Stochastic Optimization with Importance Sampling for Regularized Loss Minimization

Peilin Zhao[†,‡] ZHAOP@I2R.A-STAR.EDU.SG
Tong Zhang[‡] TZHANG@STAT.RUTGERS.EDU
[†]Data Analytics Department, Institute for Infocomm Research, A*STAR, Singapore
[‡]Department of Statistics & Biostatistics, Rutgers University, USA; and Big Data Lab, Baidu Research, China

Abstract

Uniform sampling of training data has been commonly used in traditional stochastic optimization algorithms such as Proximal Stochastic Mirror Descent (prox-SMD) and Proximal Stochastic Dual Coordinate Ascent (prox-SDCA). Although uniform sampling can guarantee that the sampled stochastic quantity is an unbiased estimate of the corresponding true quantity, the resulting estimator may have a rather high variance, which negatively affects the convergence of the underlying optimization procedure. In this paper we study stochastic optimization, including prox-SMD and prox-SDCA, with importance sampling, which improves the convergence rate by reducing the stochastic variance. We theoretically analyze the algorithms and empirically validate their effectiveness.

1. Introduction

Stochastic optimization has been extensively studied in the machine learning community (Zhang, 2004; Rakhlin et al., 2011; Shamir & Zhang, 2013; Duchi & Singer, 2009; Luo & Tseng, 1992; Mangasarian & Musicant, 1999; Hsieh et al., 2008; Shalev-Shwartz & Tewari, 2011; Lacoste-Julien et al., 2012; Nesterov, 2012b; Shalev-Shwartz & Zhang, 2012a; 2013; 2012b). At every step, a traditional stochastic optimization method will sample one training example or one dual coordinate uniformly at random from the training data, and then update the model parameter using the sampled example or dual coordinate. In this paper we focus on Proximal Stochastic Mirror Descent (prox-SMD) (Duchi & Singer, 2009; Duchi et al., 2010) and Proximal Stochastic Dual Coordinate Ascent (prox-SDCA) (Shalev-Shwartz & Zhang, 2012b) methods.

For prox-SMD, the traditional algorithms such as Stochastic Gradient Descent (SGD) sample training examples uniformly at random during the entire learning process, so that the stochastic gradient is an unbiased estimation of the true gradient (Zhang, 2004; Rakhlin et al., 2011; Shamir & Zhang, 2013; Duchi & Singer, 2009). However, the variance of the resulting stochastic gradient estimator may be large since the stochastic gradient can vary significantly over different examples. In order to improve convergence, this paper proposes a sampling distribution and the corresponding unbiased importance weighted gradient estimator that minimizes the variance. To this end, we analyze the relationship between the variance of stochastic gradient and the sampling distribution. We show that to minimize the variance, the optimal sampling distribution should be roughly proportional to the norm of the stochastic gradient. To simplify computation, we also consider the use of upper bounds for the norms. Our theoretical analysis shows that under certain conditions, the proposed sampling method can significantly improve the convergence rate, and our results include the existing theoretical results for uniformly sampled prox-SGD and SGD as special cases.

Similarly for prox-SDCA, the traditional approach such as Stochastic Dual Coordinate Ascent (SDCA) (Shalev-Shwartz & Zhang, 2013) picks a coordinate to update by sampling the training data uniformly at random (Luo & Tseng, 1992; Mangasarian & Musicant, 1999; Hsieh et al., 2008; Shalev-Shwartz & Tewari, 2011; Lacoste-Julien et al., 2012; Nesterov, 2012b; Shalev-Shwartz & Zhang, 2012a; 2013; 2012b). It was shown recently that the SDCA and prox-SDCA algorithms with uniform random sampling converge much faster than a fixed cyclic ordering (Shalev-Shwartz & Zhang, 2013; 2012b). However, this paper shows that if we employ an appropriately defined importance sampling strategy, the convergence can be further improved. To optimize sampling distribution, we analyze the connection between

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

the expected increase of dual objective and the sampling distribution, and obtain the optimal solution that depends on the smoothness or Lipschitz constants of the loss functions. Our analysis shows that under certain conditions, the proposed sampling method can significantly improve the convergence rate. In addition, our theoretical results include the existing results for uniformly sampled prox-SDCA and SDCA as special cases.

The rest of this paper is organized as follows. Section 2 reviews the related work. In section 3, we will study stochastic optimization with importance sampling. Section 4 gives our empirical evaluations. Section 5 concludes the paper. The detailed proofs of the theoretical results can be found in the full version of the paper (Zhao & Zhang, 2014).

2. Related Work

After finishing the work, we noticed that Needell et al. (2014) also considered importance sampling for stochastic gradient descent, where they suggested ideas similar to ours. Moreover Strohmer & Vershynin (2009) proposed a variant of the Kaczmarz method (an iterative method for solving systems of linear equations) which selects rows with probability proportional to their squared norms. It was pointed out in (Needell et al., 2014) that this algorithm is actually a SGD algorithm with importance sampling. Our paper studies importance sampling for more general composite objectives and the more general proximal stochastic mirror descent method, covering their algorithms as special cases. Our paper also studies prox-SDCA with importance sampling, which is not covered by previous studies. Another related work is (Xiao & Zhang, 2014), where the authors studied importance sampling for the prox-SVRG procedure, and obtained results similar to those of prox-SDCA considered in this work. The main concern of this work is on the effectiveness of importance sampling, which could be applied to many gradient based algorithms. Therefore we include the study of the standard SGD procedure for comparison, although for smooth and strongly convex objective functions it does not achieve the linear rates of SVRG, SDCA, and SAG (Roux et al., 2012).

For the primal coordinate descent procedures, some researchers have recently considered non-uniform sampling strategies (Nesterov, 2012a; Lee & Sidford, 2013). However their results cannot be directly applied to obtain duality-gap convergence for proximal SDCA which we are interested in here. In contrast, the primal-dual analysis of prox-SDCA in this paper is analogous to that of (Shalev-Shwartz & Zhang, 2013), which directly bounds the duality gap. The proof technique relies on the structure of the regularized loss minimization, which differs from the traditional primal coordinate descent analysis. The suggested distribution for the primal coordinate descent is propositional to the smoothness constant of every coordinate, while the distribution of prox-SDCA is propositional to a constant plus the smoothness constant of the primal individual loss function. These two distributions are quite different. In addition, we also provide an importance sampling distribution when the individual loss functions are Lipschitz. Finally we note that an accelerated version of prox-SDCA was proposed by Shalev-Shwartz & Zhang (2014). The procedure employs an inner-outer-iteration strategy, where the inner iteration is the standard prox-SDCA procedure. The importance sampling result of this paper can be directly applied to the accelerated prox-SDCA in that the convergence of inner iteration becomes faster than that of the uniform sampling. Therefore in this paper we only consider the unaccelerated prox-SDCA.

3. Stochastic Optimization with Importance Sampling

Let $\phi_1, \phi_2, \ldots, \phi_n$ be n vector functions from $\mathbb{R}^d$ to $\mathbb{R}$. Our goal is to find an approximate solution of the following optimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} P(\mathbf{w}) := f(\mathbf{w}) + \lambda r(\mathbf{w}), \tag{1}$$

where $f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \phi_i(\mathbf{w})$, $\lambda > 0$ is a regularization parameter, and r is a regularizer. For example, given examples $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, the Support Vector Machine problem is obtained by setting $\phi_i(\mathbf{w}) = [1 - y_i\mathbf{x}_i^\top \mathbf{w}]_+$, $[z]_+ = \max(0, z)$, and $r(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$. Regression problems also fall into the above. For example, lasso is obtained by setting $\phi_i(\mathbf{w}) = (y_i - \mathbf{x}_i^\top \mathbf{w})^2$ and $r(\mathbf{w}) = \|\mathbf{w}\|_1$.

