LEC6 DSA5103

↓

Coordinate Descent Algorithm

Recap: { ① Proximal Operator for $\begin{cases} \|\cdot\|_1 \\ \delta_H(\cdot) \end{cases}$
② Block Coordinate Descent Algo for SVM (SMO Algo)

① Proximal Operator

$$P_f(x) := \underset{y \in \mathbb{R}^n}{\operatorname{argmin}} \; f(y) + \frac{1}{2}\|y-x\|_2^2$$

{ unique existed

a) $f = \delta_C = \begin{cases} 0, & x \in C \\ +\infty, & x \notin C \end{cases}$

property:
$u = P_f(x) \Leftrightarrow x - u \in \partial f(u)$

then $P_{\delta_C}(x) = \underset{y \in \mathbb{R}^n}{\operatorname{argmin}} \; \delta_C(y) + \frac{1}{2}\|y-x\|_2^2$

$= \underset{y \in C}{\operatorname{argmin}} \; \frac{1}{2}\|y-x\|_2^2$

$= \Pi_C(x)$

then $\psi_{\delta_C}(x) = \frac{1}{2}\|\Pi_C(x) - x\|_2^2$

b) $f = \|\cdot\|_1$

then $P_{\|\cdot\|_1}(x) = \underset{y \in \mathbb{R}^n}{\operatorname{argmin}} \; \|y\|_1 + \frac{1}{2}\|y-x\|_2^2$

How to achieve this? { ① Through $u = P_f(x) \Leftrightarrow x - u \in \partial f(u)$
② Moreau-Yosida Decomposition

approach 2: $x = P_f(x) + P_{f^*}(x)$

f is positive homogeneous $\Rightarrow f^*(x) = \delta_{\partial f(0)}(x)$

$\Rightarrow P_{f^*}(x) = \Pi_{\partial f(0)}(x)$

$x \in \partial f(0) \Leftrightarrow f(y) \geq x^T(y-0) + f(0) \quad \forall y$

$\Leftrightarrow f(y) \geq \langle x, y \rangle \quad \forall y$

$\Leftrightarrow \|y\|_1 \geq \langle x, y \rangle \quad \forall y \text{ s.t. } \|y\|_1 = 1$

$\Leftrightarrow \quad \langle x, y \rangle \le 1 \quad \forall y \quad s.t \quad \|y\|_1 = 1$

$\Leftrightarrow \quad \sup_{y} \{ \langle x, y \rangle : \|y\|_1 = 1 \} \le 1$

$\Leftrightarrow \quad \|x\|_\infty \le 1 .$

$\Rightarrow P_{f^*}(x) = \Pi_{B_\infty^1}(x)$

$\Rightarrow \boxed{P_f(x) = x - \Pi_{B_\infty^1}(x)} = \begin{cases} x-1 & x \ge 1 \\ 0 & -1 \le x \le 1 \\ x+1 & x \le -1 \end{cases}$

Generalization. $\boxed{P_{\lambda \|\cdot\|_1}(x) = x - \Pi_{B_\infty^\lambda}(x)} = \begin{cases} x-\lambda & x \ge \lambda \\ 0 & -\lambda \le x \le \lambda \\ x+\lambda & x \le -\lambda \end{cases}$

② Block Coordinate Descent

Dual Formulation of SVM

$$\begin{cases} \min\limits_{\alpha} & \frac{1}{2} \sum\limits_i \sum\limits_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum\limits_{i=1}^{n} \alpha_i \\ s.t & \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \le \alpha_i \le C \quad i \in [n] \end{cases}$$

① choose (i,j) pair

② update $(\hat{\alpha}_i, \hat{\alpha}_j) = \begin{cases} \underset{\alpha_i, \alpha_j}{\text{argmin}} & \frac{1}{2} K_{ii} \alpha_i^2 + \frac{1}{2} K_{jj} \alpha_j^2 + K_{ij} y_i y_j \alpha_i \alpha_j \\ & - \alpha_i - \alpha_j \\ s.t & y_i \alpha_i + y_j \alpha_j = C \, (\hat{\alpha}_{-i,-j}) \\ & 0 \le \alpha_i \le C \\ & 0 \le \alpha_j \le C \end{cases}$

