Mirror Descent and Dual Averaging

1 Composite Convex Optimization Problem

Consider the following composite convex optimization problem:

$$\min_{x\in\mathbb{R}^d} \phi(x) \qquad \phi(x) = [f(x) + g(x)], \tag{1}$$

where $g(x)$ may be defined on the convex domain $C \subset \mathbb{R}^d$. That is, $g(x) = +\infty$ when $x \notin C$. Here we assume that $f(x)$ is a smooth convex function defined on C, with smoothness parameter L, and $g(x)$ may be nonsmooth convex function.

In the previous lectures, we have studied so-called primal methods such as gradient descent and proximal gradient descent methods. In the next few lectures, we will investigate dual and primal-dual methods for composite optimization problems. In this lecture, we will introduce the mirror descent method, which extends the gradient descent methods discussed in earlier lectures. This method serves as a bridge between primal and dual methods. We will then derive dual averaging method, which is a dual method, and we will present its analysis. The meaning of "dual" will become clear in the context of duality, which we will discuss in the next few lectures.

We consider the proximal gradient method in the last lecture, with L_1 regularization as example application. However, one drawback of proximal gradient method for L_1 regularization is that in its stochastic generalization [1], the solution produced by this method is not optimally sparse [4]. This problem can be fixed with regularized dual averaging [4], which may be considered as an extension of mirror descent, as well as the dual averaging method [3].

2 Generalized Proximal Mapping

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

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

and we assume this generalized proximal mapping can be efficiently computed. The proximal mapping discussed in previous lectures can be regarded as taking $h(z) = \|z\|_2^2/(2\eta)$.

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 any subgradient, depending on applications. The Bregman divergence of a convex function is always non-negative.

As an example, if

$$h(x) = \frac{1}{2}\|x\|_2^2,$$

then the Bregman divergence is

$$D_h(x, y) = \frac{1}{2}\|x - y\|_2^2.$$

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 $f(x)$.

We still refer to the standard smooth/strongly convex conditions discussed in previous lectures as smooth and strongly convex. In particular, if $f(x)$ is L-smooth, and $h(x)$ is 1-strongly convex, then $f(x)$ is also smooth with respect to $Lh(x)$.

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:

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

We know that $Q(y; y) = \phi(y)$. Therefore the minimizer $x^+ = \arg\min_x Q(x; y)$ can be written as

$$x^+ = \text{prox}_h(\nabla h(y) - \nabla f(y)).$$

This leads to the following general algorithm in Algorithm 1. By taking $h_t(x) = \frac{1}{2\eta_t}\|x\|_2^2$, we obtain the standard proximal gradient descent method.

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 $x_t = \text{prox}_{h_{t-1}}(\tilde{x}_t)$
Return: x_T

3 Mirror Descent

By taking $g(x) = 0$ for $x \in C$, and $h_t(x) = \eta_t^{-1} h(x)$, we obtain the mirror descent method in Algorithm 2.

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

The reason Algorithm 2 is referred to as mirror descent is that if x_t is in the interior of C, then we have the following symmetric form of update rule:

$$\nabla h(x_t) = \nabla h(x_{t-1}) - \eta_{t-1} \nabla f(x_{t-1}).$$

Example 1 *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 (unnormalized/normalized) exponentiated gradient methods.

4 Dual Averaging

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:

$$\nabla h_{t-1}(x_t) + \nabla g(x_t) = \nabla h_{t-1}(x_{t-1}) - \nabla f(x_{t-1}). \tag{2}$$

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

$$\eta_t h_t(x) = \eta_{t-1}[h_{t-1}(x) + g(x)], \tag{3}$$

then by solving the above recursion, we obtain

$$\eta_t h_t(x) = \eta_0 h_0(x) + \left(\sum_{s=0}^{t-1} \eta_s\right) g(x). \tag{4}$$

From (3) and (2), 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

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

By combining this with (4), 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.

In the algorithm, we use the notation $h(x) = \eta_0 h_0(x)$,

$$\tilde{p}_t = \eta_0 \nabla h_{t-1}(x_0) - \sum_{s=0}^{t-1} \eta_s \nabla f(x_s), \qquad \tilde{\eta}_t = \sum_{s=0}^{t-1} \eta_s.$$

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

We have the following convergence theorem.

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

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

In practice, RDA achieves better sparsity for L_1 regularization problems, because in the optimization problem defining x_t, the coefficients $\tilde{\eta}_t$ for $g(x)$ is the aggregated learning rate $\sum_{s<t} \eta_s$. In comparison, the effective coefficient is η_t for $g(x)$ in the proximal gradient method. Therefore RDA has a bigger shrinkage effect, leading to better sparsity. This point has been carefully discussed in [4]. However, empirically the convergence rate of the RDA method can be slower than that of the proximal gradient method. This can be explained from (4), where we use the Bregman divergence of a nonsmooth function $h_t(x)$ to upper bound the smoothness of $f(x)$. Since $f(x)$ is a smooth function, this upper bound is worse than the quadratic function in proximal gradient. For practical applications, this problem can be remedied using the so-called follow-the-regularized leader (FTRL) method [2]. FTRL has been widely used in the industry to solve large scale logistic regression problems.

