ECE 490: Introduction to Optimization **Fall 2018**

Lecture 16

Nonsmooth Convex Optimization, Part II

Lecturer: Bin Hu, Date:10/25/2018

In this lecture, we will talk about three things: i) the update formula for the proximal operator calculation in ISTA; ii) the convergence rate analysis of the proximal gradient method using the dissipation inequality approach; iii) the projected gradient method which can be viewed as a special case of the proximal gradient method.

16.1 Shrinkage Operator for ISTA

In the last lecture, we have talked about ISTA. The proximal update $x_{k+1} = \text{prox}_{g,\alpha}(x_k - \alpha\nabla f(x_k))$ simply requires solving the following one-dimensional subproblem

$$x_{k+1}^j = \arg\min_{x^j \in \mathbb{R}} \left\{ \frac{1}{2\alpha}(x^j - h_k^j)^2 + \mu|x^j| \right\} \tag{16.1}$$

where h_k^j is just a scalar. This subproblem yields a simple solution. For simplicity, let's consider the minimization of $q(x) = \frac{1}{2\alpha}(x-h)^2 + \mu|x|$ where both x and h are scalars. We have

$$q(x) = \begin{cases} \frac{1}{2\alpha}(x^2 - 2hx + 2\mu\alpha x + h^2) & \text{if } x \geq 0 \\ \frac{1}{2\alpha}(x^2 - 2hx - 2\mu\alpha x + h^2) & \text{if } x < 0 \end{cases} \tag{16.2}$$

If $h \geq \mu\alpha$, then $q(x)$ achieves its minimum at $h - \mu\alpha \geq 0$. If $h \leq -\mu\alpha$, then $q(x)$ achieves its minimum at $h + \mu\alpha \leq 0$. If $-\mu\alpha < h < \mu\alpha$, then $q(x)$ achieves its minimum at 0. A graphical illustration is shown in Figure 16.1.

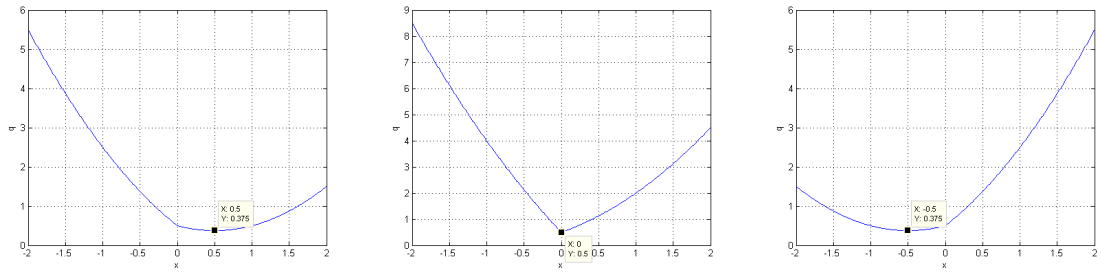

Figure 16.1. In the left plot, $h \geq \mu\alpha$ and $q(x)$ achieves its minimum at $h - \mu\alpha \geq 0$. In the right plot, $h \leq -\mu\alpha$, then $q(x)$ achieves its minimum at $h + \mu\alpha \leq 0$. The plot in the middle demonstrates the case where $-\mu\alpha < h < \mu\alpha$ and the minimum value is achieved at 0.

Consequently, (16.1) can be efficiently updated using the following so-called shrinkage operator:

$$x_{k+1}^j = \begin{cases} h_k^j - \mu\alpha & \text{if } h_k^j \geq \mu\alpha \\ 0 & \text{if } -\mu\alpha < h_k^j < \mu\alpha \\ h_k^j + \mu\alpha & \text{if } h_k^j \leq -\mu\alpha \end{cases} \tag{16.3}$$

The shrinkage operator just helps to sparsify the solution. Eventually a lot of h_k^j will end up as a value in the interval $[-\mu\alpha, \mu\alpha]$ and the solution for x will become sparse. The larger μ is, the bigger the interval $[-\mu\alpha, \mu\alpha]$ is and the more sparse the solution becomes.

