Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 13: Mirror Descent and Dual Averaging

Composite Convex Optimization

In this lecture, we consider the composite convex optimization optimization problem:

$$\min_{x \in \mathbb{R}^d} \phi(x), \qquad \phi(x) = f(x) + g(x).$$

where $g(x)$ may be defined on the convex domain $C \subset \mathbb{R}^d$. That is, $g(x) = +\infty$ when $x \notin C$.

- $f(x)$ is smooth
- $g(x)$ may be nonsmooth, such as L_1 regularization

Generalized Proximal Mapping

In this lecture, we consider the following generalization of proximal mapping

$$\mathrm{prox}_h(x) = \arg\min_z \left[-x^\top z + h(z) + g(z) \right],$$

and we assume this generalized proximal mapping can be efficiently computed.

Bregman Divergence

Give a strictly convex function $h(x)$ on C, we may define the corresponding Bregman divergence as

$$D_h(x, y) = h(x) - h(y) - \nabla h(y)^\top (x - y).$$

If $h(x)$ has more than one subgradient at y, then we may choose $\nabla h(y) \in \partial h(y)$ to be a specific subgradient, depending on applications.

The Bregman divergence of a convex function is always non-negative.

Smoothness

We say that a function is smooth with respect to a convex function $h(\cdot)$ if

$$f(x) \leq f(y) + \nabla f(y)^\top (x - y) + D_h(x, y).$$

This can be used to form an upper bound of $f(x)$.

Using the generalized proximal mapping, we may consider the following upper bound of $\phi(x)$ with any y and any h such that $f(x)$ is smooth with respect to h:

$$Q(x; y) = f(y) + \nabla f(y)^\top (x - y) + D_h(x, y) + g(x).$$

Algorithm 1: Proximal Mirror Descent

Input: $f(\cdot)$, $g(\cdot)$, x_0, and $h_1, h_2, \ldots$

Output: x_T

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

  Let $\tilde{x}_t = \nabla h_{t-1}(x_{t-1}) - \nabla f(x_{t-1})$

  Let $x_t = \mathrm{prox}_{h_{t-1}}(\tilde{x}_t)$

Return: x_T

Mirror Descent

Algorithm 2: Mirror Descent

Input: $f(\cdot)$, $g(\cdot)$, x_0, $h(x)$, $\{\eta_t\}$
Output: x_T
for $t = 1, 2, \ldots, T$ **do**
    Let $\tilde{x}_t = \nabla h(x_{t-1}) - \eta_{t-1} \nabla f(x_{t-1})$
    Let $x_t = \arg\min_{x \in C}[-\tilde{x}_t^\top x + h(x)]$
Return: x_T

Example

Example

If we take $h(x)$ as $h(x) = \sum_j (x_j \ln x_j - x_j)$, defined on $C = \mathbb{R}_+^d$. Then its gradient is $\nabla h(x) = \ln x$. Therefore the mirror update rule on C is

$$[x_t]_j = [x_t]_j \exp(-\eta_{t-1}[\nabla f(x_{t-1})]_j) \qquad j = 1, \ldots, d,$$

where $[x]_j$ denotes the j-th component of a vector x. If we take the same $h(x)$ on domain $C = \{x \in \mathbb{R}_+^d : \sum_j x_j = 1\}$, then

$$[x_t]_j = \frac{[x_t]_j \exp(-\eta_{t-1}[\nabla f(x_{t-1})]_j)}{\sum_{k=1}^d [x_t]_k \exp(-\eta_{t-1}[\nabla f(x_{t-1})]_k)}, \qquad j = 1, \ldots, d.$$

These methods are often referred to as exponentiated gradient methods.

Dual Averaging: Derivation I

We note that Algorithm 1 converges if $f(x)$ is smooth with respect to h_t for all t.

In general, the first order condition of x_t for being the solution of the general proximal mapping minimization problem is:

$$\nabla h_{t-1}(x_t) + \nabla g(x_t) = \nabla h_{t-1}(x_{t-1}) - \nabla f(x_{t-1}). \tag{1}$$

Given a sequence of positive numbers $\{\eta_t\}$, we can define

$$\eta_t h_t(x) = \eta_{t-1}[h_{t-1}(x) + g(x)], \tag{2}$$

in order to simplify the recursion in (1).

Derivaiton II

Then by solving the above recursion, we obtain

$$\eta_t h_t(x) = \eta_0 h_0(x) + \left(\sum_{s=0}^{t-1} \eta_s\right) g(x). \tag{3}$$

From (2) and (1), we obtain

$$\eta_t \nabla h_t(x_t) = \nabla[\eta_{t-1}(h_{t-1}(x_t)+g(x_t))] = \eta_{t-1}\nabla h_{t-1}(x_{t-1}) - \eta_{t-1}\nabla f(x_{t-1}).$$

Therefore by solving the recursion, we obtain

$$\eta_t \nabla h_t(x_t) = \eta_0 \nabla h_0(x_0) - \sum_{s=0}^{t-1} \eta_s \nabla f(x_s).$$

Derivation III

By combining this with (3), we obtain

$$\eta_0 \nabla h_0(x_t) + \left(\sum_{s=0}^{t-1} \eta_s\right) \nabla g(x_t) = \eta_0 \nabla h_0(x_0) - \sum_{s=0}^{t-1} \eta_s \nabla f(x_s),$$

which implies that

$$x_t = \arg\min_x \left[-\left(\eta_0 \nabla h_0(x_0) - \sum_{s=0}^{t-1} \eta_s \nabla f(x_s)\right)^\top x + \eta_0 h_0(x) + \left(\sum_{s=0}^{t-1} \eta_s\right) g(x) \right].$$

This leads to a method which is referred to as the regularized dual averaging (RDA) method.

