12. Coordinate descent methods

- theoretical justifications
- randomized coordinate descent method
- minimizing composite objectives
- accelerated coordinate descent method

Notations

consider smooth unconstrained minimization problem:

$$\underset{x \in \mathbf{R}^N}{\text{minimize}} \quad f(x)$$

- coordinate blocks: $x = (x_1, \ldots, x_n)$ with $x_i \in \mathbf{R}^{N_i}$ and $\sum_{i=1}^n N_i = N$
- more generally, partition with a permutation matrix: $U = [U_1 \cdots U_n]$

$$x_i = U_i^T x, \qquad x = \sum_{i=1}^n U_i x_i$$

- blocks of gradient:

$$\nabla_i f(x) \;=\; U_i^T \nabla f(x)$$

- coordinate update:

$$x^+ = x - t\, U_i \nabla_i f(x)$$

(Block) coordinate descent

choose $x^{(0)} \in \mathbf{R}^n$, and iterate for $k = 0, 1, 2, \ldots$

1. choose coordinate $i(k)$
2. update $x^{(k+1)} = x^{(k)} - t_k U_{i_k} \nabla_{i_k} f(x^{(k)})$

- among the first schemes for solving smooth unconstrained problems
- cyclic or round-Robin: difficult to analyze convergence
- mostly local convergence results for particular classes of problems
- does it really work (better than full gradient method)?

Steepest coordinate descent

choose $x^{(0)} \in \mathbf{R}^n$, and iterate for $k = 0, 1, 2, \ldots$

1. choose $i(k) = \underset{i \in \{1,\ldots,n\}}{\operatorname{argmax}} \|\nabla_i f(x^{(k)})\|_2$

2. update $x^{(k+1)} = x^{(k)} - t_k U_{i(k)} \nabla_{i(k)} f(x^{(k)})$

assumptions

- $\nabla f(x)$ is block-wise Lipschitz continuous

$$\|\nabla_i f(x + U_i v) - \nabla_i f(x)\|_2 \leq L_i \|v\|_2, \qquad i = 1, \ldots, n$$

- f has bounded sub-level set, in particular, define

$$R(x) = \max_y \left\{ \max_{x^\star \in X^\star} \|y - x^\star\|_2 : f(y) \leq f(x) \right\}$$

Analysis for constant step size

quadratic upper bound due to block coordinate-wise Lipschitz assumption:

$$f(x + U_i v) \leq f(x) + \langle \nabla_i f(x), v \rangle + \frac{L_i}{2}\|v\|_2^2, \qquad i = 1, \ldots, n$$

assume constant step size $0 < t \leq 1/M$, with $M \triangleq \max_{i \in \{1,\ldots,n\}} L_i$

$$f(x^+) \quad \leq \quad f(x) - \frac{t}{2}\|\nabla_i f(x)\|_2^2 \;\leq\; f(x) - \frac{t}{2n}\|\nabla f(x)\|_2^2$$

by convexity and Cauchy-Schwarz inequality,

$$\begin{aligned} f(x) - f^\star \quad &\leq \quad \langle \nabla f(x), x - x^\star \rangle \\ &\leq \quad \|\nabla f(x)\|_2 \|x - x^\star\|_2 \;\leq\; \|\nabla f(x)\|_2 R(x^{(0)}) \end{aligned}$$

therefore

$$f(x) - f(x^+) \;\geq\; \frac{t}{2nR^2}\big(f(x) - f^\star\big)^2$$

let $\Delta_k = f(x^{(k)}) - f^\star$, then

$$\Delta_k - \Delta_{k+1} \geq \frac{t}{2nR^2}\Delta_k^2$$

consider their multiplicative inverses

$$\frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k} \; = \; \frac{\Delta_k - \Delta_{k+1}}{\Delta_{k+1}\Delta_k} \; \geq \; \frac{\Delta_k - \Delta_{k+1}}{\Delta_k^2} \; \geq \; \frac{t}{2nR^2}$$

