Stochastic Gradient Descent

1 Introduction

In machine learning, we observe training data (x_i, y_i) for $i = 1, \ldots, n$, and would like to learn a model parameter w of the form

$$\min_{w \in C} \left[\frac{1}{n} \sum_{i=1}^{n} f_i(w) + g(w) \right].$$

More generally, we can write this optimization problem as:

$$\min_{w \in C} \phi(w), \quad \phi(w) = f(w) + g(w), \qquad f(w) = \mathbf{E}_{\xi \sim D} f(\xi, w), \tag{1}$$

where ξ is a random variable, drawn from a distribution D.

In the finite sample case, one may consider ξ as i, and the distribution D is to randomly choosing $\xi = i$ from $1, \ldots, n$.

Example 1 *Consider regression problem with possibly infinity training data $(x, y) \sim D$, where D is a distribution of the training data. given training point x, the prediction function is*

$$\nu(\xi, w),$$

where w is the model parameter. Let $\xi = (x, y) \sim D$, then the expected loss is

$$\mathbf{E}_{\xi \sim D}\ f(\xi, w), \qquad f(\xi, w) = (\nu(w, x) - y)^2.$$

2 Stochastic Gradient Descent

A popular method in machine learning for solving (1) is stochastic gradient descent, which picks $\xi = (x, y)$ at a time, and works with this data point. One may generalize the proximal gradient method to this situation. The algorithm is presented in Algorithm 1, where

$$\mathrm{prox}_{\eta g}(w) = \arg\min_{z} \left[\frac{1}{2\eta} \|z - w\|_2^2 + g(z) \right].$$

In practical implementations, if the number of training data is finite, one may either draw $\xi = i$ completely randomly, or use random permutation of the training data.

We note that this algorithm is a stochastic version of the proximal gradient descent. In proximal gradient descent, we use $w^{(t)} = \mathrm{prox}_{\eta_t g}(w^{(t-1)} - \eta_t \nabla_w f(w))$, and in proximal SGD, we replace $\nabla_w f(w)$ by $\nabla_w f(\xi, w^{(t-1)}))$, which is an unbiased estimate of the gradient:

$$\mathbf{E}_{\xi} \nabla_w f(\xi, w) = \nabla_w f(w).$$

In the finite sample case, each full gradient computation is

$$\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(w),$$

which requires n gradient evaluations per iteration, while SGD requires to compute

$$\nabla f_i(w)$$

per iteration, which is 1 gradient evaluation per iteration.

Algorithm 1: Proximal Stochastic Gradient Descent (Proximal SGD)

Input: $\phi(\cdot)$, learning rates $\{\eta_t\}$, $w^{(0)}$
Output: $w^{(T)}$
for $t = 1, 2, \ldots, T$ **do**
  Randomly pick $\xi \sim D$
  Let $w^{(t)} = \mathrm{prox}_{\eta_t g}(w^{(t-1)} - \eta_t \nabla_w f(\xi, w^{(t-1)}))$
Return: $w^{(T)}$

Theorem 1 *Consider proximal SGD. If $f(w)$ is convex, and for all ξ and $w \in C$:*

$$\|\nabla_w f(\xi, w)\|_2 \leq G,$$

and $g(w)$ is convex. We have for all $w \in C$:

$$\sum_{t=1}^{T}\eta_t \mathbf{E}\left[\phi(w^{(t)}) - \phi(w)\right] \leq 2G^2\sum_{t=1}^{T}\eta_t^2 + \frac{1}{2}\|w - w^{(0)}\|_2^2.$$

This bound implies that for general convex (possibly nonsmooth) problem, the number of iterations needed to achieve ϵ accuracy is

$$O(1/\epsilon^2)$$

with $\eta_t = O(1/\sqrt{T})$. This matches the complexity of proximal gradient descent for nonsmooth problems. However, proximal gradient descent requires n gradient evaluations per iteration, while SGD requires 1 gradient evaluation per iteration.

If the problem is strongly convex, then we have the following result, with $\eta_t = O(1/t)$, and the number of iterations needed to achieve ϵ is

$$O(1/\epsilon),$$

which matches matches the complexity of proximal gradient descent method in the strongly convex case, except that proximal gradient descent requires n gradient evaluations per iteration.

