EE 381V: Large Scale Optimization **Fall 2014**

Lecture 6 — September 16

Lecturer: Sujay Sanghavi *Scribe: En-Hsu Yen & Yitao Chen & Louie Wu*

6.1 Topics covered

- Coordinate Descent Method
- Steepest Descent Method

6.2 Coordinate Descent Method

For problem of huge number of variables d, the evaluation of full gradient vector $\nabla f(x)$ can be very expensive even for single iteration. This motivates a family of algorithms that minimizes a single variable at a time, termed *Coordinate Descent Methods*. This type of algorithms is applicable to objective function of the form:

$$\min_x \quad f(x) = g(x) + h(x) \tag{6.1}$$

, where $g(x)$ is a convex, differentiable function ($\nabla g(x)$ is Lipschitz-continuous), and $h(x)$ is a convex, *separable* function. A separable function $h(x)$ can be written as $h(x) = \sum_{i=1}^{d} h_i(x)$, such as ℓ_1-norm $\|x\|_1 = \sum_{i=1}^{d} |x_i|$ or box constraints

$$h_i(x) = \begin{cases} 0, & L \leq x_i \leq U \\ \infty, & o.w. \end{cases} \tag{6.2}$$

For problem with *non-separable, non-smooth* function, coordinate descent algorithm does not guarantee convergence to optimum. Figure 6.1 gives an example of non-smooth, non-separable function $f(x_1, x_2) = \max(x_1, x_2)$, for which the algorithm cannot converge to optimum since direction along any coordinate cannot get descent at a point far from optimum.

A general description of coordinate descent method comprises two steps:

- 1. Pick a coordinate $j \in \{1, ..., d\}$.
- 2. Do update $x_j^+ \leftarrow x_j + \eta^* e_j$, where $\eta^* = \underset{\eta}{argmin} f(x + \eta e_j)$.

e_j is a vector with all 0 but jth coordinate equal to 1. Figure 6.2 gives an illustration of the algorithm.

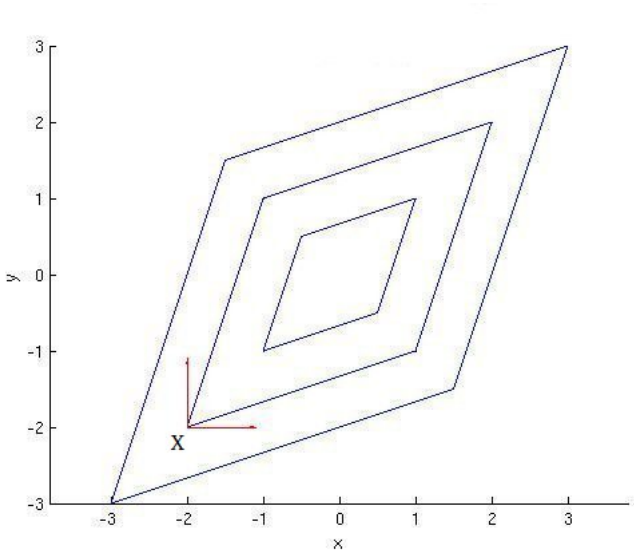

Figure 6.1. An example of non-smooth, non-separable function for which Coordinate Descent method cannot converge to optimum.

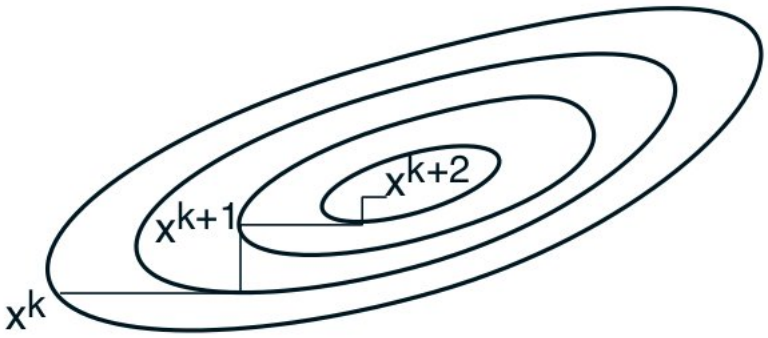

Figure 6.2. Illustration of coordinate descent method

Note the minimization in step 2 can be solved in closed form if $g(x)$ is a quadratic function. For $h(x) = 0$, the closed-form solution is simply

$$\eta^* = -\frac{\nabla_j g(x)}{\nabla_j^2 g(x)}$$

. In the presence of $h(x)$, one can derive closed-form solution simply by discussing which smooth interval of $h_j(x)$ the optimal $x_j^+ = x_j + \eta^*$ lies. For example, for box constraints