Let $\mathbf{w}^*$ be the optimal solution of (1). We say that a solution $\mathbf{w}$ is ϵ_P-sub-optimal if $P(\mathbf{w}) - P(\mathbf{w}^*) \leq \epsilon_P$. We analyze the convergence rates of the proposed algorithms with respect to the number of iterations.

3.1. prox-SMD with Importance Sampling

In this subsection, we consider the proximal stochastic mirror descent method with importance sampling. Proximal Stochastic Mirror Descent works in iterations. At each iteration $t = 1, 2, \ldots$, a sample i_t will be uniformly drawn from $\{1, 2, \ldots, n\}$, and the iterative solution will be updated by setting $\mathbf{w}^{t+1}$ as

$$\arg\min_{\mathbf{w}} \left[\langle \nabla\phi_{i_t}(\mathbf{w}^t), \mathbf{w}\rangle + \lambda r(\mathbf{w}) + \frac{1}{\eta_t}\mathcal{B}_\psi(\mathbf{w}, \mathbf{w}^t) \right], \tag{2}$$

where $\mathcal{B}_\psi$ is a Bregman divergence and $\nabla\phi_{i_t}(\mathbf{w}^t)$ denotes an arbitrary (sub-)gradient of ϕ_{i_t}. Intuitively, this method works by minimizing a first-order approximation of the function ϕ_{i_t} at the current iterate $\mathbf{w}^t$ plus the regularizer

$\lambda r(\mathbf{w})$, and forcing the next iterate $\mathbf{w}^{t+1}$ to lie close to $\mathbf{w}^t$. The step size η_t is a trade-off between these two objectives.

We assume that the exact solution of the above optimization (2) can be efficiently obtained. For example, when $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$, we have $\mathcal{B}_\psi(\mathbf{u}, \mathbf{v}) = \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2$, and the above optimization will produce the $t+1$-th iterate as: $\mathbf{w}^{t+1} = \text{prox}_{\eta_t \lambda r}(\mathbf{w}^t - \eta_t \nabla \phi_{i_t}(\mathbf{w}^t))$, where $\text{prox}_h(\mathbf{x}) = \arg\min_{\mathbf{w}} \left(h(\mathbf{w}) + \frac{1}{2}\|\mathbf{w} - \mathbf{x}\|_2^2\right)$. Furthermore, it is also assumed that the proximal mapping of $\eta_t \lambda r(\mathbf{w})$, i.e., $\text{prox}_{\eta_t \lambda r}(\mathbf{x})$, is easy to compute. For example, when $r(\mathbf{w}) = \|\mathbf{w}\|_1$, the proximal mapping of $\lambda r(\mathbf{w})$ is the shrinkage operation $\text{prox}_{\lambda r}(\mathbf{x}) = \text{sign}(\mathbf{x}) \odot [|\mathbf{x}| - \lambda]_+$, where $\odot$ is the element-wise vector product.

A disadvantage of this method is that the randomness introduces variance - this is caused by the fact that $\nabla \phi_{i_t}(\mathbf{w}^t)$ equals the gradient $\nabla f(\mathbf{w}^t)$ in expectation, but $\nabla \phi_i(\mathbf{w}^t)$ varies with i. In particular, if the stochastic gradient has a large variance, then the convergence will become slow. This paper studies prox-SMD with importance sampling to reduce the variance of stochastic gradient. The idea of importance sampling can be described as follows: at the t-th step, we assign each $i \in \{1, \ldots, n\}$ a probability $p_i^t \geq 0$ such that $\sum_{i=1}^n p_i^t = 1$; we then sample i_t from $\{1, \ldots, n\}$ based on the probability $\mathbf{p}^t = (p_1^t, \ldots, p_n^t)^\top$. If we adopt this distribution, then proximal SMD with importance sampling is obtained by setting $\mathbf{w}^{t+1}$ as the solution of

$$\min_{\mathbf{w}} \left[\langle \frac{\nabla \phi_{i_t}(\mathbf{w}^t)}{n p_{i_t}^t}, \mathbf{w} \rangle + \lambda r(\mathbf{w}) + \frac{1}{\eta_t} \mathcal{B}_\psi(\mathbf{w}, \mathbf{w}^t) \right], \quad (3)$$

which is another unbiased estimation of the optimization problem for prox-MD (or composite objective mirror descent), because $\mathbb{E}[(n p_{i_t}^t)^{-1} \nabla \phi_{i_t}(\mathbf{w}^t) | \mathbf{w}^t] = \nabla f(\mathbf{w}^t)$.

The main question is: what choice of $\mathbf{p}^t$ can optimally reduce the variance of the stochastic gradient. To answer this question, we first prove a lemma that establishes a relationship between $\mathbf{p}^t$ and the convergence rate of prox-SMD with importance sampling.

Lemma 1. *Define $\mathbf{w}^{t+1}$ by the update (3). Assume that $\psi(\cdot)$ is σ-strongly convex with respect to a norm $\|\cdot\|$ (its dual norm is $\|\cdot\|_*$), and f is μ-strongly convex and $(1/\gamma)$-smooth with respect to ψ. If $r(\mathbf{w})$ is convex and $\eta_t \in (0, \gamma]$, then $\mathbf{w}^{t+1}$ satisfies the following inequality for any $t \geq 1$,*

$$\mathbb{E}[P(\mathbf{w}^{t+1}) - P(\mathbf{w}^*)] \leq \frac{1}{\eta_t} \mathbb{E}[\mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^t) - \mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^{t+1})] - \mu \mathbb{E} \mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^t) + \frac{\eta_t}{\sigma} \mathbb{E} \mathbb{V}\left((n p_{i_t}^t)^{-1} \nabla \phi_{i_t}(\mathbf{w}^t) \right),$$

where the variance is defined as $\mathbb{V}((n p_{i_t}^t)^{-1} \nabla \phi_{i_t}(\mathbf{w}^t)) = \mathbb{E}\|(n p_{i_t}^t)^{-1} \nabla \phi_{i_t}(\mathbf{w}^t) - \nabla f(\mathbf{w}^t)\|_*^2$, and the expectation is taken with the distribution $\mathbf{p}^t$.

From the above analysis, we can observe that the smaller the variance, the more reduction on objective function we have. In the next subsection, we will study how to adopt importance sampling to reduce the variance. This observation will be made more rigorous below.