therefore

$$\frac{1}{\Delta_k} \; \geq \; \frac{1}{\Delta_0} + \frac{k}{2nL_{\max}R^2} \; \geq \; \frac{2t}{nR^2} + \frac{kt}{2nR^2}$$

finally

$$f(x^{(k)}) - f^\star \; = \; \Delta_k \; \leq \; \frac{2nR^2}{(k+4)t}$$

Bounds on full gradient Lipschitz constant

lemma: suppose $A \in \mathbf{R}^{N\times N}$ is positive semidefinite and has the partition $A = [A_{ij}]_{n\times n}$, where $A_{ij} \in \mathbf{R}^{N_i \times N_j}$ for $i, j = 1, \ldots, n$, and

$$A_{ii} \preceq L_i I_{N_i}, \qquad i = 1, \ldots, n$$

then

$$A \preceq \left(\sum_{i=1}^{n} L_i \right) I_N$$

proof:

$$\begin{aligned} x^T A x &= \sum_{i=1}^{n} \sum_{i=1}^{n} x_i^T A_{ij} x_j \leq \left(\sum_{i=1}^{n} \sqrt{x_i^T A_{ii} x_i} \right)^2 \\ &\leq \left(\sum_{i=1}^{n} L_i^{1/2} \|x_i\|_2 \right)^2 \leq \left(\sum_{i=1}^{n} L_i \right) \sum_{i=1}^{n} \|x_i\|_2^2 \end{aligned}$$

conclusion: the full gradient Lipschitz constant $L_f \leq \sum_{i=1}^{n} L_i$

Computational complexity and justifications

(steepest) coordinate descent	$O\left(\dfrac{nMR^2}{k}\right)$
full gradient method	$O\left(\dfrac{L_f R^2}{k}\right)$

in general coordinate descent has worse complexity bound

- it can happen that $M \geq O(L_f)$
- choosing $i(k)$ may rely on computing full gradient
- too expensive to do line search based on function values

nevertheless, there are justifications for *huge-scale* problems

- even computation of a function value can require substantial effort
- limits by computer memory, distributed storage, and human patience

Example

$$\underset{x\in\mathbf{R}^n}{\text{minimize}} \quad \left\{ f(x) \stackrel{\text{def}}{=} \sum_{i=1}^n f_i(x_i) + \frac{1}{2}\|Ax-b\|_2^2 \right\}$$

- f_i are convex differentiable univariate functions
- $A = [a_1 \cdots a_n] \in \mathbf{R}^{m\times n}$, and assume a_i has p_i nonzero elements

computing either function value or full gradient costs $O(\sum_{i=1}^n p_i)$ operations

computing coordinate directional derivatives: $O(p_i)$ operations

$$\begin{aligned} \nabla_i f(x) &= \nabla f_i(x_i) + a_i^T r(x), \qquad i = 1,\ldots,n \\ r(x) &= Ax - b \end{aligned}$$

- given $r(x)$, computing $\nabla_i f(x)$ requires $O(p_i)$ operations
- coordinate update $x^+ = x + \alpha e_i$ results in efficient update of residue: $r(x^+) = r(x) + \alpha a_i$, which also cost $O(p_i)$ operations

Outline

- theoretical justifications
- **randomized coordinate descent method**
- minimizing composite objectives
- accelerated coordinate descent method

Randomized coordinate descent

choose $x^{(0)} \in \mathbf{R}^n$ and $\alpha \in \mathbf{R}$, and iterate for $k = 0, 1, 2, \ldots$

1. choose $i(k)$ with probability $p_i^{(\alpha)} = \dfrac{L_i^\alpha}{\sum_{j=1}^n L_j^\alpha}$, $i = 1, \ldots, n$
2. update $x^{(k+1)} = x^{(k)} - \dfrac{1}{L_i} U_{i(k)} \nabla_{i(k)} f(x^{(k)})$

special case: $\alpha = 0$ gives uniform distribution $p_i^{(0)} = 1/n$ for $i = 1, \ldots, n$

assumptions

- $\nabla f(x)$ is block-wise Lipschitz continuous

$$\|\nabla_i f(x + U_i v_i) - \nabla_i f(x)\|_2 \leq L_i \|v_i\|_2, \qquad i = 1, \ldots, n \tag{1}$$