(6.2), the resulting solution is simply

$$x_j^+ \leftarrow \min\{\max\{x_j + \eta^*, L\}, U\}, \quad \eta^* = -\frac{\nabla_j g(x)}{\nabla_j^2 g(x)} \tag{6.3}$$

However, for $g(x)$ being other functions than quadratic, one is usually unable to solve it exactly. A common practice is finding a quadratic upper bound of the function $l(\eta) = f(x + \eta e_j)$ (just as did in the convergence analysis of gradient descent) and minimizing the quadratic upper bound to find descent step size η^*.

For step 1, there are different strategies to select the coordinate j being optimized. Three popular strategies that guarantees convergence to optimum are:

- **Cyclic Coordinate Descent**: This strategy simply go through all coordinates repeatedly using a pre-determined order.
- **Greedy Coordinate Descent**: This strategy, at each iteration, select the coordinate direction j along which maximum progress can be made. When objective function is differentiable, it simply chooses :

  $$j^* = \underset{j}{argmax} \, |\nabla_j f(x)|$$

  . This method is computationally intensive since finding coordinate j that maximizes gradient magnitude also requires evaluating the whole gradient vector $\nabla f(x)$.
- **Randomized Coordinate Descent**: This strategy picks the coordinate j from some probability distribution (usually a uniform distribution) over $\{1, ..., d\}$. Note to guarantee convergence, all coordinates should have positive probability to be drawn.

In the following, we will give a version of global convergence proof for *Cyclic Coordinate Descent* when $f(x)$ is differentiable ($h(x) = 0$). The analysis of convergence rate of coordinate descent is more involved. We will simply cite some most recent result from the literature.

6.2.1 Convergence of Coordinate Descent

Global Convergence

Theorem 6.1. *Suppose $\nabla f(x)$ is continuous. $f(x + \eta e_j)$ has attained unique minimum η^* and is monotonic between x and $x + \eta^* e_j$. Then the sequence $\{x_k\}_{k=1}^{\infty}$ produced by cyclic coordinate descent converges to a stationary point of $f(x)$.*

Proof: The sequence of function values produced by cyclic coordinate descent

$$f(x^{(k)}) \geq f(z_1^{(k)}) \geq f(z_2^{(k)}) \geq ... \geq f(z_d^{(k)}) = f(x^{(k+1)})$$

has $f(x^{(k+1)}) < f(x^{(k)})$ for $x^{k+1} \neq x^k$, since any change of coordinate must lead to a strict decrease by uniqueness of the minimum for each sub-problem $arg\min_\eta f(x + \eta e_j)$. Then since the minimization $\min_x f(x)$ is attained, there must exists a limiting function value $\bar{f} = \lim_{k\to\infty} f(x_k)$ so that $\{f(x^{(k)})\}_{k=1}^{\infty}$ not going to $-\infty$. Then we have

$$\lim_{k\to\infty} f(x^{(k)} + \eta^* e_j) - f(x^{(k)}) = \bar{f} - \bar{f} = 0$$

, which, by $f(.)$'s monotonicity between $x^{(k)}$ and $x^{(k)} + \eta^* e_j$, implies $\eta^* = 0$ and thus

$$\lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} z_1^{(k)}$$

. Apply this argument to coordinates $j = 2, ..., d$, we have limiting point $\bar{x}$:

$$\lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} z_1^{(k)} = \lim_{k\to\infty} z_2^{(k)} = ... = \lim_{k\to\infty} x^{(k+1)} = \bar{x}$$

. Then since $0 = \underset{\eta}{argmin} f(\bar{x}+\eta e_j)$ for $\forall j$, we have $\nabla_j f(\bar{x}) = 0$ for $\forall j$, that is, $\nabla f(\bar{x}) = 0$. □

