Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 22: Variance Reduction Methods for Stochastic Optimization

Finite Sum Problem in Machine Learning

We consider the following stochastic optimization problem with finite sum structure

$$\min_{w \in C} \phi(w), \quad \phi(w) = f(w), \qquad f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w). \tag{1}$$

Consider the SGD update rule with batch size of 1:

$$w^{(t)} = w^{(t-1)} - \eta_t \nabla f_i(w^{(t-1)}),$$

where i is sampled uniformly from D, which is the uniform distribution over $\{1, \ldots, n\}$.

Variance of Stochastic Gradient

Let

$$V = \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(w) - \nabla f(w)\|_2^2,$$

then we can set the optimal learning rate as

$$\eta = \frac{\|\nabla f(w)\|_2^2}{L(\|\nabla f(w)\|_2^2 + V)},$$

and obtain the following one-step convergence result:

$$\mathbf{E}_{i \sim D} f(w - \eta \nabla f_i(w)) \leq f(w) - 0.5\eta \|\nabla f(w)\|_2^2.$$

As $V \neq 0$ and $\|\nabla f(w)\|_2 \to 0$, $\eta \to 0$, and we obtain sublinear convergence.

SVRG: motivation

At each time, we keep a version of estimated w as $\tilde{w}$ that is close to the optimal w. For example, we can keep a snapshot of $\tilde{w}$ after every m SGD iterations. Moreover, we maintain the average gradient

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

and its computation requires one pass over the data using $\tilde{w}$. Note that the expectation of $\nabla f_i(\tilde{w}) - \tilde{\mu}$ over i is zero. If we define the auxiliary function

$$\tilde{f}_i(w) = f_i(w) - (\nabla f_i(\tilde{w}) - \tilde{\mu})^\top w,$$

then

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) = \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_i(w).$$

Variance Reduction

We can apply the SGD rule to the finite sum with respect to $\tilde{f}_i(w)$, and obtain the following update rule is generalized SGD:

$$w^{(t)} = w^{(t-1)} - \eta_t \nabla \tilde{f}_i(w^{(t-1)}), \quad \nabla \tilde{f}_i(w) = (\nabla f_i(w) - \nabla f_i(w) + \tilde{\mu}), \tag{2}$$

where we draw i randomly from D.
To see that the variance of the update rule (2) is reduced, we note that when both $\tilde{w}$ and $w^{(t)}$ converge to the same parameter w_*, then $\tilde{\mu} \to 0$. Therefore if $\nabla f_i(\tilde{w}) \to \nabla f_i(w_*)$, then

$$\nabla f_i(w^{(t-1)}) - \nabla f_i(\tilde{w}) + \tilde{\mu} \to \nabla f_i(w^{(t-1)}) - \nabla f_i(w_*) \to 0.$$

SVRG Algorithm

```
Algorithm 1: Stochastic Variance Reduced Gradient (SVRG)
Input: $\phi(\cdot)$, $w_0$, $\eta$, update frequency $m$
Output: $\tilde{w}^{(S)}$
Let $\tilde{w}^{(0)} = w_0$
for $s = 1, 2, \ldots, S$ do
    Let $\tilde{w} = \tilde{w}^{(s-1)}$
    Let $\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\tilde{w})$
    Let $w_0 = \tilde{w}$
    for $t = 1, \ldots, m$ do
        Select $i_t$ uniformly at random from $\{1, \ldots, n\}$
        Let $g_{i_t} = (\nabla f_{i_t}(w_{t-1}) - \nabla f_{i_t}(\tilde{w}) + \tilde{\mu}$
        Let $w_t = w_{t-1} - \eta g_{i_t}$
    option I: set $\tilde{w}^{(s)} = w_m$
    option II: set $\tilde{w}^{(s)} = w_t$ for randomly chosen $t \in \{0, \ldots, m-1\}$
Return: $\tilde{w}^{(S)}$
```

Theory

Theorem

Consider the SVRG algorithm with option II. Assume that all $f_i(w)$ are convex and L-smooth, and $f(w)$ is λ strongly convex. Let $w_* = \arg\min_w f(w)$. Assume that m is sufficiently large so that

$$\rho = \frac{1}{\lambda\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1,$$

then we have geometric convergence in expectation for SVRG:

$$\mathbf{E} f(\tilde{w}^{(s)}) \leq \mathbf{E} f(w_*) + \rho^s [f(\tilde{w}^{(0)}) - f(w_*)].$$

Proof: key ideas

It can be shown that

$$n^{-1}\sum_{i=1}^{n}\|\nabla f_i(w)-\nabla f_i(w_*)\|_2^2 \le 2L[f(w)-f(w_*)].$$

This implies the variance bound:

$$\mathbf{E}\|\nabla f_{i_t}(w_{t-1})-\nabla f_{i_t}(\tilde{w})+\tilde{\mu}\|_2^2 \le 4L[f(w_{t-1})-f(w_*)+f(\tilde{w})-f(w_*)].$$

We then obtain

$$\begin{aligned}&\mathbf{E}\,\|w_t-w_*\|_2^2\\ \le&\|w_{t-1}-w_*\|_2^2-2\eta(1-2L\eta)[f(w_{t-1})-f(w_*)]+4L\eta^2[f(\tilde{w})-f(w_*)].\end{aligned}$$

Sum over t and take expectation to obtain the desired bound.

