Exploiting Strong Convexity from Data with Primal-Dual First-Order Algorithms

Jialei Wang [1] **Lin Xiao** [2]

Abstract

We consider empirical risk minimization of linear predictors with convex loss functions. Such problems can be reformulated as convex-concave saddle point problems, and thus are well suitable for primal-dual first-order algorithms. However, primal-dual algorithms often require explicit strongly convex regularization in order to obtain fast linear convergence, and the required dual proximal mapping may not admit closed-form or efficient solution. In this paper, we develop both batch and randomized primal-dual algorithms that can exploit strong convexity from data adaptively and are capable of achieving linear convergence even without regularization. We also present dual-free variants of the adaptive primal-dual algorithms that do not require computing the dual proximal mapping, which are especially suitable for logistic regression.

1. Introduction

We consider the problem of regularized empirical risk minimization (ERM) of linear predictors. Let $a_1, \ldots, a_n \in \mathbb{R}^d$ be the feature vectors of n data samples, $\phi_i : \mathbb{R} \to \mathbb{R}$ be a convex loss function associated with the linear prediction $a_i^T x$, for $i = 1, \ldots, n$, and $g : \mathbb{R}^d \to \mathbb{R}$ be a convex regularization function for the predictor $x \in \mathbb{R}^d$. ERM amounts to solving the following convex optimization problem:

$$\min_{x \in \mathbb{R}^d} \quad \left\{ P(x) \stackrel{\text{def}}{=} \tfrac{1}{n} \sum_{i=1}^n \phi_i(a_i^T x) + g(x) \right\}. \quad (1)$$

Examples of the above formulation include many well-known classification and regression problems. For binary classification, each feature vector a_i is associated with a label $b_i \in \{\pm 1\}$. In particular logistic regression is obtained by setting $\phi_i(z) = \log(1 + \exp(-b_i z))$. For linear regression problems, each feature vector a_i is associated with a

---
[1]Department of Computer Science, The University of Chicago, Chicago, Illinois 60637, USA. [2]Microsoft Research, Redmond, Washington 98052, USA. Correspondence to: Jialei Wang <jialei@uchicago.edu>, Lin Xiao <lin.xiao@microsoft.com>.

dependent variable $b_i \in \mathbb{R}$, and $\phi_i(z) = (1/2)(z - b_i)^2$. Then we get ridge regression with $g(x) = (\lambda/2)\|x\|_2^2$, and elastic net with $g(x) = \lambda_1\|x\|_1 + (\lambda_2/2)\|x\|_2^2$.

Let $A = [a_1, \ldots, a_n]^T$ be the data matrix. Throughout this paper, we make the following assumptions:

Assumption 1. *The functions ϕ_i, g and matrix A satisfy:*

- *Each ϕ_i is δ-strongly convex and $1/\gamma$-smooth where $\gamma > 0$ and $\delta \geq 0$, and $\gamma\delta \leq 1$;*

- *g is λ-strongly convex where $\lambda \geq 0$;*

- *$\lambda + \delta\mu^2 > 0$, where $\mu = \sqrt{\lambda_{\min}(A^T A)}$.*

The strong convexity and smoothness mentioned above are with respect to the standard Euclidean norm, denoted as $\|x\| = \sqrt{x^T x}$. (See, e.g., Nesterov (2004, Sections 2.1.1 and 2.1.3) for the exact definitions.) Let $R = \max_i\{\|a_i\|\}$ and assuming $\lambda > 0$, then $R^2/(\gamma\lambda)$ is a popular definition of condition number for analyzing complexities of different algorithms. The last condition above means that the primal objective function $P(x)$ is strongly convex, even if $\lambda = 0$.

There have been extensive research activities in recent years on developing efficiently algorithms for solving problem (1). A broad class of randomized algorithms that exploit the finite sum structure in the ERM problem have emerged as very competitive both in terms of theoretical complexity and practical performance. They can be put into three categories: primal, dual, and primal-dual.

Primal randomized algorithms work with the ERM problem (1) directly. They are modern versions of randomized incremental gradient methods (e.g., Bertsekas, 2012; Nedic & Bertsekas, 2001) equipped with variance reduction techniques. Each iteration of such algorithms only process one data point a_i with complexity $O(d)$. They includes SAG (Roux et al., 2012), SAGA (Defazio et al., 2014), and SVRG (Johnson & Zhang, 2013; Xiao & Zhang, 2014), which all achieve the iteration complexity $O\left((n + R^2/(\gamma\lambda)) \log(1/\epsilon)\right)$ to find an ϵ-optimal solution. In fact, they are capable of exploiting the strong convexity from data, meaning that the condition number $R^2/(\gamma\lambda)$ in the complexity can be replaced by the more favorable one $R^2/(\gamma(\lambda + \delta\mu^2/n))$. This improvement can be achieved without explicit knowledge of μ from data.

Dual algorithms solve Fenchel dual of (1) by maximizing

$$D(y) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} -\phi_i^*(y_i) - g^* \left(-\frac{1}{n} \sum_{i=1}^{n} y_i a_i \right) \quad (2)$$

using randomized coordinate ascent algorithms. (Here ϕ_i^* and g^* denotes the conjugate functions of ϕ_i and g.) They include SDCA (Shalev-Shwartz & Zhang, 2013), Nesterov (2012) and Richtárik & Takáč (2014). They have the same complexity $O\left((n + R^2/(\gamma\lambda)) \log(1/\epsilon)\right)$, but are hard to exploit strong convexity from data.

Primal-dual algorithms solve the convex-concave saddle point problem $\min_x \max_y \mathcal{L}(x, y)$ where

$$\mathcal{L}(x, y) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \left(y_i \langle a_i, x \rangle - \phi_i^*(y_i) \right) + g(x). \quad (3)$$

In particular, SPDC (Zhang & Xiao, 2015) achieves an accelerated linear convergence rate with iteration complexity $O\left((n + \sqrt{n}R/\sqrt{\gamma\lambda}) \log(1/\epsilon)\right)$, which is better than the aforementioned non-accelerated complexity when $R^2/(\gamma\lambda) > n$. Lan & Zhou (2015) developed dual-free variants of accelerated primal-dual algorithms, but without considering the linear predictor structure in ERM. Balamurugan & Bach (2016) extended SVRG and SAGA to solving saddle point problems.

Accelerated primal and dual randomized algorithms have also been developed. Nesterov (2012), Fercoq & Richtárik (2015) and Lin et al. (2015b) developed accelerated coordinate gradient algorithms, which can be applied to solve the dual problem (2). Allen-Zhu (2016) developed an accelerated variant of SVRG. Acceleration can also be obtained using the Catalyst framework (Lin et al., 2015a). They all achieve the same $O\left((n + \sqrt{n}R/\sqrt{\gamma\lambda}) \log(1/\epsilon)\right)$ complexity. A common feature of accelerated algorithms is that they require good estimate of the strong convexity parameter. This makes hard for them to exploit strong convexity from data because the minimum singular value μ of the data matrix A is very hard to estimate in general.

In this paper, we show that primal-dual algorithms are capable of exploiting strong convexity from data if the algorithm parameters (such as step sizes) are set appropriately. While these optimal setting depends on the knowledge of the convexity parameter μ from the data, we develop adaptive variants of primal-dual algorithms that can tune the parameter automatically. Such adaptive schemes rely critically on the capability of evaluating the primal-dual optimality gaps by primal-dual algorithms.

A major disadvantage of primal-dual algorithms is that the required dual proximal mapping may not admit closed-form or efficient solution. We follow the approach of Lan & Zhou (2015) to derive dual-free variants of the primal-dual algorithms customized for ERM problems with the linear predictor structure, and show that they can also exploit strong convexity from data with correct choices of parameters or using an adaptation scheme.

Algorithm 1 Batch Primal-Dual (BPD) Algorithm

input: parameters τ, σ, θ, initial point $(\tilde{x}^{(0)} = x^{(0)}, y^{(0)})$

  for $t = 0, 1, 2, \ldots$ **do**
$$y^{(t+1)} = \text{prox}_{\sigma f^*} \left(y^{(t)} + \sigma A \tilde{x}^{(t)} \right)$$
$$x^{(t+1)} = \text{prox}_{\tau g} \left(x^{(t)} - \tau A^T y^{(t+1)} \right)$$
$$\tilde{x}^{(t+1)} = x^{(t+1)} + \theta(x^{(t+1)} - x^{(t)})$$
  end for

---

2. Batch primal-dual algorithms

Before diving into randomized primal-dual algorithms, we first consider batch primal-dual algorithms, which exhibit similar properties as their randomized variants. To this end, we consider a "batch" version of the ERM problem (1),

$$\min_{x \in \mathbb{R}^d} \left\{ P(x) \stackrel{\text{def}}{=} f(Ax) + g(x) \right\}. \quad (4)$$

where $A \in \mathbb{R}^{n \times d}$, and make the following assumption:

Assumption 2. *The functions f, g and matrix A satisfy:*

- *f is δ-strongly convex and $1/\gamma$-smooth where $\gamma > 0$ and $\delta \geq 0$, and $\gamma\delta \leq 1$;*
- *g is λ-strongly convex where $\lambda \geq 0$;*
- *$\lambda + \delta\mu^2 > 0$, where $\mu = \sqrt{\lambda_{\min}(A^T A)}$.*

For exact correspondence with problem (1), we have $f(z) = \frac{1}{n} \sum_{i=1}^{n} \phi_i(z_i)$ with $z_i = a_i^T x$. Under Assumption 1, the function $f(z)$ is δ/n-strongly convex and $1/(n\gamma)$-smooth, and $f(Ax)$ is $\delta\mu^2/n$-strongly convex and R^2/γ-smooth. However, such correspondences alone are not sufficient to exploit the structure of (1), i.e., substituting them into the batch algorithms of this section will not produce the efficient algorithms for solving problem (1) that we will present in Sections 3 and 4.2. So we do not make such correspondences explicit in this section. Rather, treat them as independent assumptions with the same notation.

Using conjugate functions, we can derive the dual of (4) as

$$\max_{y \in \mathbb{R}^n} \left\{ D(y) \stackrel{\text{def}}{=} -f^*(y) - g^*(-A^T y) \right\}, \quad (5)$$

and the convex-concave saddle point formulation is

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^n} \left\{ \mathcal{L}(x, y) \stackrel{\text{def}}{=} g(x) + y^T Ax - f^*(y) \right\}. \quad (6)$$

We consider the primal-dual first-order algorithm proposed by Chambolle & Pock (2011; 2016) for solving the saddle point problem (6), which is given as Algorithm 1. Here we call it the batch primal-dual (BPD) algorithm. Assuming that f is smooth and g is strongly convex, Chambolle & Pock (2011; 2016) showed that Algorithm 1 achieves accelerated linear convergence rate if $\lambda > 0$. However, they did not consider the case where additional or the sole source of strong convexity comes from $f(Ax)$.

In the following theorem, we show how to set the parameters τ, σ and θ to exploit both sources of strong convexity to achieve fast linear convergence.

Theorem 1. *Suppose Assumption 2 holds and $(x^\star, y^\star)$ is the unique saddle point of $\mathcal{L}$ defined in (6). Let $L = \|A\| = \sqrt{\lambda_{\max}(A^T A)}$. If we set the parameters in Algorithm 1 as*

$$\sigma = \tfrac{1}{L}\sqrt{\tfrac{\lambda+\delta\mu^2}{\gamma}}, \quad \tau = \tfrac{1}{L}\sqrt{\tfrac{\gamma}{\lambda+\delta\mu^2}}, \quad (7)$$

and $\theta = \max\{\theta_x, \theta_y\}$ where

$$\theta_x = \left(1 - \tfrac{\delta}{(\delta+2\sigma)}\tfrac{\mu^2}{L^2}\right)\tfrac{1}{1+\tau\lambda}, \quad \theta_y = \tfrac{1}{1+\sigma\gamma/2}, \quad (8)$$

then we have

$$\left(\tfrac{1}{2\tau} + \tfrac{\lambda}{2}\right)\|x^{(t)} - x^\star\|^2 + \tfrac{\gamma}{4}\|y^{(t)} - y^\star\|^2 \le \theta^t C,$$
$$\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^{(t)}) \le \theta^t C,$$

where $C = \left(\tfrac{1}{2\tau} + \tfrac{\lambda}{2}\right)\|x^{(0)} - x^\star\|^2 + \left(\tfrac{1}{2\sigma} + \tfrac{\gamma}{4}\right)\|y^{(0)} - y^\star\|^2$.

The proof of Theorem 1 is given in Appendices B and C. Here we give a detailed analysis of the convergence rate. Substituting σ and τ in (7) into the expressions for θ_y and θ_x in (8), and assuming $\gamma(\lambda + \delta\mu^2) \ll L^2$, we have

$$\theta_x \approx 1 - \tfrac{\gamma\delta\mu^2}{L^2}\left(2\tfrac{\sqrt{\gamma(\lambda+\delta\mu^2)}}{L} + \gamma\delta\right)^{-1} - \tfrac{\lambda}{L}\sqrt{\tfrac{\gamma}{\lambda+\delta\mu^2}},$$

$$\theta_y = \tfrac{1}{1+\sqrt{\gamma(\lambda+\delta\mu^2)}/(2L)} \approx 1 - \tfrac{\sqrt{\gamma(\lambda+\delta\mu^2)}}{2L}.$$

Since the overall condition number of the problem is $\frac{L^2}{\gamma(\lambda+\delta\mu^2)}$, it is clear that θ_y is an accelerated convergence rate. Next we examine θ_x in two special cases.

The case of $\delta\mu^2 = 0$ but $\lambda > 0$. In this case, we have $\tau = \tfrac{1}{L}\sqrt{\tfrac{\gamma}{\lambda}}$ and $\sigma = \tfrac{1}{L}\sqrt{\tfrac{\lambda}{\gamma}}$, and thus

$$\theta_x = \tfrac{1}{1+\sqrt{\gamma\lambda}/L} \approx 1 - \tfrac{\sqrt{\gamma\lambda}}{L}, \quad \theta_y = \tfrac{1}{1+\sqrt{\gamma\lambda}/(2L)} \approx 1 - \tfrac{\sqrt{\gamma\lambda}}{2L}.$$

Therefore we have $\theta = \max\{\theta_x, \theta_y\} \approx 1 - \tfrac{\sqrt{\lambda\gamma}}{2L}$. This indeed is an accelerated convergence rate, recovering the result of Chambolle & Pock (2011; 2016).

The case of $\lambda = 0$ but $\delta\mu^2 > 0$. In this case, we have $\tau = \tfrac{1}{L\mu}\sqrt{\tfrac{\gamma}{\delta}}$ and $\sigma = \tfrac{\mu}{L}\sqrt{\tfrac{\delta}{\gamma}}$, and

$$\theta_x = 1 - \tfrac{\gamma\delta\mu^2}{L^2} \cdot \tfrac{1}{2\sqrt{\gamma\delta}\mu/L + \gamma\delta}, \quad \theta_y \approx 1 - \tfrac{\sqrt{\gamma\delta}\mu}{2L}.$$

Notice that $\frac{1}{\gamma\delta}\frac{L^2}{\mu^2}$ is the condition number of $f(Ax)$. Next we assume $\mu \ll L$ and examine how θ_x varies with $\gamma\delta$.

- If $\gamma\delta \approx \tfrac{\mu^2}{L^2}$, meaning f is badly conditioned, then

$$\theta_x \approx 1 - \tfrac{\gamma\delta\mu^2}{L^2} \cdot \tfrac{1}{3\sqrt{\gamma\delta}\mu/L} = 1 - \tfrac{\sqrt{\gamma\delta}\mu}{3L}.$$

  Because the overall condition number is $\frac{1}{\gamma\delta}\frac{L^2}{\mu^2}$, this is an accelerated linear rate, and so is $\theta = \max\{\theta_x, \theta_y\}$.