Theorem 2 *Consider SGD. If $f(w)$ is λ strongly convex, and $g(w)$ is λ' strongly convex. Let $\eta_t = t^{-1}/(\lambda + \lambda')$. We have for all $w \in C$:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}\,\phi(w^{(t)}) \leq \phi(w) + 2G^2\frac{\ln(T+1)}{(\lambda + \lambda')T} + \frac{\lambda'}{2T}\|w - w^{(0)}\|_2^2.$$

3 Minibatch SGD

In practice, it is often inefficient to work with one data point at a time. Therefore to improve efficiency, we need to work with a minibatch B of m training samples per iteration. Here we either randomly select m training data from D to form a minibatch B, or use a random permutation, and select $m = |B|$ training data. If we let

$$f_B(w) = \frac{1}{|B|}\sum_{\xi\in B} f(\xi, w),$$

then the minibatch SGD algorithm is presented in Algorithm 2. In this case, the minibatch gradient is

$$\nabla f_B(w) = \frac{1}{|B|}\sum_{\xi\in B} \nabla_w f(\xi, w),$$

which is unbiased:

$$\mathbf{E}_B \nabla f_B(w) = \nabla f(w).$$

Algorithm 2: Proximal Minibatch Stochastic Gradient Descent (Proximal Minibatch SGD)

Input: $\phi(\cdot)$, learning rates $\{\eta_t\}$, $w^{(0)}$
Output: $w^{(T)}$
for $t = 1, 2, \ldots, T$ **do**
  Randomly pick a minibatch $B \sim D$ of size $|B| = m$
  Let $w^{(t)} = \mathrm{prox}_{\eta_t g}(w^{(t-1)} - \eta_t \nabla f_B(, w^{(t-1)}))$
Return: $w^{(T)}$

Example 2 *Consider the regression problem:*

$$\mathbf{E}_\xi f(\xi, w), \qquad f(\xi, w) = \frac{1}{2}(\nu(w, x) - y)^2 + \frac{\lambda}{2}\|w\|_2^2.$$

In this example, $g(w) = 0$ *and* $\mathrm{prox}_{\eta g}(w) = w$. *Therefore we have*

$$\begin{aligned} w^{(t)} =& w^{(t-1)} - \eta_t \left[\frac{1}{m}\sum_{\xi\in B}(\nu(w^{(t-1)}, x) - y)\nabla_w \nu(w^{(t-1)}, x) + \lambda w^{(t-1)}\right] \\ =& (1 - \eta_t\lambda)w^{(t-1)} - \frac{\eta_t}{m}\sum_{\xi\in B}(\nu(w^{(t-1)}, x) - y)\nabla_w \nu(w^{(t-1)}, x). \end{aligned}$$

Example 3 *Consider the regression problem:*

$$\mathbf{E}_\xi f(\xi, w) + g(w), \qquad g(w) = \frac{\lambda}{2}\|w\|_2^2 \qquad \textit{subject to} \;\; w \in C.$$

We assume that $g(w)$ *and* C *are convex. In this case, we have*

$$\mathrm{prox}_{\eta g}(w) = \arg\min_{z\in C}\left[\frac{1}{2\eta}\|z - w\|_2^2 + \frac{\lambda}{2}\|z\|_2^2\right] = \mathrm{proj}_C\left(\frac{1}{1+\eta\lambda}w\right).$$

The proximal gradient becomes

$$\begin{aligned} w^{(t)} =& \mathrm{prox}_{\eta_t g}\left(w^{(t-1)} - \eta_t \nabla_w f_B(w^{(t-1)})\right) \\ =& \mathrm{proj}_C\left((1-\tilde{\eta}_t\lambda)w^{(t-1)} - \tilde{\eta}_t \frac{1}{m}\sum_{\xi\in B}\nabla_w f(\xi, w^{(t-1)})\right), \end{aligned}$$

where $\tilde{\eta}_t = \eta_t/(1+\eta_t\lambda)$.

We have the following convergence result for smooth and convex problems.

