla f_i(x) + A_i^\top y \\ A_i x - \nabla g_i\left(y\right) \end{bmatrix},$$

and the full gradient of L as

$$B\left(x,y\right) := \frac{1}{n}\sum_{i=1}^{n} B_i(x,y) = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\left(\nabla f_i(x) + A_i^\top y\right) \\ \frac{1}{n}\sum_{i=1}^{n}\left(A_i x - \nabla g_i\left(y\right)\right) \end{bmatrix}.$$

A naive computation of $A_i x$ or $A_i^\top y$ takes $O(d_1 d_2)$ time. However, in many applications like policy evaluation [Du et al., 2017] and empirical risk minimization, each A_i is given as the outer product of two vectors (i.e., a rank-1 matrix), which makes $A_i x$ and $A_i^\top y$ computable in only $O(d)$ time, where $d = \max\left\{d_1, d_2\right\}$. In this case, computing an individual gradient $B_i(x,y)$ takes $O(d)$ time while computing the full gradient $B(x,y)$ takes $O(nd)$ time.

We adapt the stochastic variance reduced gradient (SVRG) method [Johnson and Zhang, 2013] to solve Problem (13). The algorithm uses two layers of loops. In an outer loop, the algorithm first computes a full gradient using a "snapshot" point $(\tilde{x}, \tilde{y})$, and then the algorithm executes N inner loops, where N is a parameter to be chosen. In each inner loop, the algorithm randomly samples an index i from $[n]$ and updates the current iterate (x,y) using a variance-reduced stochastic gradient:

$$B_i(x,y,\tilde{x},\tilde{y}) = B_i(x,y) + B(\tilde{x},\tilde{y}) - B_i(\tilde{x},\tilde{y}). \quad (14)$$

Here, $B_i(x,y)$ is the stochastic gradient at (x,y) computed using the random index i, and $B(\tilde{x},\tilde{y}) - B_i(\tilde{x},\tilde{y})$ is a term used to reduce the variance in $B_i(x,y)$ while keeping $B_i(x,y,\tilde{x},\tilde{y})$ an unbiased estimate of $B(x,y)$. The full details of the algorithm are provided in Algorithm 2. For clarity, we denote by $(\tilde{x}_t, \tilde{y}_t)$ the snapshot point in the t-th epoch (outer loop), and denote by $(x_{t,0}, y_{t,0}), (x_{t,1}, y_{t,1}), \ldots$ all the intermediate iterates within this epoch.

Algorithm 2 Primal-Dual SVRG

```
Inputs: initial points $\tilde{x}_0 \in \mathbb{R}^{d_1}, \tilde{y}_0 \in \mathbb{R}^{d_2}$, step sizes
    $\eta_1, \eta_2 > 0$, number of inner iterations $N \in \mathbb{N}$
for $t = 0, 1, \ldots$ do
    Compute $B(\tilde{x}_t, \tilde{y}_t)$
    $(x_{t,0}, y_{t,0}) = (\tilde{x}_t, \tilde{y}_t)$
    for $j = 0$ to $N - 1$ do
        Sample an index $i_j$ uniformly from $[n]$
        Compute $B_{i_j}(x_{t,j}, y_{t,j})$ and $B_{i_j}(\tilde{x}_t, \tilde{y}_t)$
        $\begin{bmatrix} x_{t,j+1} \\ y_{t,j+1} \end{bmatrix} = \begin{bmatrix} x_{t,j} \\ y_{t,j} \end{bmatrix} - \begin{bmatrix} \eta_1 I_{d_1} & 0 \\ 0 & -\eta_2 I_{d_2} \end{bmatrix} B_{i_j}(x_{t,j}, y_{t,j}, \tilde{x}_t, \tilde{y}_t),$
        where $B_{i_j}(x_{t,j}, y_{t,j}, \tilde{x}_t, \tilde{y}_t)$ is defined in (14)
    end for
    $(\tilde{x}_{t+1}, \tilde{y}_{t+1}) = (x_{t,j_t}, y_{t,j_t})$, where $j_t$ is an index
    sampled uniformly from $\{0, 1, \ldots, N-1\}$
end for
```

The following theorem establishes the linear convergence guarantee of Algorithm 2.

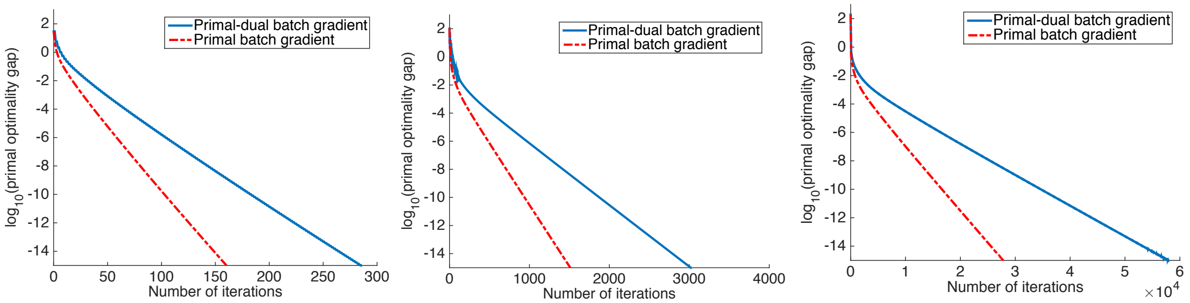
Theorem 4.1. *There exists a choice of parameters $\eta_1, \eta_2 = \text{poly}\left(\beta, \rho, M, 1/\alpha, 1/\sigma_{\min}(A)\right)^{-1}$ and $N = \text{poly}\left(\beta, \rho, M, 1/\alpha, 1/\sigma_{\min}(A)\right)$ in Algorithm 2, as well as another number $\mu = \text{poly}\left(\beta, \rho, M, 1/\alpha, 1/\sigma_{\min}(A)\right)$, such that if we define $Q_t = \mathbb{E}\left[\left\|\tilde{x}_t - x^*\right\|^2 + \mu\left\|\tilde{y}_t - \nabla g^*(A\tilde{x}_t)\right\|^2\right]$, then Algorithm 2 guarantees $Q_{t+1} \leq \frac{1}{2}Q_t$ for all t.*

Since computing a full gradient takes $O(nd)$ time and each inner loop takes $O(d)$ time, each epoch takes $O(nd + Nd)$ time in total. Therefore, the total running time of Algorithm 2 is $O\left((n+N)d\log\frac{1}{\epsilon}\right)$ in order to reach an ϵ-close solution, which is the desired running time of SVRG (note that N does not depend on n).