Algorithm 2 Adaptive Batch Primal-Dual (Ada-BPD)
input: problem constants λ, γ, δ, L and $\hat{\mu} > 0$, initial point $(x^{(0)}, y^{(0)})$, and adaptation period T.
Compute σ, τ, and θ as in (7) and (8) using $\mu = \hat{\mu}$
for $t = 0, 1, 2, \dots$ **do**
$\quad y^{(t+1)} = \text{prox}_{\sigma f^*}\left(y^{(t)} + \sigma A\tilde{x}^{(t)}\right)$
$\quad x^{(t+1)} = \text{prox}_{\tau g}\left(x^{(t)} - \tau A^T y^{(t+1)}\right)$
$\quad \tilde{x}^{(t+1)} = x^{(t+1)} + \theta(x^{(t+1)} - x^{(t)})$
$\quad$ **if** $\text{mod}(t+1, T) == 0$ **then**
$\quad\quad (\sigma, \tau, \theta) = \text{BPD-Adapt}\left(\{P^{(s)}, D^{(s)}\}_{s=t-T}^{t+1}\right)$
$\quad$ **end if**
end for

- If $\gamma\delta \approx \tfrac{\mu}{L}$, meaning f is mildly conditioned, then

$$\theta_x \approx 1 - \tfrac{\mu^3}{L^3}\tfrac{1}{2(\mu/L)^{3/2}+\mu/L} \approx 1 - \tfrac{\mu^2}{L^2}.$$

  This represents a half-accelerated rate, because the overall condition number is $\frac{1}{\gamma\delta}\frac{L^2}{\mu^2} \approx \frac{L^3}{\mu^3}$.

- If $\gamma\delta = 1$, i.e., f is a simple quadratic function, then

$$\theta_x \approx 1 - \tfrac{\mu^2}{L^2}\tfrac{1}{2\mu/L+1} \approx 1 - \tfrac{\mu^2}{L^2}.$$

  This rate does not have acceleration, because the overall condition number is $\frac{1}{\gamma\delta}\frac{L^2}{\mu^2} \approx \frac{L^2}{\mu^2}$.

In summary, the extent of acceleration in the dominating factor θ_x (which determines θ) depends on the relative size of $\gamma\delta$ and μ^2/L^2, i.e., the relative conditioning between the function f and the matrix A. In general, we have full acceleration if $\gamma\delta \le \mu^2/L^2$. The theory predicts that the acceleration degrades as the function f gets better conditioned. However, in our numerical experiments, we often observe acceleration even if $\gamma\delta$ gets closer to 1.

As explained in Chambolle & Pock (2011), Algorithm 1 is equivalent to a preconditioned ADMM. Deng & Yin (2016) characterized conditions for ADMM to obtain linear convergence without assuming both parts of the objective function being strongly convex, but they did not derive convergence rate for this case.

2.1. Adaptive batch primal-dual algorithms

In practice, it is often very hard to obtain good estimate of the problem-dependent constants, especially $\mu = \sqrt{\lambda_{\min}(A^T A)}$, in order to apply the algorithmic parameters specified in Theorem 1. Here we explore heuristics that can enable adaptive tuning of such parameters, which often lead to much improved performance in practice.

A key observation is that the convergence rate of the BPD algorithm changes monotonically with the overall strong convexity parameter $\lambda + \delta\mu^2$, regardless the extent of

---

Algorithm 3 BPD-Adapt (simple heuristic)

---

input: previous estimate $\hat{\mu}$, adaption period T, primal and dual objective values $\{P^{(s)}, D^{(s)}\}_{s=t-T}^{t}$

    if $P^{(t)} - D^{(t)} < \theta^T (P^{(t-T)} - D^{(t-T)})$ **then**
        $\hat{\mu} := \sqrt{2}\hat{\mu}$
    else
        $\hat{\mu} := \hat{\mu}/\sqrt{2}$
    end if
    Compute σ, τ, and θ as in (7) and (8) using $\mu = \hat{\mu}$

output: new parameters (σ, τ, θ)

---

acceleration. In other words, the larger $\lambda + \delta\mu^2$ is, the faster the convergence. Therefore, if we can monitor the progress of the convergence and compare it with the predicted convergence rate in Theorem 1, then we can adjust the algorithmic parameters to exploit the fastest possible convergence. More specifically, if the observed convergence is slower than the predicted convergence rate, then we should reduce the estimate of μ; if the observed convergence is better than the predicted rate, then we can try to increase μ for even faster convergence.

We formalize the above reasoning in an Adaptive BPD (Ada-BPD) algorithm described in Algorithm 2. This algorithm maintains an estimate $\hat{\mu}$ of the true constant μ, and adjust it every T iterations. We use $P^{(t)}$ and $D^{(t)}$ to represent the primal and dual objective values at $P(x^{(t)})$ and $D(y^{(t)})$, respectively. We give two implementations of the tuning procedure BPD-Adapt:

- Algorithm 3 is a simple heuristic for tuning the estimate $\hat{\mu}$, where the increasing and decreasing factor $\sqrt{2}$ can be changed to other values larger than 1;

- Algorithm 4 is a more robust heuristic. It does not rely on the specific convergence rate θ established in Theorem 1. Instead, it simply compares the current estimate of objective reduction rate $\hat{\rho}$ with the previous estimate ρ ($\approx \theta^T$). It also specifies a non-tuning range of changes in ρ, specified by the interval $[\underline{c}, \overline{c}]$.

One can also devise more sophisticated schemes; e.g., if we estimate that $\delta\mu^2 < \lambda$, then no more tuning is necessary.

The capability of accessing both the primal and dual objective values allows primal-dual algorithms to have good estimate of the convergence rate, which enables effective tuning heuristics. Automatic tuning of primal-dual algorithms have also been studied by, e.g., Malitsky & Pock (2016) and Goldstein et al. (2013), but with different goals.

Finally, we note that Theorem 1 only establishes convergence rate for the distance to the optimal point and the quantity $\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^{(t)})$, which is not quite the duality gap $P(x^{(t)}) - D(y^{(t)})$. Nevertheless, same convergence rate can also be established for the duality gap (see

---

Algorithm 4 BPD-Adapt (robust heuristic)

---

input: previous rate estimate $\rho > 0$, $\Delta = \delta\hat{\mu}^2$, period T, constants $\underline{c} < 1$ and $\overline{c} > 1$, and $\{P^{(s)}, D^{(s)}\}_{s=t-T}^{t}$

    Compute new rate estimate $\hat{\rho} = \frac{P^{(t)} - D^{(t)}}{P^{(t-T)} - D^{(t-T)}}$
    if $\hat{\rho} \leq \underline{c}\,\rho$ **then**
        $\Delta := 2\Delta, \qquad \rho := \hat{\rho}$
    else if $\hat{\rho} \geq \overline{c}\,\rho$ **then**
        $\Delta := \Delta/2, \qquad \rho := \hat{\rho}$
    else
        $\Delta := \Delta$
    end if
    $\sigma = \frac{1}{L}\sqrt{\frac{\lambda + \Delta}{\gamma}}, \quad \tau = \frac{1}{L}\sqrt{\frac{\gamma}{\lambda + \Delta}}$
    Compute θ using (8) or set $\theta = 1$

output: new parameters (σ, τ, θ)

Zhang & Xiao, 2015, Section 2.2), which can be used to better justify the adaption procedure.

3. Randomized primal-dual algorithm

In this section, we come back to the ERM problem (1), which have a finite sum structure that allows the development of randomized primal-dual algorithms. In particular, we extend the stochastic primal-dual coordinate (SPDC) algorithm (Zhang & Xiao, 2015) to exploit the strong convexity from data in order to achieve faster convergence rate.

First, we show that, by setting algorithmic parameters appropriately, the original SPDC algorithm may directly benefit from strong convexity from the loss function. We note that the SPDC algorithm is a special case of the Adaptive SPDC (Ada-SPDC) algorithm presented in Algorithm 5, by setting the adaption period $T = \infty$ (not performing any adaption). The following theorem is proved in Appendix E.

Theorem 2. *Suppose Assumption 1 holds. Let $(x^\star, y^\star)$ be the saddle point of the function $\mathcal{L}$ defined in (3), and $R = \max\{\|a_1\|, \ldots, \|a_n\|\}$. If we set $T = \infty$ in Algorithm 5 (no adaption) and let*

$$\tau = \frac{1}{4R}\sqrt{\frac{\gamma}{n\lambda + \delta\mu^2}}, \quad \sigma = \frac{1}{4R}\sqrt{\frac{n\lambda + \delta\mu^2}{\gamma}}, \qquad (9)$$

and $\theta = \max\{\theta_x, \theta_y\}$ where

$$\theta_x = \left(1 - \frac{\tau\sigma\delta\mu^2}{2n(\sigma + 4\delta)}\right)\frac{1}{1 + \tau\lambda}, \quad \theta_y = \frac{1 + ((n-1)/n)\sigma\gamma/2}{1 + \sigma\gamma/2}, \quad (10)$$

then we have

$$\left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\mathbb{E}\big[\|x^{(t)} - x^\star\|^2\big] + \frac{\gamma}{4}\mathbb{E}\big[\|y^{(t)} - y^\star\|^2\big] \leq \theta^t C,$$
$$\mathbb{E}\big[\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^{(t)})\big] \leq \theta^t C,$$

where $C = \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\|x^{(0)} - x^\star\|^2 + \left(\frac{1}{2\sigma} + \frac{\gamma}{4}\right)\|y^{(0)} - y^\star\|^2$. The expectation $\mathbb{E}[\cdot]$ is taken with respect to the history of random indices drawn at each iteration.

Algorithm 5 Adaptive SPDC (Ada-SPDC)

---

input: parameters $\sigma, \tau, \theta > 0$, initial point $(x^{(0)}, y^{(0)})$, and adaptation period T.

Set $\tilde{x}^{(0)} = x^{(0)}$

for $t = 0, 1, 2, \ldots$ **do**

    pick $k \in \{1, \ldots, n\}$ uniformly at random

    for $i \in \{1, \ldots, n\}$ **do**

        if $i == k$ **then**

$$y_k^{(t+1)} = \text{prox}_{\sigma \phi_k^*}\left(y_k^{(t)} + \sigma a_k^T \tilde{x}^{(t)}\right)$$

        else

$$y_i^{(t+1)} = y_i^{(t)}$$

        end if

    end for

$$x^{(t+1)} = \text{prox}_{\tau g}\left(x^{(t)} - \tau\left(u^{(t)} + (y_k^{(t+1)} - y_k^{(t)})a_k\right)\right)$$

$$u^{(t+1)} = u^{(t)} + \tfrac{1}{n}(y_k^{(t+1)} - y_k^{(t)})a_k$$

$$\tilde{x}^{(t+1)} = x^{(t+1)} + \theta(x^{(t+1)} - x^{(t)})$$

    if $\mod(t+1, T \cdot n) = 0$ **then**

$$(\tau, \sigma, \theta) = \text{SPDC-Adapt}\left(\{P^{(t-sn)}, D^{(t-sn)}\}_{s=0}^T\right)$$

    end if

end for

Below we give a detailed discussion on the expected convergence rate established in Theorem 2.

The cases of $\sigma \mu^2 = 0$ but $\lambda > 0$. In this case we have $\tau = \frac{1}{4R}\sqrt{\frac{\gamma}{n\lambda}}$ and $\sigma = \frac{1}{4R}\sqrt{\frac{n\lambda}{\gamma}}$, and

$$\theta_x = \frac{1}{1 + \tau\lambda} = 1 - \frac{1}{1 + 4R\sqrt{n/(\lambda\gamma)}},$$

$$\theta_y = \frac{1 + ((n-1)/n)\sigma\gamma/2}{1 + \sigma\gamma/2} = 1 - \frac{1}{n + 8R\sqrt{n/(\lambda\gamma)}}.$$

Hence $\theta = \theta_y$. These recover the parameters and convergence rate of the standard SPDC (Zhang & Xiao, 2015).

The cases of $\sigma \mu^2 > 0$ but $\lambda = 0$. In this case we have $\tau = \frac{1}{4R\mu}\sqrt{\frac{\gamma}{\delta}}$ and $\sigma = \frac{\mu}{4R}\sqrt{\frac{\delta}{\gamma}}$, and

$$\theta_x = 1 - \frac{\tau\sigma\delta\mu^2}{2n(\sigma + 4\delta)} = 1 - \frac{\gamma\delta\mu^2}{32nR^2} \cdot \frac{1}{\sqrt{\gamma\delta}\mu/(4R) + 4\gamma\delta}.$$

$$\theta_y = 1 - \frac{1}{n + 8nR/(\mu\sqrt{\gamma\delta})} \approx 1 - \frac{\sqrt{\gamma\delta}\mu}{8nR}\left(1 + \frac{\sqrt{\gamma\delta}\mu}{8R}\right)^{-1}.$$

Since the objective is R^2/γ-smooth and $\delta\mu^2/n$-strongly convex, θ_y is an accelerated rate if $\frac{\sqrt{\gamma\delta}\mu}{8R} \ll 1$ (otherwise $\theta_y \approx 1 - \frac{1}{n}$). For θ_x, we consider different situations:

- If $\mu \geq R$, then we have $\theta_x \approx 1 - \frac{\sqrt{\gamma\delta}\mu}{nR}$, which is an accelerated rate. So is $\theta = \max\{\theta_x, \theta_y\}$.

- If $\mu < R$ and $\gamma\delta \approx \frac{\mu^2}{R^2}$, then $\theta_x \approx 1 - \frac{\sqrt{\gamma\delta}\mu}{nR}$, which represents accelerated rate. The iteration complexity of SPDC is $\widetilde{O}(\frac{nR}{\mu\sqrt{\gamma\delta}})$, which is better than that of SVRG in this case, which is $\widetilde{O}(\frac{nR^2}{\gamma\delta\mu^2})$.

- If $\mu < R$ and $\gamma\delta \approx \frac{\mu}{R}$, then we get $\theta_x \approx 1 - \frac{\mu^2}{nR^2}$. This is a half-accelerated rate, because in this case SVRG would require $\widetilde{O}(\frac{nR^3}{\mu^3})$ iterations, while iteration complexity here is $\widetilde{O}(\frac{nR^2}{\mu^2})$.

- If $\mu < R$ and $\gamma\delta \approx 1$, meaning the ϕ_i's are well conditioned, then we get $\theta_x \approx 1 - \frac{\gamma\delta\mu^2}{nR^2} \approx 1 - \frac{\mu^2}{nR^2}$, which is a non-accelerated rate. The corresponding iteration complexity is the same as SVRG.

3.1. Parameter adaptation for SPDC

The SPDC-Adapt procedure called in Algorithm 5 follows the same logics as the batch adaption schemes in Algorithms 3 and 4, and we omit the details here. One thing we emphasize here is that the adaptation period T is in terms of epochs, or number of passes over the data. In addition, we only compute the primal and dual objective values after each pass or every few passes, because computing them exactly usually need to take a full pass of the data.

Another important issue is that, unlike the batch case where the duality gap usually decreases monotonically, the duality gap for randomized algorithms can fluctuate wildly. So instead of using only the two end values $P^{(t-Tn)} - D^{(t-Tn)}$ and $P^{(t)} - D^{(t)}$, we can use more points to estimate the convergence rate through a linear regression. Suppose the primal-dual values at the end of each past $T + 1$ passes are

$$\{P(0), D(0)\}, \{P(1), D(1)\}, \ldots, \{P(T), D(T)\},$$

and we need to estimate ρ (rate per pass) such that

$$P(t) - D(t) \approx \rho^t\left(P(0) - D(0)\right), \quad t = 1, \ldots, T.$$

We can turn it into a linear regression problem after taking logarithm and obtain the estimate $\hat{\rho}$ through

$$\log(\hat{\rho}) = \frac{1}{1^2 + 2^2 + \cdots + T^2} \sum_{t=1}^T t \log \frac{P(t) - D(t)}{P(0) - D(0)}.$$

The rest of the adaption procedure can follow the robust scheme in Algorithm 4. In practice, we can compute the primal-dual values more sporadically, say every few passes, and modify the regression accordingly.

4. Dual-free Primal-dual algorithms

Compared with primal algorithms, one major disadvantage of primal-dual algorithms is the requirement of computing the proximal mapping of the dual function f^* or ϕ_i^*, which may not admit closed-formed solution or efficient computation. This is especially the case for logistic regression, one of the most popular loss functions used in classification.

Lan & Zhou (2015) developed "dual-free" variants of primal-dual algorithms that avoid computing the dual proximal mapping. Their main technique is to replace the Euclidean distance in the dual proximal mapping with a Bregman divergence defined over the dual loss function itself.

Algorithm 6 Dual-Free BPD Algorithm

input: parameters $\sigma, \tau, \theta > 0$, initial point $(x^{(0)}, y^{(0)})$

Set $\tilde{x}^{(0)} = x^{(0)}$ and $v^{(0)} = (f^*)'(y^{(0)})$

for $t = 0, 1, 2, \ldots$ **do**

$\quad v^{(t+1)} = \frac{v^{(t)} + \sigma A \tilde{x}^{(t)}}{1+\sigma}, \quad y^{(t+1)} = f'(v^{(t+1)})$

$\quad x^{(t+1)} = \text{prox}_{\tau g}\left(x^{(t)} - \tau A^T y^{(t+1)}\right)$

$\quad \tilde{x}^{(t+1)} = x^{(t+1)} + \theta(x^{(t+1)} - x^{(t)})$

end for

We show how to apply this approach to solve the structured ERM problems considered in this paper. They can also exploit strong convexity from data if the algorithmic parameters are set appropriately or adapted automatically.