ISTA provides an efficient method for ℓ_1-regularized problems. However, for more general problems, the proximal gradient method may not be that useful. In many situations, g can be complicated and it is difficult to calculate the proximal operator $\text{prox}_{g,\alpha}$. The proximal gradient method is only efficient when the subproblem $\arg\min_{x^j \in \mathbb{R}} \left\{ \frac{1}{2\alpha}\|x - h\|^2 + g(x) \right\}$ can be easily solved.

16.2 Convergence Rate Analysis of Proximal Gradient

Now we modify our dissipation inequality approach to analyze the convergence rate of the proximal gradient method. Recall that the proximal gradient method iterates as

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^p} \left\{ \frac{1}{2\alpha}\|x - x_k + \alpha\nabla f(x_k)\|^2 + g(x) \right\}$$

Therefore, we can rewrite the proximal gradient method as

$$x_{k+1} = x_k - \alpha u_k - \alpha r_k \tag{16.4}$$

where $u_k = \nabla f(x_k)$ and $r_k \in \partial g(x_{k+1})$. We emphasize that r_k is a subgradient of g evaluated at x_{k+1} (not x_k)! We can still apply the dissipation inequality approach to analyze this method. Specifically, we can just follow the three-step analysis routine presented in the previous lectures.

1. Replace $u_k = \nabla f(x_k)$ and $r_k \in \partial g(x_{k+1})$ with some quadratic inequalities in the following form:

$$\begin{bmatrix} x_k - x^* \\ \nabla f(x_k) - \nabla f(x^*) \\ r_k - r^* \end{bmatrix}^{\mathsf{T}} X_j \begin{bmatrix} x_k - x^* \\ \nabla f(x_k) - \nabla f(x^*) \\ r_k - r^* \end{bmatrix} \leq 0 \tag{16.5}$$

   where r^* is a subgradient of g evaluated at x^* and satisfying $r^* = -\nabla f(x^*)$. Recall that when we analyze the gradient method with L-smooth m-strongly convex f, we just replace $u_k = \nabla f(x_k)$ with the quadratic inequality:

$$\begin{bmatrix} x_k - x^* \\ \nabla f(x_k) - \nabla f(x^*) \end{bmatrix} \begin{bmatrix} 2mLI & -(L+m)I \\ -(L+m)I & 2I \end{bmatrix} \begin{bmatrix} x_k - x^* \\ \nabla f(x_k) - \nabla f(x^*) \end{bmatrix} \leq 0$$

which can be trivially lifted as

$$\begin{bmatrix} x_k - x^* \\ \nabla f(x_k) - \nabla f(x^*) \\ r_k - r^* \end{bmatrix}^\mathsf{T} \begin{bmatrix} 2mLI & -(L+m)I & 0 \\ -(L+m)I & 2I & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_k - x^* \\ \nabla f(x_k) - \nabla f(x^*) \\ r_k - r^* \end{bmatrix} \leq 0$$

Therefore, we can just choose $X_1 = \begin{bmatrix} 2mLI & -(L+m)I & 0 \\ -(L+m)I & 2I & 0 \\ 0 & 0 & 0 \end{bmatrix}$.