The proof of Theorem 4.1 is given in Appendix A.2. It relies on the same proof idea in Section 3 as well as the standard analysis technique for SVRG by Johnson and Zhang [2013].

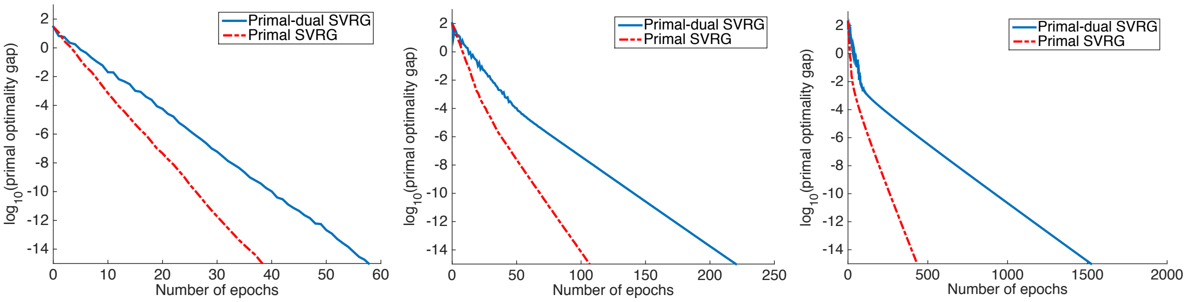
5 Preliminary Empirical Evaluation

We perform preliminary empirical evaluation for the following purposes: (i) to verify that both the primal-dual gradient method (Algorithm 1) and the primal-dual SVRG method (Algorithm 2) can indeed achieve linear convergence, (ii) to investigate the convergence rates of Algorithms 1 and 2, in comparison with their primal-only counterparts (i.e., the usual gradient descent and SVRG algorithms for the primal problem), and (iii) to



(a) Data $\sim \mathcal{N}(0, I_d)$ (b) Data $\sim \mathcal{N}(0, \Sigma)$, $\Sigma_{ij} = 2^{-|i-j|/2}$ (c) Data $\sim \mathcal{N}(0, \Sigma)$, $\Sigma_{ij} = 2^{-|i-j|/10}$

Figure 1: Comparison of batch gradient methods for smoothed-L_1-regularized regression with $d = 200, n = 500$.



(a) Data $\sim \mathcal{N}(0, I_d)$ (b) Data $\sim \mathcal{N}(0, \Sigma)$, $\Sigma_{ij} = 2^{-|i-j|/2}$ (c) Data $\sim \mathcal{N}(0, \Sigma)$, $\Sigma_{ij} = 2^{-|i-j|/10}$

Figure 2: Comparison of SVRG methods for smoothed-L_1-regularized regression with $d = 200$ and $n = 500$.

compare the convergence rates of Algorithms 1 and 2.

We consider the linear regression problem with smoothed-L_1 regularization, formulated as

$$\min_{x \in \mathbb{R}^d} \frac{1}{2n} \|Ax - b\|^2 + \lambda R_a(x), \tag{15}$$

where $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and $R_a(x) := \sum_{i=1}^d \frac{1}{a} \left(\log(1 + e^{ax_i}) + \log(1 + e^{-ax_i}) \right)$ is the *smoothed -L_1 regularization* [Schmidt et al., 2007].[5] Note that $R_a(x)$ is smooth but not strongly convex, and does not have a closed-form proximal mapping. As discussed in Section 1.2, Problem (15) admits a saddle point formulation:

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^n} \left\{ \frac{1}{n} \left(-\frac{1}{2} \|y\|^2 - b^\top y + y^\top Ax \right) + \lambda R_a(x) \right\}.$$

In this experiment we choose $a = 10$ and $\lambda = 0.01/n$.

We generate data (i.e. rows of A) from a Gaussian distribution $\mathcal{N}(0, \Sigma)$, where we consider three cases: (a) $\Sigma = I_d$, (b) $\Sigma_{ij} = 2^{-|i-j|/2}$, and (c) $\Sigma_{ij} = 2^{-|i-j|/10}$. These three choices result in small, medium, and large condition numbers of A, respectively.[6] In Figures 1 and 2, we plot the performances of batch gradient and SVRG algorithms, where we choose $d = 200$ and $n = 500$. We tune the step sizes in every case in order to observe the optimal convergence rates.

These plots show that: (i) both Algorithm 1 and Algorithm 2 can indeed achieve linear convergence, verifying our theorems; (ii) in all our examples, primal-dual methods always converge slower than the corresponding primal methods, but they are only slower by no more than 3 times; (iii) Algorithm 2 has a much faster convergence rate than Algorithm 1, especially when the condition number is large, which verifies the theoretical result that SVRG can significantly reduce the computational complexity.

6 Conclusion

We prove that the vanilla primal-dual gradient method can achieve linear convergence for convex-concave saddle point problem (1) without strong convexity in the primal variable. We develop a novel proof strategy and further use this proof strategy to show the linear convergence of the primal-dual SVRG method for saddle point problems with finite-sum structures. It would be interesting to study whether our technique can be used to analyze non-convex problems.

[5] When $a > 0$ is large we have $R_a(x) \approx \|x\|_1$ for all $x \in \mathbb{R}^d$.

[6] The condition number of a matrix A is defined as $\sigma_{\max}(A)/\sigma_{\min}(A)$.

Acknowledgements

We thank Qi Lei, Yuanzhi Li, Jialei Wang, Yining Wang and Lin Xiao for helpful discussions. Wei Hu was supported by NSF, ONR, Simons Foundation, Schmidt Foundation, Mozilla Research, Amazon Research, DARPA and SRC.

References

Zeyuan Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *STOC*, pages 1200–1205. ACM, 2017.

Francis Bach, Julien Mairal, and Jean Ponce. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*, 2008.

Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton University Press, 2009.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

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.

George HG Chen and R Tyrrell Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.

Simon S. Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *ICML*, pages 1049–1058, 2017.

