# Stochastic Gradient Dual Methods

## 1 Introduction

We consider the following stochastic optimization problem:

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

where $\xi$ is a random variable, drawn from a distribution $D$.

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

We consider dual methods, including stochastic RDA (regularized dual averaging) and stochastic ADMM (alternating direction methods of multipliers).

## 2 Stochastic Mirror Descent

Let

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

where $B \sim D$ contains $m$ independent samples from $D$.

We may extend the mirror descent method to stochastic algorithm.

**Algorithm 1:** Stochastic Mirror Descent

**Input**: $f(\cdot)$, $g(\cdot)$, $2_0$, $\{h_t(w)\}$, $\{\eta_t\}$
**Output**: $w^{(T)}$
**for** $t = 1, 2, \ldots, T$ **do**
  Randomly select a minibatch $B_t$ of $m$ independent samples from $D$
  Let $\tilde{\alpha}^{(t)} = \nabla h_t(w^{(t-1)}) - \eta_t \nabla f_{B_t}(w^{(t-1)})$
  Let $w^{(t)} = \arg\min_{w \in C}[-w^\top \tilde{\alpha}^{(t)} + h_t(w) + g(w)]$
**Return**: $w^{(T)}$

Note that if $C = \mathbb{R}^d$, $g(\cdot) = \lambda h(\cdot)$, and $h_t(\cdot) = h(\cdot)$, we have

$$(1 + \lambda)\nabla h(w^{(t)}) = \nabla h(w^{(t-1)}) - \eta_t \nabla f_B(w^{(t-1)}).$$

**Example 1** *In model combination, we want to find* $w \in C = \{w : \sum_{j=1}^d w_j = 1, \forall j \ w_j \geq 0\}$ *such that* $\sum_j w_j m_j(x)$ *fits a loss function of the form* $\mathbf{E}_\xi \ f(\xi, w)$ *with* $\xi = (x, y)$. *We assume that* $\|\nabla_w f(\xi, w)\|_\infty \leq G$. *Let* $h(w) = \sum_j w_j \log(w_j/\mu_j)$ *and* $g(w) = \lambda h(w)$.

*The algorithm becomes:*

- $(1+\lambda)\log\tilde{\alpha}^{(t)} = \lambda\log\mu + \log w^{(t-1)} - \eta_{t-1}\nabla f_B(w^{(t-1)})$
- $w^{(t)} = \tilde{\alpha}^{(t)}/\|\tilde{\alpha}^{(t)}\|_1$

In full gradient method, we optimize an upper bound of the objective function

$$Q(w, w') = f(w') + \nabla f(w')^\top (w - w') + LD_h(w; w') + g(w),$$

where we assume that $\phi(w) \le Q(w, w')$.

In stochastic method, we can only optimize

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

That is, each step of Stochastic Mirror Descent solves the following problem:

$$w^{(t)} = \arg\min_{w \in C} Q_{\eta_t, B_t}(w, w^{(t-1)}).$$

The difference of full gradient and stochastic gradient is variance:

$$Q(w; w') - Q_{\eta,B}(w, w') \le V_B(\eta/(1-\eta L), w'),$$

where we have the following definition.

**Definition 1** *We define*

$$V_B(\eta, w') = \max_w \left[ (\nabla f(w') - \nabla f_B(w'))^\top (w - w') - \eta^{-1} D_h(w, w') \right].$$

From Proposition 1, we obtain

$$V_B(\eta, w') = \eta^{-1} D_{h^*}(\nabla h(w') + \eta(\nabla f(w') - \nabla f_B(w')), \nabla h(w')).$$

If $h(\cdot)$ is 1-strongly convex, then $h^*(\cdot)$ is 1-smooth. Then

$$V_B(\eta, w') \le \frac{\eta}{2} \|\nabla f(w') - \nabla f_B(w')\|_2^2.$$

It follows that

$$\mathbf{E}_B V_B(\eta, w') \le \frac{\eta}{2m} \mathbf{E}_\xi \|\nabla f(\xi, w') - f(w')\|_2^2.$$

For the model combination example, we have the following bound. Consider random vector $\delta$ such that $\mathbf{E}\delta = \bar{\delta}$ and $\|\delta\|_\infty \le M$, then Let $\delta_B = m^{-1}\sum_{j=1}^m \delta_j$ be $m$ independent samples of $\delta$, then

$$\mathbf{E}_{\delta_B} h^*(\alpha + \delta_B - \bar{\delta}) \le \ln \sum_j \mathbf{E}_{\delta_{B,j}} \mu_j e^{\alpha_j + \delta_{B,j} - \bar{\delta}_j} \le h^*(\alpha) + M^2/2m.$$

