ECAI 2016
G.A. Kaminka et al. (Eds.)

doi:10.3233/978-1-61499-672-9-167

On Stochastic Primal-Dual Hybrid Gradient Approach for Compositely Regularized Minimization

Linbo Qiao[1,2] and **Tianyi Lin**[3] and **Yu-Gang Jiang**[4] and **Fan Yang**[5] and **Wei Liu**[6] and **Xicheng Lu**[1,2]

Abstract. We consider a wide spectrum of regularized stochastic minimization problems, where the regularization term is composite with a linear function. Examples of this formulation include graph-guided regularized minimization, generalized Lasso and a class of ℓ_1 regularized problems. The computational challenge is that the closed-form solution of the proximal mapping associated with the regularization term is not available due to the imposed linear composition. Fortunately, the structure of the regularization term allows us to reformulate it as a new convex-concave saddle point problem which can be solved using the Primal-Dual Hybrid Gradient (PDHG) approach. However, this approach may be inefficient in realistic applications as computing the full gradient of the expected objective function could be very expensive when the number of input data samples is considerably large. To address this issue, we propose a Stochastic PDHG (SPDHG) algorithm with either uniformly or non-uniformly averaged iterates. Through uniformly averaged iterates, the SPDHG algorithm converges in expectation with $O(1/\sqrt{t})$ rate for general convex objectives and $O(\log(t)/t)$ rate for strongly convex objectives, respectively. While with non-uniformly averaged iterates, the SPDHG algorithm is expected to converge with $O(1/t)$ rate for strongly convex objectives. Numerical experiments on different genres of datasets demonstrate that our proposed algorithm outperforms other competing algorithms.

1 Introduction

In this paper, we are interested in solving a class of compositely regularized convex optimization problems:

$$\min_{x\in\mathcal{X}} \mathbb{E}\left[l(x,\xi)\right] + r(Fx), \tag{1}$$

where $x \in \mathbf{R}^d$, $\mathcal{X}$ is a convex compact set with diameter D_x, $r : \mathbf{R}^l \to \mathbf{R}$ is a convex regularization function, and $F \in \mathbf{R}^{l\times d}$ is a penalty matrix (not necessarily diagonal) specifying the desired structured sparsity pattern in x. Furthermore, we denote $l(\cdot,\cdot) : \mathbf{R}^d \times \Omega \to \mathbf{R}$ as a smooth convex function when applying a prediction rule x on a sample dataset $\{\xi_i = (a_i, b_i)\}$, and the corresponding expectation is denoted by $l(x) = \mathbb{E}\left[l(x,\xi)\right]$.

When $F = I$, the above formulation accommodates quite a few classic classification and regression models including Lasso obtained by setting $l(x,\xi_i) = \frac{1}{2}\left\|a_i^\top x - b_i\right\|^2$ and $r(x) = \lambda\|x\|_1$, and linear SVM obtained by letting $l(x,\xi_i) = \max\left(0, 1 - b_i\cdot a_i^\top x\right)$ and $r(x) = (\lambda/2)\|x\|_2^2$, where $\lambda > 0$ is a parameter. Moreover, the general structure of F enables problem (1) to cover more complicated problems arising from machine learning, such as graph-guided regularized minimization [6] and the generalized Lasso model [17].

However, this modeling power also comes with a challenge in computation. In particular, when F is not diagonal, it is very likely that the proximal mapping associated with $r(F(x))$ does not admit a closed-form expression. To cope with this difficulty, we could reformulate problem (1) as a convex-concave saddle point problem by exploiting some special structure of the regularization term, and then resort to the Primal-Dual Hybrid Gradient (PDHG) approach [23]. This approach has exhibited attractive numerical performance in image processing and image restoration applications [5, 3, 23, 19]. We refer readers to [4, 8, 9] to visit convergence properties of PDHG and its variants.

In practice, $\mathbb{E}\left[l(x,\xi)\right]$ is often replaced by its empirical average on a set of training samples. In this case, the computational complexity of calling the function value or the full gradient of $l(x)$ is proportional to the number of training samples, which is extremely huge for modern data-intensive applications. This could make PDHG and linearized PDHG suffer severely from the very poor scalability. Therefore, it is promising to propose a Stochastic variant of PDHG (SPDHG). Like many well-studied incremental or stochastic gradient methods [11, 14, 10, 1, 20], we draw a sample ξ^{k+1} in random and compute a noisy gradient $\nabla l(x^k, \xi^{k+1})$ at the k-th iteration with the current iterate x^k. As a result, the proposed SPDHG method enjoys the capability of dealing with very large-scale datasets.

Another way to handle the non-diagonal F and the expected objective function $\mathbb{E}\left[l(x,\xi)\right]$ is stochastic ADMM-like methods [12, 21, 7, 15, 18, 2, 22, 16] which aim for solving the following problem after introducing an additional variable z:

$$\min_{x\in\mathcal{X}, z=Fx} l(x) + r(z), \tag{2}$$

whose augmented Lagrangian function is given by $l(x) + r(z) + \lambda^\top(z - Fx) + \frac{\gamma}{2}\|z - Fx\|_2^2$. Comparing this function with the convex-concave problem (1) in Section 3, we can see that ADMM-like methods need to update one more vector variable than PDHG-type methods in every iteration. Thus, it can be expected that the per-iteration computational cost of ADMM-like methods is higher

[1] College of Computer, National University of Defense Technology, Changsha, China. email: {qiao.linbo, xclu}@nudt.edu.cn
[2] National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Changsha, China.
[3] Research Center for Management Science and Information Analytics, Shanghai University of Finance and Economics, Shanghai, China, email: lin.tianyi@mail.shufe.edu.cn
[4] School of Computer Science, Fudan University, Shanghai, China, email: ygj@fudan.edu.cn
[5] Research Center for Management Science and Information Analytics, Shanghai University of Finance and Economics, Shanghai, China, email: fyang11@fudan.edu.cn
[6] Tencent AI Lab, China, email: wliu@ee.columbia.edu

than our proposed algorithm SPDHG, as confirmed by the numerical experiments in Section 5.

Our contribution. To the best of our knowledge, we propose in this paper a new convex-concave formulation of problem (1), as well as the first stochastic variant of the PDHG algorithm for both uniformly and non-uniformly averaged iterates with achievable iteration complexities. In particular, for uniformly averaged iterates, the proposed algorithm converges in expectation with the rate of $O(1/\sqrt{t})$ and $O(\log(t)/t)$ for convex objectives and strongly convex objectives, respectively. It is worth mentioning that the $O(1/\sqrt{t})$ convergence rate is known to be best possible for first-order stochastic algorithms under general convex objective functions [1], which has also been established for the well-known stochastic ADMM (SADMM) [12]. Moreover, when optimizing strongly convex objectives, non-uniformly averaged iterates generated by SPDHG converge with $O(1/t)$ expected rate, which is the same as that of Optimal SADMM proposed in [2]. However, as mentioned before, the significant advantage gained by SPDHG beyond SADMM is the low per-iteration complexity. The effectiveness and efficiency of the proposed SPDHG algorithm are demonstrated by encouraging empirical evaluation in graph-guided regularized minimization tasks on several real-world datasets.