- f has bounded sub-level set, and $f^\star$ is attained at some $x^\star$

Solution guarantees

convergence in expectation

$$\mathbf{E}[f(x^{(k)})] - f^\star \le \epsilon$$

high probability iteration complexity: number of iterations to reach

$$\mathbf{prob}\big(f(x^{(k)}) - f^\star \le \epsilon\big) \ge 1 - \rho$$

- confidence level $0 < \rho < 1$
- error tolerance $\epsilon > 0$

Convergence analysis

block coordinate-wise Lipschitz continuity of $\nabla f(x)$ implies for $i = 1, \ldots, n$

$$f(x + U_i v_i) \leq f(x) + \langle \nabla_i f(x), v_i \rangle + \frac{L_i}{2} \|v_i\|_2^2, \quad \forall x \in \mathbf{R}^N, \ v_i \in \mathbf{R}^{N_i}$$

coordinate update obtained by minimizing quadratic upper bound

$$\begin{aligned} x^+ &= x + U_i \hat{v}_i \\ \hat{v}_i &= \operatorname*{argmin}_{v_i} \left\{ f(x) + \langle \nabla_i f(x), v_i \rangle + \frac{L_i}{2} \|v_i\|_2^2 \right\} \end{aligned}$$

objective function is non-increasing:

$$f(x) - f(x^+) \ \geq \ \frac{1}{2L_i} \|\nabla_i f(x)\|_2^2$$

A pair of conjugate norms

for any $\alpha \in \mathbf{R}$, define

$$\|x\|_\alpha = \left(\sum_{i=1}^n L_i^\alpha \|x_i\|_2^2\right)^{1/2}, \qquad \|y\|_\alpha^* = \left(\sum_{i=1}^n L_i^{-\alpha} \|y_i\|_2^2\right)^{1/2}$$

let $S_\alpha = \sum_{i=1}^n L_i^\alpha$ (note that $S_0 = n$)

lemma (Nesterov): let f satisfy (1), then for any $\alpha \in \mathbf{R}$,

$$\|\nabla f(x) - \nabla f(y)\|_{1-\alpha}^* \le S_\alpha \|x - y\|_{1-\alpha}, \qquad \forall\, x, y \in \mathbf{R}^N$$

therefore

$$f(x) \le f(y) + \langle \nabla f(y), x - y\rangle + \frac{S_\alpha}{2}\|x - y\|_{1-\alpha}^2, \qquad \forall\, x, y \in \mathbf{R}^N$$

Convergence in expectation

theorem (Nesterov): for any $k \geq 0$,

$$\mathbf{E} f(x^{(k)}) - f^\star \;\leq\; \frac{2}{k+4} S_\alpha R_{1-\alpha}^2(x^{(0)})$$

where $R_{1-\alpha}(x^{(0)}) = \max_y \left\{ \max_{x^\star \in X^\star} \|y - x^\star\|_{1-\alpha} : f(y) \leq f(x^{(0)}) \right\}$

proof: define random variables $\xi_k = \{i(0), \ldots, i(k)\}$,

$$\begin{aligned}
f(x^{(k)}) - \mathbf{E}_{i(k)} f(x^{(k+1)}) \;&=\; \sum_{i=1}^n p_i^{(\alpha)} \big(f(x^{(k)}) - f(x^{(k)} + U_i \hat{v}_i) \big) \\
&\geq\; \sum_{i=1}^n \frac{p_i^{(\alpha)}}{2L_i} \|\nabla_i f(x^{(k)})\|_2^2 \\
&=\; \frac{1}{2S_\alpha} \big(\|\nabla f(x)\|_{1-\alpha}^* \big)^2
\end{aligned}$$