4.1. Dual-free BPD algorithm

First, we consider the batch setting. We replace the dual proximal mapping (computing $y^{(t+1)}$) in Algorithm 1 with

$$y^{(t+1)} = \arg\min_y \left\{ f^*(y) - y^T A \tilde{x}^{(t)} + \frac{1}{\sigma} \mathcal{D}(y, y^{(t)}) \right\}, \quad (11)$$

where $\mathcal{D}$ is the Bregman divergence of a strictly convex kernel function h, defined as

$$\mathcal{D}_h(y, y^{(t)}) = h(y) - h(y^{(t)}) - \langle \nabla h(y^{(t)}), y - y^{(t)} \rangle.$$

Algorithm 1 is obtained in the Euclidean setting with $h(y) = \frac{1}{2}\|y\|^2$ and $\mathcal{D}(y, y^{(t)}) = \frac{1}{2}\|y - y^{(t)}\|^2$. While our convergence results would apply for arbitrary Bregman divergence, we only focus on the case of using f^* itself as the kernel, because this allows us to compute $y^{(t+1)}$ in (11) very efficiently. The following lemma explains the details (Cf. Lan & Zhou, 2015, Lemma 1).

Lemma 1. *Let the kernel $h \equiv f^*$ in the Bregman divergence $\mathcal{D}$. If we construct a sequence of vectors $\{v^{(t)}\}$ such that $v^{(0)} = (f^*)'(y^{(0)})$ and for all $t \geq 0$,*

$$v^{(t+1)} = \frac{v^{(t)} + \sigma A \tilde{x}^{(t)}}{1+\sigma}, \quad (12)$$

then the solution to problem (11) is $y^{(t+1)} = f'(v^{(t+1)})$.

Proof. Suppose $v^{(t)} = (f^*)'(y^{(t)})$ (true for $t = 0$), then

$$\mathcal{D}(y, y^{(t)}) = f^*(y) - f^*(y^{(t)}) - v^{(t)T}(y - y^{(t)}).$$

The solution to (11) can be written as

$$y^{(t+1)} = \arg\min_y \left\{ f^*(y) - y^T A \tilde{x}^{(t)} + \frac{1}{\sigma}\left(f^*(y) - v^{(t)T}y\right)\right\}$$

$$= \arg\min_y \left\{ \left(1 + \frac{1}{\sigma}\right)f^*(y) - \left(A\tilde{x}^{(t)} + \frac{1}{\sigma}v^{(t)}\right)^T y \right\}$$

$$= \arg\max_y \left\{ \left(\frac{v^{(t)} + \sigma A \tilde{x}^{(t)}}{1+\sigma}\right)^T y - f^*(y) \right\}$$

$$= \arg\max_y \left\{ v^{(t+1)T}y - f^*(y) \right\} = f'(v^{(t+1)}),$$

where in the last equality we used the property of conjugate function when f is strongly convex and smooth. Moreover,

$$v^{(t+1)} = (f')^{-1}(y^{(t+1)}) = (f^*)'(y^{(t+1)}),$$

which completes the proof. $\qquad\square$

According to Lemma 1, we only need to provide initial points such that $v^{(0)} = (f^*)'(y^{(0)})$ is easy to compute. We do not need to compute $(f^*)'(y^{(t)})$ directly for any $t > 0$, because it is can be updated as $v^{(t)}$ in (12). Consequently, we can update $y^{(t)}$ in the BPD algorithm using the gradient $f'(v^{(t)})$, without the need of dual proximal mapping. The resulting dual-free algorithm is given in Algorithm 6.

Lan & Zhou (2015) considered a general setting which does not possess the linear predictor structure we focus on in this paper, and assumed that only the regularization g is strongly convex. Our following result shows that dual-free primal-dual algorithms can also exploit strong convexity from data with appropriate algorithmic parameters.

Theorem 3. *Suppose Assumption 2 holds and let $(x^\star, y^\star)$ be the unique saddle point of $\mathcal{L}$ defined in (6). If we set the parameters in Algorithm 6 as*

$$\tau = \frac{1}{L}\sqrt{\frac{\gamma}{\lambda + \delta\mu^2}}, \quad \sigma = \frac{1}{L}\sqrt{\gamma(\lambda + \delta\mu^2)}, \quad (13)$$

and $\theta = \max\{\theta_x, \theta_y\}$ where

$$\theta_x = \left(1 - \frac{\tau\sigma\delta\mu^2}{(4+2\sigma)}\right)\frac{1}{1+\tau\lambda}, \quad \theta_y = \frac{1}{1+\sigma/2}, \quad (14)$$

then we have

$$\left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\|x^{(t)} - x^\star\|^2 + \frac{1}{2}\mathcal{D}(y^\star, y^{(t)}) \leq \theta^t C,$$

$$\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^{(t)}) \leq \theta^t C,$$

where $C = \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\|x^{(0)} - x^\star\|^2 + \left(\frac{1}{\sigma} + \frac{1}{2}\right)\mathcal{D}(y^\star, y^{(0)})$.

Theorem 3 is proved in Appendices B and D. Assuming $\gamma(\lambda + \delta\mu^2) \ll L^2$, we have

$$\theta_x \approx 1 - \frac{\gamma\delta\mu^2}{16L^2} - \frac{\lambda}{2L}\sqrt{\frac{\gamma}{\lambda + \delta\mu^2}}, \quad \theta_y \approx 1 - \frac{\sqrt{\gamma(\lambda + \delta\mu^2)}}{4L}.$$

Again, we gain insights by consider the special cases:

- If $\delta\mu^2 = 0$ and $\lambda > 0$, then $\theta_y \approx 1 - \frac{\sqrt{\gamma\lambda}}{4L}$ and $\theta_x \approx 1 - \frac{\sqrt{\gamma\lambda}}{2L}$. So $\theta = \max\{\theta_x, \theta_y\}$ is an accelerated rate.

- If $\delta\mu^2 > 0$ and $\lambda = 0$, then $\theta_y \approx 1 - \frac{\sqrt{\gamma\delta\mu^2}}{4L}$ and $\theta_x \approx 1 - \frac{\gamma\delta\mu^2}{16L^2}$. Thus $\theta = \max\{\theta_x, \theta_y\} \approx 1 - \frac{\gamma\delta\mu^2}{16L^2}$ is not accelerated. Notice that this conclusion does not depends on the relative size of $\gamma\delta$ and μ^2/L^2, and this is the major difference from the Euclidean case discussed in Section 2.

If both $\delta\mu^2 > 0$ and $\lambda > 0$, then the extent of acceleration depends on their relative size. If λ is on the same order as $\delta\mu^2$ or larger, then accelerated rate is obtained. If λ is much smaller than $\delta\mu^2$, then the theory predicts no acceleration.

Algorithm 7 Adaptive Dual-Free SPDC (ADF-SPDC)

---

input: parameters $\sigma, \tau, \theta > 0$, initial point $(x^{(0)}, y^{(0)})$, and adaptation period T.

Set $\tilde{x}^{(0)} = x^{(0)}$ and $v_i^{(0)} = (\phi_i^*)'(y_i^{(0)})$ for $i = 1, \ldots, n$

for $t = 0, 1, 2, \ldots$ **do**

    pick $k \in \{1, \ldots, n\}$ uniformly at random

    for $i \in \{1, \ldots, n\}$ **do**

        if $i == k$ **then**

$$v_k^{(t+1)} = \frac{v_k^{(t)} + \sigma a_k^T \tilde{x}^{(t)}}{1+\sigma}, \quad y_k^{(t+1)} = \phi_k'(v_k^{(t+1)})$$

        else

$$v_i^{(t+1)} = v_i^{(t)}, \quad y_i^{(t+1)} = y_i^{(t)}$$

        end if

    end for

$$x^{(t+1)} = \text{prox}_{\tau g}\Big(x^{(t)} - \tau\big(u^{(t)} + (y_k^{(t+1)} - y_k^{(t)})a_k\big)\Big)$$

$$u^{(t+1)} = u^{(t)} + \frac{1}{n}(y_k^{(t+1)} - y_k^{(t)})a_k$$

$$\tilde{x}^{(t+1)} = x^{(t+1)} + \theta(x^{(t+1)} - x^{(t)})$$

    if $\text{mod}(t+1, T \cdot n) = 0$ **then**

        $(\tau, \sigma, \theta) = \text{SPDC-Adapt}\big(\{P^{(t-sn)}, D^{(t-sn)}\}_{s=0}^T\big)$

    end if

end for

4.2. Dual-free SPDC algorithm

The same approach can be applied to derive an Dual-free SPDC algorithm, which is described in Algorithm 7. It also includes a parameter adaption procedure, so we call it the adaptive dual-free SPDC (ADF-SPDC) algorithm. On related work, Shalev-Shwartz & Zhang (2016) and (Shalev-Shwartz, 2016) introduced dual-free SDCA.

The following theorem characterizes the choice of algorithmic parameters that can exploit strong convexity from data to achieve linear convergence (proof given in Appendix F).

Theorem 4. *Suppose Assumption 1 holds. Let $(x^\star, y^\star)$ be the saddle point of $\mathcal{L}$ defined in (3) and $R = \max\{\|a_1\|, \ldots, \|a_n\|\}$. If we set $T = \infty$ in Algorithm 7 (non adaption) and let*

$$\sigma = \frac{1}{4R}\sqrt{\gamma(n\lambda + \delta\mu^2)}, \quad \tau = \frac{1}{4R}\sqrt{\frac{\gamma}{n\lambda + \delta\mu^2}}, \quad (15)$$

and $\theta = \max\{\theta_x, \theta_y\}$ where

$$\theta_x = \Big(1 - \frac{\tau\sigma\delta\mu^2}{n(4+2\sigma)}\Big)\frac{1}{1+\tau\lambda}, \quad \theta_y = \frac{1+((n-1)/n)\sigma/2}{1+\sigma/2}, \quad (16)$$

then we have

$$\Big(\frac{1}{2\tau} + \frac{\lambda}{2}\Big)\mathbb{E}\big[\|x^{(t)} - x^\star\|^2\big] + \frac{\gamma}{4}\mathbb{E}\big[\mathcal{D}(y^\star, y^{(t)})\big] \leq \theta^t C,$$
$$\mathbb{E}\big[\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^{(t)})\big] \leq \theta^t C,$$

where $C = \big(\frac{1}{2\tau} + \frac{\lambda}{2}\big)\|x^{(0)} - x^\star\|^2 + \big(\frac{1}{\sigma} + \frac{1}{2}\big)\mathcal{D}(y^\star, y^{(0)})$.

Below we discuss the expected convergence rate established in Theorem 2 in two special cases.

The cases of $\sigma\mu^2 = 0$ but $\lambda > 0$. In this case we have $\tau = \frac{1}{4R}\sqrt{\frac{\gamma}{n\lambda}}$ and $\sigma = \frac{1}{4R}\sqrt{n\gamma\lambda}$, and

$$\theta_x = \frac{1}{1+\tau\lambda} = 1 - \frac{1}{1+4R\sqrt{n/(\lambda\gamma)}},$$
$$\theta_y = \frac{1+((n-1)/n)\sigma/2}{1+\sigma/2} = 1 - \frac{1}{n+8R\sqrt{n/(\lambda\gamma)}}.$$

These recover the convergence rate of the standard SPDC algorithm (Zhang & Xiao, 2015).

The cases of $\sigma\mu^2 > 0$ but $\lambda = 0$. In this case we have

$$\tau = \frac{1}{4R\mu}\sqrt{\frac{\gamma}{\delta}}, \quad \sigma = \frac{\mu}{4R}\sqrt{\delta\gamma},$$

and

$$\theta_x = 1 - \frac{\tau\sigma\delta\mu^2}{2n(\sigma+4)} = 1 - \frac{\gamma\delta\mu^2}{32nR^2} \cdot \frac{1}{\sqrt{\gamma\delta}\mu/(4R)+4},$$
$$\theta_y = \frac{1+((n-1)/n)\sigma/2}{1+\sigma/2} = 1 - \frac{1}{n+8nR/(\mu\sqrt{\gamma\delta})}.$$

We note that the primal function now is R^2/γ-smooth and $\delta\mu^2/n$-strongly convex. We discuss the following cases:

- If $\sqrt{\gamma\delta}\mu > R$, then we have $\theta_x \approx 1 - \frac{\sqrt{\gamma\delta}\mu}{8nR}$ and $\theta_y \approx 1 - \frac{1}{n}$. Therefore $\theta = \max\{\theta_x, \theta_y\} \approx 1 - \frac{1}{n}$.

- Otherwise, we have $\theta_x \approx 1 - \frac{\gamma\delta\mu^2}{64nR^2}$ and θ_y is of the same order. This is not an accelerated rate, and we have the same iteration complexity as SVRG.

Finally, we give concrete examples of how to compute the initial points $y^{(0)}$ and $v^{(0)}$ such that $v_i^{(0)} = (\phi_i^*)'(y_i^{(0)})$.

- For squared loss, $\phi_i(\alpha) = \frac{1}{2}(\alpha - b_i)^2$ and $\phi_i^*(\beta) = \frac{1}{2}\beta^2 + b_i\beta$. So $v_i^{(0)} = (\phi_i^*)'(y_i^{(0)}) = y_i^{(0)} + b_i$.

- For logistic regression, we have $b_i \in \{1, -1\}$ and $\phi_i(\alpha) = \log(1 + e^{-b_i\alpha})$. The conjugate function is $\phi_i^*(\beta) = (-b_i\beta)\log(-b_i\beta) + (1 + b_i\beta)\log(1 + b_i\beta)$ if $b_i\beta \in [-1, 0]$ and $+\infty$ otherwise. We can choose $y_i^{(0)} = -\frac{1}{2}b_i$ and $v_i^{(0)} = 0$ such that $v_i^{(0)} = (\phi_i^*)'(y_i^{(0)})$.

For logistic regression, we have $\delta = 0$ over the full domain of ϕ_i. However, each ϕ_i is locally strongly convex in bounded domain (Bach, 2014): if $z \in [-B, B]$, then we know $\delta = \min_z \phi_i''(z) \geq \exp(-B)/4$. Therefore it is well suitable for an adaptation scheme similar to Algorithm 4 that do not require knowledge of either δ or μ.

5. Preliminary experiments

We present preliminary experiments to demonstrate the effectiveness of our proposed algorithms. First, we consider batch primal-dual algorithms for ridge regression over a synthetic dataset. The data matrix A has sizes $n = 5000$ and $d = 3000$, and its entries are sampled from multivariate normal distribution with mean zero and covariance matrix $\Sigma_{ij} = 2^{|i-j|/2}$. We normalize all datasets
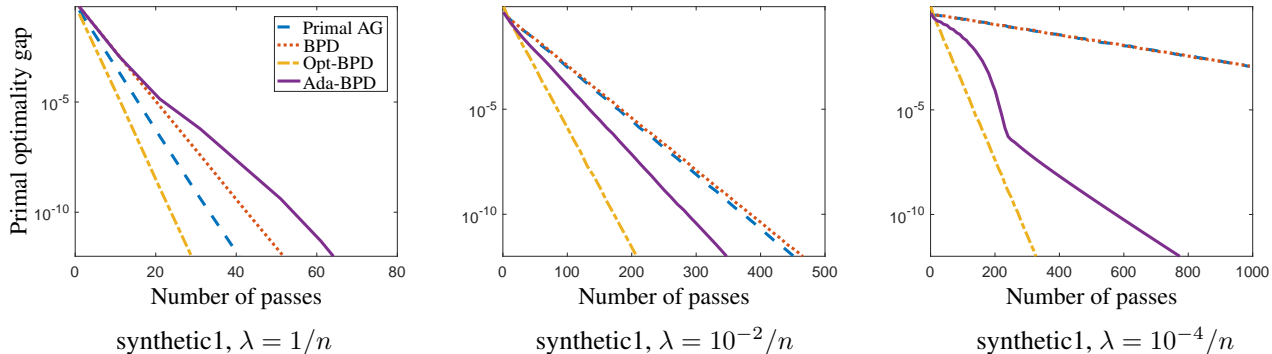
Figure 1. Comparison of batch primal-dual algorithms for a ridge regression problem with $n = 5000$ and $d = 3000$.

such that $a_i = a_i / (\max_j \|a_j\|)$, to ensure the maximum norm of the data points is 1. We use ℓ_2-regularization $g(x) = (\lambda/2)\|x\|^2$ with three choices of parameter λ: $1/n$, $10^{-2}/n$ and $10^{-4}/n$, which represent the strong, medium, and weak levels of regularization, respectively.