Theorem 3 *Consider minibatch SGD. If* $f(w)$ *is convex and* L *smooth,* $g(w)$ *is convex. Let*

$$V = \sup_{w\in C} \mathbf{E}_{\xi\sim D}\|\nabla f(\xi, w) - \nabla f(w)\|_2^2.$$

If we choose $\eta_t < 1/L$ *for all* t*, then for all* $w \in C$*:*

$$\sum_{t=1}^{T} \eta_t \mathbf{E}\left[\phi(w^{(t)}) - \phi(w)\right] \le \sum_{t=1}^{T} \frac{\eta_t^2 V}{2(1-\eta_t L)m} + \frac{1}{2}\|w - w^{(0)}\|_2^2.$$

If we take $\eta_t = \eta\sqrt{m/T}$, then

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\left[\phi(w^{(t)}) - \phi(w)\right] \le \frac{\eta\, V}{2(\sqrt{mT} - \eta m L)} + \frac{1}{2\eta\sqrt{mT}}\|w - w^{(0)}\|_2^2.$$

This means that for smooth functions, when we increase the minibatch size m, we still obtain the same convergence rate per sample, which is also the same as that of Theorem 1. However, the learning rate needs to be increased by a factor of $\sqrt{m}$. When $V \neq 0$, the number of samples needed to achieve accuracy ϵ is:

$$mT = O(V^2/\epsilon^2).$$

This can be compared to proximal gradient, which has the convergence rate of $O(1/\epsilon)$ per iteration, corresponding to the number of samples of

$$O(n/\epsilon)$$

for training data size of n.

Theorem 4 *Consider minibatch SGD. If* $f(w)$ *is* λ *strongly convex and* L *smooth,* $g(w)$ *is* λ' *strongly convex. Let* $\eta_t = 1/(2L + 0.5(t-1)(\lambda+\lambda'))$*, and*

$$V = \sup_{w\in C} V(w).$$

We have

$$\sum_{t=1}^{T}(2L - \lambda + 0.5(t-1)(\lambda+\lambda'))\mathbf{E}\left[\phi(w^{(t)}) - \phi(w)\right] \le \frac{2TV}{m} + L(2L+\lambda')\|w - w^{(0)}\|_2^2.$$

The result shows that when $V \neq 0$, in order to achieve ϵ accuracy, the number of samples needed to achieve accuracy ϵ is:

$$mT = O(V/(\lambda+\lambda')\epsilon).$$

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 proximal gradient, SDCA, to proximal SGD and proximal minibatch SGD, with various learning rate settings. The performance of SGD depends on different learning rate schedule of the form