Martin Hast, KJ Astrom, Bo Bernhardsson, and Stephen Boyd. Pid design by convex-concave optimization. In *European Control Conference (ECC)*, pages 4460–4465. IEEE, 2013.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Manuscript*, 2009.

Nicolas Le Roux, Mark Schmidt, and Francis R Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2663–2671, 2012.

Qi Lei, Ian En-Hsu Yen, Chao-yuan Wu, Inderjit S Dhillon, and Pradeep Ravikumar. Doubly greedy primal-dual coordinate descent for sparse empirical risk minimization. In *ICML*, pages 2034–2042, 2017.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *NIPS*, pages 3384–3392, 2015.

Yongchao Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Primal–dual hybrid gradient method for distributionally robust optimization problems. *Operations Research Letters*, 45(6):625–630, 2017.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *NIPS*, pages 1416–1424, 2016.

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

R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Mark Schmidt, Glenn Fung, and Rmer Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *European Conference on Machine Learning*, pages 286–297. Springer, 2007.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

Jialei Wang and Lin Xiao. Exploiting strong convexity from data with primal-dual first-order algorithms. In *ICML*, pages 3694–3702, 2017.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Lin Xiao, Adams Wei Yu, Qihang Lin, and Weizhu Chen. Dscovr: Randomized primal-dual block coordinate algorithms for asynchronous distributed optimization. *arXiv preprint arXiv:1710.05080*, 2017.

Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. In *NIPS*, pages 1537–1544, 2005.

Yuchen Zhang and Xiao Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *ICML*, pages 353–361, 2015.

Appendix

A Omitted Proofs

A.1 Finishing the Proof of Theorem 3.1

Proof of Theorem 3.1. Combining (11) and (12), we obtain

$$\begin{aligned}
&P_{t+1} = \lambda a_{t+1} + b_{t+1} \\
\le &\left(1 - \frac{\sigma_{\min}^2}{\beta}\eta_1\right)\lambda a_t + \lambda\sigma_{\max}\eta_1 b_t + \frac{\sigma_{\max}}{\alpha}\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)\eta_1 a_t + \left(1 - \alpha\eta_2 + \frac{\sigma_{\max}^2}{\alpha}\eta_1\right) b_t \\
= &\left(1 - \frac{\sigma_{\min}^2}{\beta}\eta_1 + \frac{1}{\lambda}\frac{\sigma_{\max}}{\alpha}\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)\eta_1\right)\lambda a_t + \left(1 - \alpha\eta_2 + \left(\frac{\sigma_{\max}^2}{\alpha} + \lambda\sigma_{\max}\right)\eta_1\right) b_t.
\end{aligned}$$

If we can choose λ, η_1 and η_2 such that both coefficients

$$c_1 := 1 - \frac{\sigma_{\min}^2}{\beta}\eta_1 + \frac{1}{\lambda}\frac{\sigma_{\max}}{\alpha}\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)\eta_1$$

and

$$c_2 := 1 - \alpha\eta_2 + \left(\frac{\sigma_{\max}^2}{\alpha} + \lambda\sigma_{\max}\right)\eta_1$$

are strictly less than 1, we have linear convergence.

It remains to show that our choices of parameters

$$\begin{aligned}
\lambda &= \frac{2\beta\sigma_{\max}\cdot\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)}{\alpha\sigma_{\min}^2}, \\
\eta_1 &= \frac{\alpha}{(\alpha+\beta)\left(\frac{\sigma_{\max}^2}{\alpha} + \lambda\sigma_{\max}\right)}, \\
\eta_2 &= \frac{2}{\alpha+\beta},
\end{aligned}$$

give the desired upper bound on $\max\{c_1, c_2\}$.

First we verify that our choices of η_1 and η_2 satisfy the requirements in Propositions 3.2 and 3.4. It is clear that $\eta_2 \le \frac{2}{\alpha+\beta}$ is satisfied. For η_1, we have

$$\eta_1 \le \frac{\alpha}{\beta\lambda\sigma_{\max}} = \frac{\alpha^2\sigma_{\min}^2}{2\beta^2\sigma_{\max}^2\cdot\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)} \le \frac{1}{2\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)} \le \frac{1}{\rho + \frac{\sigma_{\max}^2}{\alpha} + \frac{\sigma_{\min}^2}{\beta}}.$$

Therefore the requirements in Propositions 3.2 and 3.4 are satisfied.

Next we calculate c_1 and c_2. Since $\left(\frac{\sigma_{\max}^2}{\alpha} + \lambda\sigma_{\max}\right)\eta_1 = \left(\frac{\sigma_{\max}^2}{\alpha} + \lambda\sigma_{\max}\right)\frac{\alpha}{(\alpha+\beta)\left(\frac{\sigma_{\max}^2}{\alpha} + \lambda\sigma_{\max}\right)} = \frac{\alpha}{\alpha+\beta} = \frac{\alpha}{2}\eta_2$, we have

$$c_2 = 1 - \frac{1}{2}\alpha\eta_2 = 1 - \frac{\alpha}{\alpha+\beta}. \tag{16}$$

For c_1, since $\frac{1}{\lambda}\frac{\sigma_{\max}}{\alpha}\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right) = \frac{\alpha\sigma_{\min}^2}{2\beta\sigma_{\max}\cdot\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)}\cdot\frac{\sigma_{\max}}{\alpha}\left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right) = \frac{\sigma_{\min}^2}{2\beta}$, we have

$$c_1 = 1 - \frac{\sigma_{\min}^2}{2\beta}\eta_1 = 1 - \frac{\sigma_{\min}^2}{2\beta}\cdot\frac{\alpha}{(\alpha+\beta)\left(\frac{\sigma_{\max}^2}{\alpha} + \lambda\sigma_{\max}\right)}. \tag{17}$$