It follows that in Theorem 1, we have

$$\mathbf{E}_B V_B(\eta, w') \le \frac{\eta}{2m} G^2.$$

**Theorem 1** *Consider minibatch stochastic Mirror Descent with $h_t(\cdot) = h(\cdot)$. If $g(w)$ is convex, $f(w) - \lambda g(w)$ is convex and $Lh(w) - f(w)$ is convex. Let*

$$V = \sup_{\eta>0, w\in C} \frac{2m}{\eta} V_B(\eta, w').$$

*If we choose $\eta_t < 0.5/L$ for all $t$, then for all $w \in C$:*

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

*If we let $\eta_t = 1/(2L + 0.5(t-1)\lambda)$, then for all $w \in C$:*

$$\sum_{t=1}^{T} (2L + 0.5\lambda t)\mathbf{E}\left[\phi(w^{(t)}) - \phi(w)\right] \le \frac{2TV}{m} + 2L(2L + 0.5\lambda)D_h(w, w^{(0)}).$$

# 3 RDA

Regularized dual averaging (RDA) can be regraded as a version of mirror descent with changing $h_t(\cdot)$. We may replace full gradient in RDA by stochastic gradient. The resulting algorithm is given in Algorithm! 2. Because we choose a larger $\tilde{\eta}_t$ for proximal mapping, we have better sparsity for solving $L_1$ regularization.

**Algorithm 2:** Stochastic Regularized Dual Averaging

**Input**: $f(\cdot)$, $g(\cdot)$, $w_0$, $\eta_0, \eta_1, \eta_2, \ldots$
  $h(w)$ (default is $h(w) = \eta_0 h_0(w) = 0.5\|w\|_2^2$)
**Output**: $w^{(T)}$
Let $\tilde{\alpha}_0 \in \partial h(w^{(0)})$
Let $\tilde{\eta}_0 = \eta_0$
**for** $t = 1, 2, \ldots, T$ **do**
  Randomly select a minibatch $B$ of $m$ independent samples from $D$
  Let $\tilde{\alpha}_t = \tilde{\alpha}_{t-1} - \eta_{t-1}\nabla f_B(w^{(t-1)})$
  Let $\tilde{\eta}_t = \tilde{\eta}_{t-1} + \eta_{t-1}$
  Let $w^{(t)} = \arg\min_w \left[-\tilde{\alpha}_t^\top w + h(w) + \tilde{\eta}_t g(w)\right]$
**Return**: $w^{(T)}$

# 4 Stochastic Linearized ADMM

We may consider the following stochastic dual decomposition problem:

$$\phi(w, z) = f(w) + g(z) \qquad \text{subject to } Aw + Bz = c, . \tag{2}$$

where

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

In this case, if $f(\cdot)$ is smooth, then we may consider linearized ADMM, which works with a quadratic upper bound of $f(\cdot)$ as follows:

$$f_H(w, \tilde{w}) = f(\tilde{w}) + \nabla f(\tilde{w})^\top (w - \tilde{w}) + \frac{1}{2}\|w - \tilde{w}\|_H^2,$$

and then optimize

$$f_H(w, \tilde{w}) + \frac{\rho}{2}\|Aw + Bz - c\|_2^2.$$

For this formulation, we may take

$$H = \frac{1}{\eta} I - \rho A^\top A,$$

and the solution is

$$w = \tilde{w} - \eta \nabla f(\tilde{w}) - \eta A^\top [\alpha + \rho(A\tilde{w} + Bz - c)].$$

We can now apply the full gradient $\nabla f(\tilde{w})$ by stochastic gradient

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

and $B$ contains $m$ independent samples from $D$.

This leads to Algorithm 3.