→ closed-form solution

③ $(\alpha_i, \alpha_j) \leftarrow (\hat{\alpha}_i, \hat{\alpha}_j)$

→ It seems that, we are interested in solutions like

$$\tilde{x}_i = \underset{x_i}{\text{argmin}} \; f(x_{-i}, x_i) \quad (\text{fix } x_{-i}) \quad \forall i$$

(coordinate-wise minimizer)

Today's content

1. Coordinate-wise Minimizer
   - defn
   - in which condition, it can actually be global minimizer?

① Defn:

$\bar{x}$ is coordinate-wise minimizer

$\Leftrightarrow f(\bar{x} + d e_i) \geq f(\bar{x})$ $\quad \forall d \in \mathbb{R}, \ \forall i \in [n]$

Alternative defn:

$$\bar{x}_i \in \underset{x_i}{\operatorname{argmin}} \ f(x_i, \bar{x}_{-i}) \qquad \forall i \in [n]$$

② Question:

Global Minimizer

⟷ relation?

Coordinate-wise Minimizer

E.g. $f(x_1, x_2) = x_1^2 + x_2^2 + 2|x_1 - x_2|$

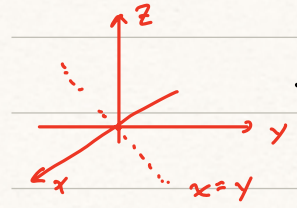
a) Coordinate-wise minimizer

$\bar{x}_1 \in \underset{x_1}{\operatorname{argmin}} \ f(x_1, \bar{x}_2)$

$= \underset{x_1}{\operatorname{argmin}} \ x_1^2 + 2|x_1 - \bar{x}_2|$ (convex)

$\Leftrightarrow 0 \in 2\bar{x}_1 + 2 \partial g_{\bar{x}_2}(\bar{x}_1)$

$g(x,y) = |x-y|$ ⇓ convex!



property:

f, g convex $\Rightarrow f+g$ convex

Pf: $\forall x, y, \lambda$

$(f+g)(\lambda x + (1-\lambda) y)$

$= f(\lambda x + (1-\lambda) y) + g(\lambda x + (1-\lambda) y)$

$\leq \lambda (f+g)(x) + (1-\lambda)(f+g)(y)$ #

Here $g_{\bar{x}_2}(x) := |x - \bar{x}_2| \Rightarrow \partial g_{\bar{x}_2}(\bar{x}_1) = \begin{cases} [-1,1] & \bar{x}_1 = \bar{x}_2 \\ -1 & \bar{x}_1 < \bar{x}_2 \\ 1 & \bar{x}_1 > \bar{x}_2 \end{cases}$

$$\Rightarrow 2\bar{x}_1 + 20\,\partial g_{\bar{x}_2}(\bar{x}_1) = \begin{cases} [-20 + 2\bar{x}_2,\ 20 + 2\bar{x}_2] & \bar{x}_1 = \bar{x}_2 \\ -20 + 2\bar{x}_1 & \bar{x}_1 < \bar{x}_2 \\ 20 + 2\bar{x}_1 & \bar{x}_1 > \bar{x}_2 \end{cases}$$

→ Suppose $\bar{x}_1 > \bar{x}_2$ then $\bar{x}_1 = -10$ so that $0 \in 2\bar{x}_1 + 20\,\partial g_{\bar{x}_2}(\bar{x}_1)$

However, when $\bar{x}_2 < \bar{x}_1$, $2\bar{x}_2 + 20\,\partial g_{\bar{x}_1}(\bar{x}_2)$