So far we have only applied the dissipation inequality to analyze iterations with the gradient evaluated at $C\xi_k$. Here r_k involves a subgradient evaluated at x_{k+1}. However, x_{k+1} is just a linear combinatin of x_k, $\nabla f(x_k)$, and r_k. We should be able to replace the relationship $r_k \in \partial g(x_{k+1})$ with some quadratic inequality in the form of (16.5). If we know g is convex, we directly have $(r_k - r^*)^\mathsf{T}(x_{k+1} - x^*) \geq 0$. This can be rewritten as a quadratic inequality:

$$\begin{bmatrix} x_{k+1} - x^* \\ r_k - r^* \end{bmatrix} \begin{bmatrix} 0 & -I \\ -I & 0 \end{bmatrix} \begin{bmatrix} x_{k+1} - x^* \\ r_k - r^* \end{bmatrix} \leq 0$$

Based on $x^*_{k+1} - x^* = x_k - x^* - \alpha(\nabla f(x_k) - \nabla f(x^*)) - \alpha(r_k - r^*)$, we can rewrite the above quadratic inequality as

$$\begin{bmatrix} x_k - x^* \\ \nabla f(x_k) - \nabla f(x^*) \\ r_k - r^* \end{bmatrix}^\mathsf{T} \begin{bmatrix} I & 0 \\ -\alpha I & 0 \\ -\alpha I & I \end{bmatrix} \begin{bmatrix} 0 & -I \\ -I & 0 \end{bmatrix} \begin{bmatrix} I & -\alpha I & -\alpha I \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} x_k - x^* \\ \nabla f(x_k) - \nabla f(x^*) \\ r_k - r^* \end{bmatrix} \leq 0$$

Therefore, we can just choose X_2 as

$$X_2 = \begin{bmatrix} I & 0 \\ -\alpha I & 0 \\ -\alpha I & I \end{bmatrix} \begin{bmatrix} 0 & -I \\ -I & 0 \end{bmatrix} \begin{bmatrix} I & -\alpha I & -\alpha I \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} 0 & 0 & -I \\ 0 & 0 & \alpha I \\ -I & \alpha I & 2\alpha I \end{bmatrix}$$

and (16.5) is satisfied.

2. With our choices of X_1 and X_2, we can apply the dissipation inequality approach. If there exists λ_1 and λ_2 such that

$$\begin{bmatrix} 1-\rho^2 & -\alpha & -\alpha \\ -\alpha & \alpha^2 & \alpha^2 \\ -\alpha & \alpha^2 & \alpha^2 \end{bmatrix} \leq \lambda_1 \begin{bmatrix} 2mL & -(L+m) & 0 \\ -(m+L) & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & \alpha \\ -1 & \alpha & 2\alpha \end{bmatrix} \tag{16.6}$$

then we have a dissipation inequality $\|x_{k+1}-x^*\|^2-\rho^2\|x_k-x^*\|^2 \leq S$ where S is given by

$$\begin{aligned} S =&\lambda_1 \begin{bmatrix} x_k-x^* \\ \nabla f(x_k)-\nabla f(x^*) \\ r_k-r^* \end{bmatrix}^{\mathsf{T}} X_1 \begin{bmatrix} x_k-x^* \\ \nabla f(x_k)-\nabla f(x^*) \\ r_k-r^* \end{bmatrix} \\ &+\lambda_2 \begin{bmatrix} x_k-x^* \\ \nabla f(x_k)-\nabla f(x^*) \\ r_k-r^* \end{bmatrix}^{\mathsf{T}} X_2 \begin{bmatrix} x_k-x^* \\ \nabla f(x_k)-\nabla f(x^*) \\ r_k-r^* \end{bmatrix} \end{aligned}$$

The key equation we use to formulate (16.6) is

$$\|x_{k+1}-x^*\|^2 = \begin{bmatrix} x_k-x^* \\ \nabla f(x_k)-\nabla f(x^*) \\ r_k-r^* \end{bmatrix}^{\mathsf{T}} \left(\begin{bmatrix} 1 & -\alpha & -\alpha \\ -\alpha & \alpha^2 & \alpha^2 \\ -\alpha & \alpha^2 & \alpha^2 \end{bmatrix} \otimes I \right) \begin{bmatrix} x_k-x^* \\ \nabla f(x_k)-\nabla f(x^*) \\ r_k-r^* \end{bmatrix}$$

3. Once we have the dissipation inequality $\|x_{k+1}-x^*\|^2-\rho^2\|x_k-x^*\|^2 \leq S$, we can immediately show the linear convergence $\|x_{k+1}-x^*\|^2 \leq \rho^2\|x_k-x^*\|^2$ using (16.5) and the non-negativity of (λ_1,λ_2).