**Algorithm 3:** Stochastic Linearized ADMM

**Input**: $\phi(\cdot)$, $A$, $B$, $c$, $\{\eta_t\}$, $G$, $\rho$, $\alpha_0$, $w_0$, $z_0$

**Output**: $w_T, z_T, \alpha_T$

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

  Let $\tilde{z}_t = z_{t-1} - \eta_t B^\top [\alpha_{t-1} + \rho(Aw_{t-1} + Bz_{t-1} - c)]$

  Let $z_t = \arg\min_z [0.5\|z - \tilde{z}_t\|_2^2 + \eta_t g(z)]$

  Let $w_t = w_{t-1} - \eta_t \nabla f_B(w_{t-1}) - \eta_t A^\top [\alpha_{t-1} + \rho(Aw_{t-1} + Bz_t - c)]$

  Let $\alpha_t = \alpha_{t-1} + \rho[Aw_t + Bz_t - c]$

**Return**: $w_T, z_T, \alpha_T$

# 5 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

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

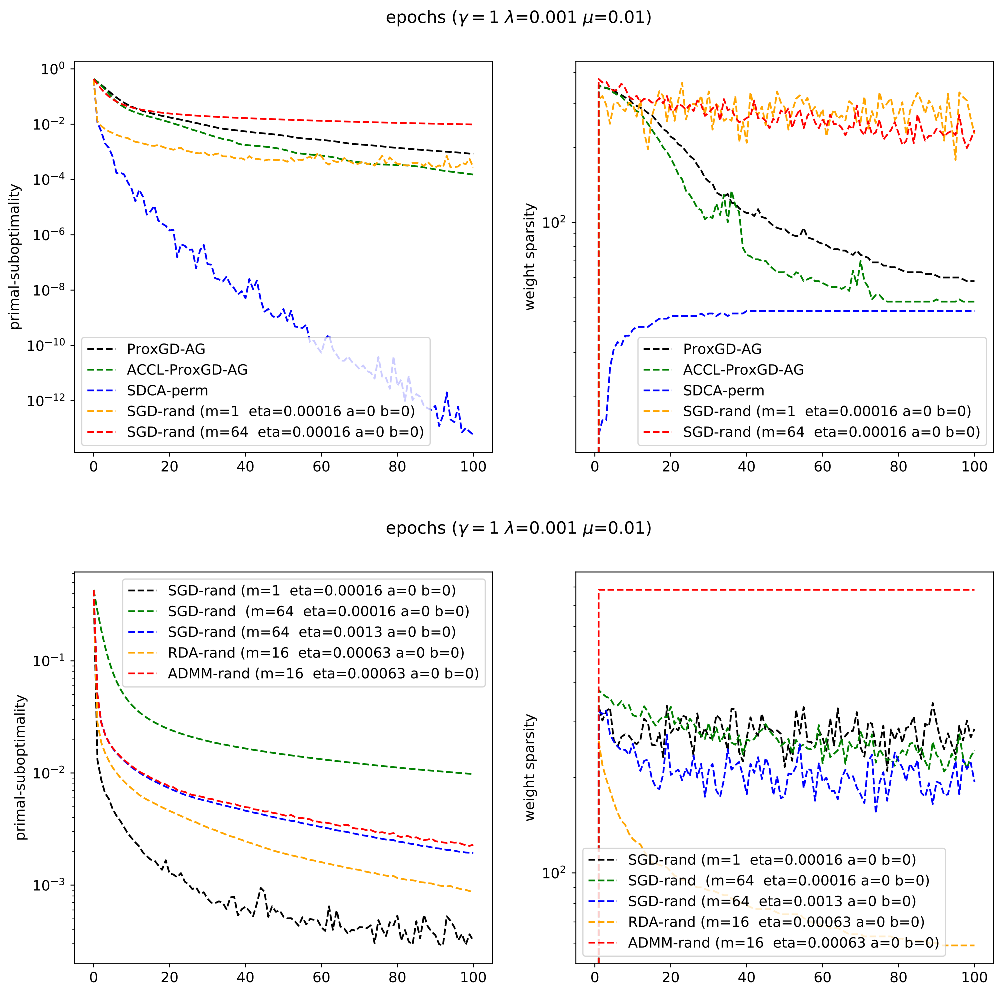


Figure 1: Comparisons of different stochastic algorithms (smooth and strongly convex)

# 6 Convergence Analysis

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

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

We have the following propositions for the minibatch SGD.

**Proposition 1** *We have*

$$\max_w \left[ -D_h(w, w') + \delta^\top (w - w') \right] = D_{h^*}(\alpha' + \delta, \alpha'),$$

*where* $\alpha' = \nabla h(w')$.