Figure 1 shows the performance of four different algorithms: the accelerated gradient algorithm for solving the primal minimization problem (Primal AG) (Nesterov, 2004) using λ as strong convexity parameter, the BPD algorithm (Algorithm 1) that uses λ as the strong convexity parameter (setting $\mu = 0$), the optimal BPD algorithm (Opt-BPD) that uses $\mu = \sqrt{\lambda_{\min}(A^T A)}$ explicitly computed from data, and the Ada-BPD algorithm (Algorithm 2) with the robust adaptation heuristic (Algorithm 4) with $T = 10$, $\underline{c} = 0.95$ and $\bar{c} = 1.5$. As expected, the performance of Primal-AG is very similar to BPD with the same strong convexity parameter. The Opt-BPD fully exploits strong convexity from data, thus has the fastest convergence. The Ada-BPD algorithm can partially exploit strong convexity from data without knowledge of μ.

Next we compare the DF-SPDC (Algorithm 5 without adaption) and ADF-SPDC (Algorithm 7 with adaption) against several state-of-the-art randomized algorithms for ERM: SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014) Katyusha (Allen-Zhu, 2016) and the standard SPDC method (Zhang & Xiao, 2015). For SVRG and Katyusha (an accelerated variant of SVRG), we choose the variance reduction period as $m = 2n$. The step sizes of all algorithms are set as their original paper suggested. For Ada-SPDC and ADF-SPDC, we use the robust adaptation scheme with $T = 10$, $\underline{c} = 0.95$ and $\bar{c} = 1.5$.

We first compare these randomized algorithms for ridge regression over the same synthetic data described above and the cpuact data from the LibSVM website[1]. The results are shown in Figure 2. With relatively strong regularization $\lambda = 1/n$, all methods perform similarly as predicted

by theory. For the synthetic dataset With $\lambda = 10^{-2}/n$, the regularization is weaker but still stronger than the hidden strong convexity from data, so the accelerated algorithms (all variants of SPDC and Katyusha) perform better than SVRG and SAGA. With $\lambda = 10^{-4}/n$, it looks that the strong convexity from data dominates the regularization. Since the non-accelerated algorithms (SVRG and SAGA) may automatically exploit strong convexity from data, they become faster than the non-adaptive accelerated methods (Katyusha, SPDC and DF-SPDC). The adaptive accelerated method, ADF-SPDC, has the fastest convergence. This shows that our theoretical results (which predict no acceleration in this case) can be further improved.

Finally we compare these randomized algorithm for logistic regression on the rcv1 dataset (from LibSVM website) and another synthetic dataset with $n = 5000$ and $d = 500$, generated similarly as before but with covariance matrix $\Sigma_{ij} = 2^{|i-j|/100}$. For the standard SPDC, we solve the dual proximal mapping using a few steps of Newton's method to high precision. The dual-free SPDC algorithms only use gradients of the logistic function. The results are presented in Figure 3. for both datasets, the strong convexity from data is very weak (or none), so the accelerated algorithms performs better.

6. Conclusions

We have shown that primal-dual first-order algorithms are capable of exploiting strong convexity from data, if the algorithmic parameters are chosen appropriately. While they may depends on problem dependent constants that are unknown, we developed heuristics for adapting the parameters on the fly and obtained improved performance in experiments. It looks that our theoretical characterization of the convergence rates can be further improved, as our experiments often demonstrate significant acceleration in cases where our theory does not predict acceleration.
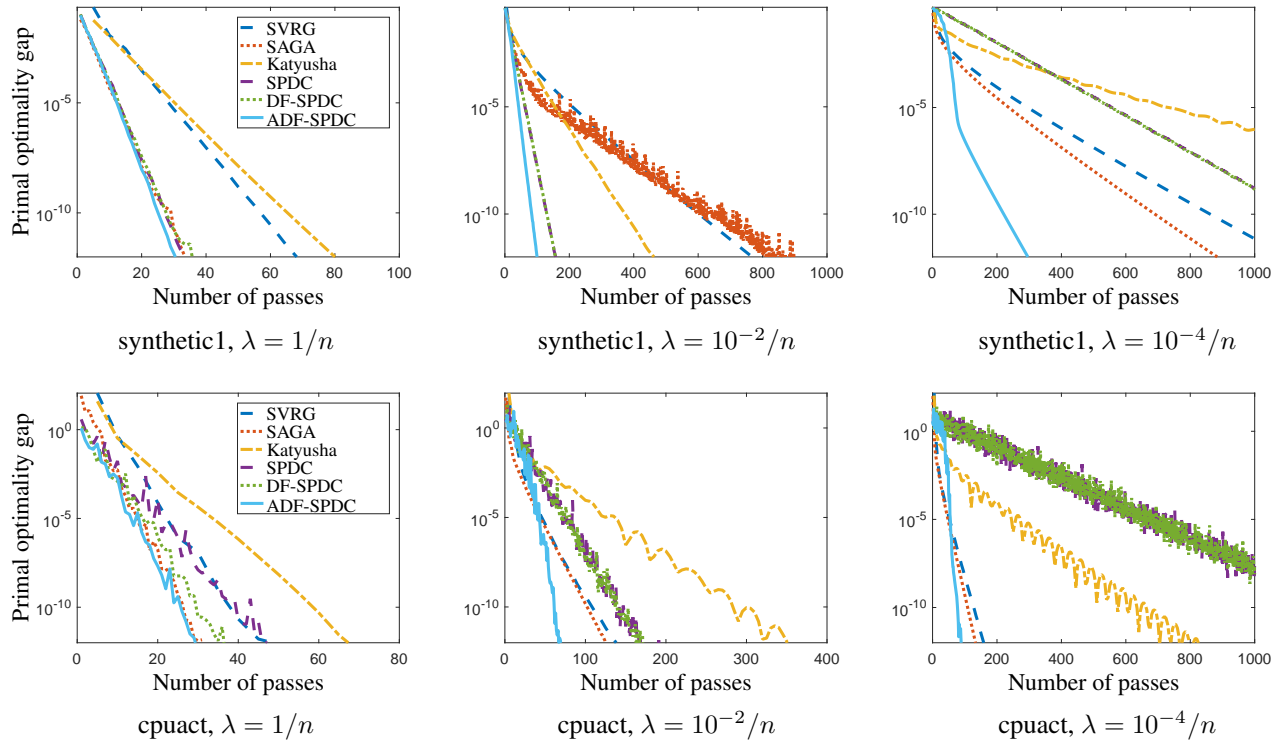
---

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Figure 2. Comparison of randomized algorithms for ridge regression problems.

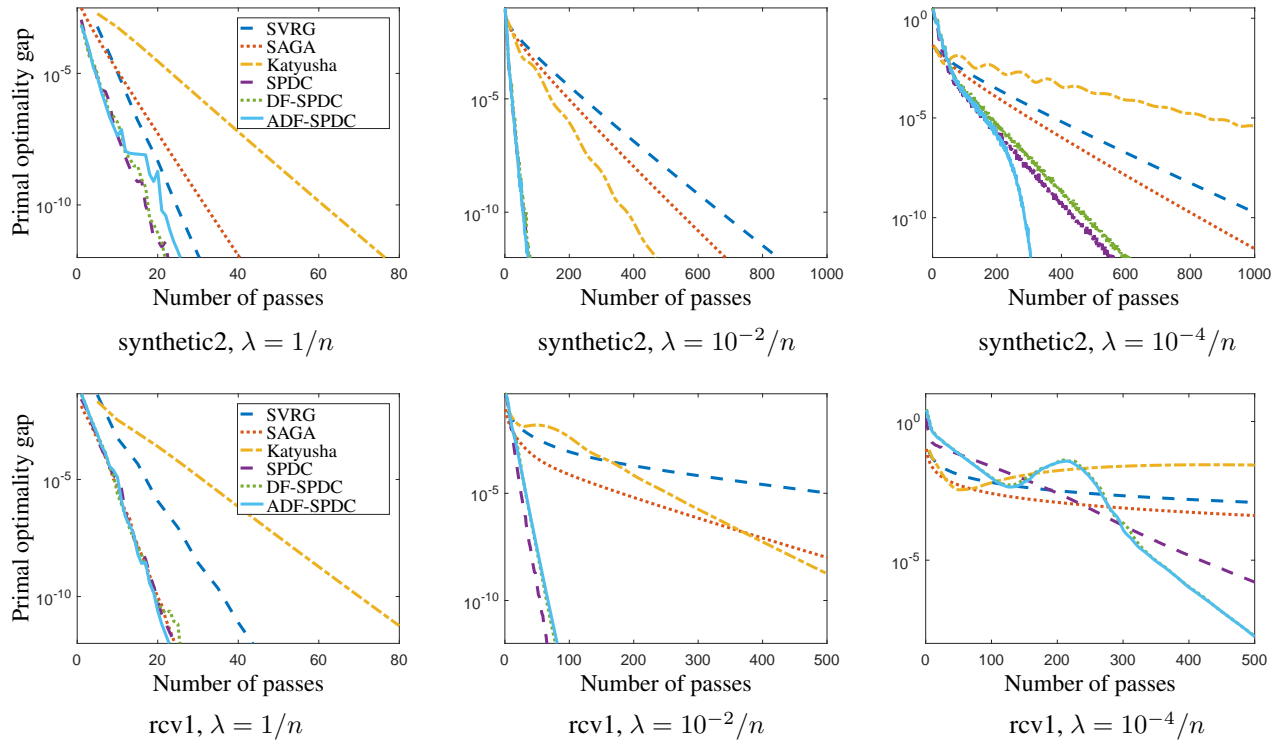*Figure 3.* Comparison of randomized algorithms for logistic regression problems.

References

Allen-Zhu, Zeyuan. Katyusha: Accelerated variance reduction for faster sgd. *ArXiv e-print 1603.05953*, 2016.

Bach, Francis. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(1):595–627, 2014.

Balamurugan, Palaniappan and Bach, Francis. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems (NIPS) 29*, pp. 1416–1424, 2016.

Bertsekas, Dimitri P. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. In Sra, Suvrit, Nowozin, Sebastian, and Wright, Stephen J. (eds.), *Optimization for Machine Learning*, chapter 4, pp. 85–120. MIT Press, 2012.

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

Chambolle, Antonin and Pock, Thomas. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming, Series A*, 159:253–287, 2016.

Defazio, Aaron, Bach, Francis, and Lacoste-Julien, Simon. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pp. 1646–1654, 2014.

Deng, Wei and Yin, Wotao. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3): 889–916, 2016.

Fercoq, Oliver and Richtárik, Peter. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

Goldstein, Tom, Li, Min, Yuan, Xiaoming, Esser, Ernie, and Baraniuk, Richard. Adaptive primal-dual hybrid gradient methods for saddle-point problems. *arXiv preprint arXiv:1305.0546*, 2013.

Johnson, Rie and Zhang, Tong. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pp. 315–323, 2013.

Lan, Guanghui and Zhou, Yi. An optimal randomized incremental gradient method. *arXiv preprint arXiv:1507.02000*, 2015.

Lin, Hongzhou, Mairal, Julien, and Harchaoui, Zaid. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pp. 3384–3392, 2015a.

Lin, Qihang, Lu, Zhaosong, and Xiao, Lin. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015b.

Malitsky, Yura and Pock, Thomas. A first-order primal-dual algorithm with linesearch. *arXiv preprint arXiv:1608.08883*, 2016.

Nedic, Angelia and Bertsekas, Dimitri P. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, Boston, 2004.

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

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

Roux, Nicolas L, Schmidt, Mark, and Bach, Francis. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in Neural Information Processing Systems*, pp. 2663–2671, 2012.

Shalev-Shwartz, Shai. Sdca without duality, regularization, and individual convexity. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 747–754, 2016.

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

Shalev-Shwartz, Shai and Zhang, Tong. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1-2): 105–145, 2016.

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

Zhang, Yuchen and Xiao, Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *Proceedings of The 32nd International Conference on Machine Learning*, pp. 353–361, 2015.

In the following appendices, we provide detailed proofs of theorems stated in the main paper. In Section A we first prove a basic inequality which is useful throughout the rest of the convergence analysis. Section B contains general analysis of the batch primal-dual algorithm that are common for proving both Theorem 1 and Theorem 3. Sections C, D, E and F give proofs for Theorem 1, Theorem 3, Theorem 2 and Theorem 4, respectively.

A. A basic lemma

Lemma 2. *Let h be a strictly convex function and $\mathcal{D}_h$ be its Bregman divergence. Suppose ψ is ν-strongly convex with respect to $\mathcal{D}_h$ and $1/\delta$-smooth (with respect to the Euclidean norm), and*

$$\hat{y} = \arg\min_{y \in C} \{\psi(y) + \eta \mathcal{D}_h(y, \bar{y})\},$$

where C is a compact convex set that lies within the relative interior of the domains of h and ψ (i.e., both h and ψ are differentiable over C). Then for any $y \in C$ and $\rho \in [0, 1]$, we have

$$\psi(y) + \eta \mathcal{D}_h(y, \bar{x}) \geq \psi(\hat{y}) + \eta \mathcal{D}_h(\hat{y}, \bar{y}) + \big(\eta + (1 - \rho)\nu\big)\mathcal{D}_h(y, \hat{y}) + \frac{\rho\delta}{2}\|\nabla\psi(y) - \nabla\psi(\hat{y})\|^2.$$

Proof. The minimizer $\hat{y}$ satisfies the following first-order optimality condition:

$$\langle \nabla\psi(\hat{y}) + \eta\nabla\mathcal{D}_h(\hat{y}, \bar{y}), \ y - \hat{y} \rangle \geq 0, \quad \forall\, y \in C.$$

Here $\nabla\mathcal{D}$ denotes partial gradient of the Bregman divergence with respect to its first argument, i.e., $\nabla\mathcal{D}(\hat{y}, \bar{y}) = \nabla h(\hat{y}) - \nabla h(\bar{y})$. So the above optimality condition is the same as

$$\langle \nabla\psi(\hat{y}) + \eta(\nabla h(\hat{y}) - \nabla h(\bar{y})), \ y - \hat{y} \rangle \geq 0, \quad \forall\, y \in C. \tag{17}$$

Since ψ is ν-strongly convex with respect to $\mathcal{D}_h$ and $1/\delta$-smooth, we have

$$\psi(y) \geq \psi(\hat{y}) + \langle \nabla\psi(\hat{y}), y - \hat{x} \rangle + \nu\mathcal{D}_h(y, \hat{y}),$$

$$\psi(y) \geq \psi(\hat{y}) + \langle \nabla\psi(\hat{y}), y - \hat{y} \rangle + \frac{\delta}{2}\|\nabla\psi(y) - \nabla\psi(\hat{y})\|^2.$$

For the second inequality, see, e.g., Theorem 2.1.5 in Nesterov (2004). Multiplying the two inequalities above by $(1 - \rho)$ and ρ respectively and adding them together, we have

$$\psi(y) \geq \psi(\hat{y}) + \langle \nabla\psi(\hat{y}), y - \hat{y} \rangle + (1 - \rho)\nu\mathcal{D}_h(y, \hat{y}) + \frac{\rho\delta}{2}\|\nabla\psi(y) - \nabla\psi(\hat{y})\|^2.$$

The Bregman divergence $\mathcal{D}_h$ satisfies the following equality:

$$\mathcal{D}_h(y, \bar{y}) = \mathcal{D}_h(y, \hat{y}) + \mathcal{D}_h(\hat{y}, \bar{y}) + \langle \nabla h(\hat{y}) - \nabla h(\bar{y}), \ y - \hat{y} \rangle.$$

We multiply this equality by η and add it to the last inequality to obtain

$$
\begin{aligned}
\psi(y) + \eta\mathcal{D}_h(y, \bar{y}) \ \geq\ & \psi(\hat{y}) + \eta\mathcal{D}_h(y, \hat{y}) + \big(\eta + (1 - \rho)\nu\big)\mathcal{D}_h(\hat{y}, \bar{y}) + \frac{\rho\delta}{2}\|\nabla\psi(y) - \nabla\psi(\hat{y})\|^2 \\
& + \langle \nabla\psi(\hat{y}) + \eta(\nabla h(\hat{y}) - \nabla h(\bar{y})), \ y - \hat{y} \rangle.
\end{aligned}
$$

Using the optimality condition in (17), the last term of inner product is nonnegative and thus can be dropped, which gives the desired inequality. $\qquad\square$

B. Common Analysis of Batch Primal-Dual Algorithms

We consider the general primal-dual update rule as:

Iteration: $(\hat{x}, \hat{y}) = \text{PD}_{\tau,\sigma}(\bar{x}, \bar{y}, \tilde{x}, \tilde{y})$

$$\hat{x} = \underset{x \in \mathbb{R}^d}{\arg\min} \left\{ g(x) + \tilde{y}^T A x + \frac{1}{2\tau} \|x - \bar{x}\|^2 \right\}, \tag{18}$$

$$\hat{y} = \underset{y \in \mathbb{R}^n}{\arg\min} \left\{ f^*(y) - y^T A\tilde{x} + \frac{1}{\sigma} \mathcal{D}(y, \bar{y}) \right\}. \tag{19}$$

Each iteration of Algorithm 1 is equivalent to the following specification of $\text{PD}_{\tau,\sigma}$:

$$\begin{aligned}
\hat{x} = x^{(t+1)}, & \quad \bar{x} = x^{(t)}, & \quad \tilde{x} = x^{(t)} + \theta(x^{(t)} - x^{(t-1)}), \\
\hat{y} = y^{(t+1)}, & \quad \bar{y} = y^{(t)}, & \quad \tilde{y} = y^{(t+1)}.
\end{aligned} \tag{20}$$

Besides Assumption 2, we also assume that f^* is ν-strongly convex with respect to a kernel function h, i.e.,

$$f^*(y') - f^*(y) - \langle \nabla f^*(y), y' - y \rangle \geq \nu \mathcal{D}_h(y', y),$$

where $\mathcal{D}_h$ is the Bregman divergence defined as

$$\mathcal{D}_h(y', y) = h(y') - h(y) - \langle \nabla h(y), y' - y \rangle.$$

We assume that h is γ'-strongly convex and $1/\delta'$-smooth. Depending on the kernel function h, this assumption on f^* may impose additional restrictions on f. In this paper, we are mostly interested in two special cases: $h(y) = (1/2)\|y\|^2$ and $h(y) = f^*(y)$ (for the latter we always have $\nu = 1$). From now on, we will omit the subscript h and use $\mathcal{D}$ denote the Bregman divergence.

