Proximal Stochastic Dual Coordinate Ascent

1 Introduction

In this lecture, we still consider the composite optimization problem, but with an added finite sum structure as follows,

$$\phi(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(X_i^\top w) + \lambda g(w), \tag{1}$$

where $w \in \mathbb{R}^d$ is the model parameter:

This formulation comes from regularized loss minimization problems for linear prediction problems, such as structured output SVM, regularized multi-class logistic regression, etc. In such applications, $X_1, \ldots, X_n$ are training data in $\mathbb{R}^{d\times k}$, $f_1, \ldots, f_n$ be a sequence of convex loss functions $\mathbb{R}^k \to \mathbb{R}$, and $g(\cdot)$ is a convex regularization function defined on $\mathbb{R}^d$. The parameter $\lambda \geq 0$ is a regularization parameter.

In this lecture, we consider the dual formulation, which can be solved using stochastic coordinate ascent methods, which optimizes one dual variable at a time [1, 2].

2 Dual Formulation

In order to derive the dual formulation of (1), we use the decomposition technique, and rewrite it as:

$$\phi(w, \{u_i\}) = \frac{1}{n}\sum_{i=1}^{n} f_i(u_i) + \lambda g(w), \quad \text{subject to } \; \forall i, X_i^\top w = u_i.$$

The Lagrangian function, using the above decomposition, can be written as:

$$L(w, \{u_i\}, \alpha) = \frac{1}{n}\sum_{i=1}^{n} f_i(u_i) + \lambda g(w) + \frac{1}{n}\sum_{i=1}^{n} \alpha_i^\top (u_i - X_i^\top w).$$

Here we have n dual variables $\{\alpha_i\}_{1,\ldots,n}$, and each $\alpha_i \in \mathbb{R}^k$.

The dual objective function is defined as:

$$\phi_D(\alpha) = \min_{w,\{u_i\}} L(w, \{u_i\}, \alpha) = \frac{1}{n}\sum_{i=1}^{n} -f_i^*(-\alpha_i) - \lambda g^*\left(\frac{1}{\lambda n}\sum_{i=1}^{n} X_i\alpha_i\right). \tag{2}$$

The strong duality holds for this problem, we have at the optimal solution:

$$w^* \in \partial g^*\left(\frac{1}{\lambda n}\sum_{i=1}^{n} X_i\alpha_i^*\right).$$

In this lecture, we consider the case that $g(\cdot)$ is strongly convex, which means that $g^*(\cdot)$ is smooth, and thus its subgradient is unique. In this case, we may define the primal solution w from the dual variables as follows:

$$w = \nabla g^* \left(\frac{1}{\lambda n} \sum_{i=1}^{n} X_i \alpha_i \right). \tag{3}$$

Example 1 *In ridge regression, we have $k = 1$ and loss is:*

$$f_i(u) = \frac{1}{2}(u - y_i)^2,$$

and regularizer is

$$g(w) = \frac{1}{2}\|w\|_2^2.$$

The primal problem is:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(x_i^\top w - y_i)^2 + \frac{\lambda}{2}\|w\|_2^2.$$

The dual problem is:

$$\frac{1}{n}\sum_{i=1}^{n} - \left[-\alpha_i^\top y_i + \frac{1}{2}\alpha_i^2 \right] - \frac{1}{2\lambda n^2} \left\| \sum_{i=1}^{n} x_i \alpha_i \right\|_2^2,$$

where each $\alpha_i \in \mathbb{R}$.

Example 2 *In structured SVM, we have the loss function*

$$f_i(X_i^\top w) = \max_{y'} \left[\delta(y', y_i) - w^\top \psi(x_i, y_i) + w^\top \psi(x_i, y') \right] \ ,$$

where $\delta(y', y)$ is the penalty of predicting a label y' when the true label is y , and the regularizer is

$$g(w) = \frac{1}{2}\|w\|_2^2.$$

Example 3 *Consider regularized multi-class logistic regression, with training data $\{(x_i, y_i)\}$. The input features are $X_i = [\psi(x_i, j)]_{j=1,\ldots,k}$, where each $\psi(x, y) \in \mathbb{R}^d$ corresponds to the feature vector of data x for class y. The label $y_i \in \{1, \ldots, k\}$ is the class label. The loss functions are*