2 Related Work

Given the importance of problem (1), various stochastic optimization algorithms have been proposed to solve problem (1) or the more general form of problem (1), which can be written into

$$\begin{aligned} \min_{x\in\mathcal{X}, y\in\mathbf{R}^d} \quad & \mathbb{E}\left[l(x,\xi)\right] + r(y), \\ \text{s.t.} \quad & Ax + By = b. \end{aligned} \tag{3}$$

It is easy to verify that problem (1) is a special case of problem (3) when $A = F$, $B = -I$ and $b = 0$.

In solving problem (3), Wang and Banerjee [18] proposed an online ADMM that requires an easy proximal map of l. However, this is difficult for many loss functions such as logistic loss function. Ouyang et al. [12], Suzuki [15], Azadi and Sra [2], Gao et al. [7], and recently Zhao et al. [21] developed several stochastic variants of ADMM, which linearize l by using its noisy subgradient or gradient and add a varying proximal term. Furthermore, Zhong and Kwok [22] and Suzuki [16] respectively proposed a stochastic averaged gradient-based ADM and a stochastic dual coordinate ascent ADM, which can both obtain improved iteration complexities. However, these methods did not explore the structure of r and need to update one more vector variable than PDHG-type methods in every iteration. We will show in the experiments that our proposed SPDHG algorithm is far more efficient than these algorithms.

It is worth mentioning that another stochastic version of the primal-dual gradient approach was also analyzed in recent work [10]. However, their convex-concave formulation is different from ours, and their algorithm cannot be applied to solve problem (1). Regarding the iteration complexity, the proposed SPDHG algorithm has accomplished the best possible one for first-order stochastic algorithms under general convex objective functions [1]. A better convergence rate of $O(1/t^2 + 1/\sqrt{t})$ can be obtained by using Nestrov's acceleration technique in [11].

The most related algorithm to our proposed SPDHG algorithm is the SPDC algorithm [20] plus its adaptive variant [24]. Similar to our SPDHG algorithm, the SPDC algorithm is also a stochastic variant of the batch primal-dual algorithm developed by Chambolle and Pock [4], which alternates between maximizing over a randomly chosen dual variable and minimizing over the primal variable. However, the SPDC algorithm does not explore the special structure of the regularization term (Assumption 3), and their convex-concave formulation is different from ours. This leads to the inability of the SPDC algorithm to solve problem (1). Specifically, [20] suggests to reformulate problem (1) as

$$\min_{x\in\mathcal{X}} \max_{y\in\mathbf{R}^d} \left\{\mathbb{E}\left[\langle y, x\rangle - l^*(y,\xi)\right] + r(Fx)\right\}, \tag{4}$$

where $l^*(y,\xi) = \sup_{\alpha\in\mathbf{R}^d}\left\{\langle\alpha, y\rangle - l(\alpha,\xi)\right\}$ is the convex conjugate of $l(x,\xi)$. Then the SPDC algorithm in solving problem (4) requires that the proximal map of l^* and $r(Fx)$ be easily computed, which is somewhat strong for a variety of application problems. In addition, the SPDC algorithm requires r to be strongly convex.

In contrast, our SPDHG algorithm only needs the smoothness of l and the convexity of r, and hence efficiently solves a wide range of graph-guided regularized optimization problems, which cannot be solved by the SPDC algorithm and its adaptive variant.

3 Preliminaries

3.1 Assumptions

We make the following assumptions (Assumption 1-4) regarding problem (1) throughout the paper:

Assumption 1 *The optimal set of problem* (1) *is nonempty.*

Assumption 2 *$l(\cdot)$ is continuously differentiable with Lipschitz continuous gradient. That is, there exists a constant $L > 0$ such that*

$$\|\nabla l(x_1) - \nabla l(x_2)\| \le L\|x_1 - x_2\|, \forall x_1, x_2 \in \mathcal{X}.$$

Many formulations in machine learning satisfy Assumption 2. The following least square and logistic functions are two commonly used ones:

$$l(x,\xi_i) = \frac{1}{2}\left\|a_i^\top x - b_i\right\|^2 \text{ or } l(x,\xi_i) = \log\left(1 + \exp\left(-b_i\cdot a_i^\top x\right)\right),$$

where $\xi_i = (a_i, b_i)$ is a single data sample.

Assumption 3 *$r(x)$ is a continuous function which is possibly non-smooth, and it can be described as follows:*

$$r(x) = \max_{y\in\mathcal{Y}}\langle y, x\rangle,$$

where $\mathcal{Y}\in\mathbf{R}^d$ is a convex compact set with diameter D_y.

Note that Assumption 3 is reasonable for the learning problems with a norm regularization such as ℓ_1-norm or nuclear norm:

$$\begin{aligned} \|x\|_1 &= \max\left\{\langle y, x\rangle \mid \|y\|_\infty \le 1\right\}, \\ \|X\|_* &= \max\left\{\langle Y, X\rangle \mid \|Y\|_2 \le 1\right\}. \end{aligned}$$

Assumption 4 *The function $l(x)$ is easy for gradient estimation. That is to say, any stochastic gradient estimation $\nabla l(\cdot,\xi)$ for $\nabla l(\cdot)$ at x satisfies*

$$\mathbb{E}\left[\nabla l(x,\xi)\right] = \nabla l(x),$$

and

$$\mathbb{E}\left[\|\nabla l(x,\xi) - \nabla l(x)\|^2\right] \le \sigma^2.$$

Algorithm 1 SPDHG

Initialize: x^0 and y^0.
for $k = 0, 1, 2, \cdots$ **do**
  Choose one data sample ξ^{k+1} randomly.
  Update y^{k+1} according to Eq. (6).
  Update x^{k+1} according to Eq. (8).
end for
Output: $\bar{x}^t = \sum_{k=0}^{t} \alpha^{k+1} x^{k+1}$ and $\bar{y}^t = \sum_{k=0}^{t} \alpha^{k+1} y^{k+1}$.

where σ is some small number, and it is used in the proof of Lemma (7).

Assumption 5 *$l(\cdot)$ is μ-**strongly convex** at x. In other words, there exists a constant $\mu > 0$ such that*

$$l(y) - l(x) - (y - x)^\top \nabla l(x) \geq \frac{\mu}{2} \|y - x\|^2, \forall y \in \mathcal{X}.$$

We remark that Assumption 5 is optional, and it is only necessary for the theoretical analysis that can lead to a lower iteration complexity.