Under the above assumptions, any solution $(x^\star, y^\star)$ to the saddle-point problem (6) satisfies the optimality condition:

$$\begin{aligned}
-A^T y^\star & \in \partial g(x^\star), \tag{21} \\
A x^\star & = \nabla f^*(y^\star). \tag{22}
\end{aligned}$$

The optimality conditions for the updates described in equations (18) and (19) are

$$-A^T \tilde{y} + \frac{1}{\tau}(\bar{x} - \hat{x}) \in \partial g(\hat{x}), \tag{23}$$

$$A\tilde{x} - \frac{1}{\sigma}\big(\nabla h(\hat{y}) - \nabla h(\bar{y})\big) = \nabla f^*(\hat{y}). \tag{24}$$

Applying Lemma 2 to the dual minimization step in (19) with $\psi(y) = f^*(y) - y^T A\tilde{x}$, $\eta = 1/\sigma$ $y = y^\star$ and $\rho = 1/2$, we obtain

$$\begin{aligned}
f^*(y^\star) - y^{\star T} A\tilde{x} + \frac{1}{\sigma}\mathcal{D}(y^\star, \bar{y}) \geq \ & f^*(\hat{y}) - \hat{y}^T A\tilde{x} + \frac{1}{\sigma}\mathcal{D}(\hat{y}, \bar{y}) \\
& + \left(\frac{1}{\sigma} + \frac{\nu}{2}\right)\mathcal{D}(y^\star, \hat{y}) + \frac{\delta}{4}\big\|\nabla f^*(y^\star) - \nabla f^*(\hat{y})\big\|^2. \tag{25}
\end{aligned}$$

Similarly, for the primal minimization step in (18), we have (setting $\rho = 0$)

$$g(x^\star) + \tilde{y}^T A x^\star + \frac{1}{2\tau}\|x^\star - \bar{x}\|^2 \geq g(\hat{x}) + \tilde{y}^T A\hat{x} + \frac{1}{2\tau}\|\hat{x} - \bar{x}\|^2 + \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^\star - \hat{x}\|^2. \tag{26}$$

Combining the two inequalities above with the definition $\mathcal{L}(x, y) = g(x) + y^T A x - f^*(y)$, we get

$$\begin{aligned}
\mathcal{L}(\hat{x}, y^\star) - \mathcal{L}(x^\star, \hat{y}) = \ & g(\hat{x}) + y^{\star T} A\hat{x} - f^*(y^\star) - g(x^\star) - \hat{y}^T A x^\star + f^*(\hat{y}) \\
\leq \ & \frac{1}{2\tau}\|x^\star - \bar{x}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^\star, \bar{y}) - \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^\star - \hat{x}\|^2 - \left(\frac{1}{\sigma} + \frac{\nu}{2}\right)\mathcal{D}(y^\star, \hat{y}) \\
& - \frac{1}{2\tau}\|\hat{x} - \bar{x}\|^2 - \frac{1}{\sigma}\mathcal{D}(\hat{y}, \bar{y}) - \frac{\delta}{4}\big\|\nabla f^*(y^\star) - \nabla f^*(\hat{y})\big\|^2 \\
& + y^{\star T} A\hat{x} - \hat{y}^T A x^\star + \tilde{y}^T A x^\star - \tilde{y}^T A\hat{x} - y^{\star T} A\tilde{x} + \hat{y}^T A\tilde{x}.
\end{aligned}$$

We can simplify the inner product terms as

$$y^{\star T} A\hat{x} - \hat{y}^T A x^\star + \tilde{y}^T A x^\star - \tilde{y}^T A\hat{x} - y^{\star T} A\tilde{x} + \hat{y}^T A\tilde{x} = (\hat{y} - \tilde{y})^T A(\hat{x} - x^\star) - (\hat{y} - y^\star)^T A(\hat{x} - \tilde{x}).$$

Rearranging terms on the two sides of the inequality, we have

$$
\begin{aligned}
\frac{1}{2\tau}\|x^\star - \bar{x}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^\star, \bar{y}) \;\geq\;\; & \mathcal{L}(\hat{x}, y^\star) - \mathcal{L}(x^\star, \hat{y}) \\
& + \frac{1}{2}\Big(\frac{1}{\tau} + \lambda\Big)\|x^\star - \hat{x}\|^2 + \Big(\frac{1}{\sigma} + \frac{\nu}{2}\Big)\mathcal{D}(y^\star, \hat{y}) \\
& + \frac{1}{2\tau}\|\hat{x} - \bar{x}\|^2 + \frac{1}{\sigma}\mathcal{D}(\hat{y}, \bar{y}) + \frac{\delta}{4}\big\|\nabla f^*(y^\star) - \nabla f^*(\hat{y})\big\|^2 \\
& + (\hat{y} - y^\star)^T A(\hat{x} - \tilde{x}) - (\hat{y} - \tilde{y})^T A(\hat{x} - x^\star).
\end{aligned}
$$

Applying the substitutions in (20) yields

$$
\begin{aligned}
\frac{1}{2\tau}\|x^\star - x^{(t)}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^\star, y^{(t)}) \;\geq\;\; & \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)}) \\
& + \frac{1}{2}\Big(\frac{1}{\tau} + \lambda\Big)\|x^\star - x^{(t+1)}\|^2 + \Big(\frac{1}{\sigma} + \frac{\nu}{2}\Big)\mathcal{D}(y^\star, y^{(t+1)}) \\
& + \frac{1}{2\tau}\|x^{(t+1)} - x^{(t)}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^{(t+1)}, y^{(t)}) + \frac{\delta}{4}\big\|\nabla f^*(y^\star) - \nabla f^*(y^{(t+1)})\big\|^2 \\
& + (y^{(t+1)} - y^\star)^T A\big(x^{(t+1)} - (x^{(t)} + \theta(x^{(t)} - x^{(t-1)}))\big). \qquad (27)
\end{aligned}
$$

We can rearrange the inner product term in (27) as

$$
\begin{aligned}
& (y^{(t+1)} - y^\star)^T A\big(x^{(t+1)} - (x^{(t)} + \theta(x^{(t)} - x^{(t-1)}))\big) \\
=\;\; & (y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)}) - \theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)}) - \theta(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)}).
\end{aligned}
$$

Using the optimality conditions in (22) and (24), we can also bound $\|\nabla f^*(y^\star) - \nabla f^*(y^{(t+1)})\|^2$:

$$
\begin{aligned}
& \big\|\nabla f^*(y^\star) - \nabla f^*(y^{(t+1)})\big\|^2 \\
=\;\; & \Big\|Ax^\star - A\big(x^{(t)} + \theta(x^{(t)} - x^{(t-1)})\big) + \frac{1}{\sigma}\big(\nabla h(y^{(t+1)}) - \nabla h(y^{(t)})\big)\Big\|^2 \\
\geq\;\; & \Big(1 - \frac{1}{\alpha}\Big)\|A(x^\star - x^{(t)})\|^2 - (\alpha - 1)\Big\|\theta A(x^{(t)} - x^{(t-1)})) - \frac{1}{\sigma}\big(\nabla h(y^{(t+1)}) - \nabla h(y^{(t)})\big)\Big\|^2,
\end{aligned}
$$

where $\alpha > 1$. With the definition $\mu = \sqrt{\lambda_{\min}(A^T A)}$, we also have $\|A(x^\star - x^{(t)})\|^2 \geq \mu^2\|x^\star - x^{(t)}\|^2$. Combining them with the inequality (27) leads to

$$
\begin{aligned}
& \frac{1}{2\tau}\|x^\star - x^{(t)}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^\star, y^{(t)}) + \theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)}) \\
\geq\;\; & \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)}) \\
& + \frac{1}{2}\Big(\frac{1}{\tau} + \lambda\Big)\|x^\star - x^{(t+1)}\|^2 + \Big(\frac{1}{\sigma} + \frac{\nu}{2}\Big)\mathcal{D}(y^\star, y^{(t+1)}) + (y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)}) \\
& + \frac{1}{2\tau}\|x^{(t+1)} - x^{(t)}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^{(t+1)}, y^{(t)}) - \theta(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)}) \\
& + \Big(1 - \frac{1}{\alpha}\Big)\frac{\delta\mu^2}{4}\|x^\star - x^{(t)}\|^2 - (\alpha - 1)\frac{\delta}{4}\Big\|\theta A(x^{(t)} - x^{(t-1)})) - \frac{1}{\sigma}\big(\nabla h(y^{(t+1)}) - \nabla h(y^{(t)})\big)\Big\|^2. \qquad (28)
\end{aligned}
$$

C. Proof of Theorem 1

Let the kernel function be $h(y) = (1/2)\|y\|^2$. In this case, we have $\mathcal{D}(y', y) = (1/2)\|y' - y\|^2$ and $\nabla h(y) = y$. Moreover, $\gamma' = \delta' = 1$ and $\nu = \gamma$. Therefore, the inequality (28) becomes

$$\frac{1}{2}\left(\frac{1}{\tau} - \left(1 - \frac{1}{\alpha}\right)\frac{\delta\mu^2}{2}\right)\|x^\star - x^{(t)}\|^2 + \frac{1}{2\sigma}\|y^\star - y^{(t)}\|^2 + \theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})$$

$$\geq \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)})$$

$$+ \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^\star - x^{(t+1)}\|^2 + \frac{1}{2}\left(\frac{1}{\sigma} + \frac{\gamma}{2}\right)\|y^\star - y^{(t+1)}\|^2 + (y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})$$

$$+ \frac{1}{2\tau}\|x^{(t+1)} - x^{(t)}\|^2 + \underline{\frac{1}{2\sigma}\|y^{(t+1)} - y^{(t)}\|^2 - \theta(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)})}$$

$$- (\alpha - 1)\frac{\delta}{4}\left\|\theta A(x^{(t)} - x^{(t-1)})) - \frac{1}{\sigma}(y^{(t+1)} - y^{(t)})\right\|^2. \tag{29}$$

Next we derive another form of the underlined items above:

$$\frac{1}{2\sigma}\|y^{(t+1)} - y^{(t)}\|^2 - \theta(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)})$$

$$= \frac{\sigma}{2}\left(\frac{1}{\sigma^2}\|y^{(t+1)} - y^{(t)}\|^2 - \frac{\theta}{\sigma}(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)})\right)$$

$$= \frac{\sigma}{2}\left(\left\|\theta A(x^{(t)} - x^{(t-1)}) - \frac{1}{\sigma}(y^{(t+1)} - y^{(t)})\right\|^2 - \theta^2\|A(x^{(t)} - x^{(t-1)})\|^2\right)$$

$$\geq \frac{\sigma}{2}\left\|\theta A(x^{(t)} - x^{(t-1)}) - \frac{1}{\sigma}(y^{(t+1)} - y^{(t)})\right\|^2 - \frac{\sigma\theta^2 L^2}{2}\|x^{(t)} - x^{(t-1)}\|^2,$$

where in the last inequality we used $\|A\| \leq L$ and hence $\|A(x^{(t)} - x^{(t-1)})\|^2 \leq L^2\|x^{(t)} - x^{(t-1)}\|^2$. Combining with inequality (29), we have

$$\frac{1}{2}\left(\frac{1}{\tau} - \left(1 - \frac{1}{\alpha}\right)\frac{\delta\mu^2}{2}\right)\|x^{(t)} - x^\star\|^2 + \frac{1}{2\sigma}\|y^{(t)} - y^\star\|^2 + \theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)}) + \frac{\sigma\theta^2 L^2}{2}\|x^{(t)} - x^{(t-1)}\|^2$$

$$\geq \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)})$$

$$+ \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^{(t+1)} - x^\star\|^2 + \frac{1}{2}\left(\frac{1}{\sigma} + \frac{\gamma}{2}\right)\|y^{(t+1)} - y^\star\|^2 + (y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)}) + \frac{1}{2\tau}\|x^{(t+1)} - x^{(t)}\|^2$$

$$+ \left(\frac{\sigma}{2} - (\alpha - 1)\frac{\delta}{4}\right)\left\|\theta A(x^{(t)} - x^{(t-1)}) - \frac{1}{\sigma}(y^{(t+1)} - y^{(t)})\right\|^2. \tag{30}$$

We can remove the last term in the above inequality as long as its coefficient is nonnegative, i.e.,

$$\frac{\sigma}{2} - (\alpha - 1)\frac{\delta}{4} \geq 0.$$

In order to maximize $1 - 1/\alpha$, we take the equality and solve for the largest value of α allowed, which results in

$$\alpha = 1 + \frac{2\sigma}{\delta}, \qquad 1 - \frac{1}{\alpha} = \frac{2\sigma}{2\sigma + \delta}.$$

Applying these values in (30) gives

$$\frac{1}{2}\left(\frac{1}{\tau} - \frac{\sigma\delta\mu^2}{2\sigma + \delta}\right)\|x^{(t)} - x^\star\|^2 + \frac{1}{2\sigma}\|y^{(t)} - y^\star\|^2 + \theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)}) + \frac{\sigma\theta^2 L^2}{2}\|x^{(t)} - x^{(t-1)}\|^2$$

$$\geq \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)})$$

$$+ \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^{(t+1)} - x^\star\|^2 + \frac{1}{2}\left(\frac{1}{\sigma} + \frac{\gamma}{2}\right)\|y^{(t+1)} - y^\star\|^2 + (y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)}) + \frac{1}{2\tau}\|x^{(t+1)} - x^{(t)}\|^2. \tag{31}$$

14

We use $\Delta^{(t+1)}$ to denote the last row in (31). Equivalently, we define

$$
\begin{aligned}
\Delta^{(t)} \;=\;& \frac{1}{2}\Big(\frac{1}{\tau}+\lambda\Big)\|x^\star - x^{(t)}\|^2 + \frac{1}{2}\Big(\frac{1}{\sigma}+\frac{\gamma}{2}\Big)\|y^\star - y^{(t)}\|^2 + (y^{(t)}-y^\star)^T A(x^{(t)} - x^{(t-1)}) + \frac{1}{2\tau}\|x^{(t)}-x^{(t-1)}\|^2 \\
\;=\;& \frac{1}{2}\Big(\frac{1}{\tau}+\lambda\Big)\|x^\star - x^{(t)}\|^2 + \frac{\gamma}{4}\|y^\star - y^{(t)}\|^2 + \frac{1}{2}\left[\begin{array}{c} x^{(t)}-x^{(t-1)} \\ y^\star - y^{(t)} \end{array}\right]^T \left[\begin{array}{cc} \frac{1}{\tau}I & -A^T \\ -A & \frac{1}{\sigma} \end{array}\right]\left[\begin{array}{c} x^{(t)}-x^{(t-1)} \\ y^\star - y^{(t)} \end{array}\right].
\end{aligned}
$$