$= 20 + 2\bar{x}_2$

$< 20 + 2\bar{x}_1$

< 0

which means $(\bar{x}_1, \bar{x}_2)$ cannot contribute to "Coordinate-wise minimizer"

→ Suppose $\bar{x}_1 = \bar{x}_2$, then we require:

$-10 \le \bar{x}_1 = \bar{x}_2 \le 10$

$\Rightarrow \{(\bar{x}_1, \bar{x}_2) : -10 \le \bar{x}_1 = \bar{x}_2 \le 10\}$ is "Coordinate-wise Minimizer"

b) Global Minimizer

obviously, $(0,0)$ is global minimizer

(From above example) → Personal Understanding

Analysis: Why Coordinate-wise minimizer cannot be Global minimizer?

$f(x_1, x_2) = x_1^2 + x_2^2 + 20|x_1 - x_2|$

$g(x) = h(x_1 - x_2)$

consider $\partial f(x) = \nabla_x (x_1^2 + x_2^2) + 20\,\partial g(x)$ ← $h(z) = |z|$

$= \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} + 20 \begin{pmatrix} 1 \\ -1 \end{pmatrix} \cdot \partial h(x_1 - x_2)$

$= \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} + \begin{pmatrix} 20 \\ -20 \end{pmatrix} \cdot \begin{cases} [-1,1] & x_1 = x_2 \\ -1, & x_1 < x_2 \\ 1, & x_1 > x_2 \end{cases}$

if $-10 \le x_1 = x_2 \le 10$, $\longrightarrow$ coordinate-wise minimizer

then $\partial f(x) = \{(2x_1 + 20\lambda, 2x_2 - 20\lambda) : \lambda \in [-1, 1]\}$

$= \{(2a + 20\lambda, 2a - 20\lambda) : \lambda \in [-1, 1], a = x_1\}$

Let's check whether $0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \partial f(x)$

if so, then there exists $\bar{\lambda} \in [-1, 1]$, s.t

$$\begin{cases} 0 = 2a + 20\lambda \\ 0 = 2a - 20\lambda \end{cases} \Rightarrow \begin{cases} \lambda = 0 \\ a = 0 = x_1 = x_2 \end{cases}$$

$\Rightarrow$ only global minimizer

Moreover, for a general function $f(\cdot)$. the main issue is:

if $\bar{x}$ is a Coordinate-wise minimizer

$\longrightarrow$ then suppose some regularity condition $\rightarrow$ differentiable

we can achieve $\nabla_i f(\bar{x}) = 0 \quad \forall i \in [n]$

$\Updownarrow$

(intuition) $\nabla f(\bar{x}) = 0 \Rightarrow \bar{x}$ global minimizer

$\longrightarrow$ However, if $f(x)$ is not smooth

then we cannot control the change of function value in all direction even if function value is well-shaped in $x(y)$ direction

$\rightarrow$ Main Issue comes from the "discontinuity" of "gradient"

$\longrightarrow$ Back to the relation between "Coordinate-wise minimzer" and "global minimizer"

we can make the Formal Description as follows

a) if f is $\begin{cases} \text{convex} \\ \text{differentiable} \end{cases}$, then Coordinate-wise Mininizer $\Updownarrow$ Global Minimizer

(Easy to prove)

b) Generally speaking **Coordinate-wise Minimizer** ⇏ **Global Minimizer**

even if $f(\cdot)$ is convex

(example is one case)

c) **Target Problem** → (when Coordinate-wise minimizer is indeed global minimizer)

$$\min_x \quad f(x) + \sum_{i=1}^n f_i(x_i)$$

Here: $\begin{cases} f(\cdot) \to \text{convex \& differentiable} \\ f_i(\cdot) \to \text{only convex} \end{cases}$

we have: Coordinate Minimizer ⇔ Global Minimizer

Pf: $\bar{x}_i \in \underset{x_i}{\operatorname{argmin}} \; f(x_i, \bar{x}_{-i}) + f_i(x_i) \qquad \forall i \in [n]$