3.1.1. ALGORITHM

According to Lemma 1 , in order to maximize the reduction on the objective value, we should choose $\mathbf{p}^t$ as the solution of the following optimization

$$\min_{\mathbf{p}^t \in \triangle^n} \mathbb{V}\left(\frac{\nabla \phi_{i_t}(\mathbf{w}^t)}{n p_{i_t}^t}\right) \Leftrightarrow \min_{\mathbf{p}^t \in \triangle^n} \frac{1}{n^2} \sum_{i=1}^n \frac{\|\nabla \phi_i(\mathbf{w}^t)\|_*^2}{p_i^t}, \quad (4)$$

where $\triangle^n$ is the n-dimensional simplex. It is easy to verify, that the solution of the above optimization is

$$p_i^t = \frac{\|\nabla \phi_i(\mathbf{w}^t)\|_*}{\sum_{j=1}^n \|\nabla \phi_j(\mathbf{w}^t)\|_*}, \quad \forall i \in \{1, 2, \ldots, n\}. \quad (5)$$

Although, this distribution can minimize the variance of the t-th stochastic gradient, it requires the calculation of n derivatives at each step, which is clearly inefficient. To solve this issue, a potential remedy is to calculate the n derivatives at some steps and then keep it for use for a relatively long time period. In addition, the true derivatives will changes every step, and thus it is beneficial to add a small constant to the sampling probability. Another practical solution is to relax the previous optimization (4) as follows

$$\min_{\mathbf{p}^t \in \triangle^n} \frac{1}{n^2} \sum_{i=1}^n \frac{\|\nabla \phi_i(\mathbf{w}^t)\|_*^2}{p_i^t} \leq \min_{\mathbf{p}^t \in \triangle^n} \frac{1}{n^2} \sum_{i=1}^n \frac{G_i^2}{p_i^t} \quad (6)$$

by introducing upperbounds

$$G_i \geq \|\nabla \phi_i(\mathbf{w}^t)\|_*, \quad \forall t.$$

Using this approach, we can approximate the distribution in (5) by solving the the right hand side of (6) as

$$p_i^t = \frac{G_i}{\sum_{j=1}^n G_j}, \quad \forall i \in \{1, 2, \ldots, n\},$$

which is independent of t.

Based on the above solution, we will suggest distributions for two kinds of loss functions - Lipschitz functions and smooth functions. First, if each $\phi_i(\mathbf{w})$ is L_i-Lipschitz w.r.t. $\|\cdot\|_*$, then $\|\nabla \phi_i(\mathbf{w})\|_* \leq L_i$ for any $\mathbf{w} \in \mathbb{R}^d$, and the suggested distribution is

$$p_i^t = \frac{L_i}{\sum_{j=1}^n L_j}, \quad \forall i \in \{1, 2, \ldots, n\}.$$

Second, if $\phi_i(\mathbf{w})$ is $(1/\gamma_i)$-smooth and $\|\mathbf{w}^t\| \leq R$ for any t (this is possible when the feasible domain is bounded), then $\|\nabla \phi_i(\mathbf{w}^t)\|_* \leq R/\gamma_i$, and the distribution becomes

$$p_i^t = \frac{\frac{1}{\gamma_i}}{\sum_{j=1}^n \frac{1}{\gamma_j}}, \quad \forall i \in \{1, 2, \ldots, n\}.$$

Finally, we can summarize the proposed Proximal SMD with importance sampling in Algorithm 1.

Algorithm 1 Proximal Stochastic Mirror Descent with Importance Sampling (Iprox-SMD)

Input: $\lambda \geq 0$, the learning rates $\eta_1, \ldots, \eta_T > 0$.
Initialize: $\mathbf{w}^1 = 0$, $p_i = \frac{L_i}{\sum_j L_j}$ or $p_i = \frac{1/\gamma_i}{\sum_j 1/\gamma_j}$, $\forall i$.
for $t = 1, \ldots, T$ **do**
Sample i_t from $\{1, \ldots, n\}$ based on $\mathbf{p}$;

$$\mathbf{w}^{t+1} = \arg\min_{\mathbf{w}} \Big[\big\langle (np_{i_t})^{-1} \nabla \phi_{i_t}(\mathbf{w}^t), \mathbf{w} \big\rangle + \lambda r(\mathbf{w}) + \frac{1}{\eta_t} \mathcal{B}_\psi(\mathbf{w}, \mathbf{w}^t) \Big];$$

end for

3.1.2. Analysis

Before presenting the results, we make some general assumptions: $r(\mathbf{0}) = 0$, and $r(\mathbf{w}) \geq 0$, for all $\mathbf{w}$. It is easy to see that these two assumptions are generally satisfied by most of the well-known regularizers.

Under the above assumptions, we first prove a convergence result for Proximal SMD with importance sampling using the previous Lemma 1.

Theorem 1. *Assume that $\psi(\cdot)$ is σ-strongly convex with respect to a norm $\|\cdot\|$, f is μ-strongly convex and $(1/\gamma)$-smooth with respect to ψ, $r(\mathbf{w})$ is convex and $\eta_t = \frac{1}{\alpha + \mu t}$ with $\alpha \geq 1/\gamma - \mu$. If we further assume $\phi_i(\mathbf{w})$ is $(1/\gamma_i)$-smooth, $\|\mathbf{w}^t\| \leq R$ for any t, and the distribution is set as $p_i^t = \frac{R/\gamma_i}{\sum_{j=1}^n R/\gamma_j}$, then the following inequality holds for any $T \geq 1$,*

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} P(\mathbf{w}^{t+1}) - P(\mathbf{w}^*) \leq O\left[\frac{(\sum_{i=1}^n R/\gamma_i)^2}{\sigma \mu n^2} \frac{\ln(\alpha + \mu T)}{T} \right].$$

In addition, if $\mu = 0$, the above bound is invalid, however if η_t is set as $\sqrt{\sigma \mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^1)}/(\sqrt{T} \frac{\sum_{i=1}^n R/\gamma_i}{n})$, we can prove the following inequality for any $T \geq 1$,

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} P(\mathbf{w}^{t+1}) - P(\mathbf{w}^*) \leq 2 \sqrt{\frac{\mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^1)}{\sigma}} \frac{\sum_{i=1}^n R/\gamma_i}{n} \frac{1}{\sqrt{T}}.$$

Remark: If $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ and $r(\mathbf{w}) = 0$, then $\mathcal{B}_\psi(\mathbf{u}, \mathbf{v}) = \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2$, and the proposed algorithm becomes SGD with importance sampling. Under these assumptions, it is known that one may get rid of the $\ln T$ factor in the convergence bound, when the objective function is strongly convex. For simplicity, we do not provide the details.

Remark. If the uniform distribution is adopted, it is easy to observe that the variance of stochastic gradient is bounded by $\frac{\sum_{i=1}^n (R/\gamma_i)^2}{n}$. Hence Theorem 1 results in an upper bound for $\frac{1}{T}\sum_{t=1}^T \mathbb{E} P(\mathbf{w}^{t+1}) - P(\mathbf{w}^*)$ of the form $O\left(\frac{\sum_{i=1}^n (R/\gamma_i)^2}{\sigma \mu n} \frac{\ln(\alpha + \mu T)}{T}\right)$ for strongly convex f, and of the form $2\sqrt{\frac{\mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^1)}{\sigma}} \frac{\sum_{i=1}^n (R/\gamma_i)^2}{n} \frac{1}{\sqrt{T}}$ for general convex f. According to the Cauchy-Schwarz inequality,

$$\frac{\sum_{i=1}^n (R/\gamma_i)^2}{n} \Big/ \left(\frac{\sum_{i=1}^n R/\gamma_i}{n}\right)^2 = \frac{n \sum_{i=1}^n (R/\gamma_i)^2}{(\sum_{i=1}^n R/\gamma_i)^2} \geq 1.$$

It implies that importance sampling always improves the convergence rate, especially when $\frac{(\sum_{i=1}^n R/\gamma_i)^2}{\sum_{i=1}^n (R/\gamma_i)^2} \ll n$.

