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

# Lecture 15

## Nonsmooth Convex Optimization, Part I

*Lecturer: Bin Hu, Date:10/23/2018*

In this lecture, we first briefly talk the convergence guarantees for the subgradient method. Next we will talk about a special class of optimization problems where objective functions are a sum of a convex smooth differentiable function and a convex non-differentiable function. This type of problems has a special sum structure and may be solved more efficiently using the proximal operator. Then we discuss the proximal gradient method. Finally, we talk about Iterative Shrinkage-Thresholding Algorithm (ISTA) which is just the the proximal gradient method applied to LASSO.

## 15.1 Properties of Subgradient Method

When the objective function $f$ is convex but non-differentiable, we can still use the subgradient method:

$$x_{k+1} = x_k - \alpha g_k$$

where $g_k \in \partial f(x_k)$. This method is quite slow and may oscillate around the optimal point. Now we present the convergence analysis for the subgradient method using the dissipation inequality approach covered in the previous lectures. Notice $f$ is convex. We have $-g_k^{\mathsf{T}}(x_k - x^*) \leq f(x^*) - f(x_k)$. This can be written as a quadratic supply rate condition:

$$\begin{bmatrix} x_k - x^* \\ g_k \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} 0 & -\frac{1}{2}I \\ -\frac{1}{2}I & 0 \end{bmatrix} \begin{bmatrix} x_k - x^* \\ g_k \end{bmatrix} \leq f(x^*) - f(x_k) \tag{15.1}$$

We further assume $\|g_k\| \leq G$. This is equivalent to another quadratic supply rate condition:

$$\begin{bmatrix} x_k - x^* \\ g_k \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x_k - x^* \\ g_k \end{bmatrix} \leq G^2 \tag{15.2}$$

If $\exists$ positive $\lambda_1$ and $\lambda_2$ such that

$$\begin{bmatrix} 0 & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} \leq \lambda_1 \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \tag{15.3}$$

then we have

$$\min_{0 \leq k \leq T} (f(x_k) - f(x^*)) \leq \frac{\|x_0 - x^*\|^2}{\lambda_1 (T+1)} + \frac{\lambda_2 G^2}{\lambda_1}. \tag{15.4}$$

To see why this is true, notice that (15.3) directly leads to

$$\begin{bmatrix} x_k - x^* \\ g_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} 0 & -\alpha I \\ -\alpha I & \alpha^2 I \end{bmatrix} \begin{bmatrix} x_k - x^* \\ g_k \end{bmatrix} \leq$$
$$\lambda_1 \begin{bmatrix} x_k - x^* \\ g_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} 0 & -\frac{1}{2} I \\ -\frac{1}{2} I & 0 \end{bmatrix} \begin{bmatrix} x_k - x^* \\ g_k \end{bmatrix} + \lambda_2 \begin{bmatrix} x_k - x^* \\ g_k \end{bmatrix}^\mathsf{T} \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x_k - x^* \\ g_k \end{bmatrix}.$$

Applying (15.1), (15.2), and the non-negativity of $(\lambda_1, \lambda_2)$, we immediately obtain the following inequality

$$\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2 \leq \lambda_1 (f(x^*) - f(x_k)) + \lambda_2 G^2$$

Summing the above inequality over $k$ leads to

$$\lambda_1 \sum_{k=0}^{T} (f(x_k) - f(x^*)) \leq \|x_0 - x^*\|^2 + \lambda_2 (T+1) G^2$$

which can be used to prove (15.4). Therefore, the behavior of the subgradient method is captured by the testing condition (15.3). Now if we choose $\lambda_1 = 2\alpha$ and $\lambda_2 = \alpha^2$, (15.3) is satisfied and the subgradient method satisfies the following bound

$$\min_{0 \leq k \leq T} f(x_k) - f(x^*) \leq \frac{\|x_0 - x^*\|^2}{2\alpha(T+1)} + \frac{\alpha G^2}{2}. \tag{15.5}$$

There is a trade-off between the convergence rate and the final error. If we choose a large stepsize $\alpha$, the first term on the right side vanishes faster but the second term (the final error) is large. If we choose small $\alpha$ to control the final error term, then the first term vanishes at a slower speed. For a fixed $T$, we can minimize the right side of the above bound by choosing $\alpha = \frac{\|x_0 - x^*\|}{G\sqrt{T+1}}$. Consequently, the right side of the above bound becomes $\frac{G\|x_0 - x^*\|}{\sqrt{T+1}}$. This is a slow sublinear rate. Also notice that the subgradient method is not a descent method. For example, consider $f(x) = |x|$. The subgradient method will start to oscillate when approaching the optimal point $x^* = 0$. We can only guarantee the sublinear convergence of $\min_{0 \leq k \leq T} f(x_k)$. We can store this value in memory and at very step $T$ we need to update this value by comparing $\min_{0 \leq k \leq T-1} f(x_k)$ with $f(x_T)$. Overall the convergence guarantees of the subgradient method is weak. More efficient methods are available for nonsmooth convex optimization problems when the objective function has additional useful structures.

