Chapter 7. Coordinate Descent (CD)

1. Cyclic CD

$$\min_{x \in \mathbb{R}^n} f(x) := f(x_1, x_2, x_3, \cdots, x_n)$$

For $k = 0, 1, \cdots$ (outer iterations)

  For $i = 1, 2, \cdots, n$ (inner iterations)

$$x_i^{k+1} \leftarrow \operatorname*{argmin}_{\xi} f(\underbrace{x_1^{k+1}, \cdots, x_{i-1}^{k+1}}_{\text{already done}}, \underbrace{\xi}_{\text{current coordinate}}, \underbrace{x_{i+1}^{k}, \cdots x_n^{k}}_{\text{haven't done}})$$

  end

end

---

initial: $x^0 = [x_1^0, x_2^0, x_3^0, \cdots x_n^0]$

$\downarrow$

$[x_1^1, x_2^0, x_3^0, \cdots, x_n^0]$

$\downarrow$

$[x_1^1, x_2^1, x_3^0, \cdots, x_n^0]$

$\downarrow$

$\vdots$

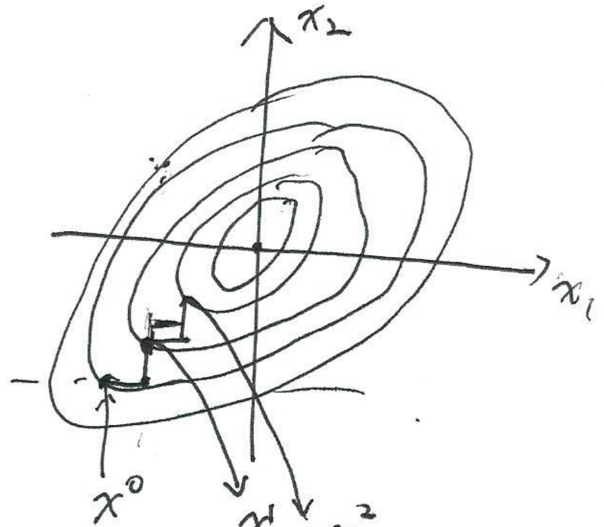
$x^1 = [x_1^1, x_2^1, x_3^1, \cdots, x_n^1]$

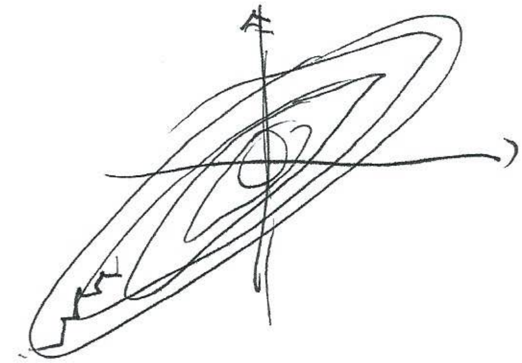
$\downarrow$

$[x_1^2, x_2^1, x_3^1, \cdots, x_n^1]$

$\downarrow$

$[\quad]$





- CD for constrained minimization

$$\min_x f(x) \quad \text{st} \quad x \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$$

$$\Downarrow$$

$$(x_1 \in \mathcal{X}_1, \; x_2 \in \mathcal{X}_2, \; \cdots, \; x_n \in \mathcal{X}_n)$$

For $k=0,1,\cdots$

For $i=1,2,\cdots,n$.

$$x_i^{k+1} \leftarrow \operatorname*{argmin}_{\xi \in \mathcal{X}_i} f(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, \xi, x_{i+1}^{k}, \cdots, x_n^{k})$$

end

end.