Note the above argument holds only if the objective function has continuous gradient $\nabla f(x)$ s.t. $\nabla_j f(x)$ exists for $\forall j$ and for all x there is uniquely defined $\nabla f(x)$. A function like $f(x_1, x_2) = \max\{x_1, x_2\}$ shown in Figure 6.1 does not have uniquely defined gradient. In addition, the above argument holds not only for *convex function*. Instead, it holds as long as $f(x + \eta e_j)$ has attained unique minimum η^* and is monotonic between x and $x + \eta^* e_j$, which is definitely true if $f(x)$ is strictly convex, and might be also true for some *non-convex* function such as $f(x_1, x_2) = (a - x_1 x_2)^2$.

Convergence Rate

The global convergence analysis does not provide a sense about how long the algorithm takes to obtain an ϵ-precise solution. However, the analysis of convergence speed is quite different for different strategies of picking coordinates. The coordinate descent with *greedy selection* strategy is simply equivalent to steepest descent in ℓ_1-norm, so we discuss its analysis in Section 6.3. For randomized and cyclic strategies, the analysis is more involved. Here we provide some very recent results from the literature.

Theorem 6.2 (A.Beck 2013 [1]). *Suppose $f(x)$ is convex and has (Lipschitz) continuous gradient $\nabla f(x)$. The cyclic coordinate descent algorithm has*

$$f(x_k) - f^* \leq \frac{c_1}{k}$$

, where f^* is the optimum. If we further assume $f(x)$ is strongly convex (with parameter m), we have

$$f(x_k) - f^* \leq \left(1 - \frac{m}{c_2}\right)^k (f(x_0) - f^*)$$

, where c_1, c_2 are constants depending on the Lipschitz constant of $\nabla f(x)$, dimension d and initial point x_0.

Theorem 6.3 (Y. E. Nesterov 2010 [2]). *Suppose $f(x)$ is convex and has (Lipschitz) continuous gradient $\nabla f(x)$. The randomized coordinate descent algorithm has*

$$f(x_k) - f^* \leq \frac{\gamma_1}{k}$$

, where f^* is the optimum. If we further assume $f(x)$ is strongly convex (with parameter m), we have

$$f(x_k) - f^* \leq \left(1 - \frac{m}{\gamma_2}\right)^k (f(x_0) - f^*)$$

, where γ_1, γ_2 are constants depending on the Lipschitz constant of $\nabla f(x)$, dimension d and initial point x_0.

The above result for randomized coordinate descent was also extended to composite function of the form (6.1) in [3]. The proof of cyclic version is actually much harder than the randomized version [2], and even in the most recent literature, the constant c_1, c_2 obtained for cyclic coordinate descent is worse than constants γ_1, γ_2 obtained for randomized coordinate descent by a factor of d.

6.3 Steepest Descent Method

The gradient descent method takes many iterations to converge for certain starting points, when the function has elongated level sets and the descent direction is slowly varying. The steepest descent method aims at choosing the best descent direction at each iteration.

Given a norm $\|\cdot\|$, a *normalized steepest descent direction* is defined as follows:

$$\Delta x_{nsd} = \arg\min_v \{\langle \nabla f(x), v\rangle, s.t. \, ||v|| \leq 1\} \tag{6.4}$$

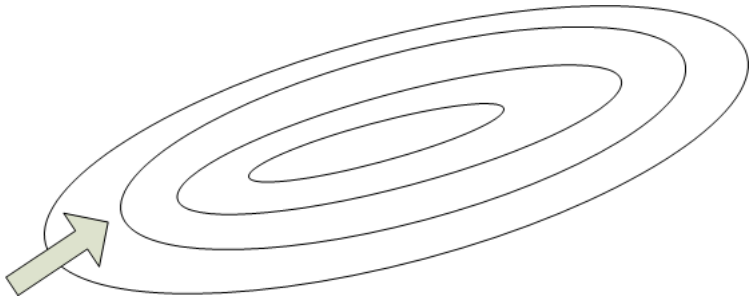

Figure 6.3. Illustration of a function having elongated level sets with slowly varying descent direction

Iteratively, the algorithm follows the following steps:

- Calculate direction of descent, Δx_{nsd}
- Calculate step size, t
- $x_+ = x + t\Delta x_{nsd}$

6.3.1 Steepest Descent for $||\cdot||_2$, $||\cdot||_1$ and $||\cdot||_\infty$

- $||\cdot||_\mathbf{2}$

  If we impose the constraint $||v||_2 \leq 1$ in Equation 6.4, then the steepest descent direction coincides with the direction of $-\nabla f(x)$, and the algorithm is the same as gradient descent.

$$\Delta x_{nsd} = \frac{-\nabla f(x)}{||\nabla f(x)||_2}$$

- $||\cdot||_\mathbf{1}$

  For $||x||_1 = \sum_i |x_i|$, a descent direction is as follows,

$$\Delta x_{nsd} = -sign\left(\frac{\partial f(x)}{\partial x_{i^*}}\right) e_{i^*}$$
$$i^* = \arg\max_i \left|\frac{\partial f}{\partial x_i}\right|$$

  In the above set of equations, e_i is the standard basis corresponding to index i. Figure 6.4 geometrically illustrates this concept.

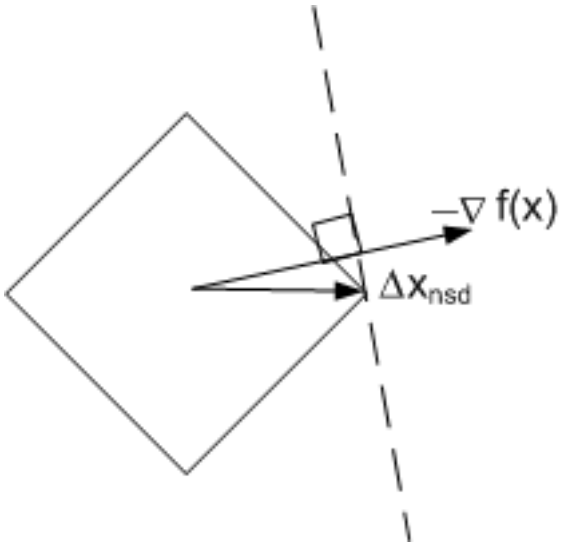

Figure 6.4. Geometric illustration of the normalized steepest descent direction for $||\cdot||_1$

- $||\cdot||_\infty$

  For $||x||_\infty = \arg\max_i |x_i|$, a descent direction is as follows,

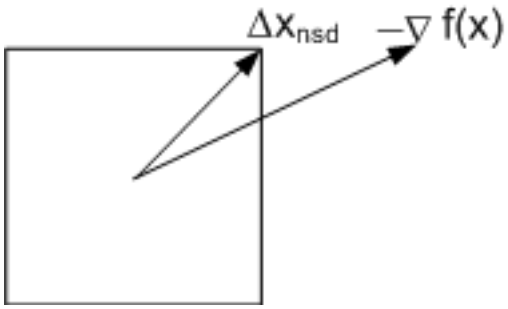

Figure 6.5. Geometric illustration of the normalized steepest descent direction for $||\cdot||_\infty$

$$\Delta x_{nsd} = sign(-\nabla f(x))$$

Figure 6.3 geometrically illustrates this solution.

Definition 1. *Given a norm* $||\cdot||$*, its dual norm* $||\cdot||_*$ *is defined by,*

$$||z||_* = sup\{\langle z, x\rangle, s.t., ||x|| \leq 1\}.$$

Having the definition of dual norm, let's see some examples.

Example 1.

$$||\cdot|| \equiv ||\cdot||_2 \Rightarrow ||\cdot||_* = ||\cdot||_2$$
$$||\cdot|| \equiv ||\cdot||_1 \Rightarrow ||\cdot||_* = ||\cdot||_\infty$$
$$||\cdot|| \equiv ||\cdot||_\infty \Rightarrow ||\cdot||_* = ||\cdot||_1$$

