Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 20: Stochastic Gradient Dual Methods

Stochastic Optimization in Machine Learning

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:

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

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

Gradient versus Stochastic Gradient

In gradient based methods, we use gradient

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

In SGD, we replace the full gradient with stochastic gradient

$$\nabla_w f(\xi, w^{(t-1)}),$$

or minibatch stochastic gradient:

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

Stochastic Mirror Descent

Algorithm 1: Stochastic Mirror Descent

Input: $f(\cdot)$, $g(\cdot)$, $\{h_t(\cdot)\}$, $\{\eta_t\}$
Output: $w^{(T)}$

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

Return: $w^{(T)}$

Example

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 $\xi = (x, y)$.
Let $h(w) = \sum_j w_j \log(w_j/\mu_j)$ and $g(w) = \lambda h(w)$.
The algorithm becomes:

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

Convergence

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

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

Then Stochastic Mirror Descent solves the following problem at each step:

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

Variance

In full gradient method, we should 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 $\phi(w) \leq Q(w, w')$.
In stochastic method, we can only optimize

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

The difference is variance: The difference of full gradient and stochastic gradient is variance:

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

Variance Bound

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

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

It follows that

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

Let $V = \mathbf{E}_\xi \|\nabla f(\xi, w') - f(w')\|_2^2$, then

$$\frac{2m}{\eta} V \leq \mathbf{E}_B V_B(\eta, w').$$

Convergence Theorem

Theorem

Consider minibatch stochastic Mirror Descent. 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 $w \in C$:

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

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

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

Stochastic Regularized Dual Averaging (RDA)

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: $w^{(T)}$

Stochastic Decomposition Problem

We may also consider the following stochastic decomposition problem:

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

where

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

ADMM

Algorithm 3: Preconditioned ADMM

Input: $\phi(\cdot)$, A, B, c, H, G, ρ, α_0, x_0, z_0

Output: w_T, z_T, α_T

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

Let $z_t =$
$\arg\min_z \left[\alpha_{t-1}^\top Bz + g(z) + \frac{\rho}{2}\|Aw_{t-1} + Bz - c\|_2^2 + \frac{1}{2}\|z - z_{t-1}\|_G^2\right]$

Let $w_t =$
$\arg\min_x \left[\alpha_{t-1}^\top Aw + f(w) + \frac{\rho}{2}\|Aw + Bz_t - c\|_2^2 + \frac{1}{2}\|w - w_{t-1}\|_H^2\right]$

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

Return: w_T, z_T, α_T

Linearization

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

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

Stochastic Linearized ADMM

Algorithm 4: Stochastic Linearized ADMM

Input: $\phi(\cdot)$, A, B, c, $\{\eta_t\}$, G, ρ, α_0, w_0, z_0

Output: w_T, z_T, α_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, α_T

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

Comparisons (smooth and strongly convex)

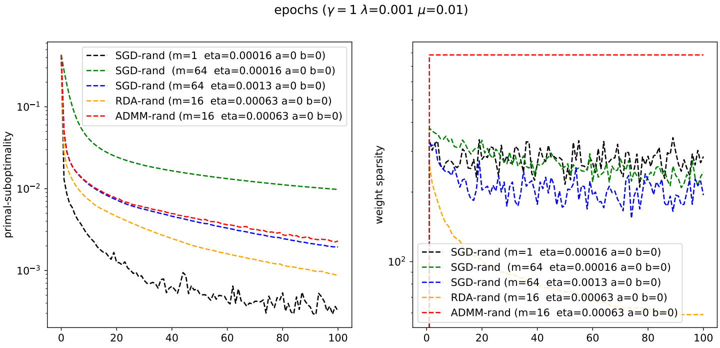


Figure: Comparisons of stochastic algorithms

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

Comparisons (near nonsmooth and strongly convex)

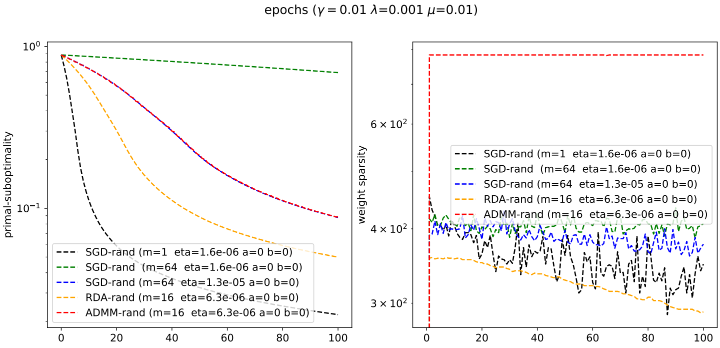
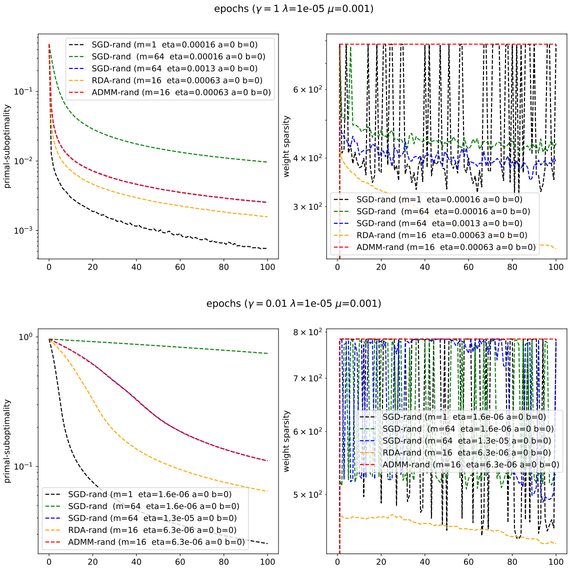


Figure: Comparisons of stochastic algorithms

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

Comparisons (near non-strongly convex)



Summary

Stochastic Optimization

- stochastic optimization

Stochastic Gradient

- unbiased estimate of the full gradient
- less computation per iteration

Convergence

- can be obtained for different cases.
- different learning rate schedule, which may depend on the minibatch size