$$
\begin{aligned}
f(x^{(k)}) - f^\star &\leq \min_{x^\star \in X^\star} \langle \nabla f(x^{(k)}), x^{(k)} - x^\star \rangle \\
&\leq \|\nabla f(x^{(k)})\|_{1-\alpha}^{*} R_{1-\alpha}(x^{(0)})
\end{aligned}
$$

therefore, with $C = 2S_\alpha R_{1-\alpha}^2(x^{(0)})$,

$$
f(x^{(k)}) - \mathbf{E}_{i(k)} f(x^{(k+1)}) \geq \frac{1}{C}\big(f(x^{(k)}) - f^\star\big)^2
$$

taking expectation of both sides with respect to $\xi_{k-1} = \{i(0), \ldots, i(k-1)\}$,

$$
\begin{aligned}
\mathbf{E} f(x^{(k)}) - \mathbf{E} f(x^{(k+1)}) &\geq \frac{1}{C}\mathbf{E}_{\xi_{k-1}}\left[\big(f(x^{(k)}) - f^\star\big)^2\right] \\
&\geq \frac{1}{C}\Big(\mathbf{E} f(x^{(k)}) - f^\star\Big)^2
\end{aligned}
$$

finally, following steps on page 12–6 to obtain desired result

Discussions

- $\alpha = 0$: $S_0 = n$ and

$$\mathbf{E}f(x^{(k)}) - f^\star \;\leq\; \frac{2n}{k+4} R_1^2(x^{(0)}) \;\leq\; \frac{2n}{k+4} \sum_{i=1}^n L_i \|x_i^{(0)} - x^\star\|_2^2$$

  corresponding rate of full gradient method: $f(x^{(k)}) - f^\star \leq \frac{\gamma}{k} R_1^2(x^{(0)})$, where γ is big enough to ensure $\nabla^2 f(x) \preceq \gamma\, \mathbf{diag}\{L_i I_{N_i}\}_{i=1}^n$

  conclusion: proportional to worst case rate of full gradient method

- $\alpha = 1$: $S_1 = \sum_{i=1}^n L_i$ and

$$\mathbf{E}f(x^{(k)}) - f^\star \;\leq\; \frac{2}{k+4} \left(\sum_{i=1}^n L_i \right) R_0^2(x^{(0)})$$

  corresponding rate of full gradient method: $f(x^{(k)}) - f^\star \leq \frac{L_f}{k} R_0^2(x^{(0)})$

  conclusion: same as worst case rate of full gradient method

but each iteration of randomized coordinate descent can be much cheaper

An interesting case

consider $\alpha = 1/2$, let $N_i = 1$ for $i = 1, \ldots, n$, and let

$$D_\infty(x^{(0)}) = \max_x \left\{ \max_{y \in X^\star} \max_{1 \le i \le n} |x_i - y_i| : f(x) \le f(x^{(0)}) \right\}$$

then $R_{1/2}^2(x^{(0)}) \le S_{1/2} D_\infty^2(x^{(0)})$ and hence

$$\mathbf{E} f(x^{(k)}) - f^\star \;\le\; \frac{2}{k+4} \left(\sum_{i=1}^n L_i^{1/2} \right)^2 D_\infty^2(x^{(0)})$$

- worst-case dimension-independent complexity of minimizing convex functions over n-dimensional box is infinite (Nemirovski & Yudin 1983)
- $S_{1/2}$ can be bounded for very big or even infinite dimension problems

conclusion: RCD can work in situations where full gradient methods have no theoretical justification