Further, given the definition of dual norm, we have,

$$\langle \nabla f(x), \Delta x_{nsd}\rangle = -||\nabla f(x)||_*$$

Fact: For any norm $||\cdot||$and its dual norm $||\cdot||_*$, there exists finite, positive constants γ and $\tilde{\gamma}$, such that for all x

$$||x|| \geq \gamma||x||_2, \quad ||x||_* \geq \tilde{\gamma}||x||_2$$

Proof: Let $S = \{(v_1, v_2, ..., v_n) \in \mathbb{R} : ||v||_2 = 1\}$. This is a closed and bounded set and thus compact. The triangle inequality implies that norms are continuous, so $||\cdot||$ must attain a minimum for some $z \in S$. Let $\gamma = ||z||$. For any vector x, let $\tilde{x} = \frac{x}{||x||_2}$, a normalized version of x under the 2-norm. Then $\tilde{x} \in S$, so $||\tilde{x}|| \geq ||z||$ and thus $||x|| = ||x||_2||\tilde{x}|| \geq ||x||_2||z|| = \gamma||x||_2$. The proof follows similarly for the dual norm. □

Theorem 6.4. *Suppose f is a strongly convex function with $mI \preceq \nabla^2 f(x) \preceq MI$. Then by using the steepest descent method with BTLS,*

$$f(x^{(k)}) - f^* \leq c^k (f(x^{(0)}) - f^*)$$

where $c = 1 - 2m\alpha\tilde{\gamma}^2 \min\{1, \frac{\beta\gamma^2}{M}\}$.

Proof: First we want to show $\eta = \frac{\gamma^2}{M}$ always satisfies the BTLS exit condition. To show this, we need to show that

$$f(x + \frac{\gamma^2}{M}\Delta x_{\text{nsd}}||\nabla f(x)||_*) \leq f(x) - \frac{1}{2}\frac{\gamma^2}{M}||\nabla f(x)||_*^2.$$

By strong convexity,

$$\begin{aligned} f(x_+) &= f(x + \eta\Delta x_{\text{nsd}}||\nabla f(x)||_*) \\ &\leq f(x) + \eta\langle \nabla f(x), \Delta x_{\text{nsd}}||\nabla f(x)||_*\rangle + \frac{M}{2}||\nabla f(x)||_*||\eta\Delta x_{\text{nsd}}||_2^2 \\ &= f(x) - \eta||\nabla f(x)||_*^2 + \frac{M}{2}\eta^2||\nabla f(x)||_*^2||\Delta x_{\text{nsd}}||_2^2. \end{aligned}$$

Since $||\Delta x_{\text{nsd}}||_2^2 \leq \frac{1}{\gamma^2}||\Delta x_{\text{nsd}}||^2 = \frac{1}{\gamma^2}$, we have

$$f(x_+) \leq f(x) - \eta||\nabla f(x)||_*^2 + \frac{M\eta^2}{2\gamma^2}||\nabla f(x)||_*^2.$$

Letting $\eta = \frac{\gamma^2}{M}$, we then have

$$f(x + \frac{\gamma^2}{M}\Delta x_{\text{sd}}) \leq f(x) - \frac{1}{2}\frac{\gamma^2}{M}||\nabla f(x)||_*^2.$$

Knowing $\eta = \frac{\gamma^2}{M}$ always satisfies the exit condition, we then have

$$\eta \geq \min\left\{1, \frac{\beta\gamma^2}{M}\right\}.$$

By the exit condition of BTLS, we have

$$\begin{aligned} f(x_+) &\leq f(x) - \alpha\eta||\nabla f(x)||_*^2 \\ &= f(x) - \alpha\min\left\{1, \frac{\beta\gamma^2}{M}\right\}||\nabla f(x)||_*^2 \\ &\leq f(x) - \alpha\min\left\{1, \frac{\beta\gamma^2}{M}\right\}\tilde{\gamma}^2||\nabla f(x)||_2^2 \\ &\leq f(x) - \alpha\min\left\{1, \frac{\beta\gamma^2}{M}\right\}\tilde{\gamma}^2 2m(f(x) - f^*). \end{aligned}$$

Equivalently,

$$f(x_+) - f^* \leq f(x) - f^* - 2m\alpha\tilde{\gamma}^2 \min\left\{1, \frac{\beta\gamma^2}{M}\right\}(f(x) - f^*).$$

So

$$c = 1 - 2m\alpha\tilde{\gamma}^2 \min\left\{1, \frac{\beta\gamma^2}{M}\right\}.$$

□

Comments on the Theorem
This theorem proves linear convergence for any norm $||\cdot||$ and sufficiently well-conditioned strongly convex function f.