$$f_i(u) = -u_{y_i} + \ln \sum_{y'=1}^{k} \exp(u_{y'}),$$

and the regularizer is

$$g(w) = \frac{1}{2}\|w\|_2^2.$$

The primal problem is

$$\frac{1}{n}\sum_{i=1}^{n} \left[-\psi(x_i, y_i)^\top w + \ln \sum_{y'=1}^{k} \exp(\psi(x_i, y')^\top w) \right] + \frac{\lambda}{2}\|w\|_2^2.$$

3 Proximal Stochastic Dual Coordinate Ascent

The dual coordinate ascent (DCA) method maximizes the dual problem (2) by optimizing one α_i at a time for a chosen i, while keeping α_j with $j \neq i$ fixed. This can be considered as an instance of alternating direction methods which we have encountered in ADMM.

We focus on a *stochastic* version of DCA, called SDCA, in which at each round we choose which dual variable α_i to optimize uniformly at random. We analyze SDCA for the case $g(\cdot)$ is 1-strongly convex. This implies that $g^*(\cdot)$ is 1-smooth.

The generic Prox-SDCA algorithm is presented in Algorithm 1. The ideas are described as follows. Consider the maximal increase of the dual objective, where we only allow to change the i'th component of α. At step t, let

$$v^{(t-1)} = (\lambda n)^{-1} \sum_i X_i \alpha_i^{(t-1)}$$

and let

$$w^{(t-1)} = \nabla g^*(v^{(t-1)}).$$

We will update the i-th dual variable $\alpha_i^{(t)} = \alpha_i^{(t-1)} + \Delta\alpha_i$, in a way that will lead to a sufficient increase of the dual objective. For primal variable, this would lead to the update $v^{(t)} = v^{(t-1)} + (\lambda n)^{-1} X_i \Delta\alpha_i$, and therefore $w^{(t)} = \nabla g^*(v^{(t)})$ can also be written as

$$w^{(t)} = \arg\max_w \left[w^\top v^{(t)} - g(w) \right] \; = \; \arg\min_w \left[-w^\top \left(n^{-1} \sum_{i=1}^n X_i \alpha_i^{(t)} \right) + \lambda g(w) \right] .$$

Note that this particular update is rather similar to the update step of proximal-gradient dual-averaging method in the SGD domain [3]. The difference is on how $\alpha^{(t)}$ is updated. Stronger results can be proved for the Prox-SDCA method when we run SDCA for $t > n$ iterations with smooth loss functions.

In order to motivate the proximal SDCA algorithm, we note that the goal of SDCA is to increase the dual objective as much as possible, and thus the optimal way to choose $\Delta\alpha_i$ would be to maximize the dual objective, namely, we shall let

$$\Delta\alpha_i = \arg \max_{\Delta\alpha_i \in \mathbb{R}^k} \left[-\frac{1}{n} f_i^*(-(\alpha_i + \Delta\alpha_i)) - \lambda g^*(v^{(t-1)} + (\lambda n)^{-1} X_i \Delta\alpha_i) \right] .$$

However, for complex $g^*(\cdot)$, this optimization problem may not be easy to solve. We will simplify this optimization problem. Instead of directly maximizing the dual objective function, we try to maximize the following proximal objective which is a lower bound of the dual objective:

$$\begin{aligned} & \arg \max_{\Delta\alpha_i \in \mathbb{R}^k} \left[-\frac{1}{n} f_i^*(-(\alpha_i + \Delta\alpha_i)) - \lambda \left(\nabla g^*(v^{(t-1)})^\top (\lambda n)^{-1} X_i \Delta\alpha_i + \frac{1}{2} \|(\lambda n)^{-1} X_i \Delta\alpha_i\|_2^2 \right) \right] \\ = \; & \arg \max_{\Delta\alpha_i \in \mathbb{R}^k} \left[-f_i^*(-(\alpha_i + \Delta\alpha_i)) - w^{(t-1)\,\top} X_i \Delta\alpha_i - \frac{1}{2\lambda n} \|X_i \Delta\alpha_i\|_2^2 \right] . \end{aligned}$$