Proximal SVRG

Algorithm 2: Proximal Stochastic Variance Reduced Gradient (Prox-SVRG)

```
Input: $\phi(\cdot)$, $w_0$, $\eta$, update frequency $m$
Output: $\tilde{w}^{(S)}$
Let $\tilde{w}^{(0)} = w_0$
Let $D$ be the distribution on $\{1, \ldots, n\}$ according to probability $Q = \{q_1, \ldots, q_n\}$
for $s = 1, 2, \ldots, S$ do
    Let $\tilde{w} = \tilde{w}^{(s-1)}$
    Let $\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{w})$
    Let $w_0 = \tilde{w}$
    for $t = 1, \ldots, m$ do
        Randomly pick $i \sim D$ and update weight
        Let $g_i = (\nabla f_i(w_{t-1}) - \nabla f_i(\tilde{w}))/(q_i n) + \tilde{\mu}$
        Let $w_t = \mathrm{prox}_{\eta g}(w_t - \eta g_i)$
    Let $\tilde{w}^{(s)} = \frac{1}{m}\sum_{t=1}^{m} w_t$
Return: $\tilde{w}^{(S)}$
```

Theory

Theorem

Assume that each $f_i(w)$ is L_i-smooth, and $g(w)$ is λ strongly-convex. Let $w_* = \arg\min_w \phi(w)$ and $L_Q = \max_i L_i/(q_i n)$. In addition, assume that $0 < \eta < 1/(4L_Q)$ and m is sufficiently large so that

$$\rho = \frac{1}{\lambda\eta(1 - 4L_Q\eta)m} + \frac{4L_Q\eta(m+1)}{(1-4L_Q\eta)m} < 1. \tag{3}$$

Then the Prox-SVRG method has geometric convergence in expectation:

$$\mathbf{E} f(\tilde{x}^{(s)}) - f(x_*) \leq \rho^s [f(\tilde{w}^{(0)}) - f(w_*)].$$

SDCA as Variance Reduction

In SDCA, we consider the following problem:

$$\min_w \frac{1}{n}\sum_{i=1}^n f_i(w) + \frac{\lambda}{2}\|w\|_2^2.$$

The optimality condition is

$$w_* = \frac{1}{\lambda n}\sum_{i=1}^n \alpha_i^*, \quad \alpha_i^* = -\nabla f_i(w_*). \tag{4}$$

We may maintain a similar dual representation as follows,

$$w^{(t)} = \frac{1}{\lambda n}\sum_{i=1}^n \alpha_i^{(t)}, \tag{5}$$

and update the primal *w* using SGD as:

$$w^{(t)} = w^{(t-1)} - \lambda\eta_t w^{(t-1)} - \eta_t \nabla f_i(w^{(t-1)}). \tag{6}$$

SDCA

Now we may use the relationship

$$-\lambda\eta_t w^{(t-1)} = -\eta_t \frac{1}{n}\sum_{i=1}^{n} \alpha_i^{(t-1)},$$

to replace the gradient $-\lambda\eta_t w^{(t-1)}$ by stochastic gradient $-\eta_t\alpha_i^{(t-1)}$. This leads to the following update

$$w^{(t)} = w^{(t-1)} - \eta_t(\alpha_i^{(t-1)} + \nabla f_i(w^{(t-1)})). \tag{7}$$

The corresponding dual update is a version of the SDCA method:

$$\alpha_\ell^{(t)} = \begin{cases} \alpha_i^{(t-1)} - \eta_t(\nabla f_i(w^{(t-1)}) + \alpha_i^{(t-1)}) & \ell = i \\ \alpha_\ell^{(t-1)} & \ell \neq i \end{cases},$$

together with primal update

$$w^{(t)} = w^{(t-1)} + \frac{1}{\lambda n}(\alpha_i^{(t)} - \alpha_i^{(t-1)}).$$

SAGA

```
Algorithm 3: SAGA
Input: $\phi(\cdot)$, $w^{(0)}$, $\eta$, update frequency $m$
Output: $\tilde{w}^{(T)}$
Initialize $\alpha_1, \ldots, \alpha_n$
Let $\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} (-\alpha_i)$
for $t = 1, 2, \ldots, T$ do
    Select $i$ uniformly at random from $\{1, \ldots, n\}$
    Let $\alpha_i' = -\nabla f_i(w^{(t-1)})$
    Let $g_i = (-\alpha_i') - (-\alpha_i) + \tilde{\mu}$
    Let $w_t = \mathrm{prox}_{\eta g}(w_{t-1} - \eta g_i)$
    Let $\tilde{\mu} = \tilde{\mu} + \frac{1}{n}(\alpha_i - \alpha_i')$
    Let $\alpha_i = \alpha_i'$
Return: $w^{(T)}$
```

