On Nesterov's Random Coordinate Descent Algorithms - Continued

Zheng Xu

University of Texas At Arlington

February 20, 2015

Revisit Random Coordinate Descent

When we assume the *partial gradient* of $f(x)$ is Lipschitz continuous, i.e.

$$\|\nabla_i f(x + \Delta x) - \nabla_i f(x)\|_{(i)} \leq L_i \|\Delta x\| \tag{1}$$

We can then introduce the random coordinate descent algorithm (hereafter RCDM),

$$i_k = \mathcal{R}_\alpha \tag{2}$$

$$x_{i_k}^{k+1} = x_{i_k}^k + \alpha_k(-\nabla_{i_k} f(x^k)) \tag{3}$$

where $\mathcal{R}_\alpha$ is a random number generator that is subject to some distribution with parameter α.

Revisit the $\mathcal{R}_\alpha$

$\mathcal{R}_\alpha \in \mathbb{N}$ is a special random counter, generating integer number with probability

$$p_\alpha^{(i)} = L_i^\alpha \cdot \left[\sum_{j=1}^{n} L_j^\alpha\right]^{-1} \tag{4}$$

where L_i is the Lipschitz constant of the *partial gradient* $\nabla_i f(x)$.
To give an example, when $\alpha = 0$, $\mathcal{R}_0$ is a uniform random number generator.

Bregman Distance of a Function

Definition

The **Bregman Distance** associated with a convex function f at the point y is

$$D_f^y(x, y) = f(x) - f(y) - \langle p, x - y \rangle \tag{5}$$

where $p \in \partial f(v)$.

Strong Convexity

Definition

A differentiable function f is called **strongly convex** with **convexity parameter** $\mu > 0$ if the following inequality holds for all points x, y in its domain:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2 \tag{6}$$

Remark

The definition of strong convexity yields a lower bound of Bregman distance of f

$$D_f^y(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{\mu}{2}\|x - y\|^2 \tag{7}$$

Strong Smoothness

Definition

A function f is **1/L-Lipschitz continuous** if

$$\|f(x) - f(y)\| \leq L\|x - y\| \tag{8}$$

for all $x, y \in \mathbf{dom} f$

In [NN04], a well-known inequality is introduced for a function with $1/L$-Lipschitz continuous gradient.

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{L}{2}\|x - y\|^2 \tag{9}$$

which yields another bound from the upper side of the Bregman distance

$$D_f^y(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{L}{2}\|x - y\|^2 \tag{10}$$

Upper and Lower Bound of Bregman Distance

From definitions, if $f(x)$ is both μ-strongly convex and L-strongly smooth, we can bound the Bregman distance or Hessian from both upper and lower side, i.e.

$$\frac{\mu}{2}\|x-y\|^2 \leq f(x) - f(y) - \langle \nabla f(y), x-y \rangle \leq \frac{L}{2}\|x-y\|^2 \tag{11}$$

Letting in the Taylor expansion, we now have the bounds for Hessian matrix of twice differentiable $f(x)$

$$\mu \leq \|\nabla^2 f(x)\| \leq L \tag{12}$$

Case when $f(x)$ is strongly convex

In [NN04], it is proven that, for deterministic numerical optimization, when objection function $f(x)$ is both $\mu-$strongly convex and $L-$strongly smooth, the optimal convergence rate for minimizing $f(x)$ is

$$f(x_k) - f^* \geq \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|x_0 - x^*\| \tag{13}$$

where $\kappa = \frac{L}{\mu}$ is the *condition number* of f. This type of convergence rate is also known as linear rate convergence.

Linear rate convergence of RCDM when f is strongly convex

Theorem

Let function $f(x)$ be strongly convex with convexity parameter $\mu > 0$. Then, for the sequence $\{x_k\}$ generated by RCDM(α, x_0) $(\alpha \neq 0)$, we have

$$\mathbb{E}f(x_k) - f^* \leq \left(1 - \frac{\mu}{\mathbb{S}}\right)^k (f(x_0) - f^*) \tag{14}$$

where $\mathbb{S} = \sum_{j=1}^{n} L_j^{\alpha}$

From this theorem, we can see the Random Coordinate Descent Method can reach the first-order optimal rate when $f(x)$ is both strongly convex and strongly smooth.