Algorithm 1: Proximal Stochastic Dual Coordinate Ascent

```
Input: $\phi(\cdot)$, $L$, $\lambda$, $\alpha^{(0)}$, and $R$ such that $\|X_i\|_2 \leq R$
Output: $\alpha^{(T)}$, $w^{(T)}$
Let $w^{(0)} = \nabla g^*(\alpha^{(0)})$
for $t = 1, 2, \ldots, T$ do
    Randomly pick $i$
    Find $\Delta\alpha_i$ such as the dual objective is no smaller than one of the following options
    Option I:
        $\Delta\alpha_i \in \arg\max_{\Delta\alpha_i} \left[-f_i^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - w^{(t-1)^\top} X_i \Delta\alpha_i - \frac{1}{2\lambda n} \|X_i \Delta\alpha_i\|_2^2\right]$
    Option II:
        Let $u$ be s.t. $-u \in \partial f_i(X_i^\top w^{(t-1)})$
        Let $z = u - \alpha_i^{(t-1)}$
        Let $s = \arg\max_{s \in [0,1]} \left[-f_i^*(-(\alpha_i^{(t-1)} + sz)) - s\, w^{(t-1)^\top} X_i z - \frac{s^2}{2\lambda n} \|X_i z\|_2^2\right]$
        Set $\Delta\alpha_i = sz$
    Option III:
        Same as Option II but replace the definition of $s$ as follows:
        Choose $s \geq \lambda n/(\lambda n + LR^2)$ so that
            $s \leq \min\left(1, \frac{f_i(X_i^\top w^{(t-1)}) + f_i^*(-\alpha_i^{(t-1)}) + w^{(t-1)^\top} X_i \alpha_i^{(t-1)} + \frac{1}{2L}\|z\|_2^2}{\|z\|_2^2 (L^{-1} + \|X_i\|_2^2/(\lambda n))}\right)$
    Let $\alpha_i^{(t)} \leftarrow \alpha_i^{(t-1)} + \Delta\alpha_i$ and $\alpha_j^{(t)} = \alpha_j^{(t-1)}$ when $j \neq i$
    Let $v^{(t)} \leftarrow v^{(t-1)} + (\lambda n)^{-1} X_i \Delta\alpha_i$
    Let $w^{(t)} \leftarrow \nabla g^*(v^{(t)})$
Return: $\alpha^{(T)}$, $w^{(T)}$
```

For example, for ridge regression, we can take option I: and solve

$$\max_{\Delta\alpha_i} \left[\Delta\alpha_i y_i - \frac{1}{2}(\alpha_i^{(t-1)} + \Delta\alpha_i)^2 - w^{(t-1)\top} x_i \Delta\alpha_i - \frac{1}{2\lambda n} \|x_i \Delta\alpha_i\|_2^2\right],$$

which leads to

$$\Delta\alpha_i = \frac{\lambda n}{\lambda n + \|x_i\|_2^2} [y_i - w^{(t-1)\top} x_i - \alpha_i^{(t-1)}].$$

We have the following convergence result for Prox-SDCA.

Theorem 1 *In Algorithm 1, assume that for all i, f_i is L-smooth. To obtain an expected duality gap of $\mathbf{E}[\phi(w^{(T)}) - \phi_D(\alpha^{(T)})] \leq \epsilon_P$, it suffices to have a total number of iterations of*

$$T \geq \left(n + \frac{R^2 L}{\lambda}\right) \log((n + \frac{R^2 L}{\lambda}) \cdot \frac{\phi(w^{(0)}) - \phi_D(\alpha^{(0)})}{\epsilon_P}).$$

The above theorem shows that in order to achieve an accuracy of ϵ, the number of data Prox-SDCA needs to process is

$$T = O\left((n + \kappa) \log \frac{1}{\epsilon}\right).$$

For Proximal Gradient Descent, since each gradient needs to process n data. Therefore the total number of data processed is

$$T = O\left((n \cdot \kappa) \log \frac{1}{\epsilon}\right).$$

This means that the linear convergence result in the above theorem is superior to those of the batch algorithms such as proximal gradient descent, or accelerated proximal gradient descent. Traditional algorithms can only achieve relatively fast convergence in constant number of iterations when the condition number $L/\lambda = O(1)$. In comparison, SDCA allows relatively fast convergence even when the condition number $L/\lambda = O(n)$, which can be a significant improvement for real applications, especially when n is large.