**Proof** Let $w$ be the maximizer. Then

$$-\nabla h(w) + \nabla h(w') + \delta = 0.$$

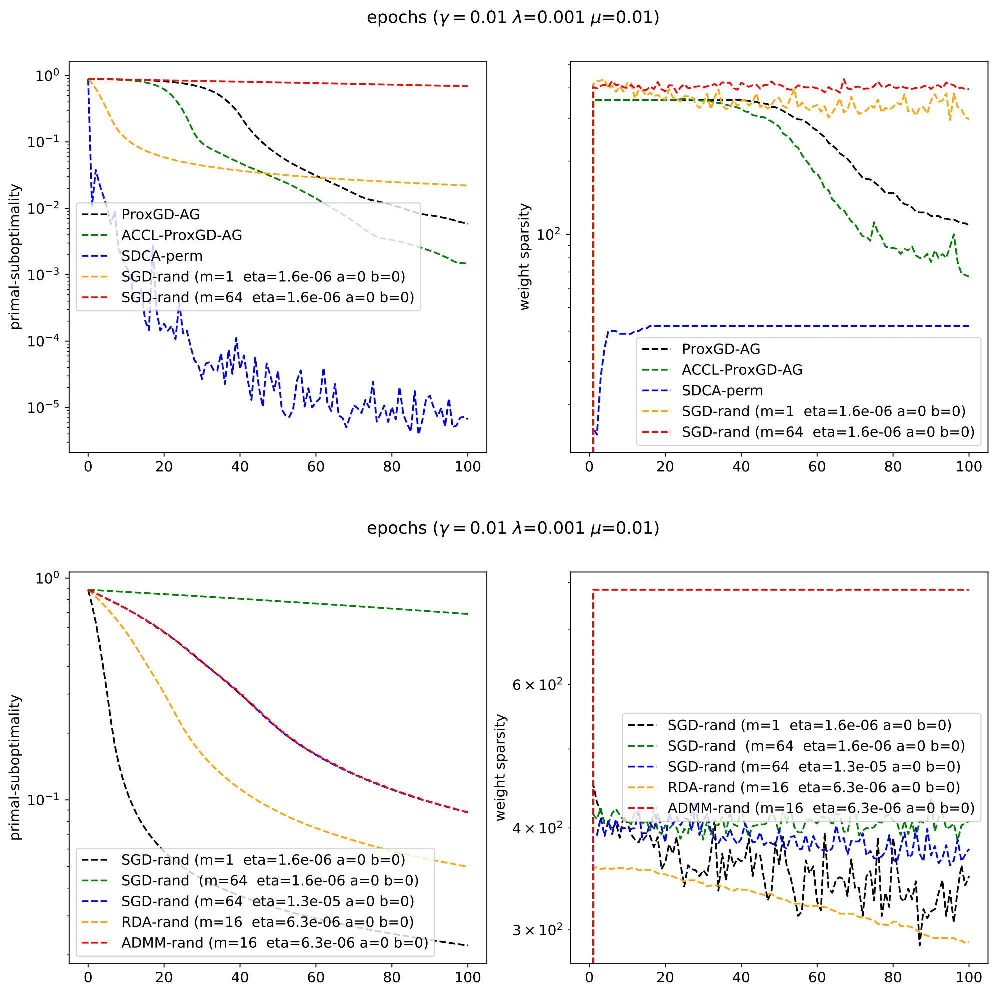


Figure 2: Comparisons of different stochastic algorithms (near non-smooth and strongly convex)

It follows that

$$-D_h(w, w') + \delta^\top (w - w') = D_h(w', w) = D_{h^*}(\alpha, \alpha'),$$

where $\alpha = \nabla h(w)$ and $\alpha' = \nabla h(w')$. This proves the result. ■

**Proposition 2** *Assume that $f(w)$ is $L \cdot h$-smooth in $C$ (that is, $Lh(w) - f(w)$ is convex). If we pick $\eta < 0.5/L$, then given any $w'$, we have*

$$\phi(w) \le Q_{\eta,B}(w; w') + 0.5\eta^{-1} D_{h^*}(\nabla h(w') + 2\eta(\nabla f_B(w') - \nabla f(w')), \nabla h(w')).$$