($f(x_i, \bar{x}_{-i}) + f_i(x_i)$ → convex)

$\Leftrightarrow$ $0 \in \nabla_i f(\bar{x}) + \partial f_i(\bar{x}_i) \qquad \forall i \in [n]$

(0 → scalar; $\partial f_i(\bar{x}_i)$ → 1-dim)

Question: $F(x) = \sum_{i=1}^n f_i(x)$

Then $\partial F(x) = \sum_{i=1}^n \partial f_i(x)$ → d-dim

Here. $f_i(x) := f_i(x_i)$ → only related to x_i

Then $\xi \in \partial f_i(x) \Leftrightarrow \forall y, \; f_i(y) \geq \xi^T (y - x) + f_i(x)$

→ choose $y_j = \begin{pmatrix} \xi_j \\ x_i \end{pmatrix}$, we have $0 \geq \xi_j \cdot \xi_j \quad \forall \xi_j$

forcing $\xi_j = 0$ for $\forall j \neq i$

$\rightarrow$ if all $\xi_j = 0$ for $\forall j \neq i$

then $f_i(y) \geq \xi^T (y - x) + f_i(x) \quad \forall y$

$\Leftrightarrow f_i(y_i) \geq \xi_i (y_i - x_i) + f_i(x_i)$

$\Leftrightarrow \xi_i \in \partial f_i(x_i)$

$\Rightarrow \partial f_i(x) = \{0\} \times \cdots \times \partial f_i(x_i) \times \cdots \times \{0\}$

i-th dimension

$\Rightarrow \partial F(x) = \sum_{i=1}^{n} \partial f_i(x) = \prod_{i=1}^{n} \partial f_i(x_i)$ (cartesian product)

$\Leftrightarrow 0 \in \nabla f(\bar{x}) + \partial F(\bar{x}) \Leftrightarrow \bar{x}$ is global minimizer

Here $\nabla f(\bar{x}) := \begin{bmatrix} \nabla_1 f(\bar{x}) \\ \vdots \\ \nabla_n f(\bar{x}) \end{bmatrix}$

$$\partial F(\bar{x}) := \begin{bmatrix} \partial f_1(\bar{x}_1) \\ \vdots \\ \partial f_n(\bar{x}_n) \end{bmatrix}$$

this part is highly-nontrivial !

$\Rightarrow$ and this is why some other general function cannot achieve this good property!

---

2. How to achieve Coordinate-wise minimizer ? (Algo)

(Brute Force)

a) $\bar{x} \rightarrow$ Coordinate-wise Minimizer

Actually like Nash Equilibrium!

$\Rightarrow$ (Necessary Condition) $\boxed{\bar{x}_i \in \underset{x_i}{\operatorname{argmin}} f(x_i, \bar{x}_{-i})}$

Consider $A_i := \{ (x_i, x_{-i}) : x_i \in \underset{\tilde{x}_i}{\operatorname{argmin}} f(\tilde{x}_i, x_{-i}) \}$

$\Rightarrow$ { Coordinate Minimizer } $= \bigcap_{i=1}^{n} A_i$

(intractable)

b) Coordinate Descent Algorithm

- → given $x^{(k)}$
- → $x_i^{(k+1)} = \underset{x_i}{\operatorname{argmin}} \; f(x_{<i}^{(k+1)}, x_i, x_{>i}^{(k)})$
- → $k \to k+1$

---

3. Application

a) $\min_x \; (x_1 - x_2)^2 + |x_1| + |x_2|$

→ Algo:

⇒ given $x_1^{(k)}, x_2^{(k)}$

⇒ $x_1^{(k+1)} = \underset{x_1}{\operatorname{argmin}} \; (x_1 - x_2^{(k)})^2 + |x_1|$ ($f(x) = |x|$)

$$= P_{\frac{1}{2}f}(x_2^{(k)})$$

$$= P_{\frac{1}{2}\|\cdot\|_1}(x_2^{(k)})$$

$$= \begin{cases} x_2^{(k)} - \frac{1}{2}, & x_2^{(k)} > 1/2 \\ 0, & \text{o/w.} \\ x_2^{(k)} + \frac{1}{2}, & x_2^{(k)} < -1/2 \end{cases}$$

⇒ $x_2^{(k+1)} = P_{\frac{1}{2}\|\cdot\|_1}(x_1^{(k+1)})$

$$= \begin{cases} x_1^{(k+1)} - 1/2, & x_1^{(k+1)} > 1/2 \\ 0, & \text{O/w} \\ x_1^{(k+1)} + 1/2, & x_1^{(k+1)} < -1/2 \end{cases}$$

b) Linear Reg with Coordinate Descent

$$\min_{\beta} \boxed{\frac{1}{2}\|X\beta - y\|_2^2}$$ differentiable + convex

$X_{\cdot i} \in \mathbb{R}^p$ $\beta_i \in \mathbb{R}$

if want to apply CD, then we should solve:

$$\hat{\beta}_i \in \arg\min_{\beta_i} \frac{1}{2}\|X_{\cdot i}\beta_i + X_{\cdot -i}\hat{\beta}_{-i} - y\|_2^2$$

$\frac{1}{2} z^T z$ $z = \cdots$

$$\Leftrightarrow \nabla_{\beta_i} L(\beta) = 0$$

$$\Leftrightarrow \nabla_{\beta_i} z \, \nabla_z L(z) = 0$$

$$\Leftrightarrow x_{\cdot i}^T \cdot (X_{\cdot i}\hat{\beta}_i + X_{\cdot -i}\hat{\beta}_{-i} - y) = 0$$

$$\Leftrightarrow \hat{\beta}_i = \frac{1}{\|x_{\cdot i}\|_2^2}\left[x_{\cdot i}^T (y - X_{\cdot -i}\hat{\beta}_{-i})\right]$$

$\Updownarrow$

"Adaptive weight GD"

$$\hat{\beta}_i \leftarrow \hat{\beta}_i - \frac{1}{\|x_{\cdot i}\|_2^2}\left(x_{\cdot i}^T (y - X_{\cdot -i}\hat{\beta}_{-i} - x_i \hat{\beta}_i)\right)$$

$$= \hat{\beta}_i - \frac{1}{\|x_{\cdot i}\|_2^2}\left(x_{\cdot i}^T (y - X\hat{\beta})\right)$$

GD update: $\hat{\beta}_i \leftarrow \hat{\beta}_i - x_{\cdot i}^T (y - X\hat{\beta})$

$$\Leftrightarrow \hat{\beta} \leftarrow \hat{\beta} - X^T(y - X\hat{\beta})$$

STILL cannot guarantee the sparsity of solution!

$\Downarrow$

if we want, then we should re-formulate the optimization problem

→ ℓ_1-norm LASSO

c) LASSO

$$\min_{\beta} \underbrace{\frac{1}{2}\|X\beta - y\|_2^2}_{\text{convex \& differentiable}} + \underbrace{\lambda \|\beta\|_1}_{\text{separable \& convex}}, \quad \lambda \sum_{i=1}^{p} |\beta_i|$$

(LASSO)

For this task, if we want to apply CD Framework, then we should solve:

$$\hat{\beta}_i \in \arg\min_{\beta_i} \frac{1}{2}\|X_{\cdot i}\beta_i + X_{\cdot -i}\hat{\beta}_{-i} - y\|_2^2 + \lambda|\beta_i|$$

$$= \arg\min_{\beta_i} \frac{1}{2}\|X_{\cdot i}\|_2^2 \beta_i^2 + \left(X_{\cdot i}^T (\underbrace{X_{\cdot -i}\hat{\beta}_{-i} - y}_{\Delta})\right)\beta_i + \lambda|\beta_i|$$