4 Empirical Studies

We study the smoothed hinge loss function $\phi_\gamma(z)$ with $\gamma = 1$, and solves the following $L_1 - L_2$ regularization problem:

$$\min_w \left[\underbrace{\frac{1}{n}\sum_{i=1}^n \phi_\gamma(w^\top x_i y_i)}_{f(w)} + \underbrace{\frac{\lambda}{2}\|w\|_2^2 + \mu\|w\|_1}_{g(w)}\right].$$

We compare SDCA to proximal gradient and accelerated proximal gradient. The results show that SDCA is superior when n is large, and especially when λn is at least $O(1)$ order. It is not as competitive as traditional algorithms when λn is much smaller than 1. This is consistent with the theory.

Moreover, it is worth noting that randomization is important in SDCA. Both random selection or using random permutation of the data work. However, the cyclic version (which only randomly permute the data at the beginning) performs rather poorly. Note that SDCA tries to maximize the dual objective function from one epoch to another epoch. This may not always decrease the primal objective function from one epoch to the next epoch, although in the limit one reduces the primal objective function to the optimal value when the dual is optimized well. This leads to some oscillation in the convergence plots.

5 Proof of Theorem 1

For convenience, we list the following simple facts about primal and dual formulations, which will be used in the proofs. For each i, we have

$$-\alpha_i^* \in \partial\phi_i(X_i^\top w^*), \quad X_i^\top w^* \in \partial\phi_i^*(-\alpha_i^*),$$

and

$$w^* = \nabla g^*(v^*), \quad v^* = \frac{1}{\lambda n}\sum_{i=1}^n X_i \alpha_i^*.$$

The key lemma is the following:

Lemma 1 *Assume that ϕ_i^* is γ-strongly-convex (where γ can be zero). Then, for any iteration t and any $s \in [0,1]$ we have*

$$\mathbf{E}[\phi_D(\alpha^{(t)}) - \phi_D(\alpha^{(t-1)})] \geq \frac{s}{n}\,\mathbf{E}\left[\phi(w^{(t-1)}) - \phi_D(\alpha^{(t-1)})\right] - \left(\frac{s}{n}\right)^2 \frac{G^{(t)}}{2\lambda}\ ,$$

where

$$G^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\left(\|X_i\|_2^2 - \frac{\gamma(1-s)\lambda n}{s}\right)\,\mathbf{E}\left[\|u_i^{(t-1)} - \alpha_i^{(t-1)}\|_2^2\right],$$

where $\|X_i\|_2$ denotes the spectral norm of X_i, and $-u_i^{(t-1)} \in \partial f_i(X_i^\top w^{(t-1)})$.

Proof Since only the i'th element of α is updated, the improvement in the dual objective can be written as

$$\begin{aligned}
&n[\phi_D(\alpha^{(t)}) - \phi_D(\alpha^{(t-1)})]\\
&= \left(-f_i^*(-\alpha_i^{(t)}) - \lambda n g^*\left(v^{(t-1)} + (\lambda n)^{-1}X_i\Delta\alpha_i\right)\right) - \left(-f_i^*(-\alpha_i^{(t-1)}) - \lambda n g^*\left(v^{(t-1)}\right)\right)\\
&\geq \underbrace{\left(-f_i^*(-\alpha_i^{(t)}) - \lambda n h\left(v^{(t-1)};(\lambda n)^{-1}X_i\Delta\alpha_i\right)\right)}_{A} - \underbrace{\left(-f_i^*(-\alpha_i^{(t-1)}) - \lambda n g^*\left(v^{(t-1)}\right)\right)}_{B},
\end{aligned}$$

where $h(v,\Delta v) = g^*(v) + \nabla g^*(v)^\top \Delta v + 0.5\|\Delta v\|_2^2$.

