Coordinate Descent Method

10:09 AM

Coordinate Descent Method:

$$\begin{pmatrix} \underset{x}{\min}\ f_0(x) \\ \text{s.t.} \\ \forall_{i\in\{1,\dots,\nu\}}\ x_i \in X_i \end{pmatrix} \quad \text{eq: 12.90} \qquad \{x = (x_1, \dots, x_\nu)\}$$

blocks of vectors

Coordinate descent method iteratively minimizes with respect to one block, while fixing the other.

$x^{(k)} = (x_1^{(k)}, \dots, x_\nu^{(k)})$ {value of the decision variable at iteration k}

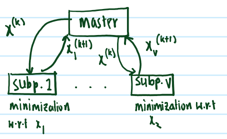
$$\underset{x_i \in X_i}{\min}\ f_0(\underbrace{x_1^{(k)}, \dots, x_{i-1}^{(k)}}_{\text{fixed}}, \underbrace{x_i}_{\text{variable}}, \underbrace{x_{i+1}^{(k)}, \dots, x_\nu^{(k)}}_{\text{fixed}}) \quad \text{is solved for } i \in \{1, \dots, \nu\}$$

[eq: Cord Dscnt Mnmzn for x_i at itrtn k]

*$\boxed{\text{The Jacobi method:}}$

[eq: Cord Dscnt Mnmzn for x_i at itrtn k] is solved, then all blocks are updated simultaneously.

$$\forall_{i\in\{1,\dots,\nu\}} \quad x_i^{(k+1)} = \underset{x_i \in X_i}{\operatorname{argmin}}\ f_0(x_1^{(k)}, \dots, x_{i-1}^{(k)}, x_i, x_{i+1}^{(k)}, \dots, x_\nu^{(k)})$$



Jacobi Algorithm

Convergence of coordinate descent is not guaranteed in general

* converges to the optimal solution under certain hypotheses of contractivity.

F {function: $Y \mapsto Y$} {contraction mapping} $\overset{\Delta}{=} \exists_{\rho\in[0,1)} \forall_{y,z\in Y} \dfrac{\|Fy - Fz\|}{\|y - z\|} \leq \rho$

Y {real vector space} {norm defined $\|\cdot\|$}

ρ: modulus of contraction

* Convergence theorem for Jacobi method

$\left(f_0\ \{\mathcal{C}_2^{\text{continuously}}\},\ F(\boxdot) = \boxdot - \gamma \nabla f_0(\boxdot)\ \{\text{contraction mapping with norm } \|\boxdot\| = \max_{i=1,\dots,\nu} \frac{\|\boxdot_i\|_2}{w_i}\}\right) \Rightarrow$ ([eq: 12.90](#)) has a unique optimum $\wedge$ $\lim_{k\to\infty} x^{(k)} = x^*$ {geometric rate}

{$\exists\ \gamma > 0$} w_i {$\in \mathbb{R}_{++}$}

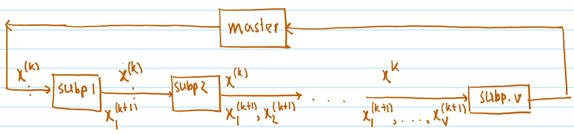
convergence rate γ_A log is geometric

*Block coordinate minimization method: (BCM)

(Also known as Gauss-Seidel method)

At each iteration blocks are updated sequentially:

$$\forall_{i\in\{1,\dots,\nu\}} \quad x_i^{(k+1)} = \underset{x_i \in X_i}{\operatorname{argmin}}\ f_0(x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_{i-1}^{(k+1)}, x_i, x_{i+1}^{(k)}, \dots, x_\nu^{(k)})$$



*$\underbrace{\text{Convergence condition}}$ for Jacobi iteration will work for BCM too

$(\boxdot - \gamma \nabla f_0(\boxdot))$ {contractive under norm $\|\boxdot\| = \max_{i=1,\dots,\nu} \|x_i\|_2 / w_i$}

$(\exists\ \gamma > 0)$ $w_i \in \mathbb{R}_{++}$

[Convergence Criterions for Block Coordinate Descent Method]

BCM will converge for following condition too:

*

$\left(f_0\ \{D, \mathcal{C}_x^{\text{continuously}}, D_{x_i}^{\text{strictly}}\ \{\text{with other blocks constant}\}\},\ \{x^{(k)}\}_{\{\text{generated by BCM}\}}\ \{\text{well-defined}\}\right) \Rightarrow \lim_{k\to\infty} x^{(k)} = x^*$

*

BCM in generally may fail to converge for non-smooth objectives, even under convexity.