5 Convergence Analysis of Algorithm 2

While the analysis leads to a result similar to that of Theorem 1, the analysis is different and conceptually simpler. It is similar to that of the proximal gradient method which we presented in Lecture 11.

Proposition 1 *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* $\forall x \in C$*:*

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

Proof Since by the definition of smoothness,

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

we have $\phi(x) = f(x) + g(x) \leq Q_t(x)$.

Moreover, we know that

$$Q_t(x) = f(x_{t-1}) - h_{t-1}(x_{t-1}) - \tilde{x}_t^\top (x - x_{t-1}) + h_{t-1}(x) + g(x).$$

Therefore by definition, the minimizer of $Q_t(x)$ is $x_t = \mathrm{prox}_{h_{t-1}}(\tilde{x}_t)$. It implies that $\exists \xi \in \partial Q_t(x)|_{x=x_t}$ such that $\xi^\top (x - x_t) \geq 0$ for all $x \in C$. Since $Q(x)$ is $\eta_t^{-1} + \lambda'$ strongly convex, we have

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

This proves the proposition. ■

Theorem 2 *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* $x \in C$*:*

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

Proof We have $h_t(x) = \eta^{-1} h(x)$.

$$\begin{aligned}\phi(x_t) \leq & Q_t(x_t) \leq Q_t(x) - D_{h_{t-1}}(x, x_t) \\ \leq & \phi(x) + D_{h_{t-1}}(x, x_{t-1}) - D_{h_{t-1}}(x, x_t).\end{aligned}$$

In the above derivation, the first two inequalities are due to Proposition 1. The third inequality is due to the convexity of $f(x)$.

By summing over from $t = 1$ to $t = T$, we have

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

This proves the result. ■

This result is similar to what we have obtained for proximal gradient methods.

6 Proof of Theorem 1

We may employ the estimate sequence method in Lecture 06 to prove the result. We consider

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

Obviously we have

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

We may define

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

We have

$$\phi_t(x) = (1 - \theta_t)\phi_{t-1}(x) + \theta_t \psi_t(x),$$

where $\theta_t = \eta_{t-1}/\tilde{\eta}_t$. The theorem follows from the following lemma and the theory of estimate sequence.

Lemma 1 *We have*

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

Proof According to our construction, we have

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

It is easy to check that $\phi(x_0) = \phi_0(x_0)$. Assume the result holds for $t-1$. Then

$$\begin{aligned}
\phi(x_t) \leq & f(x_{t-1}) + \nabla f(x_{t-1})^\top (x_t - x_{t-1}) + \frac{1}{\eta_{t-1}} D_h(x_t, x_{t-1}) + g(x_t) \\
= & \frac{\tilde{\eta}_t}{\eta_{t-1}} \phi_t(x_t) - \frac{\tilde{\eta}_{t-1}}{\eta_{t-1}} \phi_{t-1}(x_t) + \frac{1}{\eta_{t-1}} D_h(x_t, x_{t-1}) \\
\leq & \phi_t(x_t) + \frac{\tilde{\eta}_{t-1}}{\eta_{t-1}} \phi_t(x_{t-1}) - \frac{\tilde{\eta}_{t-1}}{\eta_{t-1}} \phi_{t-1}(x_t) + \frac{1}{\eta_t} D_h(x_t, x_{t-1}) \\
= & \phi_t(x_t) + \frac{\tilde{\eta}_{t-1}}{\tilde{\eta}_t} [-\phi_{t-1}(x_{t-1}) + \psi_t(x_{t-1})] - \frac{\tilde{\eta}_{t-1}}{\eta_{t-1}} [D_{\phi_{t-1}}(x_t, x_{t-1})] + \frac{1}{\eta_{t-1}} D_h(x_t, x_{t-1}) \leq \phi_t(x_t).
\end{aligned}$$

The first inequality is due to the smoothness of $f(x)$ with respect to $\eta_{t-1}^{-1} h$. The first equality uses the definition of $\phi_t(\cdot)$ and $\phi_{t-1}(\cdot)$. The second inequality is due to the assumption that x_t minimizes $\phi_t(x)$. The second equality uses

$$\begin{aligned}
\phi_t(x_{t-1}) = & \frac{\tilde{\eta}_{t-1}}{\tilde{\eta}_t} \phi_{t-1}(x_{t-1}) + \frac{\eta_{t-1}}{\tilde{\eta}_t} \psi_t(x_{t-1}), \\
\phi_{t-1}(x_t) = & \phi_{t-1}(x_{t-1}) + D_{\phi_{t-1}}(x_t, x_{t-1}),
\end{aligned}$$

and then simplify. The third inequality uses the induction hypothesis, which says that

$$-\phi_{t-1}(x_{t-1}) + \psi_t(x_{t-1}) = -\phi_{t-1}(x_{t-1}) + \phi(x_{t-1}) \leq 0,$$

and due to the fact that $\tilde{\eta}_{t-1}\phi_{t-1}(x) - h(x)$ is convex, so that $-\tilde{\eta}_{t-1} D_{\phi_{t-1}}(\cdot, \cdot) + D_h(\cdot, \cdot) \leq 0$. ■

References

[1] John Langford, Lihong Li, and Tong Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009.

[2] H Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In *AISTATS*, 2011.

[3] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Math. Program.*, 120(1):221259, April 2009.

[4] Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:25432596, December 2010.