If f is convex, we can provide the following convergence results using the analysis of (Duchi et al., 2010).

Theorem 2. *Assume that $\psi(\cdot)$ is σ-strongly convex with respect to a norm $\|\cdot\|$, f and $r(\mathbf{w})$ are convex, and $\eta_t = \eta$. If we further assume $\phi_i(\mathbf{w})$ is $(1/\gamma_i)$-smooth, $\|\mathbf{w}^t\| \leq R$ for any t, and the distribution is set as $p_i^t = \frac{R/\gamma_i}{\sum_{j=1}^n R/\gamma_j}$, then when η_t is set as $\sqrt{2\sigma \mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^1)}/(\frac{\sum_{i=1}^n R/\gamma_i}{n}\sqrt{T})$, the following inequality holds for any $T \geq 1$,*

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} P(\mathbf{w}^t) - P(\mathbf{w}^*) \leq \sqrt{\mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^1) \frac{2}{\sigma}} \left(\frac{\sum_{i=1}^n R/\gamma_i}{n}\right) \frac{1}{\sqrt{T}}.$$

If $\phi_i(\mathbf{w})$ is L_i-Lipschitz, and the distribution is set as $p_i = L_i / \sum_{j=1}^n L_j$, $\forall i$, then when η_t is set as $\sqrt{2\sigma \mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^1)}/(\frac{\sum_{i=1}^n L_i}{n}\sqrt{T})$, the following inequality holds for any $T \geq 1$,

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} P(\mathbf{w}^t) - P(\mathbf{w}^*) \leq \sqrt{\mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^1) \frac{2}{\sigma}} \left(\frac{\sum_{i=1}^n L_i}{n}\right) \frac{1}{\sqrt{T}}.$$

Remark: If the uniform distribution is adopted, it is easy to observe that the variance of stochastic gradient is bounded by $\frac{\sum_{i=1}^n (R/\gamma_i)^2}{n}$ for smooth $\phi_i(\cdot)$, and bounded by $\frac{\sum_{i=1}^n (L_i)^2}{n}$ for Lipschitz $\phi_i(\cdot)$. Theorem 2 results in an upper bound for $\frac{1}{T}\sum_{t=1}^T \mathbb{E} P(\mathbf{w}^t) - P(\mathbf{w}^*)$ of the form $\sqrt{\frac{2\mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^1) \sum_{i=1}^n (R/\gamma_i)^2}{\sigma n T}}$ for smooth ϕ_i, and of the form $\sqrt{\frac{2\mathcal{B}_\psi(\mathbf{w}^*, \mathbf{w}^1) \sum_{i=1}^n (L_i)^2}{\sigma n T}}$ for Lipschitz ϕ_i. However, according to the Cauchy-Schwarz inequality,

$$\frac{n \sum_{i=1}^n (R/\gamma_i)^2}{(\sum_{i=1}^n R/\gamma_i)^2} \geq 1, \quad \frac{n \sum_{i=1}^n L_i^2}{(\sum_{i=1}^n L_i)^2} \geq 1,$$

implies that importance sampling improves the convergence bound, especially when $\frac{(\sum_{i=1}^n R/\gamma_i)^2}{\sum_{i=1}^n (R/\gamma_i)^2} \ll n$, and when $\frac{(\sum_{i=1}^n L_i)^2}{\sum_{i=1}^n (L_i)^2} \ll n$.

3.2. prox-SDCA with Importance Sampling

In this section, we study the Proximal Stochastic Dual Coordinate Ascent method (prox-SDCA) with importance

sampling. Prox-SDCA deals with the dual problem of (1):

$$\max_{\theta} D(\theta) := \frac{1}{n}\sum_{i=1}^{n} -\phi_i^*(-\theta_i) - \lambda r^*(\frac{1}{\lambda n}\sum_{i=1}^{n}\theta_i). \quad (7)$$

We assume that $r^*(\cdot)$ is continuously differentiable; the relationship between the primal variable $\mathbf{w}$ and dual variable θ is $\mathbf{w} = \nabla r^*(\mathbf{v}(\theta))$, $\mathbf{v}(\theta) = \frac{1}{\lambda n}\sum_{i=1}^{n}\theta_i$. We also assume that $r(\mathbf{w})$ is 1-strongly convex with respect to a norm $\|\cdot\|_{P'}$, i.e., $r(\mathbf{w}+\Delta\mathbf{w}) \geq r(\mathbf{w}) + \nabla r(\mathbf{w})^\top \Delta\mathbf{w} + \frac{1}{2}\|\Delta\mathbf{w}\|_{P'}^2$, which means that $r^*(\mathbf{w})$ is 1-smooth with respect to its dual norm $\|\cdot\|_{D'}$. Namely, $r^*(\mathbf{v}+\Delta\mathbf{v}) \leq h(\mathbf{v};\Delta\mathbf{v})$, where $h(\mathbf{v};\Delta\mathbf{v}) := r^*(\mathbf{v}) + \nabla r^*(\mathbf{v})^\top\Delta\mathbf{v} + \frac{1}{2}\|\Delta\mathbf{v}\|_{D'}^2$.

At the t-th step, the Proximal Stochastic Dual Coordinate Ascent method (prox-SDCA) picks $i \in \{1,\ldots,n\}$ uniformly at random, and update the dual variable θ_i^{t-1} as:

$$\theta_i^t = \theta_i^{t-1} + \Delta\theta_i^{t-1},$$

where $\Delta\theta_i^{t-1}$ is the solution of

$$\max_{\Delta\theta_i}\left[-\phi_i^*(-(\theta_i^{t-1}+\Delta\theta_i)) - (\mathbf{w}^{t-1})^\top\Delta\theta_i - \frac{1}{2\lambda n}\|\Delta\theta_i\|_D^2\right], \quad (8)$$

which is equivalent to maximizing a lower bound of the following problem with $\mathbf{v}^{t-1} = \frac{1}{\lambda n}\sum_{i=1}^{n}\theta_i^{t-1}$

$$\max_{\Delta\theta_i}\left[-\frac{1}{n}\phi_i^*(-(\theta_i^{t-1}+\Delta\theta_i)) - \lambda r^*(\mathbf{v}^{t-1} + \frac{1}{\lambda n}\Delta\theta_i)\right].$$

However, the optimization (8) may not have a closed form solution, and in prox-SDCA we may adopt other update rules $\Delta\theta_i = s(\mathbf{u} - \theta_i^{t-1})$ for an appropriately chosen step size parameter $s > 0$ and any vector $\mathbf{u} \in \mathbb{R}^d$ such that $-\mathbf{u} \in \partial\phi_i(\mathbf{w}^{t-1})$. Note that when $r(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$, prox-SDCA is also known as SDCA.

In the following, we study prox-SDCA with importance sampling, which is to allow the algorithm to randomly pick i according to probability p_i, which is the i-th element of $\mathbf{p} \in \mathbb{R}_+^n$, $\sum_i p_i = 1$. Once we pick the coordinate i, θ_i is updated as traditional prox-SDCA. The main question we are interested in here is which $\mathbf{p} = (p_1,\ldots,p_n)^\top$ can optimally accelerate the convergence rate of prox-SDCA. To answer this question, we will introduce a lemma which will state the relationship between $\mathbf{p}$ and the convergence rate of prox-SDCA with importance sampling.