By the definition of the update we have for all $s \in [0,1]$ that

$$\begin{aligned}
A &= \max_{\Delta\alpha_i} -f_i^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - \lambda n h\left(v^{(t-1)};(\lambda n)^{-1}X_i\Delta\alpha_i\right)\\
&\geq -f_i^*(-(\alpha_i^{(t-1)} + s(u_i^{(t-1)} - \alpha_i^{(t-1)}))) - \lambda n h(v^{(t-1)};(\lambda n)^{-1}sX_i(u_i^{(t-1)} - \alpha_i^{(t-1)})). \qquad (4)
\end{aligned}$$

In the following, we drop the superscript $(t-1)$. Since f_i^* is γ-strongly convex, we have that

$$\begin{aligned}
f_i^*(-(\alpha_i + s(u_i - \alpha_i))) =& f_i^*(s(-u_i) + (1-s)(-\alpha_i)) \qquad (5)\\
\leq& s f_i^*(-u_i) + (1-s)f_i^*(-\alpha_i) - \frac{\gamma}{2}s(1-s)\|u_i - \alpha_i\|_2^2.
\end{aligned}$$

Combining this with (4) and rearranging terms we obtain that

$$\begin{aligned}
A &\geq -sf_i^*(-u_i) - (1-s)f_i^*(-\alpha_i) + \frac{\gamma}{2}s(1-s)\|u_i - \alpha_i\|_2^2 - \lambda n h(v;(\lambda n)^{-1}sX_i(u_i - \alpha_i))\\
&= -sf_i^*(-u_i) - (1-s)f_i^*(-\alpha_i) + \frac{\gamma}{2}s(1-s)\|u_i - \alpha_i\|_2^2 - \lambda n g^*(v) - sw^\top X_i(u_i - \alpha_i)\\
&\quad - \frac{s^2}{2\lambda n}\|X_i(u_i - \alpha_i)\|_2^2\\
&\geq -sf_i^*(-u_i) - (1-s)f_i^*(-\alpha_i) + \frac{\gamma}{2}s(1-s)\|u_i - \alpha_i\|_2^2 - \lambda n g^*(v) - sw^\top X_i(u_i - \alpha_i)\\
&\quad - \frac{s^2}{2\lambda n}\|X_i\|_2^2\|u_i - \alpha_i\|_2^2\\
&= \underbrace{-s(f_i^*(-u_i) + w^\top X_i u_i)}_{s\,f_i(X^\top w)} + \underbrace{(-f_i^*(-\alpha_i) - \lambda n g^*(v))}_{B} + \frac{s}{2}\left(\gamma(1-s) - \frac{s\|X_i\|_2^2}{\lambda n}\right)\|u_i - \alpha_i\|_2^2\\
&\quad + s(f_i^*(-\alpha_i) + w^\top X_i\alpha_i),
\end{aligned}$$

where we used $-u_i \in \partial f_i(X_i^\top w)$ which yields $f_i^*(-u_i) = -w^\top X_i u_i - f_i(X_i^\top w)$. Therefore

$$A - B \geq s \left[f_i(X_i^\top w) + f_i^*(-\alpha_i) + w^\top X_i \alpha_i + \left(\frac{\gamma(1-s)}{2} - \frac{s\|X_i\|_2^2}{2\lambda n} \right) \|u_i - \alpha_i\|_2^2 \right] . \tag{6}$$

Next note that with $w = \nabla g^*(v)$, we have $g(w) + g^*(v) = w^\top v$. Therefore:

$$\begin{aligned} \phi(w) - \phi_D(\alpha) &= \frac{1}{n}\sum_{i=1}^n f_i(X_i^\top w) + \lambda g(w) - \left(-\frac{1}{n}\sum_{i=1}^n f_i^*(-\alpha_i) - \lambda g^*(v) \right) \\ &= \frac{1}{n}\sum_{i=1}^n f_i(X_i^\top w) + \frac{1}{n}\sum_{i=1}^n f_i^*(-\alpha_i) + \lambda w^\top v \\ &= \frac{1}{n}\sum_{i=1}^n \left(f_i(X_i^\top w) + f_i^*(-\alpha_i) + w^\top X_i \alpha_i \right) . \end{aligned}$$