Case when $\alpha = 0$

Here, when $\alpha = 0$, the $\mathcal{R}_0$ is thus a uniform random number generator. The algorithm is thus

1. Define $v_0 = x_0$, $a_0 = \frac{1}{n}$, $b_0 = 2$
2. For $k > 0$ iterate:
   1. Solve γ_k from $\gamma_k^2 - \frac{\gamma_k}{n} = (1 - \frac{\gamma_k \sigma}{n})\frac{a_k^2}{b_k^2}$
      Set $\alpha_k = \frac{n - \gamma_k \sigma}{\gamma_k (n^2 - \sigma)}$, $\beta_k = 1 - \frac{1}{n}\gamma_k \sigma$
   2. Select $y_k = \alpha_k v_k + (1 - \alpha_k) x_k$
   3. Choose $i_k = \mathcal{R}_0$ and update
      $x_{k+1} = T_{i_k}(y_k)$, $v_{k+1} = \beta_k v_k + (1 - \beta_k) y_k - \frac{\gamma_k}{L_{i_k}} f_i'(y_k)$
   4. Set $b_{k+1} = \frac{b_k}{\sqrt{\beta_k}}$, and $a_{k+1} = \gamma_k b_{k+1}$

Convergence Analysis of Accelerated Random Coordinate Descent

Theorem

For any $k \geq 0$ we have

$$\mathbb{E}f(x_k) - f^* \leq \sigma \left[2\|x_0 - x^*\|_1^2 + \frac{1}{n^2}(f(x_0) - f^*)\right] \cdot \left[(1 + \frac{\sqrt{\sigma}}{2n})^{k+1} - (1 - \frac{\sqrt{\sigma}}{2n})^{k+1}\right]^{-2} \leq (\frac{n}{k+1})^2 \cdot \left[2\|x_0 - x^*\|_1^2 + \frac{1}{n^2}(f(x_0) - f^*)\right] \quad (15)$$

The constrained problem

Consider now the constrained minimization problem

$$\min_{x \in Q} f(x) \tag{16}$$

Here, $Q \in \mathbb{R}^N$ is a convex and closed set which constructs a constrain in this problem.

Conditional Gradient Optimization

First, we need to fix the random number generator parameter $\alpha = 0$ and thus $\mathcal{R}_\alpha$ is a uniform random number generator. The conditional coordinate descent steps are

$$u^{(i)}(x) = \arg \min_{u^{(i)} \in Q_i} \left[\langle f_i'(x), u^{(i)} - x^{(i)} \rangle + \frac{L_i}{2} \|u^{(i)} - x^{(i)}\|_{(i)}^2 \right],$$
$$V_i(x) = x + (u^{(i)}(x) - x^{(i)}),$$

Under this denotation, the algorithm is

$$i_k = \mathcal{R}_0$$
$$x_{k+1} = V_{i_k}(x_k)$$

Note this type of update is quite like the Frank-Wolfe ($\equiv$ conditional gradient) method.

Convergence Analysis for Constrained Random Coordinate Descent

Theorem

For any $k > 0$ we have

$$\mathbb{E}f(x_k) - f^* \leq \frac{n}{n+k}\left(\frac{1}{2}R_1^2(x_0) + f(x_0) - f^*\right) \tag{17}$$

if f is strongly convex in $\|\cdot\|_1$ with constant μ, then

$$\mathbb{E}f(x_k) - f^* \leq \left(1 - \frac{2\sigma}{n(1+\sigma)}\right)^k \left(\frac{1}{2}R_1^2(x_0) + f(x_0) - f^*\right) \tag{18}$$

Summary

1. We've revisited the Basic Random Coordinate Descent and the basic definition of strongly convex and strongly smooth.
2. When unconstrained objective function f is both strongly convex and strongly smooth,
   1. if $\alpha \neq 0$, the convergence rate could be accelerated to linear rate i.e. $\mathcal{O}(\theta^k)\, \theta \in (0, 1)$
   2. if $\alpha = 0$, the convergence rate could be accelerated to sublinear rate $\mathcal{O}(\frac{1}{k^2})$ where k is the iteration counter
3. Discussion of Frank-Wolfe-type constrained minimization of random coordinate descent method, could only work with $\alpha = 0$.

Thanks!


Yurii Nesterov and Nesterov.
Introductory lectures on convex optimization: A basic course, volume 87.
Springer, 2004.