Lemma 2. *Given a distribution $\mathbf{p}$, if assume ϕ_i is $(1/\gamma_i)$-smooth with norm $\|\cdot\|_P$, then for any iteration t and any s such that $s_i = s/(p_i n) \in [0,1]$, $\forall i$, we have*

$$\mathbb{E}[D(\theta^t) - D(\theta^{t-1})] \geq \frac{s}{n}\mathbb{E}[P(\mathbf{w}^{t-1}) - D(\theta^{t-1})] - \frac{sG^t}{2\lambda n^2}, \quad (9)$$

where $G^t = \frac{1}{n}\sum_{i=1}^{n}(s_i R^2 - \gamma_i(1-s_i)\lambda n)\mathbb{E}\|\mathbf{u}_i^{t-1} - \theta_i^{t-1}\|_D^2$, $R = \sup_{\mathbf{u}\neq 0}\|\mathbf{u}\|_{D'}/\|\mathbf{u}\|_D$, and $-\mathbf{u}_i^{t-1} \in \partial\phi_i(\mathbf{w}^{t-1})$.

For many interesting cases, it is easy to estimate $R = \sup_{\mathbf{u}\neq 0}\|\mathbf{u}\|_{D'}/\|\mathbf{u}\|_D$. For example, if $p > r > 0$, then $\|\mathbf{w}\|_p \leq \|\mathbf{w}\|_r \leq d^{(1/r-1/p)}\|\mathbf{w}\|_p$ for any $\mathbf{w} \in \mathbb{R}^d$.

3.2.1. Algorithm

According to Lemma 2, to maximize the dual ascent for the t-th update, we should choose s and $\mathbf{p}$ as the solution of the following optimization

$$\max_{s/(p_i n)\in[0,1],\mathbf{p}\in\triangle^n} \frac{s}{n}\mathbb{E}[P(\mathbf{w}^{t-1}) - D(\theta^{t-1})] - \frac{s}{n^2}\frac{G^t}{2\lambda}.$$

where $\triangle^n$ is the n-dimensional simplex. However, because this optimization problem is difficult to solve, we choose to relax it as follows:

$$\begin{aligned}
&\max_{\frac{s}{p_i n}\in[0,1],\mathbf{p}\in\triangle^n} \frac{s}{n}\mathbb{E}[P(\mathbf{w}^{t-1}) - D(\theta^{t-1})] - \frac{s}{n^2}\frac{G^t}{2\lambda} \\
\geq &\max_{\frac{s}{p_i n}\in[0,\frac{\lambda n\gamma_i}{R^2+\lambda n\gamma_i}],\mathbf{p}\in\triangle^n} \frac{s}{n}\mathbb{E}[P(\mathbf{w}^{t-1}) - D(\theta^{t-1})] - \frac{s}{n^2}\frac{G^t}{2\lambda} \\
\geq &\max_{\frac{s}{p_i n}\in[0,\frac{\lambda n\gamma_i}{R^2+\lambda n\gamma_i}],\mathbf{p}\in\triangle^n} \frac{s}{n}\mathbb{E}[P(\mathbf{w}^{t-1}) - D(\theta^{t-1})].
\end{aligned}$$

where the last inequality has used $G^t = \frac{1}{n}\sum_{i=1}^{n}(s_i R^2 - \gamma_i(1-s_i)\lambda n)\mathbb{E}\|\mathbf{u}_i^{t-1} - \theta_i^{t-1}\|_D^2 \leq 0$, since $s_i = s/(p_i n) \leq \frac{\lambda n\gamma_i}{R^2+\lambda n\gamma_i}$. To optimize the final relaxation, we have the following proposition

Proposition 1. *The solution to the optimization problem*

$$\max_{s,\mathbf{p}} s \quad s.t.\ s/(p_i n) \in [0, \frac{\lambda n\gamma_i}{R^2+\lambda n\gamma_i}],\ \mathbf{p} \in \triangle^n$$

is given by

$$s = \frac{n}{n + \sum_{i=1}^{n}\frac{R^2}{\lambda n\gamma_i}},\ p_i = \frac{1 + \frac{R^2}{\lambda n\gamma_i}}{n + \sum_{j=1}^{n}\frac{R^2}{\lambda n\gamma_j}}. \quad (10)$$

We omit the proof since it is simple. Given that ϕ_i is $(1/\gamma_i)$-smooth, $\forall i \in \{1,\ldots,n\}$, the sampling distribution should be set as in (10).

When $\gamma_i = 0$, the above distribution in the equation (10) is not valid. To solve this problem, we combine the facts

$$P(\mathbf{w}^{t-1}) - D(\theta^{t-1}) \geq D(\theta^*) - D(\theta^{t-1}) := \epsilon_D^{t-1},$$

where θ^* is the optimal solution of the dual problem $\max_\theta D(\theta)$, $D(\theta^t) - D(\theta^{t-1}) = \epsilon_D^{t-1} - \epsilon_D^t$, and the inequality (9), to obtain

$$\mathbb{E}[\epsilon_D^t] \leq (1 - \frac{s}{n})\mathbb{E}[\epsilon_D^{t-1}] + \frac{s}{2\lambda n^2}G^t. \quad (11)$$

According to this inequality, although every $\gamma_i = 0$, if we

further assume every ϕ_i is L_i-Lipschitz, then

$$\begin{aligned} G^t &= \frac{1}{n}\sum_{i=1}^{n}(s_i R^2 - \gamma_i(1-s_i)\lambda n)\mathbb{E}\|\mathbf{u}_i^{t-1} - \theta_i^{t-1}\|_D^2 \\ &\leq \frac{4R^2 s}{n^2}\sum_{i=1}^{n}\frac{1}{p_i}L_i^2, \end{aligned} \tag{12}$$

where we use $s_i = s/(np_i)$, $\|\mathbf{u}_i^{t-1}\| \leq L_i$ and $\|\theta_i^{t-1}\| \leq L_i$, since $-\mathbf{u}_i^{t-1}, -\theta_i^{t-1} \in \partial\phi_i(\mathbf{w}^{t-1})$. Combining the above two inequalities results in

$$\mathbb{E}[\epsilon_D^t] \leq (1-\frac{s}{n})\mathbb{E}[\epsilon_D^{t-1}] + \frac{s}{2\lambda n^2}\frac{4R^2 s}{n^2}\sum_{i=1}^{n}\frac{1}{p_i}L_i^2. \tag{13}$$

According to the above inequality, to minimize the t-th duality gap, we should choose a proper distribution to optimize the problem $\min_{\mathbf{p}\in\triangle^n}\sum_{i=1}^{n}\frac{1}{p_i}L_i^2$, for which the optimal distribution is obviously

$$p_i = L_i/\sum_{j=1}^{n}L_j.$$

Because $s_i = s/(np_i) \in [0,1]$, the above distribution further implies

$$s \in \bigcap_{i=1}^{n}[0, np_i] = \left[0, \frac{nL_{min}}{\sum_{j=1}^{n}L_j}\right] := [0,\rho],$$

where $L_{min} = \min\{L_1, L_2, \ldots, L_n\}$ and $\rho \leq 1$.

In summary, prox-SDCA with importance sampling can be described in Algorithm 2.