SAGA (continued)

$$\min_{w} \left[\psi_i(X_i^\top w) + g(w) \right]$$

Algorithm 4: SAGA

```
Input: $\phi(\cdot)$, $w^{(0)}$, $\eta$, update frequency $m$
Output: $\tilde{w}^{(T)}$
Initialize $\beta_1, \ldots, \beta_n$
Let $\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} (-X_i \beta_i)$
for $t = 1, 2, \ldots, T$ do
    Select $i$ uniformly at random from $\{1, \ldots, n\}$
    Let $\beta_i' = -\nabla \psi_i(X_i^\top w^{(t-1)})$
    Let $g_i = X_i(\beta_i - \beta_i')$
    Let $w_t = \mathrm{prox}_{\eta g}(w^{(t-1)} - \eta(g_i + \tilde{\mu}))$
    Let $\tilde{\mu} = \tilde{\mu} + \frac{1}{n} g_i$
    Let $\beta_i = \beta_i'$
Return: $w^{(T)}$
```

Minibatch

Algorithm 5: Minibatch Accelerated Prox-SVRG

```
Input: $\phi(\cdot)$, $w_0$, $\eta$, update frequency $m$
Output: $\tilde{w}^{(S)}$
Let $\tilde{w}^{(0)} = w_0$
Let $D$ be the distribution on $\{1, \ldots, n\}$ according to probability $Q = \{q_1, \ldots, q_n\}$
for $s = 1, 2, \ldots, S$ do
    Let $\tilde{w} = \tilde{w}^{(s-1)}$
    Let $\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\tilde{w})$
    Let $w_0 = \tilde{w}$
    for $t = 1, \ldots, m$ do
        Randomly pick minibatch $B \sim D$ and update weight
        Let $u_t = w_{t-1} + \beta(w_{t-1} - w_{t-2})$
        Let $g_B = (\nabla f_B(u_t) - \nabla f_B(\tilde{w}))/(q_i n) + \tilde{\mu}$
        Let $w_t = \mathrm{prox}_{\eta g}(w_t - \eta g_B)$
    Let $\tilde{w}^{(s)} = \frac{1}{m} \sum_{t=1}^{m} w_t$
Return: $\tilde{w}^{(S)}$
```

Empirical Studies

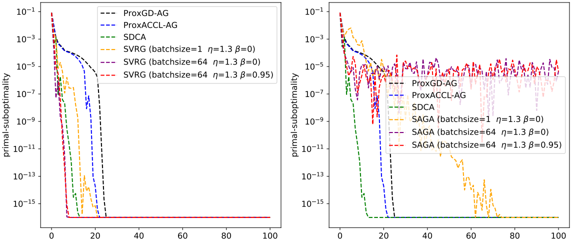
We study the smoothed hinge loss function $\phi_\gamma(z)$ with $\gamma = 1$, and solves the following $L_1 - L_2$ regularization problem:

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

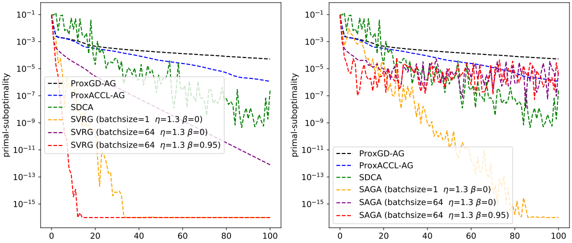
We compare different algorithms with constant learning rate.

Comparisons (smooth)

epochs ($\gamma = 1$ λ=0.001 μ=0.01)

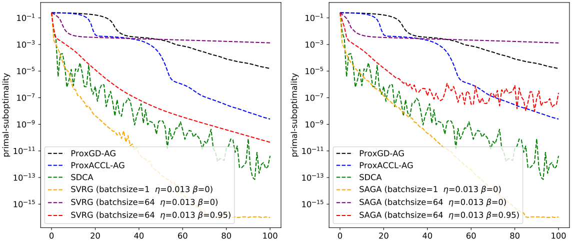


epochs ($\gamma = 1$ λ=1e-05 μ=0.001)

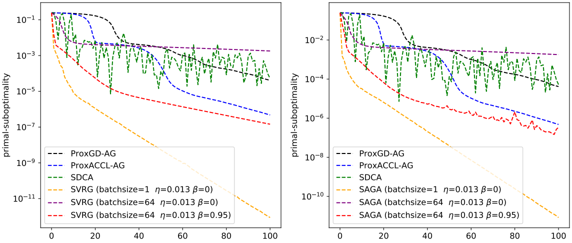


Comparisons (near non-smooth)

epochs ($\gamma = 0.01$ λ=0.001 μ=0.01)



epochs ($\gamma = 0.01$ λ=1e-05 μ=0.01)



Summary

Variance of SGD

- The SGD slows down when $t \to \infty$, $\eta \to 0$
- Variance dominates, sublinear convergence.

Variance reduction

- SVRG: reduce variance using control variate.
- Variance bounded by primal suboptimality:

$$n^{-1} \sum_{i=1}^{n} \|\nabla f_i(w) - \nabla f_i(w_*)\|_2^2 \leq 2L[f(w) - f(w_*)].$$

Fast Convergence

- $O(\kappa + n)$ instead of $O(\kappa n)$.
- Acceleration is helpful for minibatch