10-725: Optimization **Fall 2012**

Lecture 8: September 20

Lecturer: Geoff Gordon/Ryan Tibshirani *Scribes: Avinava Dubey*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

8.1 Review of Subgradient Method

One of the methods to optimize a convex but not necessarily differentiable function is to use subgradient methods. Thus the objective is to

$$\min_{x \in \mathbb{R}^n} f(x) \tag{8.1}$$

where $f(x)$ is convex, but may not be differentiable. The update step after choosing an initial point in $x^{(0)} \in \mathbb{R}^n$ is

$$x^{(k)} = x^{(k-1)} - t_k . g^{(k-1)}, \ k = 1, 2, 3, \ldots \tag{8.2}$$

where $g^{(k-1)}$ is a subgradient of the function f at $x^{(k-1)}$. This is similar to gradient descent step with the gradient replaced by sub-gradient.

A major problem with subgradient methods is the rate of convergence which is $O(1/\sqrt{k})$ when the function is Lipschitz on a bounded set containing its minimizer. A bit of the structure of the objective function can be used to get a better convergence rate. In generalized gradient decent the knowledge of certain form of the objective will lead to a faster convergence rate.

8.2 Generalized Gradient Descent

Suppose the objective function $f(x)$ has the following form

$$f(x) = g(x) + h(x) \tag{8.3}$$

such that $g(x)$ is convex and differentiable while $h(x)$ is only convex but not necessarily differentiable. When the function $f(x)$ was differentiable we made quadratic approximation to f around x to get gradient descent update. The update step is of the form $x^+ = x - t\nabla f(x)$ where x^+ is the next value of x and t is the step size. To recap the quadratic approximation around x is

$$f(x) + \nabla f(x)^T (z - x) + \frac{1}{2}(z - x)^T \nabla^2 f(x)(z - x) \tag{8.4}$$

To get to the gradient update step replace the $\nabla^2 f(x)$ by $\frac{1}{t} I$ leading to the following approximation

$$x^+ = \arg\min_z f(x) + \nabla f(x)^T (z - x) + \frac{1}{2t}\|z - x\|^2 \tag{8.5}$$

In the case when f is not differentiable but can be decomposed into sum of a convex differentiable function (g) and a convex non-differentiable function (h) then g can still be approximated by a quadratic approximation while h can be used as it is to get the following:-

$$\begin{aligned} x^+ &= \arg\min_z g(x) + \nabla g(x)^T(z-x) + \frac{1}{2t}\|z-x\|^2 + h(z) \\ &= \arg\min_z \frac{1}{2t}(\|z-x\|^2 + 2t\nabla g(x)^T(z-x) + t^2\|\nabla g(x)\|^2) + g(x) - \frac{2}{t}\|\nabla g(x)\|^2 + h(z) \\ &= \arg\min_z \frac{1}{2t}\|z-(x-t\nabla g(x))\|^2 + h(z) \quad \text{removing terms independent of } z \end{aligned} \tag{8.6}$$

The effect of the above update scheme is two fold. The first term keeps the update as close to the gradient update for g while the second term makes h small.

The above intuition leads to the following method for solving functions represented in 8.3.

1. Initialize $x^{(0)} \in \mathbb{R}^n$.
2. Let $x^+ = x^{(k-1)} - t_k \nabla g(x^{(k-1)})$
3. Define $\text{prox}_t(x) = \arg\min_{z\in\mathbb{R}^n} \frac{1}{2t}\|x-z\|^2 + h(z)$. Then
$$x^{(k)} = \text{prox}_{t_k}(x^+)$$

The update step can also be made to look similar to the update step of gradient descent by writing it as

$$x^{(k)} = x^{(k-1)} - t_k G_{t_k}(x^{(k-1)}) \tag{8.7}$$

where $G_t(x)$ is the generalized gradient and is given by

$$G_t(x) = \frac{x - \text{prox}_t(x - t\nabla g(x))}{t}$$

Discussions

1. *One optimization step has been replaced by another. So does this help?* Yes since for a number of important functions h the prox_t can be computed analytically.
2. *What if g is complicated?* Note that the prox function does not depend on the function g. For the update only the value of the gradient of g is required.