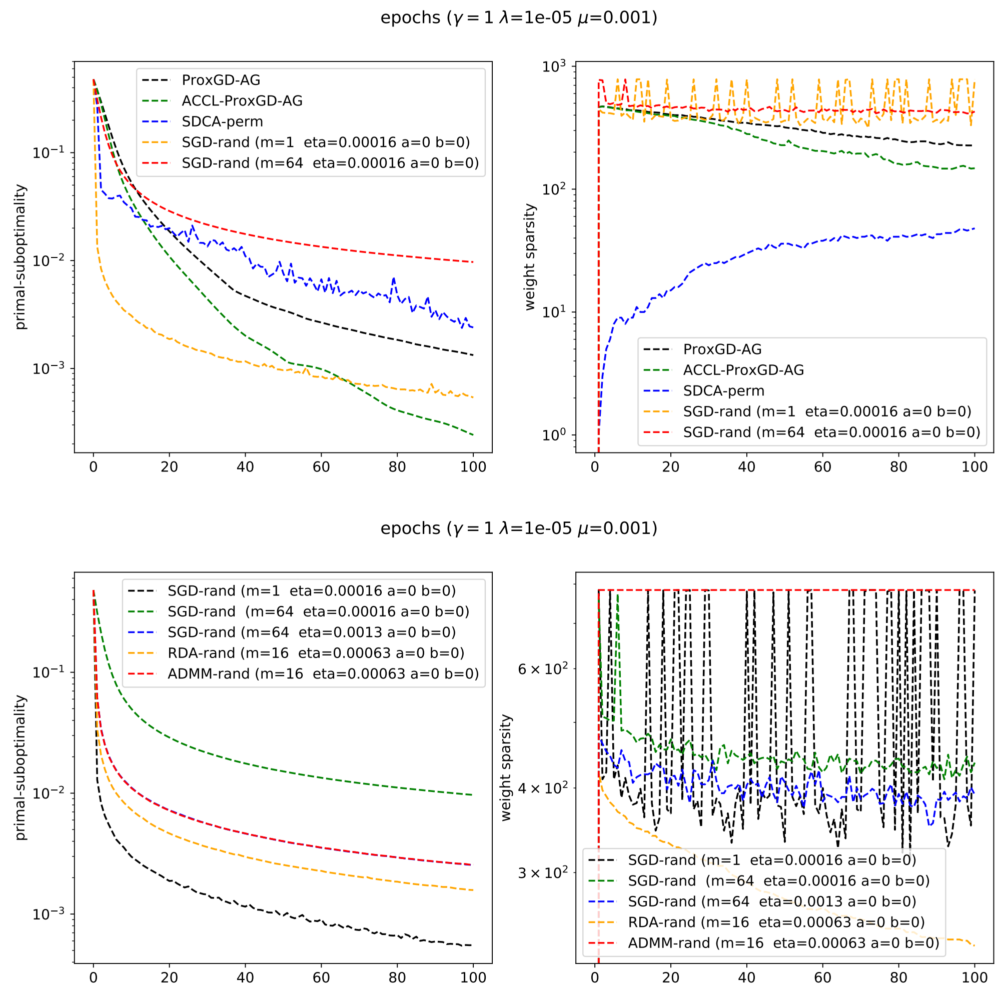


Figure 3: Comparisons of different stochastic algorithms (near non-smooth and near non-strongly convex)

**Proof** From smoothness, we have

$$
\begin{aligned}
& f(w) + g(w) \\
\leq & f(w') + \nabla f(w')^\top (w - w') + L D_h(w, w') + g(w) \\
\leq & f(w') + \nabla f_B(w')^\top (w - w') + L D_h(w, w') + (\nabla f(w') - \nabla f_B(w'))^\top (w - w') + g(w) \\
= & Q_{\eta,B}(w; w') - (\eta^{-1} - L) D_h(w, w') + (\nabla f(w') - \nabla f_B(w'))^\top (w - w') \\
\leq & Q_{\eta,B}(w; w') + (\eta^{-1} - L) D_{h^*}(\alpha' + (\nabla f_B(w') - \nabla f(w'))/(\eta^{-1} - L), \alpha'),
\end{aligned}
$$

where $\alpha' = \nabla h(w')$. The first inequality uses the smoothness of $f(x)$. The second inequality is algebra, and the third inequality uses the definition of $Q_{\eta,B}(\cdot)$. The last inequality uses Proposition 1. This proves the desired result. ∎

