① $\underbrace{\boxed{\min_{x,y} f(x,y)}}_{①} = \boxed{\min_{x}\left(\min_{y} f(x,y)\right)} = \min_{y}\left(\min_{x} f(x,y)\right)$

Suppose x^*, y^* is optimal ② attained by ②, Suppose x^{**}, y^{**} is attained by ①, then we have

$$f(x^{**}, y^{**}) \le f(x^*, y^*) = \min_{x}\left(\min_{y} f(x,y)\right) \le \min_{x}\left(f(x, y^{**})\right) \le f(x^{**}, y^{**})$$

② Latent Feature Lasso Question:

$$\min_{k \in N,\, Z \in \{0,1\}^{N \times k},\, W \in R^{k \times D}} \frac{1}{2N}\|X - ZW\|_F^2 + \frac{\lambda}{2}\|W\|_F^2$$

Consider the situation fitting k and Z

$\Rightarrow$ New problem:

$$\min_{W \in R^{k \times D}} \underbrace{\boxed{\frac{1}{2N}\|X - ZW\|_F^2}}_{\text{Say it as } L(W)} + \frac{\lambda}{2}\|W\|_F^2$$

Introducing slack variable $E = ZW$

$$\Rightarrow \min_{W \in R^{k \times D},\, E \in R^{N \times D}} \frac{1}{2N}\|X - E\|_F^2 + \frac{\lambda}{2}\|W\|_F^2 \quad \text{subject to } E = ZW \qquad Ⓐ$$

before using lagrange duality $\Rightarrow$ we check strong duality first

consider $\Omega = \begin{bmatrix} W \\ \hline E \end{bmatrix} \begin{matrix} {}_{K \times D} \\ {}_{N \times D} \end{matrix}$, $P_W \Omega = W$ 扁

$P_E \Omega = E$ 扁

the problem becomes

$$\min_{\Omega} \frac{1}{2N}\|X - P_E\Omega\|_F^2 + \frac{\lambda}{2}\|P_W\Omega\|_F^2 \quad \text{subject to}$$

$P_E\Omega = Z P_W \Omega$, which is convex problem of Ω and is feasible, since simply choosing W and let $E = ZW$

(Slater's condition)

Consider (A)'s lagrange duality :

$$L(W,E,A) = \frac{1}{2N}\|X-E\|_F^2 + \frac{\lambda}{2}\|W\|_F^2 + \langle A, E-ZW\rangle$$

$$\underset{\succeq 0}{(A)} = \min_{W,E}\max_{A} L(W,E,A) = \max_{A}\min_{W,E} L(W,E,A) \quad \text{(by strong duality)}$$

$$\max_A \min_E \Big(\underbrace{\min_W \frac{1}{2N}\|X-E\|_F^2 + \frac{\lambda}{2}\mathrm{Tr}(W^TW) + \mathrm{Tr}(A^T(E-ZW))}_{g(W)} \Big)$$

simply

$$\nabla g(W) = 0 = \lambda W + \nabla \mathrm{Tr}(-A^TZW) = \lambda W + \nabla\mathrm{Tr}(-WA^TZ) = \lambda W - Z^TA$$

W^* is given by $\frac{1}{\lambda}Z^TA$, by substituting W^*, the question becomes

$$\mathrm{Tr}(-A^TZ\frac{1}{\lambda}Z^TA)$$

$$\max_A \min_E \Big(\frac{1}{2N}\|X-E\|_F^2 + \frac{1}{2\lambda}\mathrm{Tr}(A^TZZ^TA) + \langle A,E\rangle + \langle A, -ZW\rangle \Big)$$

$$= \max_A \min_E \Big(L(E) + \frac{1}{2\lambda}\mathrm{Tr}(A^TZZ^TA) + \langle A,E\rangle - \frac{1}{\lambda}\mathrm{Tr}(A^TZZ^TA)\Big)$$

Recall convex conjugate.

$$= \max_A \min_E \Big(L(E) - \frac{1}{2\lambda}\mathrm{Tr}(A^TZZ^TA) + \langle A,E\rangle \Big), \quad f^*(y) = \max_x y^Tx - f(x)$$

$$= \max_A \Big(-\max_E \big(-L(E) - \langle A,E\rangle\big) - \frac{1}{2\lambda}\mathrm{Tr}(A^TZZ^TA)\Big)$$

$$= \max_A \Big(-\frac{1}{2\lambda}\mathrm{Tr}(A^TZZ^TA) - L^*(A) \Big) = \tilde{g}(M,A), \quad M = ZZ^T$$

when M is fixed, $\tilde{g}(M,A)$ is a convex function of A,

$g(M) \overset{\text{def}}{=} \max_A \tilde{g}(M,A)$ is still a convex function

(pointwise maximum of convex)