The quadratic form in the last term is nonnegative if the matrix

$$
M = \left[\begin{array}{cc} \frac{1}{\tau}I & -A^T \\ -A & \frac{1}{\sigma} \end{array}\right]
$$

is positive semidefinite, for which a sufficient condition is $\tau\sigma \le 1/L^2$. Under this condition,

$$
\Delta^{(t)} \;\ge\; \frac{1}{2}\Big(\frac{1}{\tau}+\lambda\Big)\|x^\star - x^{(t)}\|^2 + \frac{\gamma}{4}\|y^\star - y^{(t)}\|^2 \;\ge\; 0. \tag{32}
$$

If we can to choose τ and σ so that

$$
\frac{1}{\tau} - \frac{\sigma\delta\mu^2}{2\sigma+\delta} \le \theta\Big(\frac{1}{\tau}+\lambda\Big), \qquad \frac{1}{\sigma} \le \theta\Big(\frac{1}{\sigma}+\frac{\gamma}{2}\Big), \qquad \frac{\sigma\theta^2 L^2}{2} \le \theta\frac{1}{2\tau}, \tag{33}
$$

then, according to (31), we have

$$
\Delta^{(t+1)} + \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)}) \;\le\; \theta\Delta^{(t)}.
$$

Because $\Delta^{(t)} \ge 0$ and $\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^{(t)}) \ge 0$ for any $t \ge 0$, we have

$$
\Delta^{(t+1)} \;\le\; \theta\Delta^{(t)},
$$

which implies

$$
\Delta^{(t)} \;\le\; \theta^t \Delta^{(0)}
$$

and

$$
\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^{(t)}) \;\le\; \theta^t \Delta^{(0)}.
$$

Let θ_x and θ_y be two contraction factors determined by the first two inequalities in (33), i.e.,

$$
\begin{aligned}
\theta_x \;&=\; \Big(\frac{1}{\tau} - \frac{\sigma\delta\mu^2}{2\sigma+\delta}\Big)\Big/\Big(\frac{1}{\tau}+\lambda\Big) \;=\; \Big(1 - \frac{\tau\sigma\delta\mu^2}{2\sigma+\delta}\Big)\frac{1}{1+\tau\lambda}, \\
\theta_y \;&=\; \frac{1}{\sigma}\Big/\Big(\frac{1}{\sigma}+\frac{\gamma}{2}\Big) \;=\; \frac{1}{1+\sigma\gamma/2}.
\end{aligned}
$$

Then we can let $\theta = \max\{\theta_x, \theta_y\}$. We note that any $\theta < 1$ would satisfy the last condition in (33) provided that

$$
\tau\sigma \;=\; \frac{1}{L^2},
$$

which also makes the matrix M positive semidefinite and thus ensures the inequality (32).

Among all possible pairs τ, σ that satisfy $\tau\sigma = 1/L^2$, we choose

$$
\tau = \frac{1}{L}\sqrt{\frac{\gamma}{\lambda+\delta\mu^2}}, \qquad \sigma = \frac{1}{L}\sqrt{\frac{\lambda+\delta\mu^2}{\gamma}}, \tag{34}
$$

which give the desired results of Theorem 1.

D. Proof of Theorem 3

If we choose $h = f^*$, then

- h is γ-strongly convex and $1/\delta$-smooth, i.e., $\gamma' = \gamma$ and $\delta' = \delta$;

- f^* is 1-strongly convex with respect to h, i.e., $\nu = 1$.

For convenience, we repeat inequality (28) here:

$$
\frac{1}{2\tau}\|x^\star - x^{(t)}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^\star, y^{(t)}) + \theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})
$$
$$
\geq \quad \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)})
$$
$$
+ \frac{1}{2}\Big(\frac{1}{\tau} + \lambda\Big)\|x^\star - x^{(t+1)}\|^2 + \Big(\frac{1}{\sigma} + \frac{\nu}{2}\Big)\mathcal{D}(y^\star, y^{(t+1)}) + (y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})
$$
$$
+ \frac{1}{2\tau}\|x^{(t+1)} - x^{(t)}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^{(t+1)}, y^{(t)}) - \theta(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)})
$$
$$
+ \Big(1 - \frac{1}{\alpha}\Big)\frac{\delta\mu^2}{4}\|x^\star - x^{(t)}\|^2 - (\alpha - 1)\frac{\delta}{4}\Big\|\theta A(x^{(t)} - x^{(t-1)})) - \frac{1}{\sigma}\big(\nabla h(y^{(t+1)}) - \nabla h(y^{(t)})\big)\Big\|^2. \quad (35)
$$

We first bound the Bregman divergence $\mathcal{D}(y^{(t+1)}, y^{(t)})$ using the assumption that the kernel h is γ-strongly convex and $1/\delta$-smooth. Using similar arguments as in the proof of Lemma 2, we have for any $\rho \in [0, 1]$,

$$
D(y^{(t+1)}, y^{(t)}) \quad = \quad h(y^{(t+1)}) - h(y^{(t)}) - \langle \nabla h(y^{(t)}), y^{(t+1)} - y^{(t)}\rangle
$$
$$
\geq \quad (1 - \rho)\frac{\gamma}{2}\|y^{(t+1)} - y^{(t)}\|^2 + \rho\frac{\delta}{2}\|\nabla h(y^{(t+1)}) - \nabla h(y^{(t)})\|^2. \quad (36)
$$

For any $\beta > 0$, we can lower bound the inner product term

$$
-\theta(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)}) \geq -\frac{\beta}{2}\|y^{(t+1)} - y^{(t)}\|^2 - \frac{\theta^2 L^2}{2\beta}\|x^{(t)} - x^{(t-1)}\|^2.
$$

In addition, we have

$$
\Big\|\theta A(x^{(t)} - x^{(t-1)})) - \frac{1}{\sigma}\big(\nabla h(y^{(t+1)}) - \nabla h(y^{(t)})\big)\Big\|^2 \leq 2\theta^2 L^2\|x^{(t)} - x^{(t-1)}\|^2 + \frac{2}{\sigma^2}\|\nabla h(y^{(t+1)}) - \nabla h(y^{(t)})\|^2.
$$

Combining these bounds with (35) and (36) with $\rho = 1/2$, we arrive at

$$
\frac{1}{2}\Big(\frac{1}{\tau} - \Big(1 - \frac{1}{\alpha}\Big)\frac{\delta\mu^2}{2}\Big)\|x^\star - x^{(t)}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^\star, y^{(t)}) + \theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})
$$
$$
+ \Big(\frac{\theta^2 L^2}{2\beta} + (\alpha - 1)\frac{\delta\theta^2 L^2}{2}\Big)\|x^{(t)} - x^{(t-1)}\|^2
$$
$$
\geq \quad \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)})
$$
$$
+ \frac{1}{2}\Big(\frac{1}{\tau} + \lambda\Big)\|x^\star - x^{(t+1)}\|^2 + \Big(\frac{1}{\sigma} + \frac{1}{2}\Big)\mathcal{D}(y^\star, y^{(t+1)}) + (y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})
$$
$$
+ \Big(\frac{\gamma}{4\sigma} - \frac{\beta}{2}\Big)\|y^{(t+1)} - y^{(t)}\|^2 + \Big(\frac{\delta}{4\sigma} - \frac{(\alpha - 1)\delta}{2\sigma^2}\Big)\|\nabla h(y^{(t+1)}) - \nabla h(y^{(t)})\|^2
$$
$$
+ \frac{1}{2\tau}\|x^{(t+1)} - x^{(t)}\|^2. \quad (37)
$$

We choose α and β in (37) to zero out the coefficients of $\|y^{(t+1)} - y^{(t)}\|^2$ and $\|\nabla h(y^{(t+1)}) - \nabla h(y^{(t)})\|^2$:

$$
\alpha = 1 + \frac{\sigma}{2}, \qquad \beta = \frac{\gamma}{2\sigma}.
$$

Then the inequality (37) becomes

$$\frac{1}{2}\left(\frac{1}{\tau} - \frac{\sigma\delta\mu^2}{4+2\sigma}\right)\|x^\star - x^{(t)}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^\star, y^{(t)}) + \theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})$$

$$+ \left(\frac{\sigma\theta^2 L^2}{\gamma} + \frac{\delta\sigma\theta^2 L^2}{4}\right)\|x^{(t)} - x^{(t-1)}\|^2$$

$$\geq \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)})$$

$$+ \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^\star - x^{(t+1)}\|^2 + \left(\frac{1}{\sigma} + \frac{1}{2}\right)\mathcal{D}(y^\star, y^{(t+1)}) + (y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})$$

$$+ \frac{1}{2\tau}\|x^{(t+1)} - x^{(t)}\|^2.$$

The coefficient of $\|x^{(t)} - x^{(t-1)}\|^2$ can be bounded as

$$\frac{\sigma\theta^2 L^2}{\gamma} + \frac{\delta\sigma\theta^2 L^2}{4} = \left(\frac{1}{\gamma} + \frac{\delta}{4}\right)\sigma\theta^2 L^2 = \frac{4+\gamma\delta}{4\gamma}\sigma\theta^2 L^2 < \frac{2\sigma\theta^2 L^2}{\gamma},$$

where in the inequality we used $\gamma\delta \leq 1$. Therefore we have

$$\frac{1}{2}\left(\frac{1}{\tau} - \frac{\sigma\delta\mu^2}{4+2\sigma}\right)\|x^\star - x^{(t)}\|^2 + \frac{1}{\sigma}\mathcal{D}(y^\star, y^{(t)}) + \theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)}) + \frac{2\sigma\theta^2 L^2}{\gamma}\|x^{(t)} - x^{(t-1)}\|^2$$

$$\geq \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)})$$

$$+ \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^\star - x^{(t+1)}\|^2 + \left(\frac{1}{\sigma} + \frac{1}{2}\right)\mathcal{D}(y^\star, y^{(t+1)}) + (y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)}) + \frac{1}{2\tau}\|x^{(t+1)} - x^{(t)}\|^2.$$

We use $\Delta^{(t+1)}$ to denote the last row of the above inequality. Equivalently, we define

$$\Delta^{(t)} = \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^\star - x^{(t)}\|^2 + \left(\frac{1}{\sigma} + \frac{1}{2}\right)\mathcal{D}(y^\star, y^{(t)}) + (y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)}) + \frac{1}{2\tau}\|x^{(t)} - x^{(t-1)}\|^2.$$

Since h is γ-strongly convex, we have $\mathcal{D}(y^\star, y^{(t)}) \geq \frac{\gamma}{2}\|y^\star - y^{(t)}\|^2$, and thus

$$\Delta^{(t)} \geq \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^\star - x^{(t)}\|^2 + \frac{1}{2}\mathcal{D}(y^\star, y^{(t)}) + \frac{\gamma}{2\sigma}\|y^{(t)} - y^\star\|^2 + (y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)}) + \frac{1}{2\tau}\|x^{(t)} - x^{(t-1)}\|^2$$

$$= \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^\star - x^{(t)}\|^2 + \frac{1}{2}\mathcal{D}(y^\star, y^{(t)}) + \frac{1}{2}\begin{bmatrix} x^{(t)} - x^{(t-1)} \\ y^\star - y^{(t)} \end{bmatrix}^T \begin{bmatrix} \frac{1}{\tau}I & -A^T \\ -A & \frac{\gamma}{\sigma} \end{bmatrix}\begin{bmatrix} x^{(t)} - x^{(t-1)} \\ y^\star - y^{(t)} \end{bmatrix}.$$

The quadratic form in the last term is nonnegative if $\tau\sigma \leq \gamma/L^2$. Under this condition,

$$\Delta^{(t)} \geq \frac{1}{2}\left(\frac{1}{\tau} + \lambda\right)\|x^\star - x^{(t)}\|^2 + \frac{1}{2}\mathcal{D}(y^\star, y^{(t)}) \geq 0. \tag{38}$$

If we can to choose τ and σ so that

$$\frac{1}{\tau} - \frac{\sigma\delta\mu^2}{4+2\sigma} \leq \theta\left(\frac{1}{\tau} + \lambda\right), \qquad \frac{1}{\sigma} \leq \theta\left(\frac{1}{\sigma} + \frac{1}{2}\right), \qquad \frac{2\sigma\theta^2 L^2}{\gamma} \leq \theta\frac{1}{2\tau}, \tag{39}$$

then we have

$$\Delta^{(t+1)} + \mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^{(t+1)}) \leq \theta\Delta^{(t)}.$$

Because $\Delta^{(t)} \geq 0$ and $\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^{(t)}) \geq 0$ for any $t \geq 0$, we have

$$\Delta^{(t+1)} \leq \theta\Delta^{(t)},$$

which implies

$$\Delta^{(t)} \leq \theta^t \Delta^{(0)}$$

and

$$\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^{(t)}) \leq \theta^t \Delta^{(0)}.$$

To satisfy the last condition in (39) and also ensure the inequality (38), it suffices to have

$$\tau\sigma \leq \frac{\gamma}{4L^2}.$$

We choose

$$\tau = \frac{1}{2L}\sqrt{\frac{\gamma}{\lambda + \delta\mu^2}}, \qquad \sigma = \frac{1}{2L}\sqrt{\gamma(\lambda + \delta\mu^2)}.$$

With the above choice and assuming $\gamma(\lambda + \delta\mu^2) \ll L^2$, we have

$$\theta_y = \frac{\frac{1}{\sigma}}{\frac{1}{\sigma} + \frac{1}{2}} = \frac{1}{1 + \sigma/2} = \frac{1}{1 + \sqrt{\gamma(\lambda + \delta\mu^2)}/(4L)} \approx 1 - \frac{\sqrt{\gamma(\lambda + \delta\mu^2)}}{4L}.$$

For the contraction factor over the primal variables, we have

$$\theta_x = \frac{\frac{1}{\tau} - \frac{\sigma\delta\mu^2}{4 + 2\sigma}}{\frac{1}{\tau} + \lambda} = \frac{1 - \frac{\tau\sigma\delta\mu^2}{4 + 2\sigma}}{1 + \tau\lambda} = \frac{1 - \frac{\gamma\delta\mu^2}{4(4 + 2\sigma)L^2}}{1 + \tau\lambda} \approx 1 - \frac{\gamma\delta\mu^2}{16L^2} - \frac{\lambda}{2L}\sqrt{\frac{\gamma}{\lambda + \delta\mu^2}}.$$

This finishes the proof of Theorem 3.