Therefore, if we take expectation of (6) w.r.t. the choice of i we obtain that

$$\frac{1}{s}\,\mathbf{E}[A-B] \geq \mathbf{E}[\phi(w) - \phi_D(\alpha)] - \frac{s}{2\lambda n} \cdot \underbrace{\frac{1}{n}\sum_{i=1}^n \left(\|X_i\|_2^2 - \frac{\gamma(1-s)\lambda n}{s} \right) \|u_i - \alpha_i\|_2^2}_{=G^{(t)}} .$$

We have obtained that

$$\frac{n}{s}\,\mathbf{E}[\phi_D(\alpha^{(t)}) - \phi_D(\alpha^{(t-1)})] \geq \mathbf{E}[\phi(w^{(t-1)}) - \phi_D(\alpha^{(t-1)})] - \frac{s\,G^{(t)}}{2\lambda n} . \tag{7}$$

Multiplying both sides by s/n concludes the proof of the lemma. ■

The assumption that f_i is L-smooth implies that f_i^* is γ-strongly-convex with $\gamma = 1/L$. We will apply Lemma 1 with $s = \frac{\lambda n \gamma}{R^2 + \lambda n \gamma} \in [0,1]$. Recall that $\|X_i\|_2 \leq R$. Therefore, the choice of s implies that

$$\|X_i\|_2^2 - \frac{\gamma(1-s)\lambda n}{s} \leq R^2 - \frac{1-s}{s/(\lambda n \gamma)} = R^2 - R^2 = 0 ,$$

and hence $G^{(t)} \leq 0$ for all t. This yields,

$$\mathbf{E}[\phi_D(\alpha^{(t)}) - \phi_D(\alpha^{(t-1)})] \geq \frac{s}{n}\,\mathbf{E}[\phi(w^{(t-1)}) - \phi_D(\alpha^{(t-1)})] .$$

But since $\epsilon_D^{(t-1)} := \phi_D(\alpha^*) - \phi_D(\alpha^{(t-1)}) \leq \phi(w^{(t-1)}) - \phi_D(\alpha^{(t-1)})$ and $\phi_D(\alpha^{(t)}) - \phi_D(\alpha^{(t-1)}) = \epsilon_D^{(t-1)} - \epsilon_D^{(t)}$, we obtain that

$$\begin{aligned} \mathbf{E}[\epsilon_D^{(t)}] &\leq \left(1 - \tfrac{s}{n}\right) \mathbf{E}[\epsilon_D^{(t-1)}] \leq \left(1 - \tfrac{s}{n}\right)^t \mathbf{E}[\epsilon_D^{(0)}] \\ &\leq \left(1 - \tfrac{s}{n}\right)^t \epsilon_D^{(0)} \leq \exp(-st/n)\epsilon_D^{(0)} = \exp\left(-\frac{\lambda\gamma t}{R^2 + \lambda\gamma n} \right) \epsilon_D^{(0)} . \end{aligned}$$

This would be smaller than ϵ_D if

$$t \geq \left(n + \tfrac{R^2}{\lambda\gamma} \right) \log(\epsilon_D^{(0)}/\epsilon_D) .$$

It implies that

$$\mathbf{E}[\phi(w^{(t)}) - \phi_D(\alpha^{(t)})] \le \frac{n}{s}\mathbf{E}[\epsilon_D^{(t)} - \epsilon_D^{(t+1)}] \le \frac{n}{s}\mathbf{E}[\epsilon_D^{(t)}]. \tag{8}$$

So, requiring $\epsilon_D^{(t)} \le \frac{s}{n}\epsilon_P$ we obtain a duality gap of at most ϵ_P. This means that we should require

$$t \ge \left(n + \tfrac{R^2}{\lambda\gamma}\right) \log((n + \tfrac{R^2}{\lambda\gamma}) \cdot \tfrac{\epsilon_D^{(0)}}{\epsilon_P}) \ ,$$