However, this theorem does not give a better rate for poorly-conditioned functions.

Consider the following examples.

Example 2. Steepest Descent for $||\cdot||_1$
Suppose we use $||\cdot||_1$ to find the steepest descent direction. We know the dual norm of $||\cdot||_1$ is $||\cdot||_\infty$, and it is easy to show that for $x \in \mathbb{R}^n$

$$||x||_1 \geq ||x||_2, \quad \text{(with equality when only one } x_i \text{ is non-zero)}$$
$$||x||_\infty \geq \frac{1}{\sqrt{n}}||x||_2. \quad \text{(with equality when } x_1 = x_2 = ... = x_n)$$

Therefore, $\gamma = 1, \tilde{\gamma} = \frac{1}{\sqrt{n}}$. Compared with the convergence rate in gradient descent with BTLS, where $c = 1 - 2m\alpha \min\left\{1, \frac{\beta}{M}\right\}$, the introduction of γ and $\tilde{\gamma}$ results in an increase in c, which means a slower convergence rate.

Thus, the theorem does not give a better rate for any functions in this case.

Example 3. Change of Coordinates
For some poorly-conditioned functions, for example $f(x_1, x_2) = x_1^2 + 10x_2^2$, we hope to convert it into a well-conditioned problem by changing coordinates and then using the gradient descent method to solve it.

Specifically, let $x = Ay$ and $g(y) = f(Ay)$. Then $\nabla g(y) = A^T \nabla f(Ay)$, $\nabla^2 g(y) = A^T \nabla^2 f(Ay) A$.

We want to make sure $g(y)$ is a well-conditioned function so that gradient descent works well for $g(y)$.

We know $mI \preceq \nabla^2 f(x) \preceq MI$, so if $\nabla^2 f(x) = P \succ 0$ is a constant matrix, which means $f(x)$ is a multi-variable quadratic function like the function mentioned above, so by letting $A = P^{-\frac{1}{2}}$ we can get $\nabla^2 g(y) = I$. So if we use gradient descent for $g(y)$, we can get the best descent direction and best convergence rate. Specifically,

$$\begin{aligned} y_+ &= y - \eta \nabla g(y) \\ &= y - \eta A^T \nabla f(Ay), \\ Ay_+ &= Ay - \eta A A^T \nabla f(Ay) \\ x_+ &= x - \eta A A^T \nabla f(x). \end{aligned}$$

This is the same as using steepest descent method for $f(x)$ with the definition of norm shown below:

$$||x||_Q = (x^T Q x)^{\frac{1}{2}},$$

where $Q = (AA^T)^{-1} = P = \nabla^2 f(x)$.

This means using steepest descent method with this norm definition can get the best convergence rate.

However, if we analyze the convergence rate for this method using the theorem above, we have

$$\begin{aligned} mI \preceq & Q \preceq MI, \\ x^T Q x &\geq m||x||_2^2, \\ x^T Q^{-1} x &\geq \frac{1}{M}||x||_2^2, \end{aligned}$$

i.e.

$$\begin{aligned} \gamma &= \sqrt{m}, \\ \tilde{\gamma} &= \frac{1}{\sqrt{M}}. \end{aligned}$$

Applying the theorem directly, the convergence rate for the steepest descent with BTLS is

$$c = 1 - 2m\alpha \min\left\{\frac{1}{M}, \frac{\beta m}{M^2}\right\}.$$

This is actually worse than gradient descent with BTLS whose convergence rate is

$$c = 1 - 2m\alpha \min\left\{1, \frac{\beta}{M}\right\}.$$

Thus we see that the theorem above is not useful for better rate.

6.4 References

Bibliography

[1] A. Beck and L. Tetruashvili, On the convergence of block coordinate descent type methods, SIAM J. OPTIM, 2013.

[2] Y. E. Nesterov, Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems, CORE Discussion paper 2010/2, 2010.

[3] P. Richtarik and M. Takac. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Mathematical Programming Ser. A, 2012.