- Block Coordinate Descent $\Rightarrow$ each x_i can be a d_i-dimensional vector)

- One ~~up~~ variable update, equivalent to:

$$\left\{ \begin{array}{l} \delta \leftarrow \operatorname*{argmin}_{\delta : x_i + \delta \in \mathcal{X}_i} f(x + \delta e_i) \\ x_i \leftarrow x_i + \delta \end{array} \right.$$

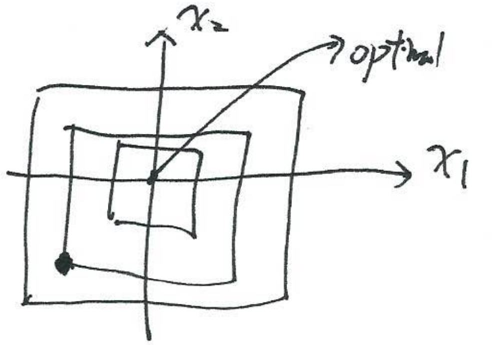
x: current solution

$e_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ ← i-th element.

~~Does it~~

Does it converge to optimal solution?

Ex 1: $\min_{x} f(x) = \max(|x_1|, |x_2|)$

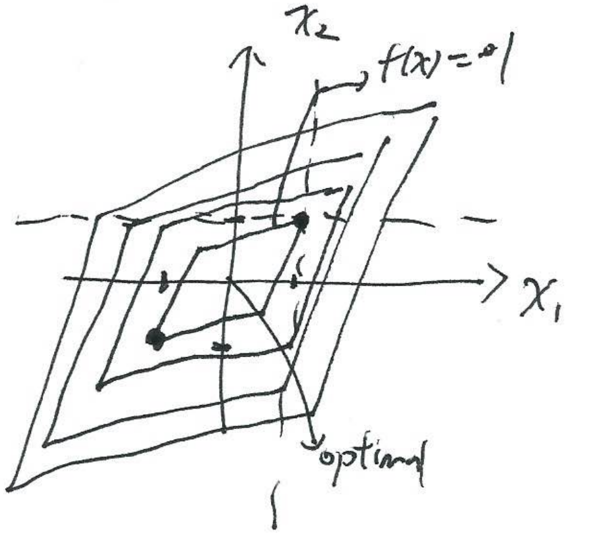


$x^0 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$: $f(x^0) = 1$

update 1 variable

$x = \begin{bmatrix} \delta \\ -1 \end{bmatrix}$ $f(x) = \max(|\delta|, 1) \geq 1$

Ex 2: $\min_{x} f(x) = \frac{1}{2}|x_1 + x_2| + |x_1 - x_2|$



$x^0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

cannot improve by changing one coordinate.

Thm: If f is ① continuous differentiable over X

② The minimum of

$$\min_{\xi} f(x_1, \cdots, x_{i-1}, \xi, x_{i+1}, \cdots, x_n)$$

is uniquely attained, $\forall i$, $\forall x$

Then every limit point of $\{x^k\}_{k=1}^{\infty}$ is a stationary point.

(Convex $\Rightarrow$ stationary point = global optimizer)

Thm: If f is continuously differentiable and there are only two blocks.

$\Rightarrow$ Every limit points are stationary points.

Examples:

— Linear regression.

$$\min_{x \in R^n} \left\{ \frac{1}{2} \|Ax - y\|^2 \right\} = f(x)$$

$$A = \overbrace{\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}}^{n} \Big\}\, m = \begin{bmatrix} \text{—}\ \bar{a}_1\ \text{—} \\ \vdots \\ \text{—}\ a_m\ \text{—} \end{bmatrix}$$

$$\min_{\delta} f(x + \delta e_i)$$

Define $g(\delta) = f(x + \delta e_i)$

$$g'(\delta) = \nabla_i f(x + \delta e_i) \quad \text{(want to be 0)}$$

$$= (A^T(A(x + \delta e_i) - y))_i$$

$$= (A^T(A(x - y) + A^T A \delta e_i))_i$$

$$= \left(\begin{bmatrix} \text{—}\ a_1\ \text{—} \\ \text{—}\ a_2\ \text{—} \\ \vdots \\ \text{—}\ a_n\ \text{—} \end{bmatrix} (Ax - y) \right)_i + \left(\begin{bmatrix} \text{—}\ a_1\ \text{—} \\ \text{—}\ a_2\ \text{—} \\ \vdots \\ \text{—}\ a_n\ \text{—} \end{bmatrix} a_i \right)$$

$$= a_i^T (Ax - y) + a_i^T a_i \cdot \delta = 0$$

$$\delta^* = a_i^T (y - Ax) / \|a_i\|^2$$

For $k=0,1,\cdots$

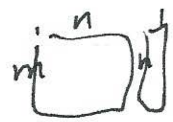
For $i=1,\cdots,n$

$$x_i \leftarrow x_i + \underbrace{\frac{a_i^T(y-Ax)}{\|a_i\|^2}}$$

end

end.