Algorithm 2 Proximal Stochastic Dual Coordinate Ascent with Importance Sampling (Iprox-SDCA)

Input: $\lambda > 0$, $R = \sup_{\mathbf{u}\neq 0}\|\mathbf{u}\|_{D'}/\|\mathbf{u}\|_D$, norms $\|\cdot\|_D$, $\|\cdot\|_{D'}$, $\gamma_1, \ldots, \gamma_n > 0$, or $L_1, \ldots, L_n \geq 0$.

Initialize: $\theta_i^0 = 0$, $\mathbf{w}^0 = \nabla r^*(0)$, $p_i = \frac{1+\frac{R^2}{\lambda n\gamma_i}}{n+\sum_{j=1}^{n}\frac{R^2}{\lambda n\gamma_j}}$, or $p_i = \frac{L_i}{\sum_{j=1}^{n}L_j}$, $\forall i \in \{1,\ldots,n\}$.

for $t = 1, \ldots, T$ **do**

Sample i_t from $\{1,\ldots,n\}$ based on $\mathbf{p}$;

$$\Delta\theta_{i_t}^{t-1} = \arg\max_{\Delta\theta_{i_t}}\left[-\phi_{i_t}^*(-(\theta_{i_t}^{t-1}+\Delta\theta_{i_t})) - (\mathbf{w}^{t-1})^\top\Delta\theta_{i_t} - \frac{1}{2\lambda n}\|\Delta\theta_{i_t}\|_{D'}^2\right];$$

$\theta_{i_t}^t = \theta_{i_t}^{t-1} + \Delta\theta_{i_t}^{t-1}$;

$\mathbf{v}^t = \mathbf{v}^{t-1} + \frac{1}{\lambda n}\Delta\theta_{i_t}^{t-1}$;

$\mathbf{w}^t = \nabla r^*(\mathbf{v}^t)$;

end for

3.2.2. Analysis

Before presenting the theoretical results, we will make several assumptions without loss of generality: a) for the loss functions: $\phi_i(0) \leq 1$, and $\forall\mathbf{w},\ \phi_i(\mathbf{w}) \geq 0$, and b) for the regularizer: $r(0) = 0$ and $\forall\mathbf{w},\ r(\mathbf{w}) \geq 0$. Then, we have the following theorem for the expected duality gap when the loss functions are smooth.

Theorem 3. *Assume ϕ_i is $(1/\gamma_i)$-smooth $\forall i \in \{1,\ldots,n\}$ and set $p_i = (1+\frac{R^2}{\lambda n\gamma_i})/(n+\sum_{j=1}^{n}\frac{R^2}{\lambda n\gamma_j})$, for all $i \in \{1,\ldots,n\}$. To obtain an expected duality gap of $\mathbb{E}[P(\mathbf{w}^t) - D(\theta^T)] \leq \epsilon_P$ for the proposed Proximal S-DCA with importance sampling, it suffices to have a total number of iterations of*

$$T \geq (n + \sum_{i=1}^{n}\frac{R^2}{\lambda n\gamma_i})\log\left((n+\sum_{i=1}^{n}\frac{R^2}{\lambda n\gamma_i})\frac{1}{\epsilon_P}\right).$$

Remark: If we employ uniform sampling, i.e., $p_i = 1/n\ \forall i$, then we have to use the same γ for all ϕ_i by choosing $\gamma_{min} = \min\{\gamma_1,\ldots,\gamma_n\}$. By replacing γ_i with γ_{min}, the theorem recovers a related result of (Shalev-Shwartz & Zhang, 2012b) under uniform sampling, i.e., $T \geq (n+\frac{R^2}{\lambda\gamma_{min}})\log\left((n+\frac{R^2}{\lambda\gamma_{min}})\frac{1}{\epsilon_P}\right)$. Since

$$\frac{n+\frac{R^2}{\lambda\gamma_{min}}}{n+\sum_{i=1}^{n}\frac{R^2}{\lambda\gamma_i n}} = \frac{n\lambda\gamma_{min}+R^2}{n\lambda\gamma_{min}+\frac{R^2}{n}\sum_{i=1}^{n}\frac{\gamma_{min}}{\gamma_i}} \geq 1,$$

the bound for importance sampling is always better, especially when $\sum_{i=1}^{n}\frac{\gamma_{min}}{\gamma_i} \ll n$.

For non-smooth loss functions, the convergence rate for Proximal SDCA with importance sampling is given below.

Theorem 4. *Consider the proposed proximal SDCA with importance sampling. Assume that ϕ_i is L_i-Lipschitz and set $p_i = L_i/\sum_{j=1}^{n}L_j$, $\forall i \in \{1,\ldots,n\}$. To obtain an expected duality gap of $\mathbb{E}[P(\bar{\mathbf{w}}) - D(\bar{\theta})] \leq \epsilon_P$ where $\bar{\mathbf{w}} = \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\mathbf{w}^{t-1}$ and $\bar{\theta} = \frac{1}{T-T_0}\sum_{t=T_0+1}^{T}\theta^{t-1}$, it suffices to have a total number of iterations of*

$$\begin{aligned} T &\geq T_0 + n/\rho + \frac{4R^2(\sum_{i=1}^{n}L_i)^2}{n^2\lambda\epsilon_P} \\ &\geq \omega + n/\rho + \frac{20R^2(\sum_{i=1}^{n}L_i)^2}{n^2\lambda\epsilon_P}, \end{aligned}$$

where $\omega = \max(0, \lceil\frac{n}{\rho}\log(\frac{\lambda n}{\rho 2R^2(\sum_{i=1}^{n}L_i)^2/n^2})\rceil)$, and $\rho = \frac{nL_{min}}{\sum_{i=1}^{n}L_i}$. Moreover, when $t \geq T_0$, we have dual sub-optimality bound of $\mathbb{E}[D(\theta^*) - D(\theta^t)] \leq \epsilon_P/2$.

Remark: If we replace all L_i by $L_{max} = \max\{L_1,\ldots,L_n\}$, the theorem is still valid, and the sampling distribution becomes the uniform distribution. In this case we recover a related result of (Shalev-Shwartz & Zhang, 2012b), i.e.,

$T \geq \max(0, 2\lceil n \log(\frac{\lambda n}{2R^2 L_{max}^2})\rceil) - n + \frac{20R^2(L_{max})^2}{\lambda \epsilon_P}$. However, the ratio of the leading terms is

$$\frac{(L_{max})^2}{(\sum_{i=1}^n L_i)^2/n^2} = (\frac{n}{\sum_{i=1}^n L_i/L_{max}})^2 \geq 1,$$

which again implies that the importance sampling bound is always better, especially when $(\sum_{i=1}^n \frac{L_i}{L_{max}})^2 \ll n^2$.

4. Experimental Results

4.1. Experimental Testbed and Setup

For simplicity, in the experiments we only consider the task of optimizing squared hinge loss based SVM with ℓ_2 regularization: $\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^n \left([1 - y_i \mathbf{w}^\top \mathbf{x}_i]_+\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$. Moreover we set $\psi = \frac{1}{2}\|\cdot\|_2^2$. In this case, we compare importance sampling versus the standard uniform sampling using Pegasos of Shalev-Shwartz et al. (2007)) for SGD, and using SDCA (Shalev-Shwartz & Zhang, 2013).