3.2 Convex-Concave Saddle Point Problem

According to Assumption 3, we are able to rewrite problem (1) as the following convex-concave saddle point problem:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \{P(y, x) = l(x) + \langle y, Fx \rangle\}. \tag{5}$$

Remark 6 *We remark here that the formulation* (5) *is greatly different from those used in [10, 20, 24], where they formulate problem* (1) *as another convex-concave saddle point problem* (4) *by using the convex conjugate of l. Therefore, their algorithms are limited to solving problem* (1) *due to the fact that the proximal mapping of $r(Fx)$ is difficult to compute.*

This problem can be solved by Linearized PDHG (LPDHG) with the following iteration scheme:

$$y^{k+1} := \operatorname*{argmax}_{y \in \mathcal{Y}} \left\{ P(y, x^k) - \frac{1}{2s} \left\| y - y^k \right\|^2 \right\}, \tag{6}$$

$$x^{k+1} := \left[x^k - \beta \left(\nabla l(x^k) + F^\top y^{k+1} \right) \right]_{\mathcal{X}}. \tag{7}$$

However, the above algorithm is inefficient since computing $\nabla l(x^k)$ in each iteration is very costly when the total number of samples n is large. This inspires us to propose a stochastic variant of PDHG, where only the noisy gradient $\nabla l(x^k, \xi^{k+1})$ is computed at each step.

4 Stochastic PDHG

In this section, we first propose our Stochastic Primal-Dual Hybrid Gradient (SPDHG) algorithms with either uniformly or non-uniformly averaged iterates for solving problem (5); and then provide the detailed convergence analysis of the proposed algorithms.

4.1 Algorithm

The SPDHG is presented in Algorithm 1, where we have addressed the following three important issues: how to apply the noisy gradient, how to choose the step-size, and how to determine the weights for the non-uniformly averaged iterates.

Stochastic Gradient: Our SPDHG algorithm shares some common features with the LPDHG algorithm. In fact, the y-subproblems for both algorithms are essentially the same, while for the x-subproblem we adopt the noisy gradient $\nabla l(x^k, \xi^{k+1})$ in SPDHG rather than the full gradient $\nabla l(x^k)$ in LPDHG, *i.e.*,

$$x^{k+1} := \left[x^k - \beta^{k+1} \left(\nabla l(x^k, \xi^{k+1}) + F^\top y^{k+1} \right) \right]_{\mathcal{X}}. \tag{8}$$

That is, in SPDHG we first maximize over the dual variable and then perform one-step stochastic gradient descent along the direction $-\nabla l(x^k, \xi^{k+1}) - F^\top y^{k+1}$ with step-size β^{k+1}.

The Step-Size β^{k+1}: The choice of the step-size β^{k+1} varies with respect to the different conditions satisfied by the objective function l. Different step-size rules also lead to different convergence rates. Note that a sequence of vanishing step-sizes is necessary since we do not adopt any technique of variance reduction in the SPDHG algorithm.

Non-uniformly Averaged Iterates: It was shown in [2] that the non-uniformly averaged iterates generated by stochastic algorithms converge with fewer iterations. Inspired by their work, through non-uniformly averaging the iterates of the SPDHG algorithm and adopting a slightly modified step-size, we manage to establish an accelerated convergence rate of $O(1/t)$ in expectation.

For the convenience of readers, we summarize the convergence properties with respect to different settings in Table 1.

Table 1: Convergence properties.

l	General Convex	Strongly Convex	
β^{k+1}	$\frac{1}{\sqrt{k+1}+L}$	$\frac{1}{\mu(k+1)+L}$	$\frac{2}{\mu(k+2)+2L}$
α^{k+1}	$\frac{1}{t+1}$		$\frac{2(k+1)}{(t+1)(t+2)}$
Rate	$O(\frac{1}{\sqrt{t}})$	$O(\frac{\log(t)}{t})$	$O(\frac{1}{t})$

4.2 Convergence of uniformly averaging under convex objective functions

In this subsection, we analyze the convergence property of the SPDHG algorithm with uniformly averaged iterates for general convex objective functions.

Lemma 7 *Let (y^{k+1}, x^{k+1}) be generated by Algorithm 1, and β^{k+1} and α^{k+1} be shown in Table 1. For any optimal solution (y^*, x^*) of problem* (5)*, it holds that*

$$\begin{aligned} 0 &\geq \mathbb{E}\left[P(y^{k+1}, x^*) - P(y^*, x^{k+1}) \right] \\ &\geq \frac{\sqrt{k+1}+L}{2} \left(\mathbb{E}\left\| x^* - x^{k+1} \right\|^2 - \mathbb{E}\left\| x^* - x^k \right\|^2 \right) \\ &\quad + \frac{1}{2s} \left(\mathbb{E}\left\| y^* - y^{k+1} \right\|^2 - \mathbb{E}\left\| y^* - y^k \right\|^2 \right) \\ &\quad - \frac{\lambda_{\max}(F^\top F) D_y^2 + \sigma^2}{\sqrt{k+1}}. \end{aligned} \tag{9}$$

Proof. For any optimal solution (y^*, x^*) of problem (5), the first-order optimality conditions for Eq. (6) and Eq. (8) are

$$0 \leq \left(y^* - y^{k+1} \right)^\top \left(-Fx^k + \frac{1}{s} \left(y^{k+1} - y^k \right) \right)$$

$$0 \leq \left(x^* - x^{k+1} \right)^\top \left[x^{k+1} - x^k + \beta^{k+1} \left(\nabla l(x^k, \xi^{k+1}) + F^\top y^{k+1} \right) \right],$$

which implies that

$$\begin{aligned}
&\left(x^* - x^{k+1}\right)^\top \nabla l(x^k, \xi^{k+1}) - \left(y^* - y^{k+1}\right)^\top F x^{k+1} \\
&+ \left(x^* - x^{k+1}\right)^\top F^\top y^{k+1} \\
\geq\ & \frac{1}{2\beta^{k+1}} \left(\left\|x^* - x^{k+1}\right\|^2 - \left\|x^* - x^k\right\|^2 + \left\|x^{k+1} - x^k\right\|^2 \right) \\
&+ \frac{1}{2s} \left(\left\|y^* - y^{k+1}\right\|^2 - \left\|y^* - y^k\right\|^2 \right) \\
&+ \left(y^* - y^{k+1}\right)^\top \left(F x^k - F x^{k+1}\right). \qquad (10)
\end{aligned}$$