compute $Ax \Rightarrow O(mn)$



compute $y-Ax \Rightarrow O(m)$

$a_i^T(y-Ax) \Rightarrow O(m)$

$\|a_i\|^2 \Rightarrow O(m)$

$\Rightarrow$ Each inner iteration: $O(mn)$

Initial x^0, $r = y - Ax^0$

For $k=0,1,\cdots$

For $i=1,\cdots,n$

$$\delta = \frac{a_i^T(r)}{\|a_i\|^2}$$

$$x_i \leftarrow x_i + \delta$$

$$\underbrace{r \leftarrow r - \delta a_i}$$

end

end

$O(m)$ $O(m)$

Each inner iteration $\to O(m)$

Each outer iteration $O(mn)$

$$\text{GD: } x \leftarrow x - \nabla f(x) = x - A^T(\underbrace{Ax}_{O(mn)} - y)$$

Same time complexity

$$x^k = \begin{bmatrix} x_1^k \\ x_2^k \\ \vdots \\ x_n^k \end{bmatrix} \xrightarrow{GD} \begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \\ \vdots \\ x_n^{k+1} \end{bmatrix}$$

$$\begin{bmatrix} x_1^k \\ x_2^k \\ \vdots \\ x_n^k \end{bmatrix} \rightarrow \cdots \rightarrow \begin{bmatrix} x_1^{k+1} \\ x_2^{k+1} \end{bmatrix}$$

Advantages:

- Each iteration is cheap (& simple)
- No step size tuning ☆

Disadvantages:

- Convergence proof is hard.
- Slow in Matlab or R. ☆
- Slow when near optimal. (depends on function)

3. Other variations:

3.1 Different ways to select variables:

General Form of CD:

For $k=0,1,\cdots$
- Pick an index $i \in \{1,\cdots,n\}$
- ~~$x_i \leftarrow \arg$~~ Update x_i (by solving the one-variable problem)

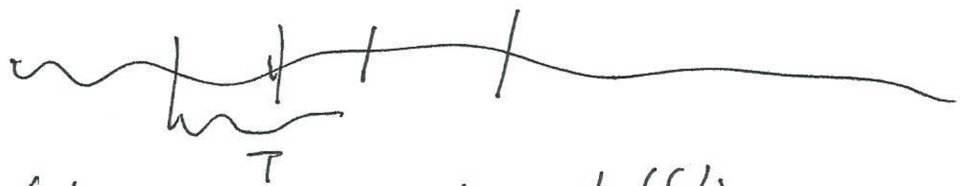
end

① Cyclic order: (in practile: very bad)

$i = 1, 2, \cdots n, 1, 2, \cdots n$

② Almost cyclic order:

each coordinate picked at least once within every $T_{>0}$ iterations



Useful example: random shuffling

For $k = 0, 1, \cdots$

  Generate a random permutation π

  For $i = 1, \cdots, n$

    Update $x_{\pi(i)}$

  end

end

③ Random Sampling (Stochatic Coordinate Descent)

For $k = 0, 1, \cdots$

  random sample $i \in \{1, \cdots, n\}$

  Update x_i

end

④ Greedy Coordinate Descent (Gauss-Southwell)

For $k = 0, 1, \cdots$

  $i \leftarrow \arg\max_i |\nabla_i f(x)|$

  Update x_i

end

Example: Quadratic minimization with constraints.