Convergence for strongly convex functions

theorem (Nesterov): if f is strongly convex with respect to the norm $\|\cdot\|_{1-\alpha}$ with convexity parameter $\sigma_{1-\alpha} > 0$, then

$$\mathbf{E} f(x^{(k)}) - f^\star \;\le\; \left(1 - \frac{\sigma_{1-\alpha}}{S_\alpha}\right)^k \left(f(x^{(0)}) - f^\star\right)$$

proof: combine consequence of strong convexity

$$f(x^{(k)}) - f^\star \;\le\; \frac{1}{\sigma_{1-\alpha}} \left(\|\nabla f(x)\|_{1-\alpha}^*\right)^2$$

with inequality on page 12–14 to obtain

$$f(x^{(k)}) - \mathbf{E}_{i(k)} f(x^{(k+1)}) \;\ge\; \frac{1}{2S_\alpha}\left(\|\nabla f(x)\|_{1-\alpha}^*\right)^2 \;\ge\; \frac{\sigma_{1-\alpha}}{S_\alpha}\left(f(x^{(k)}) - f^\star\right)$$

it remains to take expectations over $\xi_{k-1} = \{i(0), \ldots, i(k-1)\}$

High probability bounds

number of iterations to guarantee

$$\mathbf{prob}\big(f(x^{(k)}) - f^\star \le \epsilon\big) \;\ge\; 1 - \rho$$

where $0 < \rho < 1$ is confidence level and $\epsilon > 0$ is error tolerance

- for smooth convex functions

$$O\left(\frac{n}{\epsilon}\left(1 + \log\frac{1}{\rho}\right)\right)$$

- for smooth strongly convex functions

$$O\left(\frac{n}{\mu}\log\left(\frac{1}{\epsilon\rho}\right)\right)$$

Outline

- theoretical justifications
- randomized coordinate descent method
- **minimizing composite objectives**
- accelerated coordinate descent method

Minimizing composite objectives

$$\underset{x \in \mathbf{R}^N}{\text{minimize}} \quad \left\{ F(x) \triangleq f(x) + \Psi(x) \right\}$$

assumptions

- f differentiable and $\nabla f(x)$ block coordinate-wise Lipschitz continuous

$$\|\nabla_i f(x + U_i v_i) - \nabla_i f(x)\|_2 \leq L_i \|v_i\|_2, \qquad i = 1, \ldots, n$$

- Ψ is block separable:

$$\Psi(x) = \sum_{i=1}^{n} \Psi_i(x_i)$$

and each Ψ_i is convex and closed, and also *simple*

Coordinate update

use quadratic upper bound on smooth part:

$$\begin{aligned} F(x+U_i v) &= f(x+U_i v_i) + \Psi(x+U_i v_i) \\ &\le f(x) + \langle \nabla_i f(x), v_i \rangle + \frac{L_i}{2}\|v_i\|_2 + \Psi_i(x_i+v_i) + \sum_{j\neq i} \Psi_j(x_j) \end{aligned}$$

define

$$V_i(x,v_i) = f(x) + \langle \nabla_i f(x), v_i \rangle + \frac{L_i}{2}\|v_i\|_2 + \Psi_i(x_i+v_i)$$

coordinate descent takes the form

$$x^{(k+1)} = x^{(k)} + U_i \Delta x_i$$

where

$$\Delta x_i = \operatorname*{argmin}_{v_i} V(x,v_i)$$

Randomized coordinate descent for composite functions

choose $x^{(0)} \in \mathbf{R}^n$ and $\alpha \in \mathbf{R}$, and iterate for $k = 0, 1, 2, \ldots$

1. choose $i(k)$ with uniform probability $1/n$
2. compute $\Delta x_i = \operatorname{argmin}_{v_i} V(x^{(k)}, v_i)$ and update

$$x^{(k+1)} = x^{(k)} + U_i \Delta x_i$$

- similar convergence results as for the smooth case
- can only choose coordinate with uniform distribution?

(see references for details)

Outline

- theoretical justifications
- randomized coordinate descent method
- minimizing composite objectives
- **accelerated coordinate descent method**

Assumptions

restrict to unconstrained smooth minimization problem

$$\underset{x \in \Re^N}{\text{minimize}} \;\; f(x)$$

assumptions

- $\nabla f(x)$ is block-wise Lipschitz continuous

$$\|\nabla_i f(x + U_i v) - \nabla_i f(x)\|_2 \leq L_i \|v\|_2, \qquad i = 1, \ldots, n$$

- f has convexity parameter $\mu \geq 0$

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

Algorithm: ARCD(x^0)