The whole problem reduced to

$$\min_Z g(M) \text{ while } M = ZZ^T$$

We introducing a trick called atomic (nuclear) norm regularization to force M has a structure of ZZ^T ($Z \in \{0,1\}^{N\times k}$ matrix), Our problem is now

$$\min_M \{ g(M) + \mu \|M\|_S \} \quad \text{where } \|M\|_S = \min_{c\geq 0,\, a\in S} \sum^k c_a \ \text{ s.t. } M = \sum_{a\in S} c_a a$$

S is a collection of all possible atoms, $S = \{ zz^T \mid z \in \{0,1\}^N \}$

(It can be shown that all possible combination of M covers ZZ^T, so it is actually a relaxation here).

Now we use greedy coordinate descent to solve this problem

$$\min_M \{ g(M) + \lambda \|M\|_S \} = \min_{c\in \mathbb{R}_+^{\bar{k}}} \underbrace{\{ \underbrace{g(\textstyle\sum_{j=1}^{\tilde{k}} c_j z_j z_j^T)}_{f(c)} + \lambda \|c\|_1 \}}_{F(c)}, \quad \bar{k} = 2^{N-1}$$

we first choose a coordinate of c to do coordinate descent, finding correspond atom $z_j z_j^T$, and then do proximal gradient update.

$$j^* = \arg\max_j -\nabla_j f(c) = \arg\max_j \langle \nabla g(M), z_j z_j^T \rangle$$

we can find this j using MAX-CUT like problem.

Now consider the algroithm step, we have active set $\mathcal{A}$, and and new atom $z_j z_j^T$ by maxcut, we want to update corresponding $|\mathcal{A}|$ number of c_i of our all atoms.

$$c^{r+1} \leftarrow \left[c^r - \frac{\nabla f(c^r) + \lambda}{r|\mathcal{A}|} \right]_+ \quad r = 1, 2 .. T_2 \quad \left(\begin{array}{c} c > 0 \\ \text{differentiable} \end{array} \right)$$

r is Lipchiz continous of $\nabla_{c_j} f(c)$

By Danskin's theorem

$$\nabla_{c_j} f(c) = z_j A^* A^{*T} z_j / (2N_T^2)$$

Danskin's theorem statement :

$\phi : R^n \times Z \to R$, $\phi(x, z)$ is a continuous function, $Z \subset R^m$ is a compact set, Further assume $\phi(x, z)$ is convex in x for every $z \in Z$

$$f(x) = \max_{z \in Z} \phi(x, z)$$

$$Z_0(x) = \{ \bar{z} : \phi(x, \bar{z}) = \max_{z \in Z} \phi(x, z) \}$$

$\Rightarrow$ $f(x)$ is convex

$$D_y f(x) = \max_{z \in Z_0(x)} \phi'(x, z; y) \quad [\text{deriviate of direction of } y \text{ of } x]$$

for all $z \in Z$ $\frac{\partial \phi}{\partial x}$ is

$f(x)$ is differentiable at x if $Z_0(x)$ consists of a single element $\bar{z}$, in this case, the deriviative of $f(x)$

$$\frac{\partial f(x)}{\partial x} = \frac{\partial \phi(x, \bar{z})}{\partial x}$$

$$g: \underset{\downarrow \; Z}{R^{n\times k}} \times \underset{\downarrow \; A \in R^{n\times d}}{\Omega} \to \mathbb{R}$$

(compact if closed and bounded)
I believe it because numerical error ?

continous ✓

convex if fix Z ✓
also convex in A.

$$g(Z) = \max_{A\in\Omega} \tilde{g}(Z,A)$$

$$\frac{\partial\, Tr(xx^T B)}{\partial x} = Bx + B^T x.$$

convex: only one maximum [If is also true when A restricted on R_+ ?

$= AA^TZZ^T + AA^TZZ^T$

$$\nabla_{c_j} g(C) = \nabla_{c_j}\left[\max_A\left(-\frac{1}{2\lambda} Tr\left(A^T \sum_{j=1}^{k} c_j z_j z_j^T A\right) + L^*(A)\right)\right]$$

$$= \nabla_{c_j}\left[-\frac{1}{2\lambda} Tr\left(A^{*T} c_j z_j z_j^T A^*\right) - L^*(A)\right]$$

$$= -\frac{1}{\lambda} Tr\left(z_j A^* A^{*T} z_j\right) = -\frac{1}{\lambda}\left(z_j A^* A^{*T} z_j\right).$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$\frac{\partial\, Tr(cA)}{\partial c} = \frac{\partial\, c\, Tr(A)}{\partial c}.$$

Maxcut Problem:

$y = 2z - 1 \qquad z = \frac{y+1}{2}$

$$\max_{z \in \{0,1\}^N} \langle C, zz^T \rangle = \max_{z \in \{0,1\}^N} \langle C, \underbrace{\frac{y+1}{2}\left(\frac{y+1}{2}\right)^T}_{\frac{yy^T + 2y^T + 11^T}{4}} \rangle$$

$$= \max_{z \in \{0,1\}^N} \frac{1}{4}\left(\langle C, yy^T\rangle + 2\langle C, 1y^T\rangle + \langle C, 11^T\rangle\right)$$

$$= \max_{(y_0, y) \in \{-1,1\}^{N+1}} \frac{1}{4}\begin{bmatrix} y_0 \\ y \end{bmatrix}^T \begin{bmatrix} 1^T C 1 & 1^T C \\ C1 & C \end{bmatrix}\begin{bmatrix} y_0 \\ y \end{bmatrix}$$

$$\begin{bmatrix} \overset{1\times n \; n\times n \; n\times 1}{1^T C 1} & \overset{1\times n}{1^T C} \\ \overset{n\times n \; n\times 1}{C1} & \overset{n\times n}{C} \end{bmatrix}\begin{bmatrix} y_0 \\ y \end{bmatrix} = \begin{bmatrix} 1^T C 1 y_0 + 1^T C y \\ C1 y_0 + Cy \end{bmatrix}$$

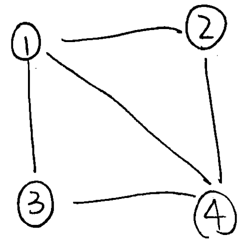
$$[y_0^T, y^T]\begin{bmatrix} 1^T C 1 y_0 + 1^T C y \\ C1 y_0 + Cy \end{bmatrix} = \begin{bmatrix} y_0 1^T C 1 y_0 + y_0^T 1^T C y \\ + \\ y_0^T C 1 y_0 + y^T C y \end{bmatrix}$$

✓

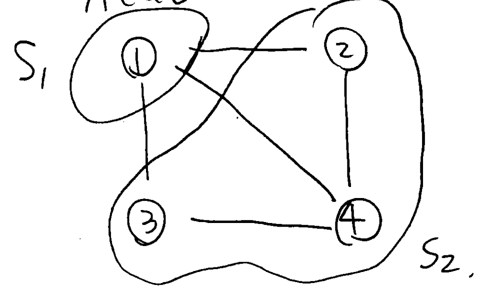
$$= \langle C, yy^T \rangle + 2\langle C, 1y^T\rangle + \langle C, 11^T\rangle.$$

Maxcut Problem:

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$



A cut example.



Suppose there are n-node, x_i characterized the cut by $x_i = 1 \in S_1$ and $x_i = -1$ if $x \in S_2$.

$$\Rightarrow x = (1, -1, -1, -1)$$

each edge value of cut edge: $\frac{1}{2} a_{ij}(1 - x_i x_j)$.

$$W = \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} (1 - x_i x_j)$$

$$= \frac{1}{4} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} x_i \overbrace{a_{ij} x_i}^{=1} - \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j \right)$$

$$\Rightarrow W = \frac{1}{4} \left(x^t \, Diag(A\mathbb{1}) x - x^t A x \right)$$

tranform (L)

$$L = Diag(A\mathbb{1}) - A$$

So the maximum cut problem can be formulated as

$$\max_{x \in \{-1,1\}^n} \frac{1}{4} x^t L x.$$

(~~[illegible]~~ similiar to spin-glass model VLSI.)

$\Rightarrow$ NP-hard problem, 2^n possibilities

$\Rightarrow$ SDP-relaxation

$\max_{x \in \{-1,1\}^n} \frac{1}{4} x^t L x$ is a integer programing

$\mathrm{Trace}(x^t L x) = \mathrm{Trace}(L x x^T) = L \bullet X$, X is real symmetric

$X \geq 0$, $\mathrm{diag}(X) \Leftrightarrow x_i^2 = 1$

$\left(\begin{array}{l} v^T x x^T v \\ = (x^T v)^T (x^T v) \geq 0 \end{array} \right)$

$\Downarrow \Uparrow$

reformulation :

Maximize $L \bullet X$
subject to $\mathrm{diag}(X) = e$, $\mathrm{rank}(X) = 1$ and $X \geq 0$

$\Downarrow$ SDP - relaxation

$\underset{X}{\text{Maximize}}$ $L \circ X$
subject to $\mathrm{diag}(X) = e$ and $X \geq 0$

↳ many posible solver!

(standard form : min : $L \circ X$
subject to $A_i \circ X = b_i$
$X \geq 0$)

pick $A_i = \begin{cases} a_{ii} = 1 \\ 0 \end{cases}$