Furthermore, for any $\gamma > 0$ we have

$$\begin{aligned}
&\left(y^* - y^{k+1}\right)^\top \left(F x^k - F x^{k+1}\right) \\
\geq\ & -\frac{\lambda_{\max}(F^\top F) D_y^2}{\gamma} - \frac{\gamma}{4} \left\|x^k - x^{k+1}\right\|^2, \qquad (11)
\end{aligned}$$

and

$$\begin{aligned}
&\left(x^* - x^{k+1}\right)^\top \nabla l(x^k, \xi^{k+1}) \\
=\ & \left(x^* - x^{k+1}\right)^\top \nabla l(x^k) + \left(x^* - x^{k+1}\right)^\top \delta^{k+1} \\
\leq\ & l(x^*) - l(x^{k+1}) + \frac{L}{2} \left\|x^k - x^{k+1}\right\|^2 + \left(x^* - x^{k+1}\right)^\top \delta^{k+1} \\
\leq\ & l(x^*) - l(x^{k+1}) + \left(x^* - x^k\right)^\top \delta^{k+1} \\
&+ \frac{L + \sqrt{k+1}/2}{2} \left\|x^k - x^{k+1}\right\|^2 + \frac{1}{\sqrt{k+1}} \left\|\delta^{k+1}\right\|^2,
\end{aligned}$$

where the first inequality holds due to Lemma 6.2 [7], and $\delta^{k+1} = \nabla l(x^k, \xi^{k+1}) - \nabla l(x^k)$. Then by letting $\gamma = \sqrt{k+1}$ in Eq. (11), we obtain

$$\begin{aligned}
&l(x^*) - l(x^{k+1}) + \begin{pmatrix} y^* - y^{k+1} \\ x^* - x^{k+1} \end{pmatrix}^\top \begin{pmatrix} -F x^{k+1} \\ F^\top y^{k+1} \end{pmatrix} \\
\geq\ & \frac{1}{2\beta^{k+1}} \left(\left\|x^* - x^{k+1}\right\|^2 - \left\|x^* - x^k\right\|^2 \right) \\
&+ \frac{1}{2s} \left(\left\|y^* - y^{k+1}\right\|^2 - \left\|y^* - y^k\right\|^2 \right) - \frac{\lambda_{\max}(F^\top F) D_y^2}{\sqrt{k+1}} \\
&- \left(x^* - x^k\right)^\top \delta^{k+1} - \frac{\left\|\delta^{k+1}\right\|^2}{\sqrt{k+1}}.
\end{aligned}$$

Since x^k is independent of ξ^{k+1}, we take the expectation on both sides of the above inequality conditioning on x^k, y^k, and conclude that

$$\begin{aligned}
&\mathbb{E}\left[P(y^{k+1}, x^*) - P(y^*, x^{k+1})\right] \\
\geq\ & \frac{1}{2\beta^{k+1}} \left(\mathbb{E}\left\|x^* - x^{k+1}\right\|^2 - \left\|x^* - x^k\right\|^2 \right) - \frac{\mathbb{E}\left\|\delta^{k+1}\right\|^2}{\sqrt{k+1}} \\
&+ \frac{1}{2s} \left(\mathbb{E}\left\|y^* - y^{k+1}\right\|^2 - \left\|y^* - y^k\right\|^2 \right) - \frac{\lambda_{\max}(F^\top F) D_y^2}{\sqrt{k+1}}.
\end{aligned}$$

Finally, Eq. (9) follows from the above inequality and Assumption 4. □

We present the main result for uniformly averaged iterates under general convex objective functions in the following theorem.

Theorem 8 *Denote β^{k+1}, α^{k+1} and $\left(\bar{y}^t, \bar{x}^t\right)$ as shown in Table 1. For any optimal solution (y^*, x^*) of problem* (5)*, $\left(\bar{y}^t, \bar{x}^t\right)$ converges to (y^*, x^*) with $O(1/\sqrt{t})$ rate in expectation.*

Proof. Because $(y^k, x^k) \in \mathcal{Y} \times \mathcal{X}$, it holds true that $(\bar{y}^t, \bar{x}^t) \in \mathcal{Y} \times \mathcal{X}$ for all $t \geq 0$. By invoking the convexity of function $l(\cdot)$ and using Eq. (10), we have

$$\begin{aligned}
&\mathbb{E}\left[P(\bar{y}^t, x^*) - P(y^*, \bar{x}^t)\right] \\
\geq\ & \frac{1}{t+1} \sum_{k=0}^{t} \left[\frac{1}{2s} \left(\mathbb{E}\left\|y^* - y^{k+1}\right\|^2 - \mathbb{E}\left\|y^* - y^k\right\|^2 \right) \right. \\
&+ \frac{\sqrt{k+1} + L}{2} \left(\mathbb{E}\left\|x^* - x^{k+1}\right\|^2 - \mathbb{E}\left\|x^* - x^k\right\|^2 \right) \\
&\left. - \frac{\lambda_{\max}(F^\top F) D_y^2}{\sqrt{k+1}} - \frac{\sigma^2}{\sqrt{k+1}} \right] \\
\geq\ & -\frac{D_y^2}{2s(t+1)} - \frac{L D_x^2}{2(t+1)} - \frac{D_x^2 + 2\lambda_{\max}(F^\top F) D_y^2 + 2\sigma^2}{\sqrt{t+1}}.
\end{aligned}$$

This together with the fact that $\mathbb{E}\left[P(\bar{y}^t, x^*) - P(y^*, \bar{x}^t)\right] \leq 0$ implies the conclusion in Theorem 8. □

4.3 Convergence of uniformly averaging under strongly convex objective functions

In this subsection, we analyze the convergence property of the SPDHG algorithm with uniformly averaged iterates for strongly convex objective functions.

Lemma 9 *Let (y^{k+1}, x^{k+1}) be generated by Algorithm 1, and β^{k+1} and α^{k+1} be shown in Table 1. For any optimal solution (y^*, x^*) of problem* (5)*, it holds that*

$$\begin{aligned}
0 \geq\ & \mathbb{E}\left[P(y^{k+1}, x^*) - P(y^*, x^{k+1})\right] \qquad (12) \\
\geq\ & \frac{\mu(k+1) + L}{2} \mathbb{E}\left\|x^* - x^{k+1}\right\|^2 + \frac{1}{2s} \mathbb{E}\left\|y^* - y^{k+1}\right\|^2 \\
&- \frac{\mu k + L}{2} \mathbb{E}\left\|x^* - x^k\right\|^2 - \frac{1}{2s} \mathbb{E}\left\|y^* - y^k\right\|^2 \\
&- \frac{\lambda_{\max}(F^\top F) D_y^2 + \sigma^2}{\mu(k+1)}.
\end{aligned}$$

Proof. By using the same argument as Lemma 7 and the strongly convexity of l, we have

$$\begin{aligned}
&\left(x^* - x^{k+1}\right)^\top \nabla l(x^k, \xi^{k+1}) \qquad (13) \\
\leq\ & l(x^*) - l(x^k) - \frac{\mu}{2} \left\|x^* - x^k\right\|^2 + l(x^k) - l(x^{k+1}) \\
&+ \frac{L}{2} \left\|x^k - x^{k+1}\right\|^2 + \left(x^* - x^{k+1}\right)^\top \delta^{k+1} \\
\leq\ & l(x^*) - l(x^{k+1}) + \left(x^* - x^k\right)^\top \delta^{k+1} - \frac{\mu}{2} \left\|x^* - x^k\right\|^2 \\
&+ \frac{L}{2} \left\|x^k - x^{k+1}\right\|^2 + \frac{\kappa}{4} \left\|x^k - x^{k+1}\right\|^2 + \frac{1}{\kappa} \left\|\delta^{k+1}\right\|^2.
\end{aligned}$$