An important exception for which BCM converges:

$f_0(x) = \underbrace{\phi(x)}_{\text{differentiable in } x, D} + \sum_{i=1}^{\nu} \psi_i(x_i)$ — each $\psi_i(x_i)$ {D, maybe nonsmooth}

independent convex constraint $x_i \in X$ can be

written as $I_{X}(x_i) = \psi_i(x_i)$, so $\underset{x_i\in X}{\min} f_0(x)$ can

be covered in this setup

in $x, \mathcal{D}$ # be covered in this setup $x_i \in \mathcal{X}$

* Theorem 12.5.

$$\left(x^{(0)} \{ \in \mathcal{X}, \text{[initial point for BCM]} \}, \; \underbrace{S_0 = \{ x : f_0(x) \le f_0(x^0) \}}_{\text{[sublevel set at } f_0(x^0)]} \{ [\le] \}, \; f_0(x) = \underbrace{\phi(x)}_{\{\mathcal{D}, \, 6_\phi \text{ on } S_0\}} + \underbrace{\sum_{i=1}^{N} \psi_i(x_i)}_{\{\mathcal{D}\}} \right) \Rightarrow \lim_{k \to \infty} x^{(k)} = x^*$$

Theorem 12.5 applies to Lasso problem or other ℓ_1 norm regularized problem, thus guranting convergence of BCM in those cases.

* Power iteration and block coordinate descent method:

- Power iteration {class of methods applied to specific eigenvalue, singular value problems}

  = {coordinate descent}

Problem: Finding rank-one approximation of a matrix A:

$$\underset{x,y}{\gamma} \; \|A - xy^T\|_F^2 \qquad \# \; \|A\|_F^2 = \sum_{i=1}^{n} \sigma_i^2$$

this is convex in x because, $\|X\|_F^2 \; \mathcal{D} \xrightarrow[\text{''}]{\square + A} \|X + \square\|_F^2 \; \mathcal{D}$
however if you consider x, y seperately then it is bilinear and not convex in (x, y)

A stationary point in x an y seperately (where the derivative is 0) finding:

$$\|A - xy^T\|_F = \operatorname{tr}\left(\underbrace{(A - xy^T)(A - xy^T)^T}_{} \right) \qquad [\because \|X\|_F = \sqrt{\operatorname{tr}(AA^T)}]$$

$$\# \; (A - xy^T)(A^T - (xy^T)^T)$$
$$= (A - xy^T)(A^T - yx^T)$$
$$= AA^T - Ayx^T - xy^TA^T + xy^Tyx^T$$
$$= AA^T - Ayx^T - xy^TA^T + \|y\|_2^2 \, xx^T \; \#$$

$$= \operatorname{tr}(AA^T - Ayx^T - xy^TA^T + \|y\|_2^2 xx^T)$$

[eq: expression of ||A-x y^T||_F^2 in terms trace]

$\operatorname{tr}(x^Tx) = \operatorname{tr}(\|x\|_2^2) = \|x\|_2^2$

$$= \operatorname{tr}(AA^T) - \operatorname{tr}(Ayx^T) - \operatorname{tr}(xy^TA^T) + \|y\|_2^2 \operatorname{tr}(xx^T) \qquad [\because \operatorname{tr} \text{ is a linear operator}]$$

$$\nabla_x \|A - xy^T\|_F^2 = \nabla_x (\operatorname{tr}(AA^T)) - \nabla_x \operatorname{tr}(Ayx^T) - \nabla_x \operatorname{tr}(xy^TA^T) + \|y\|_2^2 \nabla_x \|x\|_2^2$$

$$\# \; \frac{\partial}{\partial X} \operatorname{tr}(AX^T) = \frac{\partial}{\partial X} \operatorname{tr}(X^TA) = A$$
$$\frac{\partial}{\partial X} \operatorname{tr}(XA) = A^T \; \#$$

→ # mnemonic: transpose 있으면 transpose 없애기
〃 없으면 transpose 생기게.

$$= 0 - Ay - \underbrace{(y^TA^T)^T}_{Ay} + \|y\|_2^2 \cdot 2x$$

$$= -2Ay + \|y\|_2^2 \, 2x$$

For finding a stationary point in x:

$$\nabla_x \|A - xy^T\|_F^2 = -2Ay + \|y\|_2^2 \, 2x = 0$$

$$\Leftrightarrow 2\|y\|_2^2 x = 2Ay$$

$$\Leftrightarrow \boxed{x\|y\|_2^2 = Ay}$$

[eq: stationary point 1]

$$\text{Similarly, } \nabla_y \|A - xy^T\|_F^2 = \underbrace{\nabla_y(\operatorname{tr}(AA^T))}_{=0} - \nabla_y \overbrace{\operatorname{tr}(Ayx^T)}^{\operatorname{tr}(x^TAy)} - \nabla_y \overbrace{(xy^TA^T)}^{\operatorname{tr}(y^TA^Tx)} + \|x\|_2^2 \nabla_y \|y\|_2^2$$