Set $v^0 = x^0$, choose $\gamma_0 > 0$ arbitrarily, and repeat for $k = 0, 1, 2, \ldots$

1. Compute $\alpha_k \in (0, n]$ from the equation

$$\alpha_k^2 = \left(1 - \tfrac{\alpha_k}{n}\right)\gamma_k + \tfrac{\alpha_k}{n}\mu$$

   and set $\gamma_{k+1} = \left(1 - \frac{\alpha_k}{n}\right)\gamma_k + \frac{\alpha_k}{n}\mu$

2. Compute $y^k = \frac{1}{\frac{\alpha_k}{n}\gamma_k + \gamma_{k+1}}\left(\frac{\alpha_k}{n}\gamma_k v^k + \gamma_{k+1} x^k\right)$

3. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random, and update

$$x^{k+1} = y^k - \tfrac{1}{L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k)$$

4. Set $v^{k+1} = \frac{1}{\gamma_{k+1}}\left(\left(1 - \frac{\alpha_k}{n}\right)\gamma_k v^k + \frac{\alpha_k}{n}\mu y^k - \frac{\alpha_k}{L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k)\right)$

Algorithm: ARCD(x^0)

Set $v^0 = x^0$, choose $\alpha_{-1} \in (0, n]$, and repeat for $k = 0, 1, 2, \ldots$

1. Compute $\alpha_k \in (0, n]$ from the equation

$$\alpha_k^2 = \left(1 - \frac{\alpha_k}{n}\right) \alpha_{k-1}^2 + \frac{\alpha_k}{n} \mu,$$

   and set $\theta_k = \frac{n\alpha_k - \mu}{n^2 - \mu}$, $\beta_k = 1 - \frac{\mu}{n\alpha_k}$

2. Compute $y^k \ = \ \theta_k v^k + (1 - \theta_k) x^k$

3. Choose $i_k \in \{1, \ldots, n\}$ uniformly at random, and update

$$x^{k+1} = y^k - \frac{1}{L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k)$$

4. Set $v^{k+1} = \beta_k v^k + (1 - \beta_k) y^k - \frac{1}{\alpha_k L_{i_k}} U_{i_k} \nabla_{i_k} f(y^k)$

Convergence analysis

theorem: Let $x^\star$ be an solution of $\min_x f(x)$ and $f^\star$ be the optimal value. If $\{x^k\}$ is generated by ARCD method, then for any $k \geq 0$

$$\mathbf{E}[f(x^k)] - f^\star \;\leq\; \lambda_k \left(f(x^0) - f^\star + \frac{\gamma_0}{2} \|x^0 - x^\star\|_L^2 \right),$$

where $\lambda_0 = 1$ and $\lambda_k = \prod_{i=0}^{k-1} \left(1 - \frac{\alpha_i}{n}\right)$. In particular, if $\gamma_0 \geq \mu$, then

$$\lambda_k \;\leq\; \min \left\{ \left(1 - \frac{\sqrt{\mu}}{n}\right)^k, \; \left(\frac{n}{n + k\frac{\sqrt{\gamma_0}}{2}} \right)^2 \right\}.$$

- when $n = 1$, recovers results for accelerated full gradient methods
- efficient implementation possible using change of variables

Randomized estimate sequence

definition: $\{(\phi_k(x), \lambda_k)\}_{k=0}^{\infty}$ is a randomized estimate sequence of $f(x)$ if