Note that $\lambda\sigma_{\max} = \frac{2\beta\sigma_{\max}^2 \cdot \left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)}{\alpha\sigma_{\min}^2} \ge \frac{2\beta\sigma_{\max}^4}{\alpha^2\sigma_{\min}^2} \ge \frac{2\sigma_{\max}^2}{\alpha}$. Then (17) implies

$$\begin{aligned} c_1 &\le 1 - \frac{\alpha\sigma_{\min}^2}{2\beta(\alpha+\beta)\left(\frac{1}{2}\lambda\sigma_{\max} + \lambda\sigma_{\max}\right)} = 1 - \frac{\alpha\sigma_{\min}^2}{3\beta(\alpha+\beta)\lambda\sigma_{\max}} \\ &= 1 - \frac{\alpha^2\sigma_{\min}^4}{6\beta^2(\alpha+\beta)\sigma_{\max}^2 \cdot \left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)} \le 1 - \frac{\alpha^2\sigma_{\min}^4}{12\beta^3\sigma_{\max}^2 \cdot \left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)}. \end{aligned} \tag{18}$$

Combining (16) and (18), we obtain

$$\max\{c_1, c_2\} \le \max\left\{1 - \frac{\alpha}{\alpha+\beta}, 1 - \frac{\alpha^2\sigma_{\min}^4}{12\beta^3\sigma_{\max}^2 \cdot \left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)}\right\} = 1 - \frac{\alpha^2\sigma_{\min}^4}{12\beta^3\sigma_{\max}^2 \cdot \left(\rho + \frac{\sigma_{\max}^2}{\alpha}\right)}.$$

Therefore the proof is completed. □

A.2 Proof of Theorem 4.1

Proof of Theorem 4.1. Denote $\sigma_{\min} := \sigma_{\min}(A)$. Let

$$\begin{aligned} \varphi_i(x,y) &:= \nabla_x L_i(x,y) = \nabla f_i(x) + A_i^\top y, \\ \varphi(x,y) &:= \nabla_x L(x,y) = \nabla f(x) + A^\top y, \\ \psi_i(x,y) &:= \nabla_y L_i(x,y) = A_i x - \nabla g_i(y), \\ \psi(x,y) &:= \nabla_y L(x,y) = Ax - \nabla g(y), \end{aligned}$$

and define $\theta : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ as $\theta(x) := \nabla g^*(Ax)$. Note that we have $\theta(x^*) = y^*$ from (6).

Step 1: Bound for the Primal Variable. Consider one iteration of the inner loop $x_{t,j+1} = x_{t,j} - \eta_1\left(\varphi_{i_j}(x_{t,j}, y_{t,j}) + \varphi(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(\tilde{x}_t, \tilde{y}_t)\right)$. We now only consider the randomness in i_j, conditioned on everything in previous iterations. Because we have $\mathbb{E}[x_{t,j+1}] = x_{t,j} - \eta_1\varphi(x_{t,j}, y_{t,j})$, using the equation $\mathbb{E}\|\xi\|^2 = \|\mathbb{E}\xi\|^2 + \mathbb{E}\|\xi - \mathbb{E}\xi\|^2$ we have:

$$\begin{aligned} \mathbb{E}\|x_{t,j+1} - x^*\|^2 &= \|\mathbb{E}[x_{t,j+1} - x^*]\|^2 + \mathbb{E}\left[\|(x_{t,j+1} - x^*) - \mathbb{E}[x_{t,j+1} - x^*]\|^2\right] \\ &= \|x_{t,j} - \eta_1\varphi(x_{t,j}, y_{t,j}) - x^*\|^2 + \eta_1^2\mathbb{E}\left[\left\|\varphi_{i_j}(x_{t,j}, y_{t,j}) + \varphi(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \varphi(x_{t,j}, y_{t,j})\right\|^2\right]. \end{aligned} \tag{19}$$

For the first term in (19), we note that it has the same form as the update rule in Algorithm 1 for the primal variable. Therefore, we can apply Proposition 3.2 and get (noticing $\sigma_{\max}(A) \le M$, and assuming η_1 is sufficiently small)

$$\begin{aligned} &\|x_{t,j} - \eta_1\varphi(x_{t,j}, y_{t,j}) - x^*\|^2 \\ \le& \left(\left(1 - \frac{\sigma_{\min}^2}{\beta}\eta_1\right)\|x_{t,j} - x^*\| + M\eta_1\|y_{t,j} - \theta(x_{t,j})\|\right)^2 \\ \le& \left[\left(1 - \frac{\sigma_{\min}^2}{\beta}\eta_1\right) + \frac{\sigma_{\min}^2}{\beta}\eta_1\right] \cdot \left[\left(1 - \frac{\sigma_{\min}^2}{\beta}\eta_1\right)\|x_{t,j} - x^*\|^2 + \frac{\beta M^2}{\sigma_{\min}^2}\eta_1\|y_{t,j} - \theta(x_{t,j})\|^2\right] \\ =& \left(1 - \frac{\sigma_{\min}^2}{\beta}\eta_1\right)\|x_{t,j} - x^*\|^2 + \frac{\beta M^2}{\sigma_{\min}^2}\eta_1\|y_{t,j} - \theta(x_{t,j})\|^2, \end{aligned} \tag{20}$$

where the second inequality is due to Cauchy-Schwarz inequality.

Next, we bound the second term in (19). First, using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ we get

$$\begin{aligned} &\eta_1^2\mathbb{E}\left[\left\|\varphi_{i_j}(x_{t,j}, y_{t,j}) + \varphi(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \varphi(x_{t,j}, y_{t,j})\right\|^2\right] \\ \le& 2\eta_1^2\mathbb{E}\left[\left\|\varphi_{i_j}(x_{t,j}, y_{t,j}) - \varphi(x_{t,j}, y_{t,j}) - \varphi_{i_j}(x^*, y^*) + \varphi(x^*, y^*)\right\|^2\right] \end{aligned}$$

$$+ 2\eta_1^2 \mathbb{E}\left[\left\|\varphi(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(\tilde{x}_t, \tilde{y}_t) + \varphi_{i_j}(x^*, y^*) - \varphi(x^*, y^*)\right\|^2\right].$$

Note that $\mathbb{E}\left[\varphi_{i_j}(x_{t,j}, y_{t,j}) - \varphi_{i_j}(x^*, y^*)\right] = \varphi(x_{t,j}, y_{t,j}) - \varphi(x^*, y^*)$ and $\mathbb{E}\left[\varphi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(x^*, y^*)\right] = \varphi(\tilde{x}_t, \tilde{y}_t) - \varphi(x^*, y^*)$. Using $\mathbb{E}\left\|\xi - \mathbb{E}\xi\right\|^2 \le \mathbb{E}\left\|\xi\right\|^2$ we get

$$\begin{aligned}
&\eta_1^2 \mathbb{E}\left[\left\|\varphi_{i_j}(x_{t,j}, y_{t,j}) + \varphi(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \varphi(x_{t,j}, y_{t,j})\right\|^2\right] \\
\le& 2\eta_1^2 \mathbb{E}\left[\left\|\varphi_{i_j}(x_{t,j}, y_{t,j}) - \varphi_{i_j}(x^*, y^*)\right\|^2 + \left\|\varphi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(x^*, y^*)\right\|^2\right].
\end{aligned}$$