$$= -\nabla_y(\operatorname{tr}(x^TAy)) - \nabla_y(y^TA^Tx) + \|x\|_2^2 \nabla_y \|y\|_2^2$$

$$\# \; \frac{\partial}{\partial X} \operatorname{tr}(AX^T) = \frac{\partial}{\partial X} \operatorname{tr}(X^TA) = A$$
$$\frac{\partial}{\partial X} \operatorname{tr}(XA) = \frac{\partial}{\partial X} \operatorname{tr}(AX) = A^T \; \#$$

$$= -(x^TA)^T - A^Tx + \|x\|_2^2 \, 2y = -2A^Tx + 2\|x\|_2^2 y$$

Finding a stationary point in y:

$$\nabla_x \|A - xy^T\|_F^2 = -2A^Tx + 2\|x\|_2^2 y = 0$$

$$\Leftrightarrow \boxed{\|x\|_2^2 y = A^Tx}$$

[eq: stationary point 2]

$$\left. \begin{aligned} x\|y\|_2^2 &= Ay, \\ y\|x\|_2^2 &= A^Tx \end{aligned} \right\}$$

lets normalize the vectors as follows: $u = \dfrac{x}{\|x\|_2}$, $v = \dfrac{y}{\|y\|_2}$, $\sigma = \|x\|_2 \|y\|_2$

$$\text{then } xy^T = u \|x\|_2 v^T \|y\|_2 = \sigma uv^T$$

then: $$\left(\left(\underset{x,y}{\gamma} \; \|A - xy^T\|_F^2 \right) = \left(\underset{u,v,\sigma}{\gamma} \; \|A - \sigma uv^T\|_F^2 \right) \right.$$
$\quad \gamma \quad \|u\|_2 = 1$

[problem: dyadic version of rank 1 approximation problem] σ

then: $\left(\begin{array}{l} \min\limits_{x,y} \|A - xy^T\|_F^2 \end{array} \right) = \left(\begin{array}{l} \min\limits_{u,v,\sigma} \|A - \sigma u v^T\|_F^2 \\ \text{s.t.} \\ \quad \|u\|_2 = 1 \\ \quad \|v\|_2 = 1 \\ \quad \sigma \geq 0 \end{array} \right)$ [problem: dyadic version of rank 1 approximation problem]

$$\|x\|_2^2 \|y\|_2^2 = (\underbrace{\|x\|_2 \|y\|_2}_{\sigma})^2 = \sigma^2$$

$$\mathrm{tr}(x^T x) = \mathrm{tr}(\|x\|_2^2) = \|x\|_2^2$$

now note that in σ this is convex $\quad x = \|x\|_2 u \quad y = \|y\|_2 v$

[eq: expression of | |A-x y^T থেকে: $\|A - xy^T\|_F^2 = \mathrm{tr}(AA^T) - \mathrm{tr}(A y x^T) - \mathrm{tr}(x y^T A^T) + \|y\|_2^2 \, \mathrm{tr}(x x^T)$

$$= \mathrm{tr}(AA^T) - \mathrm{tr}(A \|y\|_2 v \, \|x\|_2 u^T) - \mathrm{tr}(\|x\|_2 u \, \|y\|_2{}^T v^T A^T) + \sigma^2$$

$$= \mathrm{tr}(AA^T) - \mathrm{tr}(A \sigma v u^T) - \mathrm{tr}(\sigma u v^T A^T) + \sigma^2$$

$$= \mathrm{tr}(AA^T) - \sigma \left(\underbrace{\mathrm{tr}(A \sigma v u^T) + \mathrm{tr}(u v^T A^T)}_{\mathrm{tr}(A v u^T + u v^T A^T)} \right) + \sigma^2$$

// tr(·) is linear operator

$$\nabla_\sigma \|A - xy^T\|_F^2 = 0 - \mathrm{tr}(A v u^T + u v^T A^T) + 2\sigma = 0$$

$\therefore \sigma = \frac{1}{2} \underbrace{\mathrm{tr}(A v u^T + u v^T A^T)}$ $= \frac{1}{2} 2 u^T A v = u^T A v$, plugging this value ([problem: dyadic version of rank 1 approximation problem) we get $\left(\begin{array}{l} \min\limits_{u,v,\sigma} \|A - \sigma u v^T\|_F^2 \\ \text{s.t.} \\ \quad \|u\|_2 = 1, \ \|v\|_2 = 1 \end{array} \right)$

\# $\mathrm{tr}(A v u^T) + \mathrm{tr}(u v^T A^T)$

$= \mathrm{tr}(u^T A v) + \mathrm{tr}(v^T A^T u)$

\# $(u^T (A v))^T = (Av)^T (u^T)^T = v^T A^T u$

$= u^T A v + (u^T A v)^T = u^T A v + u^T A v = 2 u^T A v$ #

\#number so trace

\# [illegible]

this is a number so transpose will be itsels