8.2.1 ISTA

Lets consider a case when $h(x)$ is $L1$ penalty on x ie $h(x) = \lambda\|x\|_1$ where $\lambda \geq 0$. Then the prox function is defined as

$$\begin{aligned} \text{prox}_t(x) &= \arg\min_{z\in\mathbb{R}^n} \frac{1}{2t}\|x-z\|^2 + \lambda\|z\|_1 \\ &= S_{\lambda t}(x) \qquad \text{refer to lecture 7 slide 16} \end{aligned} \tag{8.8}$$

where $S_{\lambda t}(x)$ is the soft-thresholding operator.

$$[S_\lambda(x)]_i = \begin{cases} x_i - \lambda & \text{if } x_i > \lambda \\ 0 & \text{if } -\lambda \leq x_i \geq \lambda \\ x_i + \lambda & \text{if } x_i < -\lambda \end{cases}$$

8.2.1.1 LASSO

In lasso the objective function can be written as

$$\begin{aligned} f(x) &= \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1 \\ &= g(x) + h(x) \end{aligned} \tag{8.9}$$

Where $g(x) = \frac{1}{2}\|y - Ax\|^2$. Thus $\nabla g(x) = -A^T(y - Ax)$. Hence the generalized gradient update is

$$x^+ = S_{\lambda t}(x + tA^T(y - Ax)) \tag{8.10}$$

The resulting algorithm is faster than sub-gradient method as shown in the figure 8.1.

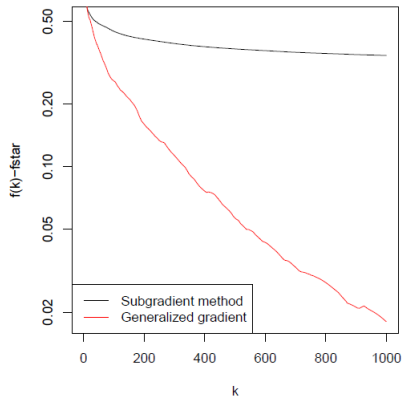

Figure 8.1: ISTA vs subgradient decent

8.2.2 Convergencce Analysis

For $f(x) = g(x) + h(x)$, lets assume that

- g is convex, differentiable, ∇g is Lipshitz continuous with constant $L > 0$.
- h is convex, $\text{prox}_t(x) = \arg\min_z\{\frac{\|x-z\|^2}{2t} + h(z)\}$ can be evaluated

Then the following holds

Theorem 8.1 *Generalized gradient descent with fixed step size $t \leq \frac{1}{L}$ satisfies*

$$f(x^{(k)}) - f(x^*) \leq \frac{\|x^{(0)} - x^*\|^2}{2tk}$$

where x^* is the optimal solution.

Discussions

1. This has the same convergence rate $O(1/k)$ as that of gradient descent but *this counts the number of iteration but not the number of operations.*
2. Why does generalized gradient descent converge at $O(1/k)$ where as subgradient converges at $O(1/\sqrt{k})$. It is because of the additional knowledge about ∇g being Lipschitz continuous.

Proof: We begin by first showing that

$$f(y) \leq g(x) + \nabla g(x)^T(y-x) + \frac{L}{2}\|y-x\|^2 + h(y) \quad \forall x, y$$

Since ∇g is Lipschitz with constant L, $\nabla^2 g \preceq LI$, we have $\forall x, y$

$$(x-y)^T(\nabla^2 g(x) - LI)(x-y) \leq 0$$

which implies that

$$L\|x-y\|^2 \geq (x-y)^T \nabla^2 g(x)(x-y)$$

By Taylor's expansion

$$\begin{aligned} g(y) &= g(x) + \nabla g(x)^T(y-x) + \frac{1}{2}(x-y)^T\nabla^2 g(x)(x-y) \\ &\leq g(x) + \nabla g(x)^T(y-x) + \frac{L}{2}\|y-x\|^2 \end{aligned}$$

therefore

$$f(y) \leq g(x) + \nabla g(x)^T(y-x) + \frac{L}{2}\|y-x\|^2 + h(y) \quad \forall x, y \tag{8.11}$$

substituting $y = x^+ = x - tG_t(x)$ we have

$$f(x^+) \leq g(x) - t\nabla g(x)^T G_t(x) + \frac{Lt}{2}\|G_t\|^2 + h(x - tG_t(x)) \tag{8.12}$$

Now

$$\begin{aligned} x - tG_t(x) &= \arg\min_z \frac{1}{2t}\|z - (x - t\nabla g(x))\|^2 + h(z) \\ &= \arg\min_z \nabla g(x)^T(z-x) + \frac{1}{2t}\|z-x\|^2 + h(z) \end{aligned}$$