$$= \arg\min_{\beta_i} \frac{1}{2}\left(\beta_i^2 + \frac{X_{\cdot i}^T \Delta}{\|X_{\cdot i}\|_2^2}\beta_i\right) + \frac{\lambda}{\|X_{\cdot i}\|_2^2}|\beta_i|$$

$$= \arg\min_{\beta_i} \frac{1}{2}\left(\beta_i + \frac{X_{\cdot i}^T \Delta}{\|X_{\cdot i}\|_2^2}\right)^2 + \frac{\lambda}{\|X_{\cdot i}\|_2^2}|\beta_i|$$

$$= \arg\min_{\beta_i} \frac{\lambda}{\|X_{\cdot i}\|_2^2}|\beta_i| + \frac{1}{2}\left(\beta_i - \left(-\frac{X_{\cdot i}^T \Delta}{\|X_{\cdot i}\|_2^2}\right)\right)^2$$

$$= P_{\frac{\lambda}{\|X_{\cdot i}\|_2^2}|\cdot|}\left(-\frac{X_{\cdot i}^T \Delta}{\|X_{\cdot i}\|_2^2}\right)$$

$$= P_{\frac{\lambda}{\|X_{\cdot i}\|_2^2}|\cdot|}\left(\frac{X_{\cdot i}^T (y - X_{\cdot -i}^T \hat{\beta}_{-i})}{\|X_{\cdot i}\|_2^2}\right)$$

↓

the update for LR without regularization through **Coordinate Descent Method**

LR: $\hat{\beta}_i \leftarrow \frac{X_{\cdot i}^T (y - X_{\cdot -i}^T \hat{\beta}_{-i})}{\|X_{\cdot i}\|_2^2}$

↓

LASSO: $\hat{\beta}_i = \text{soft-threshold}(\hat{\beta}_i)$

d) Box-constraint Regression

$$\begin{cases} \min\limits_{\beta} & \frac{1}{2}\|X\beta - y\|_2^2 \\ \text{s.t} & l \le \beta \le u \end{cases}$$

unconstrained optimization

$$\Leftrightarrow \boxed{\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|X\beta - y\|_2^2 + \delta_C(\beta)}$$

$$C = \{\beta : l \le \beta \le u\} = C_1 \times C_2 \times \cdots \times C_p$$

$$\Leftrightarrow \boxed{\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|X\beta - y\|_2^2 + \sum_{i=1}^{p} \delta_{C_i}(\beta_i)}$$

$$C_i = \{\beta_i : l_i \le \beta_i \le u_i\}$$

if want to Apply Coordinate Descent Algo,
then we need to solve:

$$\hat{\beta}_i = \arg\min_{\beta_i} \frac{1}{2}\|X_{\cdot i}\beta_i + X_{\cdot -i}\hat{\beta}_{-i} - y\|_2^2 + \delta_{C_i}(\beta_i)$$

$$= \arg\min_{\beta_i} \frac{1}{2}\|x_{\cdot i}\|_2^2 \beta_i^2 + \underbrace{x_{\cdot i}^T (X_{\cdot -i}^T \hat{\beta}_{-i} - y)}_{\textcircled{A}} \beta_i + \delta_{C_i}(\beta_i)$$

$$= \arg\min_{\beta_i} \frac{1}{2}\left[\beta_i - \left(\frac{x_{\cdot i}^T (y - X_{\cdot -i}^T \hat{\beta}_{-i})}{\|x_{\cdot i}\|_2^2}\right)\right]^2 + \frac{1}{\|x_{\cdot i}\|_2^2}\delta_{C_i}(\beta_i)$$

$$= P_{\frac{1}{\|x_{\cdot i}\|^2}\delta_{C_i}(\cdot)}\left(\frac{x_{\cdot i}^T (y - X_{\cdot -i}^T \hat{\beta}_{-i})}{\|x_{\cdot i}\|_2^2}\right)$$

$$= P_{\delta_{C_i}(\cdot)}\left(\frac{x_{\cdot i}^T (y - x_{\cdot -i}^T \hat{\beta}_{-i})}{\|x_{\cdot i}\|_2^2}\right)$$

$$= \Pi_{C_i}\left(\frac{x_{\cdot i}^T (y - x_{\cdot -i}^T \hat{\beta}_{-i})}{\|x_{\cdot i}\|_2^2}\right)$$