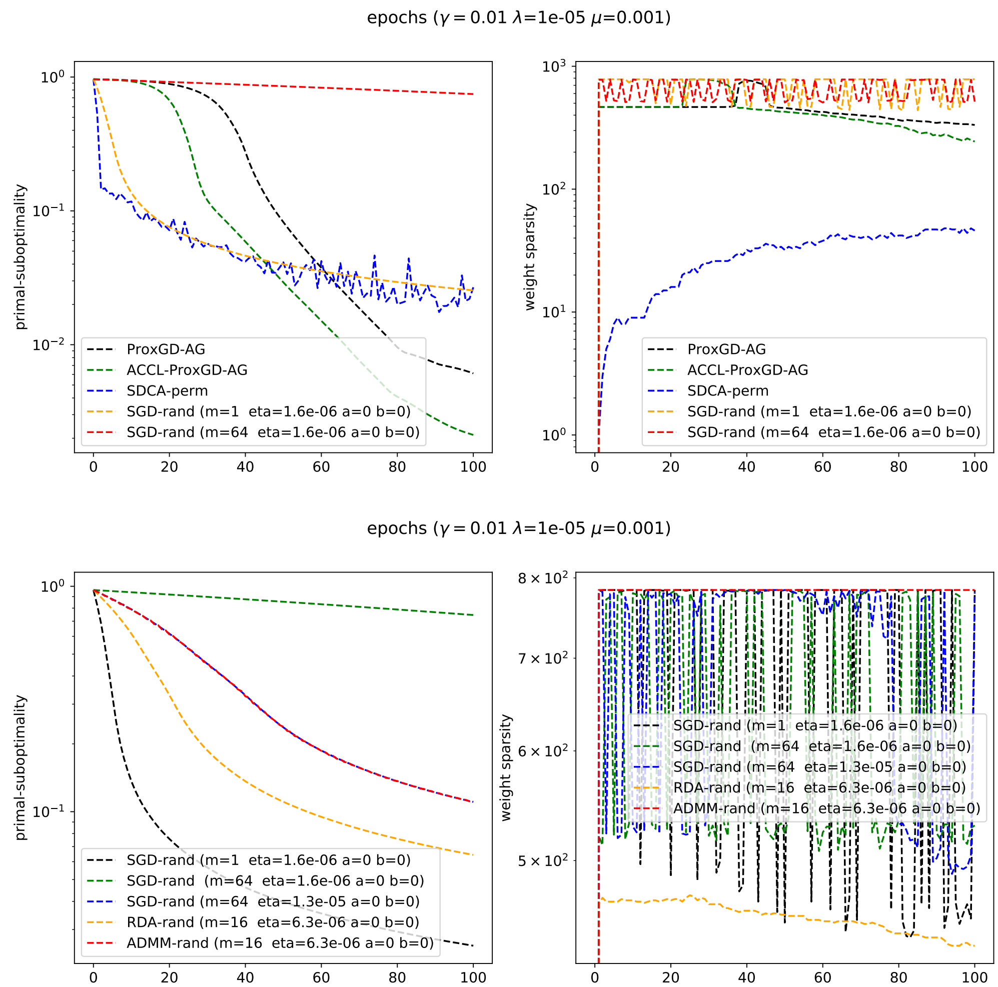


Figure 4: Comparisons of different stochastic algorithms (near non-smooth and near non-strongly convex)

**Proposition 3** *IF $f(w)$ is $\lambda$ with respect to $h(\cdot)$ and strongly convex, and $g(w) - \lambda h(w)$ is convex. We have for all $w$:*

$$Q_B(w^{(t)}; w^{(t-1)}) \leq \phi(w) + \eta_t^{-1} D_h(w, w^{(t-1)}) + \left(\eta_t^{-1} + \lambda\right) D_h(w, w^{(t)}).$$

**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 $w$:

$$\begin{aligned} Q_B(w^{(t)}; w^{(t-1)}) \leq & Q_B(w; w^{(t-1)}) - \left(\eta_t^{-1} + \lambda\right) D_h(w, w^{(t)}) \\ \leq & \phi(w) + \eta_t^{-1} D_h(w, w^{(t-1)}) - \left(\eta_t^{-1} + \lambda\right) D_h(w, w^{(t)}). \end{aligned}$$

This proves the result. ■

### 6.1 Proof of Theorem 1

Using Proposition 2, we obtain for minibatch $B_t$:

$$\phi(w^{(t)}) \leq \phi(w) + \frac{\eta_t V}{2m(1-\eta_t L)} + \eta_t^{-1} D_h(w, w^{(t-1)}) - \left(\eta_t^{-1} + \lambda\right) D_h(w, w^{(t)}).$$

For $\lambda = 0$, take expectation, we obtain

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

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

For $\lambda > 0$, we divide by $\eta_t$ and obtain

$$\eta_{t+1}^{-1}\phi(w^{(t)}) \leq \phi(w) + \frac{2V}{m} + \eta_t^{-1}\eta_{t+1}^{-1} D_h(w, w^{(t-1)}) - \eta_{t+1}^{-1}\eta_{t+2}^{-1} D_h(w, w^{(t)}).$$

Taking expectation with $\lambda = 0$, and by summing over $t = 1$ to $T$, we obtain the desired bound.