## 15.2 Proximal Gradient Method

In many applications, the objective function is a sum of two terms:

$$f(x) + g(x) \tag{15.6}$$

where $f$ is a smooth convex differentiable function and $g$ is a non-differentiable convex function. The proximal gradient method is developed for such problems. The proximal gradient method can be viewed as a direct extension of the gradient method. If both $f$ and $g$ are differentiable, then the gradient method can be directly applied and iterates as follows

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

Basically, we just make a first-order Taylor expansion of $f+g$ around $x_k$ and add a quadratic regularization term. When $g$ is not differentiable, it does not make sense to approximate $g$ with its Taylor expansion. We can just expand the differentiable part $f$ and keep $g$ as it is. This leads to the proximal gradient method which iterates as

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^p} \left\{ f(x_k) + \nabla f(x_k)^\mathsf{T}(x - x_k) + \frac{1}{2\alpha}\|x - x_k\|^2 + g(x) \right\} \tag{15.7}$$

At every step $k$, one needs to solve the above subproblem. When $g = 0$, the subproblem at every step is just a positive definite quadratic minimization problem that yields simple analytical solution $x_{k+1} = x_k - \alpha\nabla f(x_k)$. For the proximal gradient method, the complexity of the subproblem really depends on $g$. When $g$ is simple, e.g. $\ell_1$-regularizaer, the subproblem yields simple analytical solution and the proximal gradient method becomes efficient. For general $g$, the subproblem can be difficult and the proximal gradient method may not be applicable. Therefore, the proximal gradient method is less general than the subgradient method but can be more efficient for certain applications when (15.7) yields simple solutions.

To see why (15.7) is called the "proximal gradient" method, just notice that (15.7) can be equivalently rewritten as

$$x_{k+1} = \text{prox}_{g,\alpha}(x_k - \alpha\nabla f(x_k)) \tag{15.8}$$

where the proximal operator is defined as $\text{prox}_{g,\alpha}(x) = \arg\min_{y\in\mathbb{R}^p} \left\{ \frac{1}{2\alpha}\|y - x\|^2 + g(y) \right\}$. By definition, (15.8) is equivalent to

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^p} \left\{ \frac{1}{2\alpha}\|x - x_k + \alpha\nabla f(x_k)\|^2 + g(x) \right\} \tag{15.9}$$

One can verify (15.9) and (15.7) are the same.

If one can perform the proximal operator calculation at low cost, then one can apply the proximal gradient method efficiently and obtain some iteration complexity guarantees similar to the gradient method. We will talk about the convergence rate of the proximal gradient method in the next lecture. Now we present one important application where the proximal gradient method is useful.

## 15.3 Sparsity-Induced Optimization: LASSO and ISTA

Here we present least absolute shrinkage and selection operator (LASSO) as one important example for (15.6). LASSO involves the following objective function

$$\frac{1}{n}\sum_{i=1}^{n}(a_i^{\mathsf{T}}x - b_i)^2 + \mu\|x\|_1 \tag{15.10}$$

where $\|\cdot\|_1$ denotes the $\ell_1$-norm. Specifically, suppose $x \in \mathbb{R}^p$ and

$$x = \begin{bmatrix} x^1 \\ x^2 \\ \vdots \\ x^p \end{bmatrix},$$

one has $\|x\|_1 = |x^1| + |x^2| + \ldots + |x^n|$. The $\ell_1$ regularizar is used to induce sparsity. The optimal solution for (15.10) will have a lot of zero entries due to the use of the term $\|x\|_1$. In LASSO, one perform the regression analysis and the feature selection simultaneously.

When applied to (15.10), the proximal gradient method becomes the so-called Iterative Shrinkage-Thresholding Algorithm (ISTA). Based on (15.9), ISTA iterates as follows

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^p}\left\{\frac{1}{2\alpha}\|x - h_k\|^2 + \mu\|x\|_1\right\} = \arg\min_{x\in\mathbb{R}^p}\sum_{j=1}^{p}\left(\frac{1}{2\alpha}(x^j - h_k^j)^2 + \mu|x^j|\right) \tag{15.11}$$

where $h_k = x_k - \alpha\nabla f(x_k)$ is just some vector. On the right side, the only term that involves $x^j$ is $\frac{1}{2\alpha}(x^j - h_k^j)^2 + \mu|x^j|$. So there is no coupling between different entries of $x_{k+1}$ and we have

$$x_{k+1}^j = \arg\min_{x^j\in\mathbb{R}}\left(\frac{1}{2\alpha}(x^j - h_k^j)^2 + \mu|x^j|\right) \tag{15.12}$$

where $x^j$ and $h_k^j$ are both scalars. The above subproblem yields a simple analytical solution that can be calculated by the shrinkage operator. This is how ISTA gets its name. Due to the simplicity of (15.12), we can solve LASSO efficiently using ISTA. Actually similar ideas can be applied to any convex $\ell_1$-regularized problems. We will talk about the solution for (15.12) in next lecture.

One can also incorporate momentum terms into the proximal gradient method. For example, if one combines Nesterov's accelerate method with ISTA, one obtain a new algorithm called FISTA (Fast Iterative Shrinkage-Thresholding Algorithm). One can also combine Heavy-ball method with the proximal gradient method and the resultant method is called iPiano.