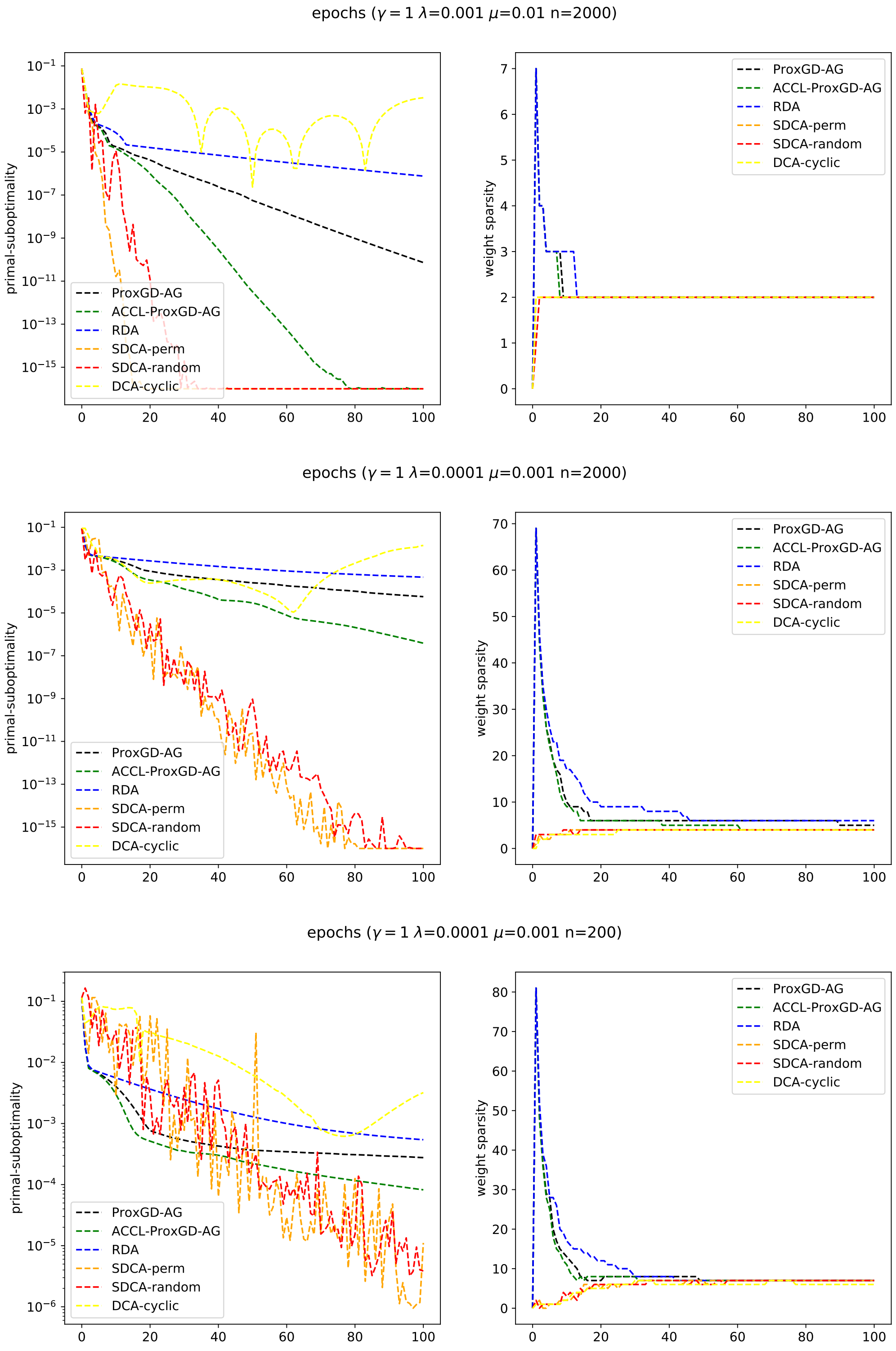
which proves Theorem 1.

References

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

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

[3] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:25432596, December 2010.



epochs ($\gamma = 1$ λ=0.001 μ=0.01 n=2000)

epochs ($\gamma = 1$ λ=0.0001 μ=0.001 n=2000)

epochs ($\gamma = 1$ λ=0.0001 μ=0.001 n=200)

Figure 1: Comparisons of Proximal Gradient, RDA, DCA and SDCA