Then from

$$\begin{aligned}
\left\|\varphi_i(x, y) - \varphi_i(x^*, y^*)\right\| &= \left\|\nabla f_i(x) + A_i^\top y - \nabla f_i(x^*) - A_i^\top y^*\right\| \\
&\le \left\|\nabla f_i(x) - \nabla f_i(x^*)\right\| + \left\|A_i^\top y - A_i^\top y^*\right\| \\
&\le \rho\left\|x - x^*\right\| + M\left\|y - y^*\right\| \\
&\le \rho\left\|x - x^*\right\| + M\left(\left\|y - \theta(x)\right\| + \left\|\theta(x) - y^*\right\|\right) \\
&= \rho\left\|x - x^*\right\| + M\left(\left\|y - \theta(x)\right\| + \left\|\theta(x) - \theta(x^*)\right\|\right) \\
&\le \rho\left\|x - x^*\right\| + M\left(\left\|y - \theta(x)\right\| + \frac{M}{\alpha}\left\|x - x^*\right\|\right) \\
&= \left(\rho + M^2/\alpha\right)\left\|x - x^*\right\| + M\left\|y - \theta(x)\right\|
\end{aligned} \tag{21}$$

we obtain

$$\begin{aligned}
&\eta_1^2 \mathbb{E}\left[\left\|\varphi_{i_j}(x_{t,j}, y_{t,j}) + \varphi(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \varphi(x_{t,j}, y_{t,j})\right\|^2\right] \\
\le& 2\eta_1^2\left[\left(\rho + M^2/\alpha\right)\left\|x_{t,j} - x^*\right\| + M\left\|y_{t,j} - \theta(x_{t,j})\right\|\right]^2 \\
&+ 2\eta_1^2\left[\left(\rho + M^2/\alpha\right)\left\|\tilde{x}_t - x^*\right\| + M\left\|\tilde{y}_t - \theta(\tilde{x}_t)\right\|\right]^2.
\end{aligned} \tag{22}$$

Plugging (20) and (22) into (19), we have

$$\begin{aligned}
&\mathbb{E}\|x_{t,j+1} - x^*\|^2 \\
\le& \left(1 - \frac{\sigma_{\min}^2}{\beta}\eta_1\right)\left\|x_{t,j} - x^*\right\|^2 + \frac{\beta M^2}{\sigma_{\min}^2}\eta_1\left\|y_{t,j} - \theta(x_{t,j})\right\|^2 \\
&+ 2\eta_1^2\left[\left(\rho + M^2/\alpha\right)\left\|x_{t,j} - x^*\right\| + M\left\|y_{t,j} - \theta(x_{t,j})\right\|\right]^2 \\
&+ 2\eta_1^2\left[\left(\rho + M^2/\alpha\right)\left\|\tilde{x}_t - x^*\right\| + M\left\|\tilde{y}_t - \theta(\tilde{x}_t)\right\|\right]^2 \\
\le& \left(1 - \eta_1/c_1\right)\left\|x_{t,j} - x^*\right\|^2 + c_2\eta_1\left\|y_{t,j} - \theta(x_{t,j})\right\|^2 + c_3\eta_1^2\left\|\tilde{x}_t - x^*\right\|^2 + c_4\eta_1^2\left\|\tilde{y}_t - \theta(\tilde{x}_t)\right\|^2,
\end{aligned} \tag{23}$$

where $c_1, \ldots, c_4$ all have the form $\operatorname{poly}\left(\beta, \rho, M, 1/\alpha, 1/\sigma_{\min}\right)$. Here we assume η_1 is sufficiently small.

Step 2: Bound for the Dual Variable. The dual update takes the form $y_{t,j+1} = y_{t,j} + \eta_2\left(\psi_{i_j}(x_{t,j}, y_{t,j}) + \psi(\tilde{x}_t, \tilde{y}_t) - \psi_{i_j}(\tilde{x}_t, \tilde{y}_t)\right)$. Same as before, We first only consider the randomness in i_j, conditioned on everything in previous iterations.

We have

$$\begin{aligned}
&\mathbb{E}\left[\left\|y_{t,j+1} - \theta(x_{t,j+1})\right\|^2\right] \\
\le& \mathbb{E}\left[\left(\left\|y_{t,j+1} - \theta(x_{t,j})\right\| + \left\|\theta(x_{t,j}) - \theta(x_{t,j+1})\right\|\right)^2\right] \\
=& \mathbb{E}\left[\left\|y_{t,j+1} - \theta(x_{t,j})\right\|^2 + \left\|\theta(x_{t,j}) - \theta(x_{t,j+1})\right\|^2 + 2\left\|y_{t,j+1} - \theta(x_{t,j})\right\| \cdot \left\|\theta(x_{t,j}) - \theta(x_{t,j+1})\right\|\right] \\
\le& \mathbb{E}\left\|y_{t,j+1} - \theta(x_{t,j})\right\|^2 + \mathbb{E}\left\|\theta(x_{t,j}) - \theta(x_{t,j+1})\right\|^2 + 2\sqrt{\mathbb{E}\left\|y_{t,j+1} - \theta(x_{t,j})\right\|^2 \cdot \mathbb{E}\left\|\theta(x_{t,j}) - \theta(x_{t,j+1})\right\|^2} \\
=& A + B + 2\sqrt{AB},
\end{aligned} \tag{24}$$

where $A := \mathbb{E}\left\|y_{t,j+1} - \theta(x_{t,j})\right\|^2$ and $B := \mathbb{E}\left\|\theta(x_{t,j}) - \theta(x_{t,j+1})\right\|^2$. The second inequality above is due to Cauchy-Schwarz inequality. Thus it remains to bound A and B.