Substituting Eq. (11) with $\gamma = \mu(k+1)$ and Eq. (13) with $\kappa =$

$\mu(k+1)$ into Eq. (10) yields that

$$\begin{aligned} & l(x^*) - l(x^{k+1}) + \begin{pmatrix} y^* - y^{k+1} \\ x^* - x^{k+1} \end{pmatrix}^\top \begin{pmatrix} -Fx^{k+1} \\ F^\top y^{k+1} \end{pmatrix} \\ \geq & \frac{1}{2s}\left\|y^* - y^{k+1}\right\|^2 - \frac{1}{2s}\left\|y^* - y^k\right\|^2 - \frac{\left\|\delta^{k+1}\right\|^2}{\mu(k+1)} \\ & + \frac{\mu(k+1)+L}{2}\left\|x^* - x^{k+1}\right\|^2 - \frac{\mu k + L}{2}\left\|x^* - x^k\right\|^2 \\ & + \left(\frac{1}{2\beta^{k+1}} - \frac{L+\mu(k+1)}{2}\right)\left\|x^k - x^{k+1}\right\|^2. \\ & - \frac{\lambda_{\max}(F^\top F)D_y^2}{\mu(k+1)} - \left(x^* - x^k\right)^\top \delta^{k+1}. \end{aligned}$$

Then we obtain Eq. (12) as the same as that in Lemma 7. □

We present the main result in the following theorem when the objective function is further assumed to be strongly convex.

Theorem 10 *Denote β^{k+1}, α^{k+1} and $\left(\bar{y}^t, \bar{x}^t\right)$ as shown in Table 1. For any optimal solution (y^*, x^*) of problem* (5)*, $\left(\bar{y}^t, \bar{x}^t\right)$ converges to (y^*, x^*) with $O(\log(t)/t)$ rate in expectation.*

Proof. Because $(y^k, x^k) \in \mathcal{Y} \times \mathcal{X}$, it holds that $(\bar{y}^t, \bar{x}^t) \in \mathcal{Y} \times \mathcal{X}$ for all $t \geq 0$. By invoking the convexity of function $l(\cdot)$ and using Eq. (12), we have

$$\begin{aligned} & \mathbb{E}\left[P(\bar{y}^t, x^*) - P(y^*, \bar{x}^t)\right] \\ \geq & \frac{1}{t+1}\sum_{k=0}^{t}\left[\frac{1}{2s}\left(\mathbb{E}\left\|y^* - y^{k+1}\right\|^2 - \mathbb{E}\left\|y^* - y^k\right\|^2\right)\right. \\ & + \frac{\mu(k+1)+L}{2}\left\|x^* - x^{k+1}\right\|^2 - \frac{\mu k + L}{2}\left\|x^* - x^k\right\|^2 \\ & \left. - \frac{\lambda_{\max}(F^\top F)D_y^2 + \sigma^2}{\mu(k+1)}\right] \\ \geq & -\frac{D_y^2}{2s(t+1)} - \frac{LD_x^2}{2(t+1)} - \frac{\left(\lambda_{\max}(F^\top F)D_y^2 + \sigma^2\right)\log(t+1)}{\mu(t+1)}. \end{aligned}$$

This together with the fact that $\mathbb{E}\left[P(\bar{y}^t, x^*) - P(y^*, \bar{x}^t)\right] \leq 0$ implies the conclusion in Theorem 10. □

4.4 Convergence of non-uniformly averaging under strongly convex objective functions

In this subsection, we analyze the convergence property of the SPDHG algorithm with non-uniformly averaged iterates for strongly convex objective functions.

Lemma 11 *Let (y^{k+1}, x^{k+1}) be generated by Algorithm 1, and β^{k+1} and α^{k+1} be shown in Table 1. For any optimal solution (y^*, x^*) of problem* (5)*, it holds that*

$$\begin{aligned} 0 \geq & \mathbb{E}\left[P(y^{k+1}, x^*) - P(y^*, x^{k+1})\right] \qquad (14) \\ \geq & \frac{\mu(k+2)+2L}{4}\mathbb{E}\left\|x^* - x^{k+1}\right\|^2 + \frac{1}{2s}\mathbb{E}\left\|y^* - y^{k+1}\right\|^2 \\ & - \frac{\mu k + 2L}{4}\mathbb{E}\left\|x^* - x^k\right\|^2 - \frac{1}{2s}\mathbb{E}\left\|y^* - y^k\right\|^2 \\ & - \frac{2\lambda_{\max}(F^\top F)D_y^2 + 2\sigma^2}{\mu(k+1)}. \end{aligned}$$

Proof. By substituting Eq. (11) with $\gamma = \frac{\mu(k+1)}{2}$ and Eq. (13) with $\kappa = \frac{\mu(k+1)}{2}$ into Eq. (10), we have

$$\begin{aligned} & \left(y^* - y^{k+1}\right)^\top \left(Fx^k - Fx^{k+1}\right) \\ \geq & -\frac{2\lambda_{\max}(F^\top F)D_y^2}{\mu(k+1)} - \frac{\mu(k+1)}{8}\left\|x^k - x^{k+1}\right\|^2, \end{aligned}$$

and

$$\begin{aligned} & \left(x^* - x^{k+1}\right)^\top \nabla l(x^k, \xi^{k+1}) \\ \leq & l(x^*) - l(x^{k+1}) + \left(x^* - x^k\right)^\top \delta^{k+1} - \frac{\mu}{2}\left\|x^* - x^k\right\|^2 \\ & + \frac{L}{2}\left\|x^k - x^{k+1}\right\|^2 + \frac{\mu(k+1)}{8}\left\|x^k - x^{k+1}\right\|^2 \\ & + \frac{2}{\mu(k+1)}\left\|\delta^{k+1}\right\|^2. \end{aligned}$$

Then we plug the above two inequalities into Eq. (10), and then follow the same argument as Lemma 9 to obtain the desired inequality in Eq. (14). □

We present the main result for non-uniformly averaged iterates under strongly convex functions in the following theorem.