For Iprox-SGD, using the inequality $P(\mathbf{w}^*) = D(\theta^*)$, we can get $\|\mathbf{w}^*\|_2 \leq 1/\sqrt{\lambda}$. Thus the theoretical analysis is still valid if we project the iterative solutions onto $\{\mathbf{w} \in \mathbb{R}^d | \|\mathbf{w}\|_2 \leq 1/\sqrt{\lambda}\}$ using Euclidean distance. Setting $\phi_i(\mathbf{w}) = \left([1 - y_i \mathbf{w}^\top \mathbf{x}_i]_+\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$ so that $\nabla\phi_i(\mathbf{w}) = -2[1 - y_i \mathbf{w}^\top \mathbf{x}_i]_+ y_i \mathbf{x}_i + \lambda \mathbf{w}$. Because $\|\nabla\phi_i(\mathbf{w})\|_2 \leq 2(1 + \|\mathbf{x}_i\|_2/\sqrt{\lambda})\|\mathbf{x}_i\|_2 + \sqrt{\lambda}$, according to our analysis, the optimal distribution is $p_i = \frac{2(1+\|\mathbf{x}_i\|_2/\sqrt{\lambda})\|\mathbf{x}_i\|_2 + \sqrt{\lambda}}{\sum_{j=1}^n [2(1+\|\mathbf{x}_j\|_2/\sqrt{\lambda})\|\mathbf{x}_j\|_2 + \sqrt{\lambda}]}$. Finally, $r(\mathbf{w}) = 0$ and $\text{prox}_{\lambda r}(\mathbf{x}) = \mathbf{x}$.

For Iprox-SDCA, we set $r(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$, which is 1-strongly convex with $\|\cdot\|_{P'} = \|\cdot\|_2$; we also have $\phi_i(\mathbf{w}) = \left([1 - y_i \mathbf{w}^\top \mathbf{x}_i]_+\right)^2$, which is $(2\|\mathbf{x}_i\|_2^2)$-smooth with respect to $\|\cdot\|_P = \|\cdot\|_2$. As a result, the optimal distribution for proximal SDCA with importance sampling should be $p_i = (1 + \frac{2\|\mathbf{x}_i\|_2^2}{\lambda n})/(n + \sum_{j=1}^n \frac{2\|\mathbf{x}_j\|_2^2}{\lambda n})$, where we used the fact $R = \sup_{\mathbf{u}\neq 0} \|\mathbf{u}\|_{D'}/\|\mathbf{u}\|_D = 1$. It can be derived that the dual function of $\phi(\cdot)$ is

$$\phi_i^*(-\theta) = \begin{cases} -\alpha + \alpha^2/4 & \theta = \alpha y_i \mathbf{x}_i, \alpha \geq 0 \\ \infty & \text{otherwise} \end{cases}.$$

The Iprox-SDCA method may employ the closed-form solution: $\Delta\theta_i = \max\left(\frac{1 - y_i \mathbf{w}^\top \mathbf{x}_i - \alpha_i/2}{1/2 + \|\mathbf{x}_i\|_2^2/(\lambda n)}, -\alpha_i\right) y_i \mathbf{x}_i$.

To evaluate the performance of our algorithms, the experiments were performed on several real world datasets downloaded from the LIBSVM website www.csie.ntu.edu.tw/~cjlin/libsvmtools/. The dataset characteristics are provided in the Table 1.

Table 1. Datasets used in the experiments.

Dataset	Dataset Size	Features
ijcnn1	49990	22
kdd2010(algebra)	8407752	20216830
w8a	49749	300

For fair comparison, all algorithms use the same setup in our experiments. In particular, the regularization parameter λ of SVM is set to 10^{-4}, 10^{-6}, 10^{-4} for ijcnn1, kdd2010(algebra), and w8a, respectively. For prox-SGD and Iprox-SGD, the step size is set to $\eta_t = 1/(\lambda t)$ for all the datasets.

Given these parameters, we estimated the ratios between the constants in the convergence bounds for uniform sampling and the proposed importance sampling strategies, which are listed in Table 2. These ratios imply that the importance sampling will be effective for SGD on kdd2010 and w8a, but not very effective for ijcnn1, which will be verified by empirical results. In addition, these ratios imply that importance sampling accelerates SDCA for all the datasets, which will also be demonstrated empirically.

Table 2. Theoretical Constant Ratios for The Datasets.

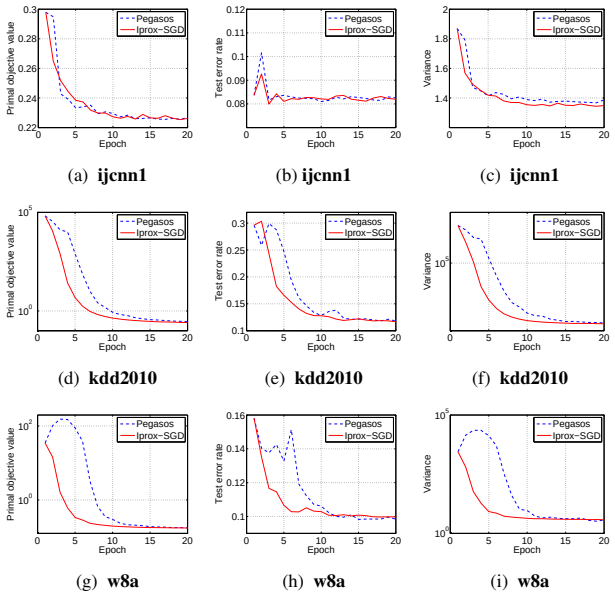
Constant Ratio	ijcnn1	kdd2010	w8a
$\frac{n\sum_{i=1}^n (G_i)^2}{(\sum_{i=1}^n G_i)^2}$	1.0643	1.4667	1.9236
$\frac{n\lambda\gamma_{min} + R^2}{n\lambda\gamma_{min} + \frac{R^2}{n}\sum_{i=1}^n \frac{\gamma_{min}}{\gamma_i}}$	1.1262	1.1404	1.3467

All experiments were conducted by fixing five different random seeds for each dataset, and the reported results were averaged over these five runs. We evaluated the learning performance by measuring the primal objective value ($P(\mathbf{w}^t)$) for SGD, and the duality gap ($P(\mathbf{w}^t) - D(\theta^t)$) for SDCA. In addition, to examine the generalization ability of the learning algorithms, we evaluated the test error rates. Moreover, we report the variances of the stochastic gradients of the two algorithms to check the effectiveness of importance sampling. Finally, for Iprox-SGD and Iprox-SDCA, the uniform sampling is adopted at the first epoch, so that the performance is the same with SGD and SDCA at the first epoch, respectively.

4.2. Evaluation on Iprox-SGD

Figure 1 summarizes results in terms of primal objective values, test error rates and variances of the stochastic gradients varying over the learning process on all the datasets for SGD and Iprox-SGD. Epoch for the horizontal axis is the number of iterations divided by dataset size.