$$\eta_t = \eta/(1 + a\sqrt{t} + bt).$$

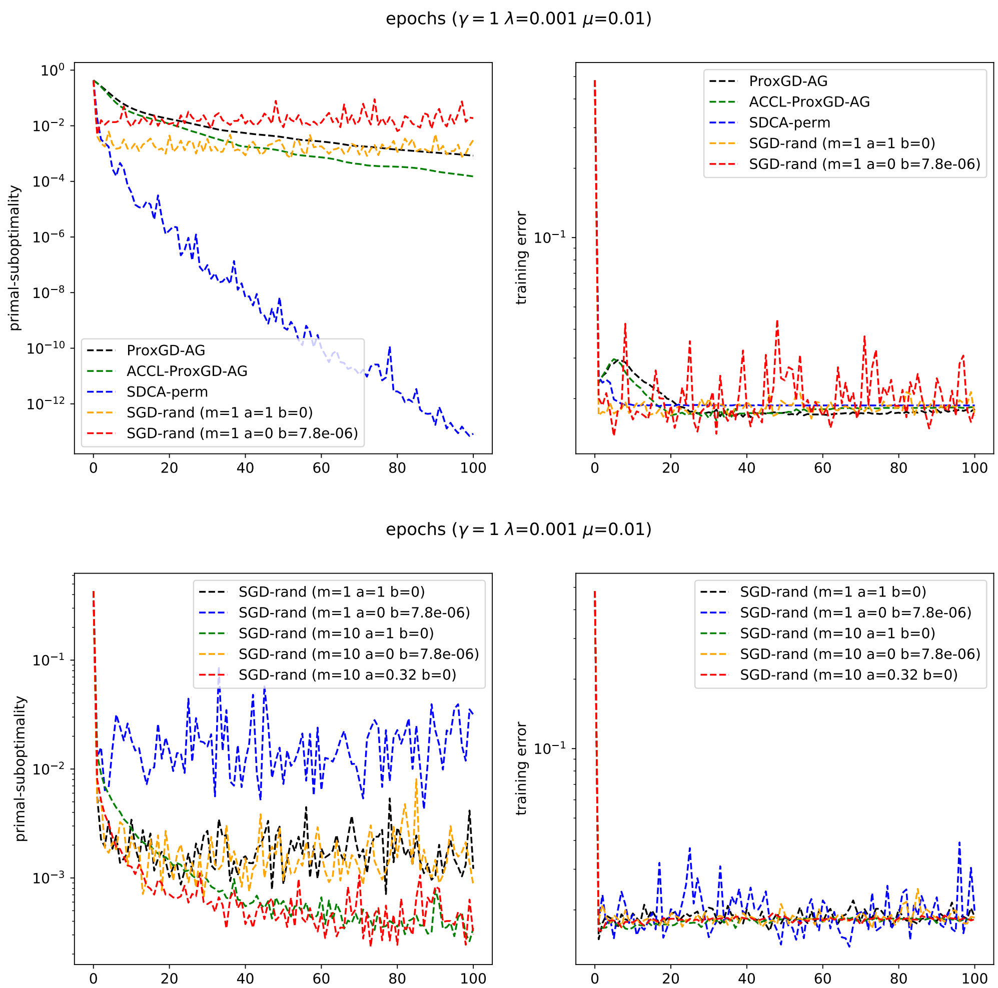

Figure 1: Comparisons of Proximal Gradient, SDCA and SGD (smooth and strongly convex)