Theorem 12 *Denote β^{k+1}, α^{k+1} and $\left(\bar{y}^t, \bar{x}^t\right)$ as shown in Table 1. For any optimal solution (y^*, x^*) of problem* (5)*, $\left(\bar{y}^t, \bar{x}^t\right)$ converges to (y^*, x^*) with $O(1/t)$ rate in expectation.*

Proof. We have $(\bar{y}^t, \bar{x}^t) \in \mathcal{Y} \times \mathcal{X}$ for all $t \geq 0$. By invoking the convexity of function $l(\cdot)$ and using Eq. (14), we have

$$\begin{aligned} & \mathbb{E}\left[P(\bar{y}^t, x^*) - P(y^*, \bar{x}^t)\right] \\ \geq & \frac{2}{(t+1)(t+2)}\sum_{k=0}^{t}(k+1)\left[-\frac{2\lambda_{\max}(F^\top F)D_y^2 + 2\sigma^2}{\mu(k+1)}\right. \\ & + \frac{\mu(k+2)+2L}{4}\left\|x^* - x^{k+1}\right\|^2 - \frac{\mu k + 2L}{4}\left\|x^* - x^k\right\|^2 \\ & \left.\frac{1}{2s}\left(\mathbb{E}\left\|y^* - y^{k+1}\right\|^2 - \mathbb{E}\left\|y^* - y^k\right\|^2\right)\right] \\ \geq & -\frac{D_y^2}{s(t+2)} - \frac{LD_x^2}{t+2} - \frac{4\lambda_{\max}(F^\top F)D_y^2 + 4\sigma^2}{\mu(t+2)} \\ & + \frac{\mu}{2(t+1)(t+2)}\sum_{k=0}^{t}\left[(k+2)(k+1)\left\|x^* - x^{k+1}\right\|^2\right. \\ & \left. - (k+1)k\left\|x^* - x^k\right\|^2\right]. \end{aligned}$$

Therefore, we conclude that

$$\begin{aligned} 0 \geq & \mathbb{E}\left[P(\bar{y}^t, x^*) - P(y^*, \bar{x}^t)\right] \\ \geq & -\frac{D_y^2}{s(t+2)} - \frac{LD_x^2}{t+2} - \frac{4\lambda_{\max}(F^\top F)D_y^2 + 4\sigma^2}{\mu(t+2)}, \end{aligned}$$

which implies the conclusion in Theorem 12. □

5 Experiments

We conduct experiments by evaluating two models: graph-guided logistic regression (GGLR) (15) and graph-guided regularized logistic regression (GGRLR) (16) [22],

$$\min_{x \in \mathcal{X}} l(x) + \lambda\|Fx\|_1 \qquad (15)$$

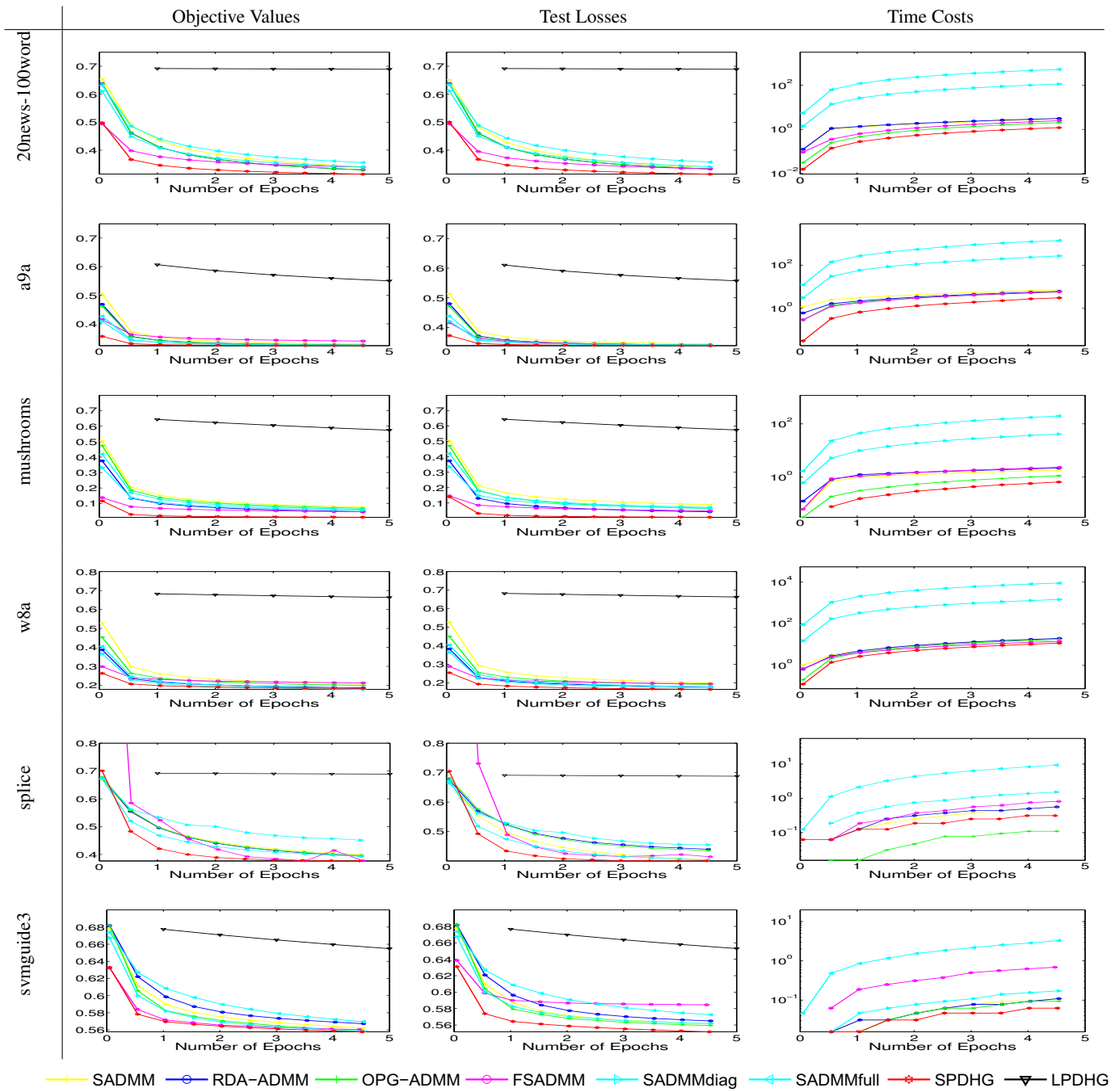

Figure 1: Comparison of SPDHG with STOC-ADMM (SADMM), RDA-ADMM, OPG-ADMM, Fast-SADMM (FSADMM), Ada-SADMMdiag, Ada-SADMMfull and LPDHG on **Graph-Guided Logistic Regression** Task. Epoch for the horizontal axis is the number of iterations divided by the dataset size. **Left Panels**: Averaged objective values. **Middle Panels**: Averaged test losses. **Right Panels**: Averaged time costs (in seconds).

and

$$\min_{x\in\mathcal{X}} l(x) + \frac{\gamma}{2}\|x\|_2^2 + \lambda\|Fx\|_1. \tag{16}$$

Here $l(x) = \frac{1}{N}\left[\sum_{i=1}^{N} l(x,\xi_i)\right]$ is empirical average of $l(x,\xi_i)$ on a set of samples, and $l(x,\xi_i)$ is logistic function $\log\left(1+\exp\left(-b_i\cdot a_i^\top x\right)\right)$, where $\xi_i = (a_i, b_i)$. λ is the regularization parameter. F is a penalty matrix promoting the desired sparse structure of x, which is generated by sparse inverse covariance selection [13]. To proceed, we reformulate problems (15) and (16) into the convex-concave saddle point problem (5) and apply our proposed SPDHG algorithm. On the other hand, we can reformulate problems (15) and (16) into problem (2) by introducing an additional variable $z = Fx$ and then apply stochastic ADMM algorithms.