First, the left column summarized the primal objective values of Iprox-SGD in comparison to SGD with uniform sampling on all the datasets. On the last two datasets, the proposed Iprox-SGD algorithm achieved the fastest convergence rates. Because these two algorithms adopted the same learning rates, this observation implies that the proposed importance sampling does sampled more informative stochastic gradient during the learning process. Second, the central column summarized the test error rates of the two algorithms, where Iprox-SGD achieves significantly smaller test error rates than those of SGD on the last two dataset. This indicates that the proposed importance sampling ap-

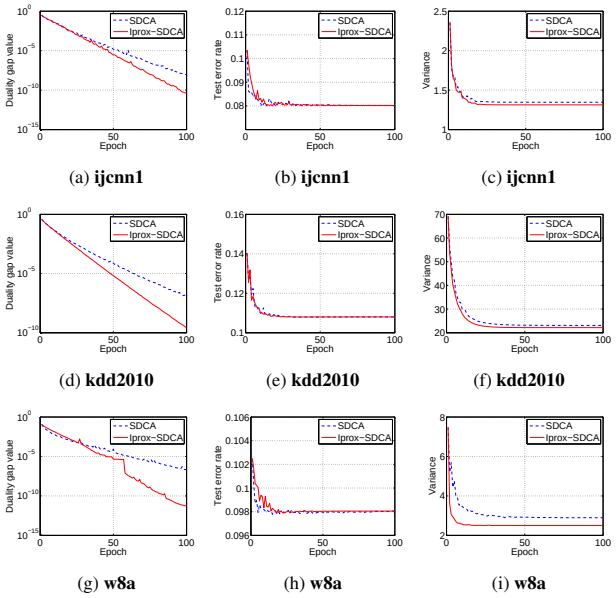


(a) **ijcnn1** (b) **ijcnn1** (c) **ijcnn1**

(d) **kdd2010** (e) **kdd2010** (f) **kdd2010**

(g) **w8a** (h) **w8a** (i) **w8a**

Figure 1. Comparison between Pegasos with Iprox-SGD.



(a) **ijcnn1** (b) **ijcnn1** (c) **ijcnn1**

(d) **kdd2010** (e) **kdd2010** (f) **kdd2010**

(g) **w8a** (h) **w8a** (i) **w8a**

Figure 2. Comparison between SDCA with Iprox-SDCA.

proach is effective in improving generalization ability. In addition, the right column shows the variances of stochastic gradients for the Iprox-SGD and SGD algorithms, where we can observe Iprox-SGD enjoys much smaller variances than SGD on the last two dataset. This again demonstrates that the proposed importance sampling strategy is effective in reducing the variance of the stochastic gradients. Finally, on the first dataset, the proposed Iprox-SGD algorithm achieved a convergence rate comparable to that of the traditional prox-SGD, which indicates that Iprox-SGD may degenerate into the traditional prox-SGD when the variance is not reduced.

4.3. Evaluation on Iprox-SDCA

Figure 2 summarizes experimental results in terms of duality gap values, test error rates and variances of the stochastic gradients varying over the learning process on all the datasets for SDCA and Iprox-SDCA.

We have several observations from these empirical results. First, the left column summarized the dual gap values of Iprox-SDCA in comparison to SDCA with uniform sampling on all the datasets. According to the dual gap values on all the datasets, the proposed Iprox-SDCA algorithm converges faster than the standard SDCA algorithm, which indicates that the proposed importance sampling strategy is more effective than uniform sampling. Second, the central column summarized the test error rates of the two algorithms, where the test error rates of Iprox-SDCA is comparable with those of SDCA on all the dataset. The results indicate that SDCA is quite fast at the first few epochs so that the importance sampling does not improve the test accuracy, although importance sampling can accelerate the minimization of duality gap. In addition, the right column shows the variances of stochastic gradients for the Iprox-SDCA and SDCA algorithms, where we can observe Iprox-SDCA enjoys slightly smaller variances than those of SDCA on all datasets. However the improvement is not large enough to significantly reduce the test error. This is because SDCA by itself is already a stochastic variance reduction gradient method (Johnson & Zhang, 2013). We believe if distributed prox-SDCA adopts this importance sampling strategy, then the corresponding improvement can be more significant.

5. Conclusion

This paper studies stochastic optimization with importance sampling, including importance sampling strategies for prox-SMD and prox-SDCA. For prox-SMD with importance sampling, our analysis shows that in order to reduce variance, the sample distribution should depend on the norms of the gradients of the loss functions, which can be relaxed to the smooth constants or the Lipschitz constants of all the loss functions; for prox-SDCA with importance sampling, our analysis shows that the sampling distribution should rely on the smooth constants or Lipschitz constants of all the loss functions. Compared to the traditional prox-SGD and prox-SDCA methods, we have shown that the proposed importance sampling methods can significantly improve the convergence rate under suitable conditions. Finally, we performed a set of empirical experiments to confirm the theoretical analysis.

Acknowledgments

The research of Peilin Zhao and Tong Zhang is partially supported by NSF-IIS-1407939 and NSF-IIS 1250985.

References

Duchi, John and Singer, Yoram. Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research*, 10:2899–2934, 2009.

Duchi, John C., Shalev-Shwartz, Shai, Singer, Yoram, and Tewari, Ambuj. Composite objective mirror descent. In *COLT*, pp. 14–26, 2010.

Hsieh, Cho-Jui, Chang, Kai-Wei, Lin, Chih-Jen, Keerthi, S Sathiya, and Sundararajan, Sellamanickam. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th international conference on Machine learning*, pp. 408–415. ACM, 2008.

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

Lacoste-Julien, Simon, Jaggi, Martin, Schmidt, Mark W., and Pletscher, Patrick. Stochastic block-coordinate frank-wolfe optimization for structural svms. *CoRR*, abs/1207.4747, 2012.

Lee, Yin Tat and Sidford, Aaron. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. *arXiv preprint arXiv:1305.1922*, 2013.

Luo, Zhi-Quan and Tseng, Paul. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

Mangasarian, Olvi L and Musicant, David R. Successive overrelaxation for support vector machines. *Neural Networks, IEEE Transactions on*, 10(5):1032–1037, 1999.

Needell, Deanna, Srebro, Nathan, and Ward, Rachel. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *arXiv preprint arXiv:1310.5715v3*, 2014.

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

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

Rakhlin, Alexander, Shamir, Ohad, and Sridharan, Karthik. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Roux, Nicolas L, Schmidt, Mark, and Bach, Francis R. 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 and Tewari, Ambuj. Stochastic methods for l_1-regularized loss minimization. *Journal of Machine Learning Research*, 12:1865–1892, 2011.

Shalev-Shwartz, Shai and Zhang, Tong. Proximal stochastic dual coordinate ascent. *CoRR*, abs/1211.2717, 2012a.

Shalev-Shwartz, Shai and Zhang, Tong. Proximal stochastic dual coordinate ascent. *arXiv preprint arXiv:1211.2717*, 2012b.

Shalev-Shwartz, Shai and Zhang, Tong. Stochastic dual coordinate ascent methods for regularized loss minimization. *JMLR*, pp. 567–599, 2013.

Shalev-Shwartz, Shai and Zhang, Tong. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *ICML*, 2014.

Shalev-Shwartz, Shai, Singer, Yoram, and Srebro, Nathan. Pegasos: Primal estimated sub-gradient solver for svm. In *ICML*, pp. 807–814, 2007.

Shamir, Ohad and Zhang, Tong. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 71–79, 2013.

Strohmer, Thomas and Vershynin, Roman. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2): 262–278, 2009.

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

Zhang, Tong. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *ICML*, 2004.

Zhao, Peilin and Zhang, Tong. Stochastic optimization with importance sampling. *arXiv preprint arXiv:1401.2753*, 2014.