4. Block CD & Application

$$\begin{cases} \min\limits_{x \in \mathcal{X}} & F(x) = f(x) + \sum\limits_{i=1}^{m} r_i(x_i) \\ \text{s.t} & x_i \in \mathcal{X}_i \qquad \bigcup\limits_{i=1}^{m} \mathcal{X}_i = \mathcal{X} \\ & r_i : \mathcal{X}_i \longrightarrow (-\infty, +\infty] \end{cases}$$

$r_i(x_i)$: function in block of coordinates

- CD (differentiable)
- ↓ guarantee
- $\nabla f(\bar{x}) = 0$ (stationary point)

→ Algorithm is quite similar to CD Algo

Application (Matrix Factorization)

e) $$\begin{cases} \min\limits_{W,H} & \frac{1}{2} \| V - WH \|_F^2 \\ \text{s.t} & W \geq 0 \qquad H \geq 0 \end{cases}$$

$\frac{1}{2} \| V - WH \|_F^2$ → non-convex

⇒ CD method can only guarantee Stationary point (local minimizer) (element-wise, not PSD)

$W \geq 0$, $H \geq 0$ → separable

$V \in \mathbb{R}^{m \times n}$, $W \in \mathbb{R}^{m \times r}$, $H \in \mathbb{R}^{r \times n}$

Calculation:

① $\frac{1}{2} \| V - WH \|_F^2$

$$\begin{cases} W = \begin{pmatrix} W_{\cdot 1} & \cdots & W_{\cdot r} \end{pmatrix} \in \mathbb{R}^{m \times r} \\ H = \begin{pmatrix} H_{1\cdot} \\ \vdots \\ H_{r\cdot} \end{pmatrix} \in \mathbb{R}^{r \times n} \end{cases}$$

$$= \frac{1}{2} \| V - \sum_{i=1}^{r} W_{\cdot i} \cdot H_{i\cdot} \|_F^2$$

$$= \frac{1}{2} \| V - W_{\cdot i} H_{i\cdot} - W_{\cdot -i} H_{-i\cdot} \|_F^2$$

$$:= \frac{1}{2} \| \Delta - W_{\cdot i} H_{i\cdot} \|_F^2 \quad (\text{where } \Delta := V - W_{\cdot -i} H_{-i\cdot} \in \mathbb{R}^{m \times n})$$

$$= \frac{1}{2} \langle \Delta - W_{\cdot i} H_{i\cdot}, \Delta - W_{\cdot i} H_{i\cdot} \rangle_F$$

(where $\langle A, B \rangle_F := \mathrm{Tr}(A^T B)$)

$$= \frac{1}{2} \| \Delta \|_F^2 + \frac{1}{2} \| W_{\cdot i} H_{i\cdot} \|_F^2 - \langle \Delta, W_{\cdot i} H_{i\cdot} \rangle_F$$

$$= \frac{1}{2}\|\Delta\|_F^2 + \frac{1}{2}\mathrm{Tr}\left(W_{\cdot i} H_{i\cdot} \cdot H_{i\cdot}^T W_{\cdot i}^T\right) - \cdots$$

$$= \frac{1}{2}\|\Delta\|_F^2 + \frac{1}{2}\mathrm{Tr}\left(H_{i\cdot}^T W_{\cdot i}^T W_{\cdot i} H_{i\cdot}\right) - \cdots$$