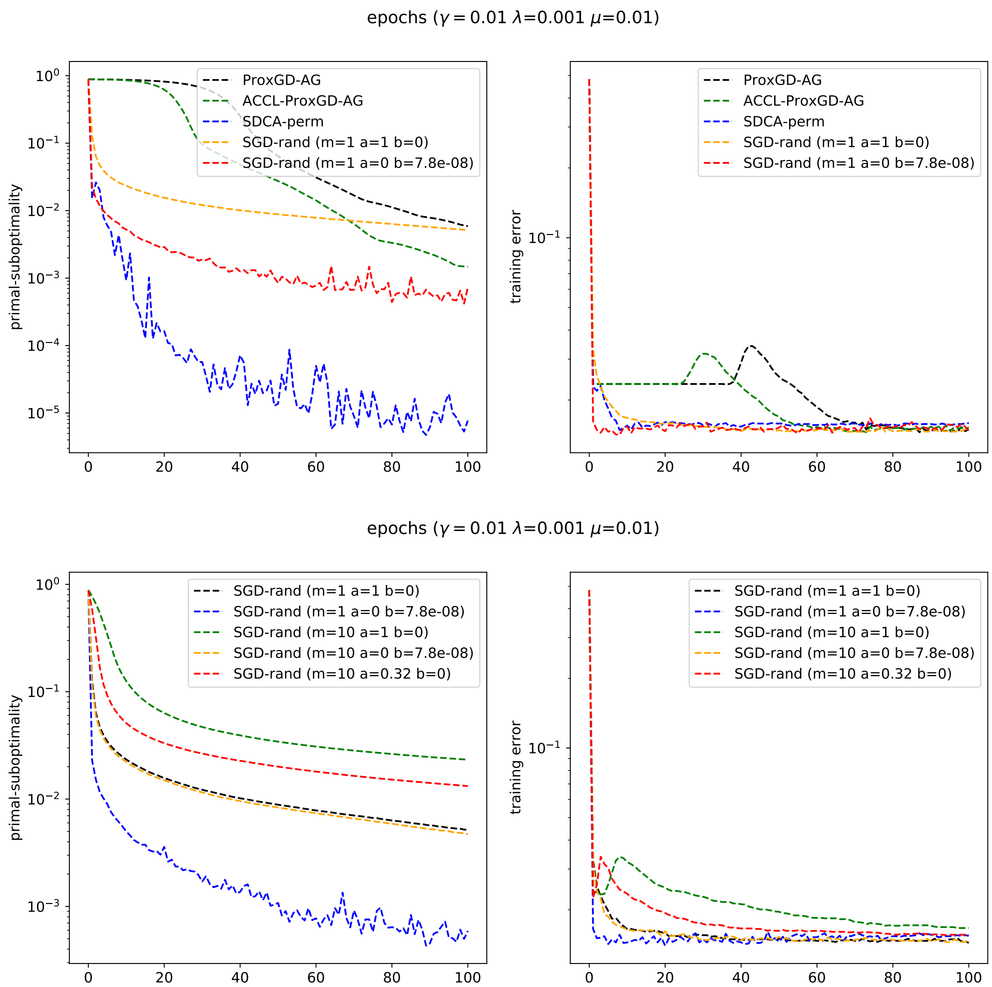

Figure 2: Comparisons of Proximal Gradient, SDCA and SGD (near non-smooth and strongly convex)

5 Convergence Analysis

Consider a minibatch B, and define for $\eta > 0$,

$$Q_{\eta,B}(w;w') = f(w') + \nabla f_B(w')^\top (w - w') + \frac{1}{2\eta}\|w - w'\|_2^2 + g(w).$$

We have the following propositions for the minibatch SGD.

Proposition 1 *Assume that $f(w)$ is L-smooth in C. If we pick $\eta < 1/L$, then given any w', we have*

$$\phi(w) \le Q_{\eta,B}(w;w') + \frac{\eta}{2(1-\eta L)}\|\nabla f_B(w') - \nabla f(w')\|_2^2.$$