In the experiments, we compare our SPDHG algorithm with the LPDHG algorithm, and six existing stochastic ADMM algo-

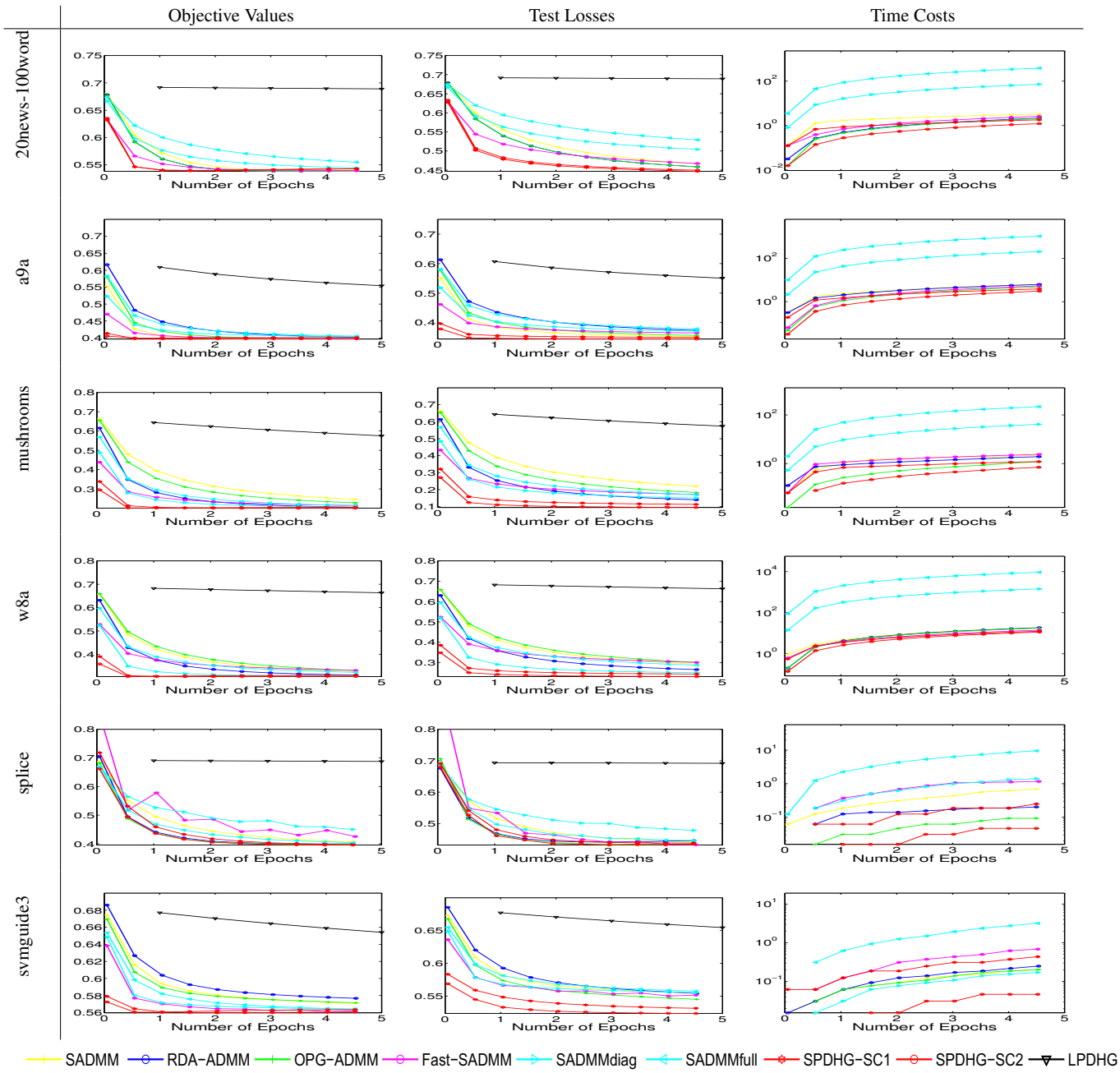

Figure 2: Comparison of SPDHG-SC1 (Uniformly Averaged) and SPDHG-SC2 (Non-Uniformly Averaged) with STOC-ADMM (SADMM), RDA-ADMM, OPG-ADMM, Fast-SADMM (FSADMM), Ada-SADMMdiag, Ada-SADMMfull and LPDHG on **Graph-Guided Regularized Logistic Regression** Task. Epoch for the horizontal axis is the number of iterations divided by the dataset size. **Left Panels**: Averaged objective values. **Middle Panels**: Averaged test losses. **Right Panels**: Averaged time costs (in seconds).

rithms[7]: SADMM [12], OPG-ADMM [15], RDA-ADMM [15], FSADMM[22], and two variants of adaptive SADMM (*i.e.*, SADMMdiag and SADMMfull) [21]. We do not include online ADMM [18] and SDCA-ADMM [16] since [15] has shown that RDA-ADMM performs better than online ADMM while [20] has shown that the performance of FSADMM is comparable to that of SDCA-ADMM. Finally, SPDC and Adaptive SPDC are excluded from the experiments since they cannot solve problem (15) and problem (16), as clarified in Section 2.

The experiments are conducted on six binary classification datasets: *20news*[8], *a9a*, *mushrooms*, *w8a*, *splice* and *svmguide3*[9]. On

[7] We use the code of SADMM, OPG-ADMM, RDA-ADMM and FSADMM provided by the authors while implementing two variants of adaptive SADMM according to [21].

[8] www.cs.nyu.edu/roweis/data.html.