We can bound A similar to (19):

$$\begin{aligned}
A &= \mathbb{E}\|y_{t,j+1} - \theta(x_{t,j})\|^2 \\
&= \|\mathbb{E}\left[x_{t,j+1} - \theta(x_{t,j})\right]\|^2 + \mathbb{E}\left[\|(x_{t,j+1} - \theta(x_{t,j})) - \mathbb{E}\left[x_{t,j+1} - \theta(x_{t,j})\right]\|^2\right] \\
&= \|y_{t,j} + \eta_2 \psi(x_{t,j}, y_{t,j}) - \theta(x_{t,j})\|^2 + \eta_2^2 \mathbb{E}\left[\left\|\psi_{i_j}(x_{t,j}, y_{t,j}) + \psi(\tilde{x}_t, \tilde{y}_t) - \psi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \psi(x_{t,j}, y_{t,j})\right\|^2\right].
\end{aligned} \tag{25}$$

For the first term in (25), we can directly apply (10) and get (assuming η_2 to be sufficiently small)

$$\|y_{t,j} + \eta_2 \psi(x_{t,j}, y_{t,j}) - \theta(x_{t,j})\| \le (1 - \alpha\eta_2) \|y_{t,j} - \theta(x_{t,j})\|. \tag{26}$$

The second term in (25) can be bounded in the same way as we did for the second term in (19):

$$\begin{aligned}
&\eta_2^2 \mathbb{E}\left[\left\|\psi_{i_j}(x_{t,j}, y_{t,j}) + \psi(\tilde{x}_t, \tilde{y}_t) - \psi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \psi(x_{t,j}, y_{t,j})\right\|^2\right] \\
\le\, &2\eta_2^2 \mathbb{E}\left[\left\|\psi_{i_j}(x_{t,j}, y_{t,j}) - \psi(x_{t,j}, y_{t,j}) - \psi_{i_j}(x^*, y^*) + \psi(x^*, y^*)\right\|^2\right] \\
&+ 2\eta_2^2 \mathbb{E}\left[\left\|\psi(\tilde{x}_t, \tilde{y}_t) - \psi_{i_j}(\tilde{x}_t, \tilde{y}_t) + \psi_{i_j}(x^*, y^*) - \psi(x^*, y^*)\right\|^2\right] \\
\le\, &2\eta_2^2 \mathbb{E}\left[\left\|\psi_{i_j}(x_{t,j}, y_{t,j}) - \psi_{i_j}(x^*, y^*)\right\|^2\right] + 2\eta_2^2 \mathbb{E}\left[\left\|\psi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \psi_{i_j}(x^*, y^*)\right\|^2\right] \\
\le\, &2\eta_2^2 \left(M\left(1 + \beta/\alpha\right) \|x_{t,j} - x^*\| + \beta \|y_{t,j} - \theta(x_{t,j})\|\right)^2 \\
&+ 2\eta_2^2 \left(M\left(1 + \beta/\alpha\right) \|\tilde{x}_t - x^*\| + \beta \|\tilde{y}_t - \theta(\tilde{x}_t)\|\right)^2,
\end{aligned} \tag{27}$$

where we have used

$$\begin{aligned}
\|\psi_i(x, y) - \psi_i(x^*, y^*)\| &= \|A_i x - \nabla g_i(y) - A_i x^* + \nabla g_i(y^*)\| \\
&\le \|A_i x - A_i x^*\| + \|\nabla g_i(y) - \nabla g_i(y^*)\| \\
&\le M\|x - x^*\| + \beta\|y - y^*\| \\
&\le M\|x - x^*\| + \beta\left(\|y - \theta(x)\| + \|\theta(x) - y^*\|\right) \\
&= M\|x - x^*\| + \beta\left(\|y - \theta(x)\| + \|\theta(x) - \theta(x^*)\|\right) \\
&\le M\|x - x^*\| + \beta\left(\|y - \theta(x)\| + \frac{M}{\alpha}\|x - x^*\|\right) \\
&= M\left(1 + \beta/\alpha\right)\|x - x^*\| + \beta\|y - \theta(x)\|.
\end{aligned}$$

Plugging (26) and (27) into (25) we get

$$\begin{aligned}
A &\le (1 - \alpha\eta_2)^2 \|y_{t,j} - \theta(x_{t,j})\|^2 + 2\eta_2^2 \left(M\left(1 + \beta/\alpha\right)\|x_{t,j} - x^*\| + \beta\|y_{t,j} - \theta(x_{t,j})\|\right)^2 \\
&\quad + 2\eta_2^2 \left(M\left(1 + \beta/\alpha\right)\|\tilde{x}_t - x^*\| + \beta\|\tilde{y}_t - \theta(\tilde{x}_t)\|\right)^2 \\
&\le \left(1 - \eta_2/c_5\right)\|y_{t,j} - \theta(x_{t,j})\|^2 + c_6\eta_2^2\|x_{t,j} - x^*\|^2 + c_7\eta_2^2\|\tilde{x}_t - x^*\|^2 + c_8\eta_2^2\|\tilde{y}_t - \theta(\tilde{x}_t)\|^2,
\end{aligned} \tag{28}$$

where $c_5, \ldots, c_8$ all have the form $\text{poly}\left(\beta, M, 1/\alpha\right)$. Here we assume η_1 is sufficiently small.

For B, we have

$$\begin{aligned}
B &= \mathbb{E}\|\theta(x_{t,j}) - \theta(x_{t,j+1})\|^2 \\
&\le \left(M/\alpha\right)^2 \mathbb{E}\|x_{t,j+1} - x_{t,j}\|^2 \\
&= \left(M/\alpha\right)^2 \eta_1^2 \mathbb{E}\left\|\varphi_{i_j}(x_{t,j}, y_{t,j}) + \varphi(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(\tilde{x}_t, \tilde{y}_t)\right\|^2 \\
&\le 2\left(M/\alpha\right)^2 \eta_1^2 \left(\mathbb{E}\left\|\varphi_{i_j}(x_{t,j}, y_{t,j}) + \varphi(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \varphi(x_{t,j}, y_{t,j})\right\|^2 + \mathbb{E}\|\varphi(x_{t,j}, y_{t,j})\|^2\right) \\
&= 2\left(M/\alpha\right)^2 \eta_1^2 \left(\mathbb{E}\left\|\varphi_{i_j}(x_{t,j}, y_{t,j}) + \varphi(\tilde{x}_t, \tilde{y}_t) - \varphi_{i_j}(\tilde{x}_t, \tilde{y}_t) - \varphi(x_{t,j}, y_{t,j})\right\|^2 + \mathbb{E}\|\varphi(x_{t,j}, y_{t,j}) - \varphi(x^*, y^*)\|^2\right).
\end{aligned}$$