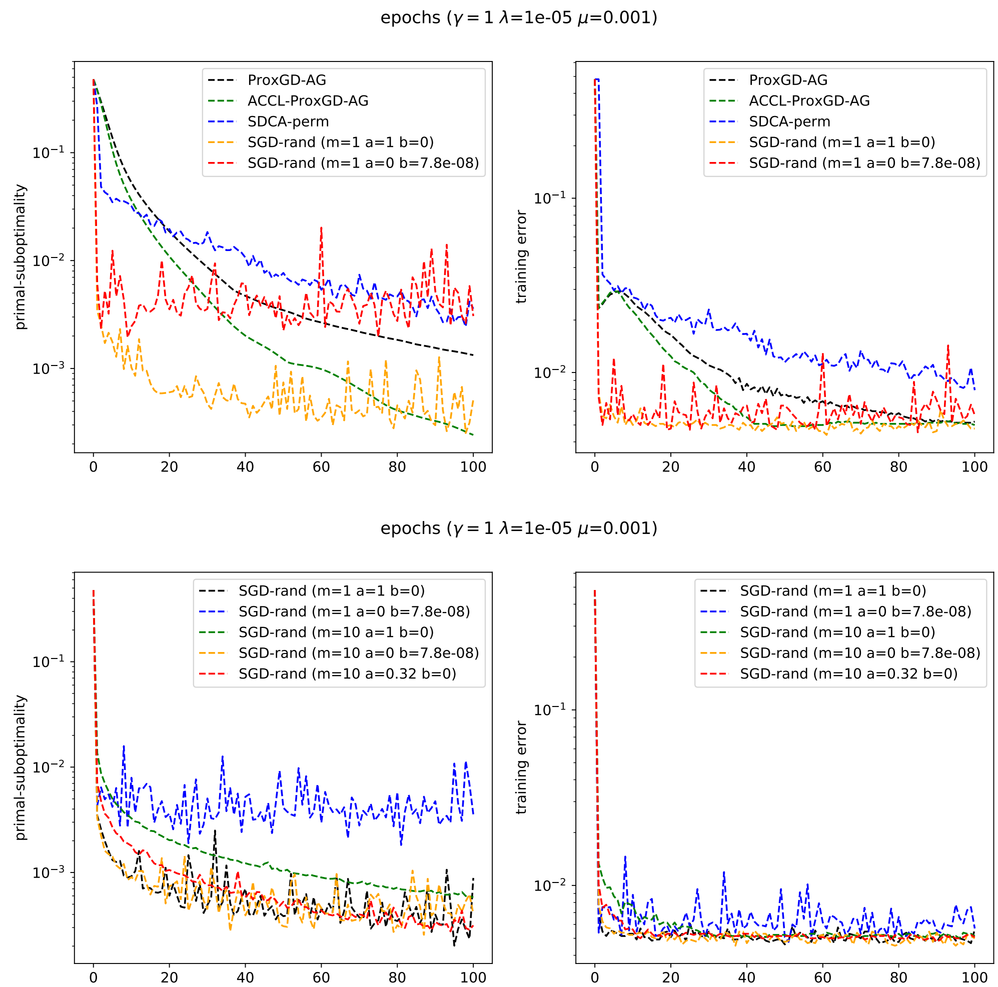

Figure 3: Comparisons of Proximal Gradient, SDCA and SGD (smooth and near non-strongly convex)

Proof From smoothness, we have

$$
\begin{aligned}
& f(w) + g(w) \\
\leq & f(w') + \nabla f(w')^\top (w - w') + \frac{L}{2}\|w - w'\|_2^2 + g(w) \\
\leq & f(w') + \nabla f_B(w')^\top (w - w') + \frac{L}{2}\|w - w'\|_2^2 + |(\nabla f_B(w') - \nabla f(w'))^\top (w - w')| + g(w) \\
= & Q_{\eta,B}(w; w') - \frac{\eta^{-1} - L}{2}\|w - w'\|_2^2 + |(\nabla f_B(w') - \nabla f(w'))^\top (w - w')| \\
\leq & Q_{\eta,B}(w; w') + \frac{1}{2(\eta^{-1} - L)}\|\nabla f_B(w') - \nabla f(w')\|_2^2.
\end{aligned}
$$

The first inequality uses the smoothness of $f(x)$. The second inequality is algebra, and the third