E. Proof of Theorem 2

We consider the SPDC algorithm in the Euclidean case with $h(x) = (1/2)\|x\|^2$. The corresponding batch case analysis is given in Section C. For each $i = 1, \ldots, n$, let $\tilde{y}_i$ be

$$\tilde{y}_i = \arg\min_y \left\{ \phi_i^*(y) + \frac{(y - y_i^{(t)})^2}{2\sigma} - y\langle a_i, \tilde{x}^{(t)}\rangle \right\}.$$

Based on the first-order optimality condition, we have

$$\langle a_i, \tilde{x}^{(t)}\rangle - \frac{(\tilde{y}_i - y_i^{(t)})}{\sigma} \in \phi_i^{*'}(\tilde{y}_i).$$

Also, since $y_i^\star$ minimizes $\phi_i^*(y) - y\langle a_i, x^*\rangle$, we have

$$\langle a_i, x^*\rangle \in \phi_i^{*'}(y_i^\star).$$

By Lemma 2 with $\rho = 1/2$, we have

$$-y_i^\star\langle a_i, \tilde{x}^{(t)}\rangle + \phi_i^*(y_i^\star) + \frac{(y_i^{(t)} - y_i^\star)^2}{2\sigma} \geq \left(\frac{1}{\sigma} + \frac{\gamma}{2}\right)\frac{(\tilde{y}_i - y_i^\star)^2}{2} + \phi_i^*(\tilde{y}_i) - \tilde{y}_i\langle a_i, \tilde{x}^{(t)}\rangle$$
$$+ \frac{(\tilde{y}_i - y_i^{(t)})^2}{2\sigma} + \frac{\delta}{4}(\phi_i^{*'}(\tilde{y}_i) - \phi_i^{*'}(y_i^\star))^2,$$

and re-arranging terms, we get

$$\frac{(y_i^{(t)} - y_i^\star)^2}{2\sigma} \geq \left(\frac{1}{\sigma} + \frac{\gamma}{2}\right)\frac{(\tilde{y}_i - y_i^\star)^2}{2} + \frac{(\tilde{y}_i - y_i^{(t)})^2}{2\sigma} - (\tilde{y}_i - y_i^\star)\langle a_i, \tilde{x}^{(t)}\rangle + (\phi_i^*(\tilde{y}_i) - \phi_i^*(y_i^\star))$$
$$+ \frac{\delta}{4}(\phi_i^{*'}(\tilde{y}_i) - \phi_i^{*'}(y_i^\star))^2. \tag{40}$$

Notice that

$$\mathbb{E}[y_i^{(t+1)}] = \frac{1}{n} \cdot \tilde{y}_i + \frac{n-1}{n} \cdot y_i^{(t)},$$

$$\mathbb{E}[(y_i^{(t+1)} - y_i^\star)^2] = \frac{(\tilde{y}_i - y_i^\star)^2}{n} + \frac{(n-1)(y_i^{(t)} - y_i^\star)^2}{n},$$

$$\mathbb{E}[(y_i^{(t+1)} - y_i^{(t)})^2] = \frac{(\tilde{y}_i - y_i^{(t)})^2}{n},$$

$$\mathbb{E}[\phi_i^*(y_i^{(t+1)})] = \frac{1}{n} \cdot \phi_i^*(\tilde{y}_i) + \frac{n-1}{n} \cdot \phi_i^*(y_i^{(t)}).$$

Plug the above relations into (40) and divide both sides by n, we have

$$\left(\frac{1}{2\sigma} + \frac{(n-1)\gamma}{4n}\right)(y_i^{(t)} - y_i^\star)^2 \geq \left(\frac{1}{2\sigma} + \frac{\gamma}{4}\right)\mathbb{E}[(y_i^{(t+1)} - y_i^\star)^2] + \frac{1}{2\sigma}\mathbb{E}[(y_i^{(t+1)} - y_i^{(t)})^2]$$

$$- \left(\mathbb{E}[(y_i^{(t+1)} - y_i^{(t)})] + \frac{1}{n}(y_i^{(t)} - y_i^\star)\right)\langle a_i, \tilde{x}^{(t)}\rangle$$

$$+ \mathbb{E}[\phi_i^*(y_i^{(t+1)})] - \phi_i^*(y_i^{(t)}) + \frac{1}{n}(\phi_i^*(y_i^{(t)}) - \phi_i^*(y_i^\star))$$

$$+ \frac{\delta}{4n}\left(\langle a_i, \tilde{x}^{(t)} - x^\star\rangle - \frac{(\tilde{y}_i - y_i^{(t)})}{\sigma}\right)^2,$$

and summing over $i = 1, \ldots, n$, we get

$$\left(\frac{1}{2\sigma} + \frac{(n-1)\gamma}{4n}\right)\|y^{(t)} - y^\star\|^2 \geq \left(\frac{1}{2\sigma} + \frac{\gamma}{4}\right)\mathbb{E}[\|y^{(t+1)} - y^\star\|^2] + \frac{\mathbb{E}[\|y^{(t+1)} - y^{(t)}\|^2]}{2\sigma}$$

$$+ \phi_k^*(y_k^{(t+1)}) - \phi_k^*(y_k^{(t)}) + \frac{1}{n}\sum_{i=1}^n (\phi_i^*(y_i^{(t)}) - \phi_i^*(y_i^\star))$$

$$- \left\langle n(u^{(t+1)} - u^{(t)}) + (u^{(t)} - u^*), \tilde{x}^{(t)}\right\rangle$$

$$+ \frac{\delta}{4n}\left\|A(x^* - \tilde{x}^{(t)}) + \frac{(\tilde{y} - y^{(t)})}{\sigma}\right\|^2,$$

where

$$u^{(t)} = \frac{1}{n}\sum_{i=1}^n y_i^{(t)} a_i, \quad u^{(t+1)} = \frac{1}{n}\sum_{i=1}^n y_i^{(t+1)} a_i, \quad \text{and} \quad u^* = \frac{1}{n}\sum_{i=1}^n y_i^\star a_i.$$

On the other hand, since $x^{(t+1)}$ minimizes the $\frac{1}{\tau} + \lambda$-strongly convex objective

$$g(x) + \left\langle u^{(t)} + n(u^{(t+1)} - u^{(t)}), x\right\rangle + \frac{\|x - x^{(t)}\|^2}{2\tau},$$

we can apply Lemma 2 with $\rho = 0$ to obtain

$$g(x^\star) + \langle u^{(t)} + n(u^{(t+1)} - u^{(t)}), x^\star\rangle + \frac{\|x^{(t)} - x^\star\|^2}{2\tau}$$

$$\geq g(x^{(t+1)}) + \langle u^{(t)} + n(u^{(t+1)} - u^{(t)}), x^{(t+1)}\rangle + \frac{\|x^{(t+1)} - x^{(t)}\|^2}{2\tau} + \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\|x^{(t+1)} - x^\star\|^2,$$

and re-arranging terms we get

$$\frac{\|x^{(t)} - x^\star\|^2}{2\tau} \geq \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\mathbb{E}[\|x^{(t+1)} - x^\star\|^2] + \frac{\mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2]}{2\tau} + \mathbb{E}[g(x^{(t+1)}) - g(x^\star)]$$

$$+ \mathbb{E}[\langle u^{(t)} + n(u^{(t+1)} - u^{(t)}), x^{(t+1)} - x^\star\rangle].$$

Also notice that

$$\mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^\star) + n(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t+1)})) - (n-1)(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t)}))$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\phi_i^*(y_i^{(t)}) - \phi_i^*(y^\star)) + (\phi_k^*(y_k^{(t+1)}) - \phi_k^*(y_k^{(t)})) + g(x^{(t+1)}) - g(x^\star)$$

$$+ \langle u^\star, x^{(t+1)} \rangle - \langle u^{(t)}, x^\star \rangle + n\langle u^{(t)} - u^{(t+1)}, x^\star \rangle.$$

Combining everything together, we have

$$\frac{\|x^{(t)} - x^\star\|^2}{2\tau} + \left(\frac{1}{2\sigma} + \frac{(n-1)\gamma}{4n}\right)\|y^{(t)} - y^\star\|^2 + (n-1)(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t)}))$$

$$\geq \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\mathbb{E}[\|x^{(t+1)} - x^\star\|^2] + \left(\frac{1}{2\sigma} + \frac{\gamma}{4}\right)\mathbb{E}[\|y^{(t+1)} - y^\star\|^2] + \frac{\mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2]}{2\tau} + \frac{\mathbb{E}[\|y^{(t+1)} - y^{(t)}\|^2]}{2\sigma}$$

$$+ \mathbb{E}[\mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^\star) + n(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t+1)}))]$$

$$+ \mathbb{E}[\langle u^{(t)} - u^* + n(u^{(t+1)} - u^{(t)}), x^{(t+1)} - \bar{x}^t \rangle] + \frac{\delta}{4n}\left\|A(x^* - \tilde{x}^{(t)}) + \frac{(\tilde{y} - y^{(t)})}{\sigma}\right\|^2.$$

Next we notice that

$$\frac{\delta}{4n}\left\|A(x^* - \tilde{x}^{(t)}) + \frac{n(\mathbb{E}[y^{(t+1)}] - y^{(t)})}{\sigma}\right\|^2 = \frac{\delta}{4n}\left\|A(x^* - x^{(t)}) - \theta A(x^{(t)} - x^{(t-1)}) + \frac{(\tilde{y} - y^{(t)})}{\sigma}\right\|^2$$

$$\geq \left(1 - \frac{1}{\alpha}\right)\frac{\delta}{4n}\left\|A(x^\star - x^{(t)})\right\|^2$$

$$- (\alpha - 1)\frac{\delta}{4n}\left\|\theta A(x^{(t)} - x^{(t-1)}) + \frac{(\tilde{y} - y^{(t)})}{\sigma}\right\|^2,$$

for some $\alpha > 1$ and

$$\left\|A(x^* - x^{(t)})\right\|^2 \geq \mu^2\|x^\star - x^{(t)}\|^2,$$

and

$$\left\|\theta A(x^{(t)} - x^{(t-1)}) + \frac{(\tilde{y} - y^{(t)})}{\sigma}\right\|^2 \geq -2\theta^2\|A(x^{(t)} - x^{(t-1)})\|^2 - \frac{2}{\sigma^2}\|\tilde{y} - y^{(t)}\|^2$$

$$\geq -2\theta^2 L^2\|x^{(t)} - x^{(t-1)}\|^2 - \frac{2n}{\sigma^2}\mathbb{E}[\|y^{(t+1)} - y^{(t)}\|^2].$$

We follow the same reasoning as in the standard SPDC analysis,

$$\langle u^{(t)} - u^* + n(u^{(t+1)} - u^{(t)}), x^{(t+1)} - \tilde{x}^{(t)} \rangle = \frac{(y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})}{n} - \frac{\theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})}{n}$$

$$+ \frac{(n-1)}{n}(y^{(t+1)} - y^{(t)})^T A(x^{(t+1)} - x^{(t)}) - \theta(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)}),$$

and using Cauchy-Schwartz inequality, we have

$$|(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)})| \leq \frac{\|(y^{(t+1)} - y^{(t)})^T A\|^2}{1/(2\tau)} + \frac{\|x^{(t)} - x^{(t-1)}\|^2}{8\tau}$$

$$\leq \frac{\|y^{(t+1)} - y^{(t)}\|^2}{1/(2\tau R^2)},$$

and

$$|(y^{(t+1)} - y^{(t)})^T A(x^{(t+1)} - x^{(t)})| \leq \frac{\|(y^{(t+1)} - y^{(t)})^T A\|^2}{1/(2\tau)} + \frac{\|x^{(t+1)} - x^{(t)}\|^2}{8\tau}$$

$$\leq \frac{\|y^{(t+1)} - y^{(t)}\|^2}{1/(2\tau R^2)}.$$

Thus we get

$$\langle u^{(t)} - u^* + n(u^{(t+1)} - u^{(t)}), x^{(t+1)} - \tilde{x}^{(t)} \rangle \geq \frac{(y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})}{n} - \frac{\theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})}{n}$$
$$- \frac{\|y^{(t+1)} - y^{(t)}\|^2}{1/(4\tau R^2)} - \frac{\|x^{(t+1)} - x^{(t)}\|^2}{8\tau} - \frac{\theta\|x^{(t)} - x^{(t-1)}\|^2}{8\tau}.$$

Putting everything together, we have

$$\left(\frac{1}{2\tau} - \frac{(1-1/\alpha)\delta\mu^2}{4n}\right)\|x^{(t)} - x^\star\|^2 + \left(\frac{1}{2\sigma} + \frac{(n-1)\gamma}{4n}\right)\|y^{(t)} - y^\star\|^2 + \theta(\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^\star))$$

$$+ (n-1)(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t)})) + \theta\left(\frac{1}{8\tau} + \frac{(\alpha-1)\theta\delta L^2}{2n}\right)\|x^{(t)} - x^{t-1}\|^2 + \frac{\theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})}{n}$$

$$\geq \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\mathbb{E}[\|x^{(t+1)} - x^\star\|^2] + \left(\frac{1}{2\sigma} + \frac{\gamma}{4}\right)\mathbb{E}[\|y^{(t+1)} - y^\star\|^2] + \frac{\mathbb{E}[(y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})]}{n}$$

$$+ \mathbb{E}[\mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^\star) + n(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t+1)}))]$$

$$+ \left(\frac{1}{2\tau} - \frac{1}{8\tau}\right)\mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2]$$

$$+ \left(\frac{1}{2\sigma} - 4R^2\tau - \frac{(\alpha-1)\delta}{2\sigma^2}\right)\mathbb{E}[\|y^{(t+1)} - y^{(t)}\|^2].$$

If we choose the parameters as

$$\alpha = \frac{\sigma}{4\delta} + 1, \quad \sigma\tau = \frac{1}{16R^2},$$

then we know

$$\frac{1}{2\sigma} - 4R^2\tau - \frac{(\alpha-1)\delta}{2\sigma^2} = \frac{1}{2\sigma} - \frac{1}{4\sigma} - \frac{1}{8\sigma} > 0,$$

and

$$\frac{(\alpha-1)\theta\delta L^2}{2n} \leq \frac{\sigma L^2}{8n^2} \leq \frac{\sigma R^2}{8} \leq \frac{1}{256\tau},$$

thus

$$\frac{1}{8\tau} + \frac{(\alpha-1)\theta\delta L^2}{2n} \leq \frac{3}{8\tau}.$$

In addition, we have

$$1 - \frac{1}{\alpha} = \frac{\sigma}{\sigma + 4\delta}.$$

Finally we obtain

$$\left(\frac{1}{2\tau} - \frac{\sigma\delta\mu^2}{4n(\sigma + 4\delta)}\right)\|x^{(t)} - x^\star\|^2 + \left(\frac{1}{2\sigma} + \frac{(n-1)\gamma}{4n}\right)\|y^{(t)} - y^\star\|^2 + \theta(\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^\star))$$

$$+ (n-1)(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t)})) + \theta \cdot \frac{3}{8\tau}\|x^{(t)} - x^{(t-1)}\|^2 + \frac{\theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})}{n}$$

$$\geq \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\mathbb{E}[\|x^{(t+1)} - x^\star\|^2] + \left(\frac{1}{2\sigma} + \frac{\gamma}{4}\right)\mathbb{E}[\|y^{(t+1)} - y^\star\|^2] + \frac{\mathbb{E}[(y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})]}{n}$$

$$+ \mathbb{E}[\mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^\star) + n(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t+1)}))] + \frac{3}{8\tau}\mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2].$$

Now we can define θ_x and θ_y as the ratios between the coefficients in the x-distance and y-distance terms, and let $\theta = \max\{\theta_x, \theta_y\}$ as before. Choosing the step-size parameters as

$$\tau = \frac{1}{4R}\sqrt{\frac{\gamma}{n\lambda + \delta\mu^2}}, \quad \sigma = \frac{1}{4R}\sqrt{\frac{n\lambda + \delta\mu^2}{\gamma}}$$

gives the desired result.