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

Table 2: Statistics of datasets.

dataset	number of samples	dimensionality
svmguide3	1243	21
splice	1000	60
a9a	32,561	123
w8a	64,700	300
20news	16,242	100
mushrooms	8,124	112

each dataset, we use 80% samples for training and 20% for testing, and calculate the lipschitz constant L as its classical upper bound $\hat{L} = 0.25\max_{1\leq i\leq n}\|a_i\|^2$. The regularization parameters are set to $\lambda = 10^{-5}$ and $\gamma = 10^{-2}$. To reduce statistical variability, experimental results are averaged over 10 repetitions. We set the parameters of SPDHG exactly following our theory while using cross validation to select the parameters of the other algorithms. Additionally, we use the metrics in [22] to compare our algorithm with the other algorithms, including objective values, test losses and time costs to compare our algorithm with the other. The "test loss" means the value of the empirically averaged loss evaluated on a test dataset, while the "objective value" means the sum of the empirically averaged loss and regularized terms evaluated on a training dataset, and the "time cost" means the computational time consumption of each algorithm. Specifically, we use test losses (*i.e.*, $l(x)$) on test datasets, objective values (*i.e.*, $l(x)+\lambda\|Fx\|_1$ on the GGLR task and $l(x) + \frac{\gamma}{2}\|x\|_2^2 + \lambda\|Fx\|_1$ on the GGRLR task) on training datasets, and computational time costs on training datasets.

Figure 1 shows the objective values, test losses and time costs as the function of the number of epochs on the GGLR task, where the objective function is convex but not necessarily strongly convex. We observe that our algorithm SPDHG mostly achieves the best performance, surpassing six stochastic ADMM algorithms, all of which outperform LPDHG by a significant margin. FSADMM sometimes achieves better solutions but consumes much more computational time than SPDHG. In fact, our algorithm requires the least iterations and computational time among all the evaluated algorithms. Therefore, the performance of our algorithm SPDHG on four datasets is most stable and effective among all algorithms.

We further compare our algorithm against the other algorithms on the GGRLR task, where the objective function is strongly convex. The experimental results are displayed in Figure 2. Our algorithm still outperforms the other algorithms consistently, which supports our analysis in the previous sections. We also find that the difference between uniformly averaging and non-uniformly averaging shown in Figure 2 is not significant. One reason is that our algorithm converges within only one or two effective epochs. In this case, non-uniformly averaging will not exhibit its advantage.

6 Conclusions

In this paper, we proposed a novel convex-concave saddle point formulation to resolve problem (1) as well as the first stochastic variant of the PDHG algorithm, namely SPDHG. The new algorithm can tackle a variety of real-world problems which cannot be solved by the existing stochastic primal-dual algorithms proposed in [10, 20, 24]. We further proved that the proposed SPDHG algorithm converges in expectation with the rate of $O(1/\sqrt{t})$ and $O(\log(t)/t)$ for general and strongly convex objectives, respectively. By averaging iterates non-uniformly, the SPDHG algorithm converges in expectation with the rate of $O(1/t)$ for strongly convex objectives.

The SPDHG algorithm is well-suited for addressing compositely regularized minimization problems when the penalty matrix F is non-diagonal. The experiments in performing graph-guided logistic regression and graph-guided regularized logistic regression tasks demonstrated that our SPDHG algorithm outperforms the other competing stochastic algorithms.

Acknowledgments

The work was partially supported by the National Natural Science Foundation of China under Grant No. 61303264.

REFERENCES

[1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright, 'Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization', *IEEE Transactions on Information Theory*, **58**(5), 32–35, (2012).

[2] S. Azadi and S. Sra, 'Towards an optimal stochastic alternating direction method of multipliers', in *ICML*, pp. 620–628, (2014).

[3] S. Bonettini and V. Ruggiero, 'On the convergence of primal–dual hybrid gradient algorithms for total variation image restoration', *Journal of Mathematical Imaging and Vision*, **44**(3), 236–253, (2012).

[4] A. Chambolle and T. Pock, 'A first-order primal-dual algorithm for convex problems with applications to imaging', *Journal of Mathematical Imaging and Vision*, **40**(1), 120–145, (2011).

[5] E. Esser, X. Zhang, and T. F. Chan, 'A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science', *SIAM Journal on Imaging Sciences*, **3**(4), 1015–1046, (2010).

[6] J. Friedman, T. Hastie, and R. Tibshirani, 'The elements of statistical learning: Data mining, inference, and prediction', *Springer Series in Statistics*, (2009).

[7] X. Gao, B. Jiang, and S. Zhang, 'On the information-adaptive variants of the admm: an iteration complexity perspective', *Optimization Online*, (2014).

[8] T. Goldstein, M. Li, X. Yuan, E. Esser, and R. Baraniuk, 'Adaptive primal-dual hybrid gradient methods for saddle-point problems', *ArXiv Preprint 1305.0546*, (2013).

[9] B. He and X. Yuan, 'Convergence analysis of primal-dual algorithms for a saddle-point problem: from contraction perspective', *SIAM Journal on Imaging Sciences*, **5**(1), 119–149, (2012).

[10] G. Lan, 'An optimal randomized incremental gradient method', *ArXiv Preprint 1507.02000*, (2015).

[11] X. Lin, 'Dual averaging methods for regularized stochastic learning and online optimization', *Journal of Machine Learning Research*, **11**, 2543–2596, (2010).

[12] H. Ouyang, N. He, L. Tran, and A. Gray, 'Stochastic alternating direction method of multipliers', in *ICML*, pp. 80–88, (2013).

[13] K. Scheinberg, S. Ma, and D. Goldfarb, 'Sparse inverse covariance selection via alternating linearization methods', in *NIPS*, pp. 2101–2109, (2010).

[14] M. Schmidt, N. L. Roux, and F. Bach, 'Minimizing finite sums with the stochastic average gradient', *ArXiv Preprint 1309.2388*, (2013).

[15] T. Suzuki, 'Dual averaging and proximal gradient descent for online alternating direction multiplier method', in *ICML*, pp. 392–400, (2013).

[16] T. Suzuki, 'Stochastic dual coordinate ascent with alternating direction method of multipliers', in *ICML*, pp. 736–744, (2014).

[17] R. J. Tibshirani and J. Taylor, 'The solution path of the generalized lasso', *Annals of Statistics*, **39**(3), 1335–1371, (2011).

[18] H. Wang and A. Banerjee, 'Online alternating direction method', in *ICML*, pp. 1119–1126, (2012).

[19] X. Zhang, M. Burger, and S. Osher, 'A unified primal-dual algorithm framework based on bregman iteration', *Journal of Scientific Computing*, **46**(1), 20–46, (2011).

[20] Y. Zhang and L. Xiao, 'Stochastic primal-dual coordinate method for regularized empirical risk minimization', in *ICML*, pp. 353–361, (2015).

[21] P. Zhao, J. Yang, T. Zhang, and P. Li, 'Adaptive stochastic alternating direction method of multipliers', in *ICML*, pp. 69–77, (2015).

[22] W. Zhong and J. Kwok, 'Fast stochastic alternating direction method of multipliers', in *ICML*, pp. 46–54, (2014).

[23] M. Zhu and T. F. Chan, 'An efficient primal-dual hybrid gradient algorithm for total variation image restoration', *UCLA CAM Report*, 8–34, (2008).

[24] Z. Zhu and A. J. Storkey, 'Adaptive stochastic primal-dual coordinate descent for separable saddle point problems', in *Machine Learning and Knowledge Discovery in Databases*, 645–658, Springer, (2015).