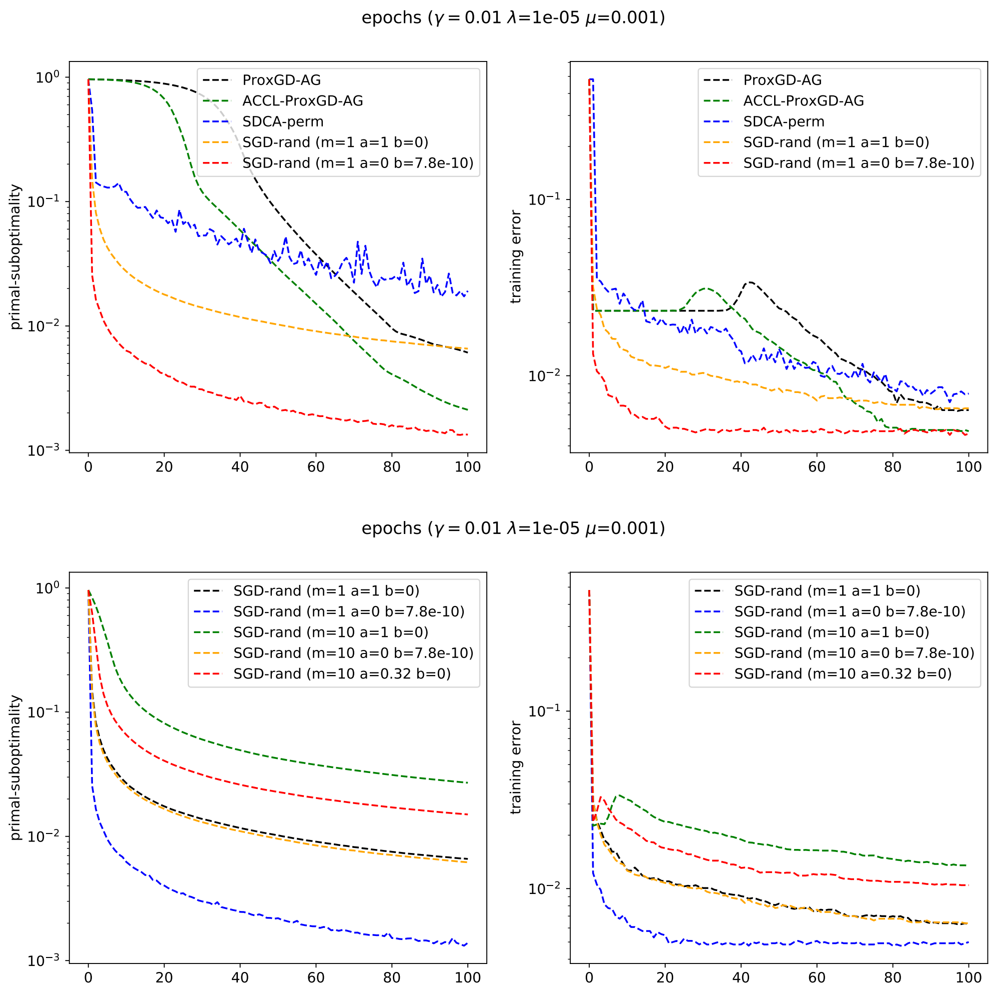

Figure 4: Comparisons of Proximal Gradient, SDCA and SGD (near non-smooth and near non-strongly convex)

inequality uses the definition of $Q_{\eta,B}(\cdot)$. The last inequality uses

$$-\frac{\eta^{-1}-L}{2}\|u\|_2^2 + |u^\top v| \leq \frac{1}{2(\eta^{-1}-L)}\|v\|_2^2.$$

This proves the desired result. ■

Proposition 2 *Assume that $f(w)$ is convex. Then given any w', we have*

$$\phi(w) \leq Q_{\eta,B}(w;w') + \frac{\eta}{2}\|\nabla f_B(w') - \nabla f(w)\|_2^2.$$

Proof We have

$$\begin{aligned}f(w)+g(w) \leq & f(w')+\nabla f(w)^\top(w-w')+g(w)\\ \leq & f(w')+\nabla f_B(w')^\top(w-w')+|(\nabla f_B(w')-\nabla f(w))^\top(w-w')|+g(w)\\ = & Q_{\eta,B}(w;w')-\frac{1}{2\eta}\|w-w'\|_2^2+|(\nabla f_B(w')-\nabla f(w))^\top(w-w')|\\ \leq & Q_{\eta,B}(w;w')+\frac{\eta}{2}\|\nabla f_B(w')-\nabla f(w)\|_2^2.\end{aligned}$$

This proves the result. ■

The following result is straight forward.

Proposition 3 *Given w, define*

$$V(w)=\mathbf{E}_{\xi\sim D}\|\nabla f(\xi,w)-\nabla f(w)\|_2^2.$$

If minibatch B has m independent samples from D, then

$$\mathbf{E}_{B\sim D}\ \|\nabla f_B(w')-\nabla f(w')\|_2^2 \leq \frac{1}{m}V(w').$$

Proposition 4 *IF $f(w)$ is λ and strongly convex, and $g(w)$ is λ' strongly convex. We have for all w:*

$$Q_B(w^{(t)};w^{(t-1)}) \leq \phi(w)+\frac{\eta_t^{-1}-\lambda}{2}\|w-w^{(t-1)}\|_2^2-\frac{\eta_t^{-1}+\lambda'}{2}\|w-w^{(t)}\|_2^2.$$

Proof We have

$$w^{(t)}=\arg\min_w Q_{\eta_t,B}(w;w^{(t-1)}).$$

Therefore using the strong convexity of Q, we have for all w:

$$\begin{aligned}Q_B(w^{(t)};w^{(t-1)}) \leq & Q_B(w;w^{(t-1)})-\left(\frac{1}{2\eta_t}+\frac{\lambda'}{2}\right)\|w-w^{(t)}\|_2^2\\ \leq & \phi(w)+\left(\frac{1}{2\eta_t}-\frac{\lambda}{2}\right)\|w-w^{(t-1)}\|_2^2-\left(\frac{1}{2\eta_t}+\frac{\lambda'}{2}\right)\|w-w^{(t)}\|_2^2.\end{aligned}$$

The second inequality uses the strong convexity of $f(\cdot)$. This proves the result. ■

5.1 Proof of Theorem 1

Using Proposition 2 (with $\lambda=\lambda'=0$) and Proposition 4, we obtain for minibatch $B_t=\{\xi\}$:

$$\phi(w^{(t)}) \leq \phi(w)+2\eta_t G^2+\frac{1}{2\eta_t}\|w-w^{(t-1)}\|_2^2-\frac{1}{2\eta_t}\|w-w^{(t)}\|_2^2.$$

Taking expectation, we have:

$$\eta_t\mathbf{E}\left[\phi(w^{(t)})-\phi(w)\right] \leq 2\eta_t^2G^2+\frac{1}{2}[\|w-w^{(t-1)}\|_2^2-\|w-w^{(t)}\|_2^2].$$

By summing over $t=1$ to T, we obtain the desired bound.