Then using (22) and the smoothness of φ ((21) holds for φ as well) we obtain

$$\begin{aligned}
B \leq & 4\left(M/\alpha\right)^2 \eta_1^2 \left[\left(\rho + M^2/\alpha\right)\left\|x_{t,j} - x^*\right\| + M\left\|y_{t,j} - \theta(x_{t,j})\right\|\right]^2 \\
& + 4\left(M/\alpha\right)^2 \eta_1^2 \left[\left(\rho + M^2/\alpha\right)\left\|\tilde{x}_t - x^*\right\| + M\left\|\tilde{y}_t - \theta(\tilde{x}_t)\right\|\right]^2 \\
& + 2\left(M/\alpha\right)^2 \eta_1^2 \left[\left(\rho + M^2/\alpha\right)\left\|x_{t,j} - x^*\right\| + M\left\|y_{t,j} - \theta(x_{t,j})\right\|\right]^2 \\
= & 6\left(M/\alpha\right)^2 \eta_1^2 \left[\left(\rho + M^2/\alpha\right)\left\|x_{t,j} - x^*\right\| + M\left\|y_{t,j} - \theta(x_{t,j})\right\|\right]^2 \\
& + 4\left(M/\alpha\right)^2 \eta_1^2 \left[\left(\rho + M^2/\alpha\right)\left\|\tilde{x}_t - x^*\right\| + M\left\|\tilde{y}_t - \theta(\tilde{x}_t)\right\|\right]^2 \\
\leq & c_9\eta_1^2\left\|y_{t,j} - \theta(x_{t,j})\right\|^2 + c_{10}\eta_1^2\left\|x_{t,j} - x^*\right\|^2 + c_{11}\eta_1^2\left\|\tilde{x}_t - x^*\right\|^2 + c_{12}\eta_1^2\left\|\tilde{y}_t - \theta(\tilde{x}_t)\right\|^2,
\end{aligned} \tag{29}$$

where $c_9, \ldots, c_{12}$ all have the form $\operatorname{poly}(\beta, M, 1/\alpha)$.

Therefore, plugging (28) and (29) into (24), we get

$$\begin{aligned}
& \mathbb{E}\left[\left\|y_{t,j+1} - \theta(x_{t,j+1})\right\|^2\right] \\
\leq & A + B + 2\sqrt{AB} \\
\leq & A + B + \eta_1 A + \frac{B}{\eta_1} \\
\leq & (1+\eta_1)\left(\left(1 - \eta_2/c_5\right)\left\|y_{t,j} - \theta(x_{t,j})\right\|^2 + c_6\eta_2^2\left\|x_{t,j} - x^*\right\|^2 + c_7\eta_2^2\left\|\tilde{x}_t - x^*\right\|^2 + c_8\eta_2^2\left\|\tilde{y}_t - \theta(\tilde{x}_t)\right\|^2\right) \\
& + (1 + 1/\eta_1)\left(c_9\eta_1^2\left\|y_{t,j} - \theta(x_{t,j})\right\|^2 + c_{10}\eta_1^2\left\|x_{t,j} - x^*\right\|^2 + c_{11}\eta_1^2\left\|\tilde{x}_t - x^*\right\|^2 + c_{12}\eta_1^2\left\|\tilde{y}_t - \theta(\tilde{x}_t)\right\|^2\right) \\
\leq & \left(1 - \eta_2/c_{13}\right)\left\|y_{t,j} - \theta(x_{t,j})\right\|^2 + c_{14}\eta_2^2\left\|x_{t,j} - x^*\right\|^2 + c_{15}\eta_2^2\left\|\tilde{x}_t - x^*\right\|^2 + c_{16}\eta_2^2\left\|\tilde{y}_t - \theta(\tilde{x}_t)\right\|^2,
\end{aligned} \tag{30}$$

where $c_{13}, \ldots, c_{16}$ all have the form $\operatorname{poly}(\beta, M, 1/\alpha)$. Here we assume η_1 is chosen sufficiently small *given* η_2.

Step3: Putting Things Together. Let $p_t := \mathbb{E}\left\|\tilde{x}_t - x^*\right\|^2$ and $q_t := \mathbb{E}\left\|\tilde{y}_t - \theta(\tilde{x}_t)\right\|^2$. Taking expectation with respect to everything we were conditioned on, we can write (23) as

$$\mathbb{E}\|x_{t,j+1} - x^*\|^2 \leq \left(1 - \eta_1/c_1\right)\mathbb{E}\left\|x_{t,j} - x^*\right\|^2 + c_2\eta_1\mathbb{E}\left\|y_{t,j} - \theta(x_{t,j})\right\|^2 + c_3\eta_1^2 p_t + c_4\eta_1^2 q_t.$$

Taking sum over $j = 0, 1, \ldots, N-1$, and noticing that $\tilde{x}_t = x_{t,0}$ and that $\tilde{x}_{t+1} = x_{t,j_t}$ for a random $j_t \in \{0, 1, \ldots, N-1\}$, we obtain

$$\frac{1}{N}\left(\mathbb{E}\|x_{t,N} - x^*\|^2 - \mathbb{E}\|\tilde{x}_t - x^*\|^2\right) \leq -\frac{\eta_1}{c_1}\mathbb{E}\left\|\tilde{x}_{t+1} - x^*\right\|^2 + c_2\eta_1\mathbb{E}\left\|\tilde{y}_{t+1} - \theta(x_{t+1})\right\|^2 + c_3\eta_1^2 p_t + c_4\eta_1^2 q_t,$$

which implies

$$-\frac{1}{N}p_t \leq -\frac{\eta_1}{c_1}p_{t+1} + c_2\eta_1 q_{t+1} + c_3\eta_1^2 p_t + c_4\eta_1^2 q_t,$$

i.e.