RDA Algorithm

Algorithm 3: Regularized Dual Averaging

Input: $f(\cdot)$, $g(\cdot)$, x_0, $\eta_0, \eta_1, \eta_2, \ldots$
$h(x)$ (default is $h(x) = \eta_0 h_0(x) = 0.5\|x\|_2^2$)

Output: x_T

Let $\tilde{\alpha}_0 \in \partial h(x_0)$

Let $\tilde{\eta}_0 = \eta_0$

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

  Let $\tilde{\alpha}_t = \tilde{\alpha}_{t-1} - \eta_{t-1} \nabla f(x_{t-1})$

  Let $\tilde{\eta}_t = \tilde{\eta}_{t-1} + \eta_{t-1}$

  Let $x_t = \arg\min_x \left[-\tilde{\alpha}_t^\top x + h(x) + \tilde{\eta}_t g(x) \right]$

Return: x_T

L_1 Regularization: Proximal Gradient versus RDA

Solving

$$f(x) + \mu \|x\|_1.$$

Proximal Mapping:

$$\mathrm{prox}_\eta(x) = \arg\min_z \left[\frac{1}{2\eta} \|z - x\|_2^2 + \mu \|z\|_1 \right] = [\mathrm{sign}(x_j)(|x_j| - \mu\eta)_+]_{j=1,\ldots,d}$$

Proximal Gradient

- $\alpha_t = x_{t-1} - \eta \nabla f(x_{t-1})$
- $x_t = \mathrm{prox}_\eta(\alpha_t)$

RDA

- $\alpha_t = \alpha_{t-1} - \eta \nabla f(x_{t-1})$
- $x_t = \mathrm{prox}_{\eta t}(\alpha_t)$

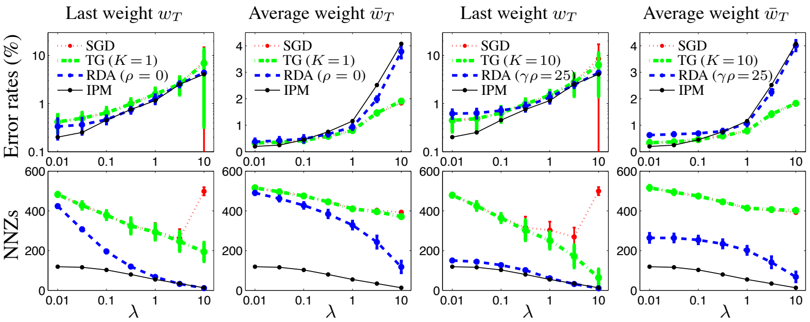

Figure 3: Tradeoffs between testing error rates and NNZs in solutions (for classifying 6 and 7).

Convergence Theorem of RDA

We have the following convergence theorem.

Theorem

Consider Algorithm 3. Assume that $f(x)$ is smooth with respect to $\eta_t^{-1} h(\cdot)$ for all t. Then for all $x \in C$:

$$\phi(x_t) \leq \phi(x) + \frac{1}{\tilde{\eta}_t}[\tilde{\eta}_0(\phi(x_0) - \phi(x)) + D_h(x, x_0)].$$

Proof Sketch

We employ the estimate sequence method

$$\psi_t(x) = f(x_{t-1}) + \nabla f(x_{t-1})^\top (x - x_{t-1}) + g(x).$$

Obviously we have

$$\psi_t(x) \leq \phi(x).$$

Define

$$\phi_t(x) = \frac{1}{\tilde{\eta}_t} \left[\tilde{\eta}_0 \phi(x_0) - h(x_0) - \tilde{\alpha}_0^\top (x - x_0) + \sum_{s \leq t} \eta_{s-1} \psi_s(x) + h(x) \right]$$

then

$$\phi(x_t) \leq \phi_t(x_t).$$

Convergence Analysis of Mirror Descent

Proposition

Assume that in Algorithm 1, $f(x)$ is h_t-smooth for all t. If we let

$$Q_t(x) = f(x_{t-1}) + \nabla f(x_{t-1})^\top (x - x_{t-1}) + D_{h_{t-1}}(x; x_{t-1}) + g(x),$$

then $\phi(x) \leq Q_t(x)$ and

$$x_t = \arg\min_x Q_t(x).$$

Moreover, if $g(x)$ is λ'-strongly convex, then $\forall x \in C$:

$$Q_t(x) - Q_t(x_t) \geq D_{h_{t-1}}(x; x_t) + \frac{\lambda'}{2}\|x - x_t\|_2^2.$$

Theorem

Assume that we take $\eta_t = \eta$ in Algorithm 2, and $f(x)$ is smooth with respect to $\eta^{-1} h(\cdot)$. Then we have for all $x \in C$:

$$\frac{1}{T}\sum_{t=1}^{T} \phi(x_t) \leq \phi(x) + \frac{1}{\eta T} D_h(x, x_0).$$

Summary

Composite convex optimization problem

$$\phi(x) := \underbrace{f(x)}_{\text{smooth}} + \underbrace{g(x)}_{\text{nonsmooth}}$$

- primal methods: proximal gradients

From primal to dual methods:

- Mirror Descent
- Regularized Dual Averaging