5.2 Proof of Theorem 2

Using Proposition 4 and Proposition 2, we obtain for minibatch $B_t = \{\xi\}$:

$$\phi(w^{(t)}) \leq \phi(w) + 2\eta_t G^2 + \frac{\eta_t^{-1} - \lambda}{2}\|w - w^{(t-1)}\|_2^2 - \frac{\eta_{t+1}^{-1} - \lambda}{2}\|w - w^{(t)}\|_2^2.$$

By summing over $t = 1$ to T, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\phi(w^{(t)}) \leq \phi(w) + 2G^2\frac{\ln(T+1)}{(\lambda+\lambda')T} + \frac{\lambda'}{2T}\|w - w^{(0)}\|_2^2.$$

5.3 Proof of Theorem 3

Using Proposition 2 (with $\lambda = \lambda' = 0$) and Proposition 1, we obtain for minibatch B_t:

$$\phi(w^{(t)}) \leq \phi(w) + \frac{\eta_t}{2(1-\eta_t L)}\|\nabla f_{B_t}(w^{(t-1)}) - \nabla f(w^{(t-1)})\|_2^2 + \frac{1}{2\eta_t}\|w - w^{(t-1)}\|_2^2 - \frac{1}{2\eta_t}\|w - w^{(t)}\|_2^2.$$

Taking expectation, we have:

$$\eta_t \mathbf{E}\,[\phi(w^{(t)}) - \phi(w)] \leq \frac{\eta_t^2 V}{2(1-\eta_t L)m} + \frac{1}{2}\mathbf{E}\,[\|w - w^{(t-1)}\|_2^2 - \|w - w^{(t)}\|_2^2].$$

By summing over $t = 1$ to T, we obtain the desired bound.

5.4 Proof of Theorem 4

Using Proposition 2 and Proposition 1, we obtain for minibatch B_t:

It implies that with minibatch $B_t \sim D$:

$$\mathbf{E}\,\phi(w^{(t)}) \leq \phi(w) + \frac{\eta_t}{m}V(w^{(t-1)}) + \frac{\eta_t^{-1} - \lambda}{2}\|w - w^{(t-1)}\|_2^2 - \frac{\eta_{t+2}^{-1} - \lambda}{2}\|w - w^{(t)}\|_2^2.$$

Multiply by $\eta_{t+1}^{-1} - \lambda$, and let $\rho_t = (\eta_t^{-1} - \lambda)(\eta_{t+1}^{-1} - \lambda)$, we have

$$(\eta_{t+1}^{-1} - \lambda)[\mathbf{E}\,\phi(w^{(t)}) - \phi(w)] \leq \frac{\eta_t(\eta_{t+1}^{-1} - \lambda)}{m}V(w^{(t-1)}) + \frac{\rho_t}{2}\|w - w^{(t-1)}\|_2^2 - \frac{\rho_{t+1}}{2}\|w - w^{(t)}\|_2^2.$$

By summing over $t = 1$ to $t = T$, we obtain

$$\begin{aligned}&\sum_{t=1}^{T}(2L - \lambda + 0.5(t-1)(\lambda+\lambda'))\mathbf{E}\,[\phi(w^{(t)}) - \phi(w)] \\ \leq &\sum_{t=1}^{T}\frac{\eta_t(\eta_{t+1}^{-1} - \lambda)V}{m} + \frac{(2L-\lambda)(2L+\lambda')}{2}\|w - w^{(0)}\|_2^2.\end{aligned}$$

We obtain the result by observation that $\eta_t(\eta_{t+1}^{-1} - \lambda) \leq 2$.