$$p_{t+1} \leq \left(\frac{c_1}{\eta_1 N} + c_1 c_3 \eta_1\right)p_t + c_1 c_2 q_{t+1} + c_1 c_4 \eta_1 q_t.$$

Similarly, from (30) we can get

$$q_{t+1} \leq \left(\frac{c_{13}}{\eta_2 N} + c_{13} c_{16} \eta_2\right)q_t + c_{13} c_{14} \eta_2 p_{t+1} + c_{13} c_{15} \eta_2 p_t.$$

Then it is easy to see that one can choose μ and N to be sufficiently large $\operatorname{poly}(\beta, \rho, M, 1/\alpha, 1/\sigma_{\min})$ and choose η_1 and η_2 to be sufficiently small $\operatorname{poly}(\beta, \rho, M, 1/\alpha, 1/\sigma_{\min})^{-1}$ such that

$$p_{t+1} + \mu q_{t+1} \leq \frac{1}{2}\left(p_t + \mu q_t\right).$$

This completes the proof of Theorem 4.1. □

B Linear Convergence of the Primal-Dual Gradient Method When Both f and g are Smooth and Strongly Convex

In this section we show that if both f and g are smooth and strongly convex, Algorithm 1 can achieve linear convergence for Problem (1). Note that this proof is much simpler than that of Theorem 3.1.

We denote $\sigma_{\max} := \sigma_{\max}(A)$.

Theorem B.1. *Suppose f is β_1-smooth and α_1-strongly convex, and g is β_2-smooth and α_2-strongly convex. If we choose $\eta_1 = \min\left\{\frac{1}{\alpha_1+\beta_1}, \frac{\alpha_2}{4\sigma_{\max}^2}\right\}$ and $\eta_2 = \min\left\{\frac{1}{\alpha_2+\beta_2}, \frac{\alpha_1}{4\sigma_{\max}^2}\right\}$ in Algorithm 1 and let $R_t = \eta_2 \left\|x_t - x^*\right\|^2 + \eta_1 \left\|y_t - y^*\right\|^2$, then we have*

$$R_{t+1} \le \left(1 - \frac{1}{2}\min\left\{\frac{\alpha_1}{\alpha_1+\beta_1}, \frac{\alpha_2}{\alpha_2+\beta_2}, \frac{\alpha_1\alpha_2}{4\sigma_{\max}^2}\right\}\right) R_t.$$

Proof. From the update rule $x_{t+1} = x_t - \eta_1 \nabla_x L(x_t, y_t)$ we have

$$\left\|x_{t+1} - x^*\right\|^2 = \left\|x_t - x^*\right\|^2 - 2\eta_1 \left\langle x_t - x^*, \nabla_x L(x_t, y_t)\right\rangle + \eta_1^2 \left\|\nabla_x L(x_t, y_t)\right\|^2. \tag{31}$$

The inner product term in (31) can be bounded as

$$\begin{aligned}
\langle x_t - x^*, \nabla_x L(x_t, y_t)\rangle &= \langle x_t - x^*, \nabla_x L(x_t, y_t) - \nabla_x L(x^*, y^*)\rangle \\
&= \langle x_t - x^*, \nabla f(x_t) - \nabla f(x^*)\rangle + \langle x_t - x^*, A^\top (y_t - y^*)\rangle \\
&\ge \frac{\alpha_1\beta_1}{\alpha_1+\beta_1}\left\|x_t - x^*\right\|^2 + \frac{1}{\alpha_1+\beta_1}\left\|\nabla f(x_t) - \nabla f(x^*)\right\|^2 + \langle x_t - x^*, A^\top (y_t - y^*)\rangle,
\end{aligned}$$

where we have used Lemma 3.11 in [Bubeck, 2015].

The third term in (31) can be bounded as

$$\begin{aligned}
\left\|\nabla_x L(x_t, y_t)\right\|^2 &= \left\|\nabla_x L(x_t, y_t) - \nabla_x L(x^*, y^*)\right\|^2 \\
&= \left\|\nabla f(x_t) - \nabla f(x^*) + A^\top (y_t - y^*)\right\|^2 \\
&\le 2\left(\left\|\nabla f(x_t) - \nabla f(x^*)\right\|^2 + \left\|A^\top (y_t - y^*)\right\|^2\right) \\
&\le 2\left(\left\|\nabla f(x_t) - \nabla f(x^*)\right\|^2 + \sigma_{\max}^2 \left\|y_t - y^*\right\|^2\right).
\end{aligned}$$

By symmetry, for the dual variable, we have an inequality similar to (31). Combining everything together, and using $\eta_1 \le \frac{1}{\alpha_1+\beta_1}$ and $\eta_2 \le \frac{1}{\alpha_2+\beta_2}$, with some routine calculations we can get

$$\begin{aligned}
&\eta_2 \left\|x_{t+1} - x^*\right\|^2 + \eta_1 \left\|y_{t+1} - y^*\right\|^2 \\
&\le \left(1 - 2\alpha_1\eta_1 + 2\alpha_1^2\eta_1^2 + 2\sigma_{\max}^2\eta_1\eta_2\right)\eta_2\left\|x_t - x^*\right\|^2 + \left(1 - 2\alpha_2\eta_2 + 2\alpha_2^2\eta_2^2 + 2\sigma_{\max}^2\eta_1\eta_2\right)\eta_1\left\|y_t - y^*\right\|^2.
\end{aligned}$$

Then, from our choices of η_1 and η_2, the above inequality implies

$$\begin{aligned}
R_{t+1} &\le \left(1 - 2\alpha_1\eta_1 + \alpha_1\eta_1 + \frac{1}{2}\alpha_1\eta_1\right)\eta_2\left\|x_t - x^*\right\|^2 + \left(1 - 2\alpha_2\eta_2 + \alpha_2\eta_2 + \frac{1}{2}\alpha_2\eta_2\right)\eta_1\left\|y_t - y^*\right\|^2 \\
&\le \left(1 - \frac{1}{2}\min\{\alpha_1\eta_1, \alpha_2\eta_2\}\right) R_t \\
&= \left(1 - \frac{1}{2}\min\left\{\frac{\alpha_1}{\alpha_1+\beta_1}, \frac{\alpha_2}{\alpha_2+\beta_2}, \frac{\alpha_1\alpha_2}{4\sigma_{\max}^2}\right\}\right) R_t.
\end{aligned}$$

□