$\mathrm{Tr}\left(H_{i\cdot}\Delta^T W_{\cdot i}\right)$

$$= \frac{1}{2}\|\Delta\|_F^2 + \frac{1}{2}\|H_{i\cdot}^T\|_2^2 \cdot \|W_{\cdot i}\|_2^2 - \mathrm{Tr}\left(\Delta^T W_{\cdot i} H_{i\cdot}\right)$$

$$= \frac{1}{2}\|\Delta\|_F^2 + \frac{1}{2}\|H_{i\cdot}^T\|_2^2 \|W_{\cdot i}\|_2^2 - \begin{cases} \langle \Delta^T W_{\cdot i}, H_{i\cdot} \rangle \\ \langle \Delta H_{i\cdot}^T, W_{\cdot i} \rangle \end{cases}$$

$\Delta \in \mathbb{R}^{m \times n}$

② a) fix H & W_{-i}, update W_i by:

$$\rightarrow \tilde{W}_{\cdot i} = \underset{w_i}{\mathrm{argmin}} \ \frac{1}{2}\|w_i\|_2^2 - \frac{\langle W_i, \tilde{\Delta}\tilde{H}_{i\cdot}^T \rangle}{\|\tilde{H}_{i\cdot}^T\|_2^2} + \delta_{\mathbb{R}_+^m}(w_i)$$

$$= \underset{w_i}{\mathrm{argmin}} \ \frac{1}{2}\left\| w_i - \frac{\tilde{\Delta}\tilde{H}_i^T}{\|\tilde{H}_{i\cdot}\|_2^2} \right\|_2^2 + \delta_{\mathbb{R}_+^m}(w_i)$$

$$= \rho_{\delta_{\mathbb{R}_+^m}(\cdot)}\left(\frac{\tilde{\Delta}\,\tilde{H}_{i\cdot}^T}{\|\tilde{H}_{i\cdot}\|_2^2}\right)$$

$$= \Pi_{\mathbb{R}_+^m}\left(\frac{(V - \tilde{W}_{\cdot(-i)}\tilde{H}_{(-i)\cdot})\tilde{H}_{i\cdot}^T}{\|\tilde{H}_{i\cdot}\|_2^2}\right)$$

$$\rightarrow \tilde{H}_{i\cdot}^T = \underset{H_i}{\mathrm{argmin}} \ \frac{1}{2}\|H_i\|_2^2 - \frac{\langle \tilde{\Delta}^T \tilde{W}_{\cdot i}, H_i \rangle}{\|\tilde{W}_{\cdot i}\|_2^2} + \delta_{\mathbb{R}_+^n}(H_i)$$

$$= \underset{H_i}{\mathrm{argmin}} \ \frac{1}{2}\left\| H_i - \frac{\tilde{\Delta}^T \tilde{W}_{\cdot i}}{\|\tilde{W}_{\cdot i}\|_2^2} \right\|_2^2 + \delta_{\mathbb{R}_+^n}(H_i)$$

$$= \rho_{\delta_{\mathbb{R}_+^n}(\cdot)}\left(\frac{\tilde{\Delta}^T \tilde{W}_{\cdot i}}{\|\tilde{W}_{\cdot i}\|_2^2}\right)$$

$$= \Pi_{\mathbb{R}_+^n}\left(\frac{(V - \tilde{W}_{\cdot(-i)}\tilde{H}_{(-i)\cdot})^T \tilde{W}_{\cdot i}}{\|\tilde{W}_{\cdot i}\|_2^2}\right)$$

Rmk: ① $\min_x : f(x)$

if f is non-convex, but differentiable

then we still have $\nabla f(\bar{x}) = 0$ (necessary condition for optimality)

↓

which means $\bar{x}$ will be a stationary point,

not a global optimizer

② $\min_x : f(x) + \sum_{i=1}^{n} r_i(x_i) := F(x)$

if f is non-convex, but differentiable

then we still have $0 \in \partial F(\bar{x})$

↓

"stationary point"