Therefore, we only need to find λ_1 and λ_2 such that (16.6) holds. If we choose $\lambda_2=\alpha$, (16.6) becomes

$$\begin{bmatrix} 1-\rho^2 & -\alpha & 0 \\ -\alpha & \alpha^2 & 0 \\ 0 & 0 & -\alpha^2 \end{bmatrix} \leq \lambda_1 \begin{bmatrix} 2mL & -(L+m) & 0 \\ -(m+L) & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

To ensure the above inequality holds, we only need

$$\begin{bmatrix} 1-\rho^2 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} \leq \lambda_1 \begin{bmatrix} 2mL & -(L+m) \\ -(m+L) & 2 \end{bmatrix}$$

which is exactly the testing condition for the gradient method. So we can obtain the same convergence rate bounds for the proximal gradient method by solving the above testing condition.

We see that the dissipation inequality approach is general enough to handle the proximal operator. Consider FISTA that iterates as

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_{g,\alpha}(y_k-\alpha\nabla f(y_k)) \\ y_k &= (1+\beta)x_k-\beta x_{k-1} \end{aligned}$$

The above iteration can be rewritten as

$$\begin{aligned} x_{k+1} &= y_k-\alpha\nabla f(y_k)-\alpha r_k \\ y_k &= (1+\beta)x_k-\beta x_{k-1} \end{aligned}$$

where $r_k \in \partial g(x_{k+1})$. Similarly we can replace $r_k \in \partial g(x_{k+1})$ with some quadratic supply rate condition and then perform the rate analysis. It is worth mentioning that proving convergence in $V(\xi_k) + f(x_k) + g(x_k) - f(x^*) - f(x^*)$ requires some extra technical details. We omit these details here.

16.3 Projected Gradient Method

Proximal gradient method can also be used to solve some relatively simple constrained optimization problem.

Consider the constrained optimization

$$\min_{x \in X} f(x)$$

where the feasible set X is a convex set. We can reformulate the above problem as an unconstrained optimization problem $\min_{x \in \mathbb{R}^p}\{f(x) + g(x)\}$ where $g(x)$ is an indicator function:

$$g(x) = \begin{cases} 0 & \text{if } x \in X \\ +\infty & \text{if } x \notin X \end{cases} \tag{16.7}$$

Clearly, for $x \in X$, we have $f(x) + g(x) = f(x)$. It is straightforward to verify the two formulations are equivalent (verify this yourself!). Now we can apply the proximal gradient method $x_{k+1} = \text{prox}_{g,\alpha}(x_k - \alpha \nabla f(x_k))$ to solve the original constrained optimization problem. The proximal update is

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^p} \left\{ \frac{1}{2\alpha} \|x - h_k\|^2 + g(x) \right\}$$

where $h_k = x_k - \alpha \nabla f(x_k)$. By definition, one can show the above update is equivalent to

$$x_{k+1} = \arg\min_{x \in X} \|x - h_k\|^2 \tag{16.8}$$

The above operation is called projection. Therefore, in this case, the proximal gradient method just becomes the projected gradient method. Sometimes when X is simple and nice, the projection can be easily computed and the projected gradient method is very efficient. For example, consider $X = \{x : x^i \geq 0, \forall 1 \leq i \leq p\}$. Then (16.8) can be updated using the simple analytical formula

$$x_{k+1}^i = \begin{cases} h_k^i & \text{if } h_k^i \geq 0 \\ 0 & \text{if } h_k^i < 0 \end{cases} \tag{16.9}$$

Similarly, other box constraints can also be easily handled by the projected gradient method. When X is more complicated, the projected gradient method may not be that useful.