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}x^TQx + b^Tx \quad \text{st} \quad \overset{(x>0)}{x \geq 0} \quad (x_i \geq 0 \ \forall i)$$

Stochatic CD:

$$\underset{\delta}{\text{argmin}} \ f(x + \delta e_i) \quad \text{st} \quad x_i + \delta \geq 0 \qquad \boxed{\delta \geq -x_i}$$

$$e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$= \frac{1}{2}(x+\delta e_i)^T Q (x + \delta e_i) + b^T(x + \delta e_i)$$

$$= \frac{1}{2}x^TQx + x^TQ\delta e_i + \frac{1}{2}\delta^2 e_i^TQe_i + b^Tx + \delta b^Te_i$$

$= Q_{ii}$

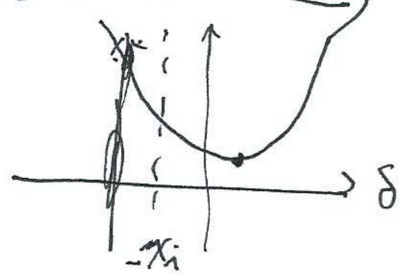
$$= \frac{1}{2}Q_{ii}\delta^2 + (x^Tq_i)\delta + b_i\delta + \text{const}$$

$$= \frac{1}{2}Q_{ii}\delta^2 + (b_i + x^Tq_i)\delta + \text{const}.$$

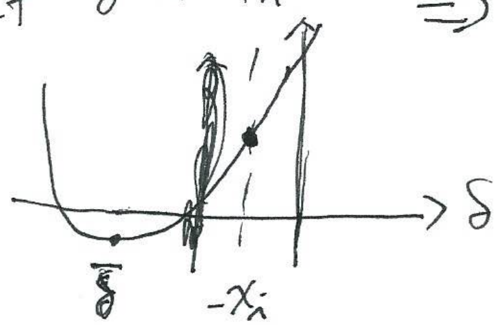
If no constraint:

$$\delta = -(b_i + x^Tq_i)/Q_{ii} = \bar{\delta}$$

If $\bar{\delta} \geq -x_i \Rightarrow \delta^* = \bar{\delta}$



If $\bar{\delta} < -x_i \Rightarrow \delta^* = -x_i$



SCD for solving this

For $k=0,1,\dots$

randomly choose $i \in \{1,\dots n\}$

Compute $\bar{\delta} = -(b_i + x^T q_i)/Q_{ii}$

If $\bar{\delta} \geq -x_i$

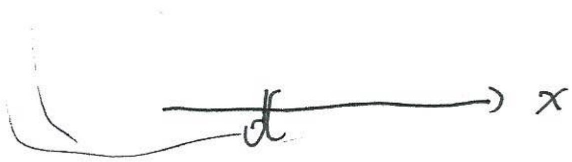
$x_i \leftarrow x_i + \bar{\delta}$

If $\bar{\delta} < -x_i$

$x_i \leftarrow x_i - x_i = 0$

end

$\min\limits_{x} x^2$ st $x > 0$ $(x \geq 0)$



Convergence for stochastic CD. $(x^k$

Thm: If f is twice differentiable and (actually $L = \max_i \nabla_{ii}^2 f(x)$ $\forall x$)

$\checkmark$ $mI \preceq \nabla^2 f(x) \preceq LI$, $m, L > 0$, $\forall x$

then $\{x^k\}_{k=1}^{\infty}$ generated by SCD:

$$E[f(x^{k+1})] - f(x^*) \le \left(1 - \frac{m}{nL}\right)\left(E[f(x^k)] - f(x^*)\right)$$

Iter $k+1$ — iteration k

Pf: Strategy: ϕ^k, ϕ^{k+1}

① $E[f(x^k) - f(x^{k+1})] \ge \frac{1}{2Ln} E[\|\nabla f(x^k)\|^2]$

② $f(x^k) - f(x^*) \le \frac{1}{2m}\|\nabla f(x^k)\|^2$.