- $\lambda_k \to 0$ (assume λ_k independent of $\xi_k = \{i_0, \ldots, i_k\}$)
- $\mathbf{E}_{\xi_{k-1}}[\phi_k(x)] \;\leq\; (1-\lambda_k)f(x) + \lambda_k\phi_0(x), \;\; \forall\, x \in \Re^N$

lemma: if $\{x^{(k)}\}$ satisfies $\mathbf{E}_{\xi_{k-1}}[f(x^k)] \;\leq\; \min_x \mathbf{E}_{\xi_{k-1}}[\phi_k(x)]$, then

$$\mathbf{E}_{\xi_{k-1}}[f(x^k)] - f^\star \;\leq\; \lambda_k\left(\phi_0(x^\star) - f^\star\right) \;\to\; 0$$

proof:

$$\begin{aligned}
\mathbf{E}_{\xi_{k-1}}[f(x^k)] &\leq \min_x \mathbf{E}_{\xi_{k-1}}[\phi_k(x)] \\
&\leq \min_x \{(1-\lambda_k)f(x) + \lambda_k\phi_0(x)\} \\
&\leq (1-\lambda_k)f(x^\star) + \lambda_k\phi_0(x^\star) \\
&= f^\star + \lambda_k(\phi_0(x^\star) - f^\star)
\end{aligned}$$

Construction of randomized estimate sequence

lemma: if $\{\alpha_k\}_{k\geq 0}$ satisfies $\alpha_k \in (0,n)$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$, then

$$\lambda_{k+1} = \left(1 - \frac{\alpha_k}{n}\right)\lambda_k, \quad \text{with } \lambda_0 = 1$$

$$\phi_{k+1}(x) = \left(1 - \frac{\alpha_k}{n}\right)\phi_k(x) + \frac{\alpha_k}{n}\left(f(y^k) + n\langle \nabla_{i_k} f(y^k), x_{i_k} - y^k_{i_k}\rangle + \frac{\mu}{2}\|x - y^k\|_L^2\right)$$

is a pair of randomized estimate sequence

proof: for $k=0$, $\mathbf{E}_{\xi_{-1}}[\phi_0(x)] = \phi_0(x) = (1-\lambda_0)f(x) + \lambda_0\phi_0(x)$; then

$$\begin{aligned}
\mathbf{E}_{\xi_k}[\phi_{k+1}(x)] &= \mathbf{E}_{\xi_{k-1}}\left[\mathbf{E}_{i_k}[\phi_{k+1}(x)]\right] \\
&= \mathbf{E}_{\xi_{k-1}}\left[\left(1-\frac{\alpha_k}{n}\right)\phi_k(x) + \frac{\alpha_k}{n}\left(f(y^k) + \langle \nabla f(y^k), x - y^k\rangle + \frac{\mu}{2}\|x-y^k\|_L^2\right)\right] \\
&\leq \mathbf{E}_{\xi_{k-1}}\left[\left(1 - \tfrac{\alpha_k}{n}\right)\phi_k(x) + \tfrac{\alpha_k}{n} f(x)\right] \\
&\leq \left(1 - \frac{\alpha_k}{n}\right)\left[(1-\lambda_k)f(x) + \lambda_k\phi_0(x)\right] + \frac{\alpha_k}{n} f(x) \qquad \ldots
\end{aligned}$$

Derivation of APCD

- let $\phi_0(x) = \phi_0^\star + \frac{\gamma_0}{2}\|x - v^0\|_L^2$, then for all $k \geq 0$,

$$\phi_k(x) = \phi_k^\star + \frac{\gamma_k}{2}\|x - v^k\|_L^2$$

  can derive expressions for $\phi_k^\star$, γ_k and v^k explicitly

- follow the same steps as in deriving accelerated full gradient method

- actually use a strong condition

$$\mathbf{E}_{\xi_{k-1}} f(x^k) \;\leq\; \mathbf{E}_{\xi_{k-1}}[\min_x \phi_k(x)]$$

  which implies

$$\mathbf{E}_{\xi_{k-1}} f(x^k) \;\leq\; \min_x \mathbf{E}_{\xi_{k-1}}[\phi_k(x)]$$

References

- Yu. Nesterov, *Efficiency of coordinate descent methods on huge-scale optimization problems*, SIAM Journal on Optimization, 2012
- P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Mathematical Programming, 2014
- Z. Lu and L. Xiao, *On the complexity analysis of randomized block-coordinate descent methods*, MSR Tech Report, 2013
- Y. T. Lee and A. Sidford, *Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems*, arXiv, 2013
- P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, 2009