This implies that there exist a $v \in \delta h(z)$ such that at minima $\nabla g(x) + \frac{1}{t}(z-x) + v = 0$, but the minimum occurs at $z = x - tG_t(x)$, thus

$$\begin{aligned} &\nabla g(x) - G_t(x) + v = 0, \quad v \in \delta h(x - tG_t(x)) \\ &\Rightarrow G_t(x) - \nabla g(x) \in \delta h(x - tG_t(x)) \end{aligned} \tag{8.13}$$

Since h is convex,

$$\begin{aligned} &h(x) \geq h(x - tG_t(x)) + (G_t(x) - \nabla g(x))^T tG_t(x) \\ &h(x - tG_t(x)) \leq h(x) - t(G_t(x) - \nabla g(x))^T G_t(x) \end{aligned} \tag{8.14}$$

Substituting equation 8.14 in 8.12 we get

$$f(x^+) \leq f(x) - (1 - \frac{Lt}{2})t\|G_t(x)\|^2 \tag{8.15}$$

Since f is convex, $f(x) \leq f(x^*) + G_t(x)^T(x - x^*)$, substituting it above we get

$$\begin{aligned} f(x^+) &\leq f(x^*) + G_t(x)^T(x - x^*) - (1 - \frac{Lt}{2})t\|G_t(x)\|^2 \\ &\leq f(x^*) + G_t(x)^T(x - x^*) - \frac{t}{2}\|G_t(x)\|^2 \quad \text{since } t < 1/L \\ &= f(x^*) + \frac{1}{2t}(\|x - x^*\|^2 - \|x^+ - x^*\|^2) \end{aligned} \tag{8.16}$$

Summing over iterations we have

$$\begin{aligned} \sum_{i=1}^{k}(f(x^{(i)} - f(x^*)) &\leq \frac{1}{2t}(\|x^{(0)} - x^*\|^2 - \|x^{(k)} - x^*\|^2) \\ &\leq \frac{1}{2t}\|x^{(0)} - x^*\|^2 \end{aligned} \tag{8.17}$$

Since the difference is non-increasing we have

$$f(x^{(k)}) - f(x^*) \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)} - f(x^*)) \leq \frac{1}{2tk}\|x^{(0)} - x^*\|^2 \tag{8.18}$$

■

8.2.3 Backtracking Line Search

The procedure for back tracking is similar to gradient descent:

- Fix $0 < \beta < 1$
- Then at each iteration start with $t = 1$, and while

  $$f(x - tG_t(x)) \geq f(x) - \frac{t}{2}\|G_t(x)\|^2$$

  update $t = \beta t$

Theorem 8.2 *Generalized gradient descent with backtracking line search satisfies*

$$f(x^{(k)}) - f(x^*) \leq \frac{\|x^{(0)} - x^*\|^2}{2t_{min}k}$$

where $t_{min} = \min\{1, \beta/L\}$

8.2.4 Example: Matrix Completion

Given matrix A, $m \times n$, only some entries are observed $A_{ij},\ (i,j) \in \Omega$, the objective is to fill in the missing entries. This can be used to predict user preferences such as user rating for unseen movies. The objective is to

$$\min_{X \in \mathbb{R}^{m \times m}} \frac{1}{2}\sum_{(i,j)\in\Omega}(A_{ij} - X_{ij})^2 + \lambda\|X\|_* \tag{8.19}$$

where $\|X\|_*$ is the nuclear norm of X and is given by

$$\|X\|_* = \sum_{i=1}^{r} \sigma_i(X)$$

where $r = rank(X)$ and σ_i is the ith singular value.

To solve this first lets define a projection operator onto the observed set

$$[P_\Omega(X)]_{ij} = \begin{cases} X_{ij} & (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

The objective is to minimize the following function

$$f(X) = \frac{1}{2}\|P_\Omega(A) - P_\Omega(X)\|_F^2 + \lambda\|X\|_*$$

which has the same form as $g(x) + h(x)$. Now projection function is convex and the Frobenius norm is differentiable and convex and the nuclear norm is convex but not differentiable. Here we can apply generalized gradient descent. The gradient is $\nabla g(X) = -(P_\Omega(A) - P_\Omega(X))$ and the prox function is

$$\text{prox}_t(X) = \arg\min_{Z \in \mathbb{R}^{m\times n}} \frac{1}{2t}\|X - Z\|_F^2 + \lambda\|Z\|_*$$

If we select $\text{prox}_t(X) = Z$ then Z should satisfy

$$0 \in Z - X + \lambda t.\delta\|Z\|_* \tag{8.20}$$

Note that if $Z = U\Sigma V^T$ then

$$\delta\|Z\|_* = \{UV^T + W : W \in \mathbb{R}^{m\times n}, \|W\| \leq 1, U^TW = 0, WV = 0\}$$