① $g(\delta) = f(x + \delta e_i)$, $\left(\begin{array}{l}\delta^* = \arg\min_\delta g(\delta) \\ x_i^{k+1} = x_i^k + \delta^*\end{array}\right)$

$g'(\delta) = \nabla_i f(x + \delta e_i)$

$g''(\delta) = \nabla_{ii}^2 f(x + \delta e_i) \le L$

$\Rightarrow m \le g''(\delta) \le L$

Taylor expansion.

$$g(\delta) \le g(0) + g'(0)\cdot\delta + \frac{L}{2}\delta^2 \qquad \left(\begin{array}{l} g'(0) + L\delta = 0 \\ \delta = -\frac{g'(0)}{L}\end{array}\right)$$

take $\tilde{\delta} = -\frac{g'(0)}{L}$

$$g(\delta^*) \le g(\tilde{\delta}) \le g(0) - \frac{g'(0)^2}{L} + \frac{L}{2}\left(\frac{g'(0)^2}{L^2}\right) = g(0) - \frac{g'(0)^2}{2L}$$

$$\rightarrow f(x^{k+1}) \le f(x^k) - \frac{(\nabla_i f(x^k))^2}{2L}$$

$$\underbrace{E_i\left[f(x^{k}) - f(x^{k+1})\right]} \geq E_i\left[\frac{\nabla_i f(x^k)^2}{2L}\right]$$

$$= \frac{1}{2L}\cdot\left(\frac{1}{n}\sum_i \nabla_i f(x^k)^2\right)$$

$$\underbrace{= \frac{1}{2Ln}\cdot \|\nabla_i f(x^k)\|^2}$$

② Taylor Expansion

$$f(y) \geq \underbrace{f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2}_{\substack{\shortparallel \\ h(y)}} \quad \forall y, x$$

$$\tilde{y} = \underset{y}{\operatorname{argmin}}\ h(y) \Rightarrow \nabla f(x) + m(\tilde{y} - x) = 0$$

$$\tilde{y} = x - \frac{1}{m}\nabla f(x)$$

$$\cancel{h(\tilde{y}) \Rightarrow}\ h(\tilde{y}) = f(x) - \frac{\|\nabla f(x)\|^2}{m} + \frac{\|\nabla f(x)\|^2}{2m}$$

$$= f(x) - \frac{\|\nabla f(x)\|^2}{2m}$$

$$f(x^*) \geq h(x^*) \geq h(\tilde{y}) = f(x) - \frac{\|\nabla f(x)\|^2}{2m}$$

$$\Rightarrow f(x^k) - f(x^*) \leq \frac{\|\nabla f(x)\|^2}{2m}$$

Combining ①, ②,

$$\phi^k = E[f(x^k)],\ \phi^* = f(x^*),$$

$$\phi^{k+1} - \phi^* = \phi^{k+1} - \phi^k + \phi^k - \phi^*$$

$$\leq -\frac{1}{2Ln}E[\|\nabla f(x^k)\|^2] + \phi^k - \phi^*$$

$$\leq -\frac{1}{2Ln}\cdot 2m\cdot\big(\underbrace{E[f(x^k)]}_{\phi^k} - \underbrace{f(x^*)}_{\phi^*}\big) + \phi^k - \phi^*$$

$$= \left(1 - \frac{m}{Ln}\right)(\phi^k - \phi^*) \ \#$$

$GD: f(x^{k+1}) - f(x^*) \le (1 - \frac{m}{L})(f(x^k) - f(x^*))$

$SCD: E[f(x^{k+1})] - f(x^*) \le (1 - \frac{m}{nL})(E[f(x^k)] - f(x^*))$

n iterations of SCD $\approx$ 1 iteration for GD

$$\left(1 - \frac{m}{nL}\right)^n \underset{\ge}{\approx} \left(1 - \frac{m}{L}\right)$$

$$\ge 1 - n\frac{m}{nL} = 1 - \frac{m}{L} \quad // \quad \Rightarrow CD \approx GD$$

if $a < 1$

$(1-a)^n \ge 1 - na$

$(1-0.5)^2 =$ $(1-a)^2 = 1 - 2a + a^2$

$E_{i^{k+1}}[f(x^k) - f(x^{k+1})] \ge O \cdot (f(x^k) - f^*)$

$E_{i^{k+1}} E_{i^k} E_{i^{k-1}} \cdots E_{i^0} (f(x) - f^*)$

$\quad //$

$E(f(x^k))$