F. Proof of Theorem 4

In this setting, for i-th coordinate of the dual variables y we choose $h = \phi_i^*$, let

$$\mathcal{D}_i(y_i, y_i') = \phi_i^*(y_i) - \phi_i^*(y_i') + \langle (\phi_i^*)'(y_i'), y_i - y_i' \rangle,$$

and define

$$\mathcal{D}(y, y') = \sum_{i=1}^{n} \mathcal{D}_i(y_i, y_i').$$

For $i = 1, \ldots, n$, let $\tilde{y}_i$ be

$$\tilde{y}_i = \arg\min_y \left\{ \phi_i^*(y) + \frac{\mathcal{D}_i(y, y_i^{(t)})}{\sigma} - y \langle a_i, \tilde{x}^{(t)} \rangle \right\}.$$

Based on the first-order optimality condition, we have

$$\langle a_i, \tilde{x}^{(t)} \rangle - \frac{(\phi_i^*)'(\tilde{y}_i) - (\phi_i^*)'(y_i^{(t)})}{\sigma} \in (\phi_i^*)'(\tilde{y}_i).$$

Also since $y_i^\star$ minimizes $\phi_i^*(y) - y \langle a_i, x^* \rangle$, we have

$$\langle a_i, x^* \rangle \in (\phi_i^*)'(y_i^\star).$$

Using Lemma 2 with $\rho = 1/2$, we obtain

$$-y_i^\star \langle a_i, \tilde{x}^{(t)} \rangle + \phi_i^*(y_i^\star) + \frac{\mathcal{D}_i(y_i^\star, y_i^{(t)})}{\sigma} \geq \left(\frac{1}{\sigma} + \frac{1}{2} \right) \mathcal{D}_i(y_i^\star, \tilde{y}_i) + \phi_i^*(\tilde{y}_i) - \tilde{y}_i \langle a_i, \tilde{x}^{(t)} \rangle$$

$$+ \frac{\mathcal{D}_i(\tilde{y}_i, y_i^{(t)})}{\sigma} + \frac{\delta}{4} ((\phi_i^*)'(\tilde{y}_i) - (\phi_i^*)'(y_i^\star))^2,$$

and rearranging terms, we get

$$\frac{\mathcal{D}_i(y_i^\star, y_i^{(t)})}{\sigma} \geq \left(\frac{1}{\sigma} + \frac{1}{2} \right) \mathcal{D}_i(y_i^\star, \tilde{y}_i) + \frac{\mathcal{D}_i(\tilde{y}_i, y_i^{(t)})}{\sigma} - (\tilde{y}_i - y_i^\star) \langle a_i, \tilde{x}^{(t)} \rangle + (\phi_i^*(\tilde{y}_i) - \phi_i^*(y_i^\star))$$

$$+ \frac{\delta}{4} ((\phi_i^*)'(\tilde{y}_i) - (\phi_i^*)'(y_i^\star))^2. \tag{41}$$

With i.i.d. random sampling at each iteration, we have the following relations:

$$\mathbb{E}[y_i^{(t+1)}] = \frac{1}{n} \cdot \tilde{y}_i + \frac{n-1}{n} \cdot y_i^{(t)},$$

$$\mathbb{E}[\mathcal{D}_i(y_i^{(t+1)}, y_i^\star)] = \frac{\mathcal{D}_i(\tilde{y}_i, y_i^\star)}{n} + \frac{(n-1)\mathcal{D}_i(y_i^{(t)}, y_i^\star)}{n},$$

$$\mathbb{E}[\mathcal{D}_i(y_i^{(t+1)}, y_i^{(t)})] = \frac{\mathcal{D}_i(\tilde{y}_i, y_i^{(t)})}{n},$$

$$\mathbb{E}[\phi_i^*(y_i^{(t+1)})] = \frac{1}{n} \cdot \phi_i^*(\tilde{y}_i) + \frac{n-1}{n} \cdot \phi_i^*(y_i^{(t)}).$$

Plugging the above relations into (41) and dividing both sides by n, we have

$$\left(\frac{1}{\sigma} + \frac{(n-1)}{2n} \right) \mathcal{D}_i(y_i^{(t)}, y_i^\star) \geq \left(\frac{1}{\sigma} + \frac{1}{2} \right) \mathcal{D}_i(y_i^{(t+1)}, y_i^\star) + \frac{1}{\sigma} \mathbb{E}[\mathcal{D}_i(y_i^{(t+1)}, y_i^{(t)})]$$

$$- \left(\mathbb{E}[(y_i^{(t+1)} - y_i^{(t)})] + \frac{1}{n} (y_i^{(t)} - y_i^\star) \right) \langle a_i, \tilde{x}^{(t)} \rangle$$

$$+ \mathbb{E}[\phi_i^*(y_i^{(t+1)})] - \phi_i^*(y_i^{(t)}) + \frac{1}{n} (\phi_i^*(y_i^{(t)}) - \phi_i^*(y_i^\star))$$

$$+ \frac{\delta}{4n} \left(\langle a_i, \tilde{x}^{(t)} - x^\star \rangle - \frac{((\phi_i^*)'(\tilde{y}_i) - (\phi_i^*)'(y_i^{(t)}))}{\sigma} \right)^2,$$

and summing over $i = 1, \ldots, n$, we get

$$\left(\frac{1}{\sigma} + \frac{(n-1)}{2n}\right)\mathcal{D}(y^{(t)}, y^\star) \geq \left(\frac{1}{\sigma} + \frac{1}{2}\right)\mathbb{E}[\mathcal{D}(y^{(t+1)}, y^\star)] + \frac{\mathbb{E}[\mathcal{D}(y^{(t+1)}, y^{(t)})]}{\sigma}$$

$$+ \phi_k^*(y_k^{(t+1)}) - \phi_k^*(y_k^{(t)}) + \frac{1}{n}\sum_{i=1}^{n}(\phi_i^*(y_i^{(t)}) - \phi_i^*(y_i^\star))$$

$$- \left\langle n(u^{(t+1)} - u^{(t)}) + (u^{(t)} - u^*), \tilde{x}^{(t)} \right\rangle$$

$$+ \frac{\delta}{4n}\left\| A(x^* - \tilde{x}^{(t)}) + \frac{(\phi^{*'}(\tilde{y}) - \phi^{*'}(y^{(t)}))}{\sigma}\right\|^2,$$

where $\phi^{*'}(y^{(t)})$ is a n-dimensional vector such that the i-th coordinate is

$$[\phi^{*'}(y^{(t)})]_i = (\phi_i^*)'(y_i^{(t)}),$$

and

$$u^{(t)} = \frac{1}{n}\sum_{i=1}^{n}y_i^{(t)}a_i, \quad u^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}y_i^{(t+1)}a_i, \quad \text{and} \quad u^* = \frac{1}{n}\sum_{i=1}^{n}y_i^\star a_i.$$

On the other hand, since $x^{(t+1)}$ minimizes a $\frac{1}{\tau} + \lambda$-strongly convex objective

$$g(x) + \left\langle u^{(t)} + n(u^{(t+1)} - u^{(t)}), x \right\rangle + \frac{\|x - x^{(t)}\|^2}{2\tau},$$

we can apply Lemma 2 with $\rho = 0$ to obtain

$$g(x^\star) + \langle u^{(t)} + n(u^{(t+1)} - u^{(t)}), x^\star \rangle + \frac{\|x^{(t)} - x^\star\|^2}{2\tau}$$

$$\geq g(x^{(t+1)}) + \langle u^{(t)} + n(u^{(t+1)} - u^{(t)}), x^{(t+1)} \rangle + \frac{\|x^{(t+1)} - x^{(t)}\|^2}{2\tau} + \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\|x^{(t+1)} - x^\star\|^2,$$

and rearranging terms, we get

$$\frac{\|x^{(t)} - x^\star\|^2}{2\tau} \geq \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\mathbb{E}[\|x^{(t+1)} - x^\star\|^2] + \frac{\mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2]}{2\tau} + \mathbb{E}[g(x^{(t+1)}) - g(x^\star)]$$

$$+ \mathbb{E}[\langle u^{(t)} + n(u^{(t+1)} - u^{(t)}), x^{(t+1)} - x^\star \rangle].$$

Notice that

$$\mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^\star) + n(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t+1)})) - (n-1)(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t)}))$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\phi_i^*(y_i^{(t)}) - \phi_i^*(y^\star)) + (\phi_k^*(y_k^{(t+1)}) - \phi_k^*(y_k^{(t)})) + g(x^{(t+1)}) - g(x^\star)$$

$$+ \langle u^\star, x^{(t+1)} \rangle - \langle u^{(t)}, x^\star \rangle + n\langle u^{(t)} - u^{(t+1)}, x^\star \rangle,$$

so

$$\frac{\|x^{(t)} - x^\star\|^2}{2\tau} + \left(\frac{1}{\sigma} + \frac{(n-1)}{2n}\right)\mathcal{D}(y^{(t)}, y^\star) + (n-1)(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t)}))$$

$$\geq \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\mathbb{E}[\|x^{(t+1)} - x^\star\|^2] + \left(\frac{1}{\sigma} + \frac{1}{2}\right)\mathbb{E}[\mathcal{D}(y^{(t+1)}, y^\star)] + \frac{\mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2]}{2\tau} + \frac{\mathbb{E}[\mathcal{D}(y^{(t+1)}, y^{(t)})]}{\sigma}$$

$$+ \mathbb{E}[\mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^\star) + n(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t+1)}))]$$

$$+ \mathbb{E}[\langle u^{(t)} - u^* + n(u^{(t+1)} - u^{(t)}), x^{(t+1)} - \bar{x}^t \rangle] + \frac{\delta}{4n}\left\| A(x^* - \tilde{x}^{(t)}) + \frac{(\phi^{*'}(\tilde{y}) - \phi^{*'}(y^{(t)}))}{\sigma}\right\|^2.$$

23

Next, we have

$$\frac{\delta}{4n}\left\|A(x^* - \tilde{x}^{(t)}) + \frac{(\phi^{*'}(\tilde{y}) - \phi^{*'}(y^{(t)}))}{\sigma}\right\|^2 = \frac{\delta}{4n}\left\|A(x^* - x^{(t)}) - \theta A(x^{(t)} - x^{(t-1)}) + \frac{(\phi^{*'}(\tilde{y}) - \phi^{*'}(y^{(t)}))}{\sigma}\right\|^2$$

$$\geq \left(1 - \frac{1}{\alpha}\right)\frac{\delta}{4n}\left\|A(x^* - x^{(t)})\right\|^2$$

$$- (\alpha - 1)\frac{\delta}{4n}\left\|\theta A(x^{(t)} - x^{(t-1)}) + \frac{(\phi^{*'}(\tilde{y}) - \phi^{*'}(y^{(t)}))}{\sigma}\right\|^2,$$

for any $\alpha > 1$ and

$$\left\|A(x^* - x^{(t)})\right\|^2 \geq \mu^2\|x^* - x^{(t)}\|^2,$$

and

$$\left\|\theta A(x^{(t)} - x^{(t-1)}) + \frac{(\phi^{*'}(\tilde{y}) - \phi^{*'}(y^{(t)}))}{\sigma}\right\|^2 \geq -2\theta^2\|A(x^{(t)} - x^{(t-1)})\|^2 - \frac{2}{\sigma^2}\|\phi^{*'}(\tilde{y}) - \phi^{*'}(y^{(t)})\|^2]$$

$$\geq -2\theta^2 L^2\|x^{(t)} - x^{(t-1)}\|^2 - \frac{2n}{\sigma^2}\mathbb{E}[\|\phi^{*'}(y^{(t+1)}) - \phi^{*'}(y^{(t)})\|^2].$$

Following the same reasoning as in the standard SPDC analysis, we have

$$\langle u^{(t)} - u^* + n(u^{(t+1)} - u^{(t)}), x^{(t+1)} - \tilde{x}^{(t)}\rangle = \frac{(y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})}{n} - \frac{\theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})}{n}$$

$$+ \frac{(n-1)}{n}(y^{(t+1)} - y^{(t)})^T A(x^{(t+1)} - x^{(t)}) - \theta(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)}),$$

and using Cauchy-Schwartz inequality, we have

$$|(y^{(t+1)} - y^{(t)})^T A(x^{(t)} - x^{(t-1)})| \leq \frac{\|(y^{(t+1)} - y^{(t)})^T A\|^2}{1/(2\tau)} + \frac{\|x^{(t)} - x^{(t-1)}\|^2}{8\tau}$$

$$\leq \frac{\|y^{(t+1)} - y^{(t)}\|^2}{1/(2\tau R^2)},$$

and

$$|(y^{(t+1)} - y^{(t)})^T A(x^{(t+1)} - x^{(t)})| \leq \frac{\|(y^{(t+1)} - y^{(t)})^T A\|^2}{1/(2\tau)} + \frac{\|x^{(t+1)} - x^{(t)}\|^2}{8\tau}$$

$$\leq \frac{\|y^{(t+1)} - y^{(t)}\|^2}{1/(2\tau R^2)}.$$

Thus we get

$$\langle u^{(t)} - u^* + n(u^{(t+1)} - u^{(t)}), x^{(t+1)} - \tilde{x}^{(t)}\rangle \geq \frac{(y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})}{n} - \frac{\theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})}{n}$$

$$- \frac{\|y^{(t+1)} - y^{(t)}\|^2}{1/(4\tau R^2)} - \frac{\|x^{(t+1)} - x^{(t)}\|^2}{8\tau} - \frac{\theta\|x^{(t)} - x^{(t-1)}\|^2}{8\tau}.$$

Also we can lower bound the term $\mathcal{D}(y^{(t+1)}, y^{(t)})$ using Lemma 2 with $\rho = 1/2$:

$$\mathcal{D}(y^{(t+1)}, y^{(t)}) = \sum_{i=1}^n \left(\phi_i^*(y_i^{(t+1)}) - \phi_i^*(y_i^{(t)}) - \langle(\phi_i^*)'(y_i^{(t)}), y_i^{(t+1)} - y_i^{(t)}\rangle\right)$$

$$\geq \sum_{i=1}^n \left(\frac{\gamma}{2}(y_i^{(t+1)} - y_i^{(t)})^2 + \frac{\delta}{2}((\phi_i^*)'(y_i^{(t+1)}) - (\phi_i^*)'(y_i^{(t)}))^2\right)$$

$$= \frac{\gamma}{2}\|y^{(t+1)} - y^{(t)}\|^2 + \frac{\delta}{2}\|\phi^{*'}(y^{(t+1)}) - \phi^{*'}(y^{(t)})\|^2.$$

Combining everything above together, we have

$$\left(\frac{1}{2\tau} - \frac{(1-1/\alpha)\delta\mu^2}{4n}\right)\|x^{(t)} - x^\star\|^2 + \left(\frac{1}{\sigma} + \frac{(n-1)}{2n}\right)\mathcal{D}(y^{(t)}, y^\star) + \theta(\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^\star))$$

$$+ (n-1)(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t)})) + \theta\left(\frac{1}{8\tau} + \frac{(\alpha-1)\theta\delta L^2}{2n}\right)\|x^{(t)} - x^{t-1}\|^2 + \frac{\theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})}{n}$$

$$\geq \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\mathbb{E}[\|x^{(t+1)} - x^\star\|^2] + \left(\frac{1}{\sigma} + \frac{1}{2}\right)\mathbb{E}[\mathcal{D}(y^{(t+1)}, y^\star)] + \frac{\mathbb{E}[(y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})]}{n}$$

$$+ \mathbb{E}[\mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^\star) + n(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t+1)}))]$$

$$+ \left(\frac{1}{2\tau} - \frac{1}{8\tau}\right)\mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2] + \left(\frac{\gamma}{2\sigma} - 4R^2\tau\right)\mathbb{E}[\|y^{(t+1)} - y^{(t)}\|^2]$$

$$+ \left(\frac{\delta}{2\sigma} - \frac{(\alpha-1)\delta}{2\sigma^2}\right)\mathbb{E}[\|\phi^{*'}(y^{(t+1)}) - \phi^{*'}(y^{(t)})\|^2].$$

If we choose the parameters as

$$\alpha = \frac{\sigma}{4} + 1, \quad \sigma\tau = \frac{\gamma}{16R^2},$$

then we know

$$\frac{\gamma}{2\sigma} - 4R^2\tau = \frac{\gamma}{2\sigma} - \frac{\gamma}{4\sigma} > 0,$$

and

$$\frac{\delta}{2\sigma} - \frac{(\alpha-1)\delta}{2\sigma^2} = \frac{\delta}{2\sigma} - \frac{\delta}{8\sigma} > 0$$

and

$$\frac{(\alpha-1)\theta\delta L^2}{2n} \leq \frac{\sigma\delta L^2}{8n^2} \leq \frac{\delta\sigma R^2}{8} \leq \frac{\delta\gamma}{256\tau} \leq \frac{1}{256\tau},$$

thus

$$\frac{1}{8\tau} + \frac{(\alpha-1)\theta\delta L^2}{2n} \leq \frac{3}{8\tau}.$$

In addition, we have

$$1 - \frac{1}{\alpha} = \frac{\sigma}{\sigma+4}.$$

Finally we obtain

$$\left(\frac{1}{2\tau} - \frac{\sigma\delta\mu^2}{4n(\sigma+4)}\right)\|x^{(t)} - x^\star\|^2 + \left(\frac{1}{\sigma} + \frac{(n-1)}{2n}\right)\mathcal{D}(y^{(t)}, y^\star) + \theta(\mathcal{L}(x^{(t)}, y^\star) - \mathcal{L}(x^\star, y^\star))$$

$$+ (n-1)(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t)})) + \theta \cdot \frac{3}{8\tau}\|x^{(t)} - x^{(t-1)}\|^2 + \frac{\theta(y^{(t)} - y^\star)^T A(x^{(t)} - x^{(t-1)})}{n}$$

$$\geq \left(\frac{1}{2\tau} + \frac{\lambda}{2}\right)\mathbb{E}[\|x^{(t+1)} - x^\star\|^2] + \left(\frac{1}{\sigma} + \frac{1}{2}\right)\mathbb{E}[\|y^{(t+1)} - y^\star\|^2] + \frac{\mathbb{E}[(y^{(t+1)} - y^\star)^T A(x^{(t+1)} - x^{(t)})]}{n}$$

$$+ \mathbb{E}[\mathcal{L}(x^{(t+1)}, y^\star) - \mathcal{L}(x^\star, y^\star) + n(\mathcal{L}(x^\star, y^\star) - \mathcal{L}(x^\star, y^{(t+1)}))] + \frac{3}{8\tau}\mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2].$$

As before, we can define θ_x and θ_y as the ratios between the coefficients in the x-distance and y-distance terms, and let $\theta = \max\{\theta_x, \theta_y\}$. Then choosing the step-size parameters as

$$\tau = \frac{1}{4R}\sqrt{\frac{\gamma}{n\lambda + \delta\mu^2}}, \quad \sigma = \frac{1}{4R}\sqrt{\gamma(n\lambda + \delta\mu^2)}$$

gives the desired result.