If we let $Z = U\Sigma_\lambda V^T$ then equation 8.20 holds, here $X = U\Sigma V^T$ is the SVD and Σ_λ is given by the diagonal matrix

$$(\Sigma_\lambda)_{ii} = \max\{\Sigma_{ii} - \lambda, 0\}$$

This is true because $X - Z = \lambda UV^T \in \delta\|Z\|_*$. Thus the prox function can be written as

$$\text{prox}_t(X) = S_{\lambda t}(X) = U\Sigma_\lambda V^T$$

and the generalized gradient update step is

$$X^+ = S_{\lambda t}(X + t(P_\Omega(A) - P_\Omega(X)))$$

since $\|\nabla g(Y) - \nabla g(X)\|_F = \|P_\Omega(A) - P_\Omega(X)\|_F \leq \|Y - X\|_F$ the Lipschitz constant is $L = 1$ thus the step size can be picked as $t = 1$ leading to

$$X^+ = S_\lambda(P_\Omega(A) + X - P_\Omega(X)) = S_\lambda(P_\Omega(A) + P_\Omega^\perp(X))$$

where $P_\Omega(X) + P_\Omega^\perp(X) = X$ This is called the **soft-impute** algorithm for matrix completion.

Discussions

1. *Why is this method called "generalized gradient descent"?* Note that the function to be minimized is $f = g + h$ and if we let $h = 0$ its equivalent to gradient descent, if we set $g = 0$ its proximal minimization algorithm and if we set $h = I_C$ where I_C denotes projection onto a set then it becomes projected gradient descent.
2. Since all the three methods are specific cases of this they have $O(1/k)$ convergence rate.

8.2.5 Projected Gradient Descent

The problem of minimizing over a closed convex ($min_{x \in C} g(x)$) set can be shown to be equivalent to $min_x g(x) + I_C(x)$ where

$$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

The prox function for this problem is

$$\begin{aligned} prox_t(x) &= \arg\min_z \frac{1}{2t}\|x - z\|^2 + I_C(z) \\ &= \arg\min_{z \in C} \|x - z\|^2 \\ &= P_C(x) \quad \text{projection operator on } C \end{aligned} \tag{8.21}$$

Thus the update step is

$$x^+ = P_C(x - t\nabla(x))$$

which implies that it first performs the usual gradient update and then its projected onto the constrained set as shown in the figure 8.2 below:-

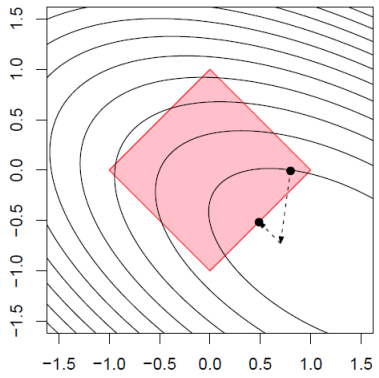

Figure 8.2: The gradient step may take the solution out of the set C shown shaded. It is then projected back on the set.

Some sets that are easy to project on are:-

- Affine images $C = \{Ax + b : x \in \mathbb{R}^n\}$
- Solution set of linear system $C = \{x \in \mathbb{R}^n : Ax = b\}$
- Nonnegative orthant $C = \{x \; in \mathbb{R}^n : x \succeq 0\}$
- Norm balls $C = \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$, for $p = 1, 2, \infty$

Note that it may be hard to project on arbitrary polyhedron.

8.2.6 Proximal Minimization

If $g = 0$ then the objective is just

$$\min_x h(x)$$

and can be solved by using just the prox update

$$x^+ = \arg\min_z \frac{1}{2t}\|x - z\|^2 + h(z)$$

Note that if the prox function can not be solved in closed form then this method is not implementable.

Discussion

1. Generalized gradient descent assumes that the prox can be evaluated exactly.
2. There is in general no convergence guarantee if the prox fucntion cannot be calculated exactly.
3. There are some exceptions to this(eg. partial proximal minimization)

8.3 To be Covered in Next Lecture

In the next lecture we would deal with cutting edge first order methods that can achieve an optimal rate of $O(1/k^2)$

References

[LSU11] E. Candes, "Lecture Notes for Math 301," *Stanford University*, Winter 2010-2011.

[LSU11] L. Vandenberghe, "Lecture Notes for EE 236C," *UCLA*, Spring 2010-2011.