UNIVERSITY OF CALIFORNIA, BERKELEY

HW 8: EE127

Oscar Ortega

April 5, 2019

1 AUXILIARY PROBLEM: PSD MATRICES

Let $A, B \in \mathbb{S}_+^n$

$$
\begin{aligned}
\langle A, B \rangle &= \operatorname{tr}(A^T B) \\
&= \operatorname{tr}(AB) \\
&= \operatorname{tr}((V\Lambda V^T)_A (V\Lambda V^T)_B) \\
&= \operatorname{tr}(\Lambda_A \Lambda_B) \\
\sum_{i=1}^{n} \lambda_{i,a} \lambda_{i,b} &\geq 0
\end{aligned}
$$

2 Auxiliary Problem: Matrices as Eigenvectors

- a: Let $A \in \mathbf{S}_+^m$, $B \in \mathbf{S}_+^n$, $X \in \mathbf{R}^{m,n}$

  $g(X) = AXB$ is a linear mapping.

  Proof:

  $$g(0) = A0B = (A0)B = 0B = 0$$

  Let $X_1, X_2 \in \mathbf{R}^{m,n}$

  $$\begin{aligned} g(X_1 + X_2) &= A(X_1 + X_2)B \\ &= (AX_1 + AX_2)B \\ &= AX_1B + AX_2B \\ &= g(X_1) + g(X_2) \end{aligned}$$

  Let $\alpha \in \mathbf{R}$, $X \in \mathbf{R}^{m,n}$

  $$\begin{aligned} g(\alpha X) &= A\alpha XB \\ &= \alpha AXB \\ &= \alpha g(X) \end{aligned}$$

- b: Let u and v be eigenvectors of A and B, with eigenvalues λ_a and λ_b

  $$\begin{aligned} g(uv^T) &= Auv^TB \\ &= Auv^TB^T \\ &= \lambda_a u \lambda_b v^T \\ &= \lambda_a \lambda_b uv^T \end{aligned}$$

  We know that because $A \in \mathbf{S}_+^m$, $B \in \mathbf{S}_+^n$, $\exists u_1, ..., u_m \in \mathbf{R}^m$ and $\exists v_1, ..., v_n \in \mathbf{R}^n$ s.t

  $$g(u_i v_j^T) = \lambda_{i,a} \lambda_{j,b} u_i v_j^T$$

  Because this mapping is linear, all eigenvector-eigenvalue pairings will be of the form $\alpha u_i v_j^T : \alpha \in \mathbf{R}$, with eigenvalues $\lambda_{i,a}\lambda_{b,j}$

- c: $h(X)$ is symmetric:
  Proof: Let $X, Y \in \mathbf{R}^{m,n}$

$$\begin{aligned}
&\langle h(X), Y\rangle \\
&= \operatorname{tr}(h(X)^T Y) \\
&= \operatorname{tr}((\sum_{i=1}^{L} A_i X B_i)^T Y) \\
&= tr(Y^T(\sum_{i=1}^{L} A_i X B_i)) \\
&= tr(Y^T h(X)) \\
&= \langle Y, h(X)\rangle
\end{aligned}$$

  $h(X)$ is positive semi-definite:
  Proof: Let $X \in \mathbf{R}^{m,n}$

$$\begin{aligned}
&\langle h(X), X\rangle \\
&= \operatorname{tr}(X^T h(X)) \\
&= \operatorname{tr}(X^T \sum_{i=1}^{L} A_i X B_i)
\end{aligned}$$

  We know that because A_i, B_i are positive semi-definite, $\exists (V\lambda V^T)_{Ai}$ and $(V\lambda V^T)_{Bi}$ s.t $A_i = (V\Lambda V^T)_{Ai}$ and $B_i = (V\Lambda V^T)_{Bi}$

$$\begin{aligned}
&= \operatorname{tr}(X^T \sum_{i=1}^{L} (V\Lambda V^T)_{Ai} X (V\Lambda V^T)_{Bi}) \\
&= \operatorname{tr}(X^T \sum_{i=1}^{L} \Lambda_{Ai} X \Lambda_{Bi}) \\
&= \sum_{i=1}^{L} \operatorname{tr}(X^T \Lambda_{Ai} X \Lambda_{Bi}) \\
&= \sum_{i=1}^{L} \sum_{j=1}^{n} \lambda_{ai}\lambda_{bi} X_i^T X_i \\
&= \sum_{i=1}^{L} \sum_{j=1}^{n} \lambda_{ai}\lambda_{bi} \|X_i\|_2^2 \geq 0
\end{aligned}$$

- d:

$$\begin{aligned}
\lambda_{\max}(h) &\geq \frac{\langle X, h(X)\rangle}{\langle X, X\rangle} \\
&= \frac{\operatorname{tr}(X^T h(X))}{\operatorname{tr}(X^T X)}
\end{aligned}$$

If we let $X = uv^T$

$$= \frac{\mathrm{tr}((uv^T)^T h(uv^T))}{\mathrm{tr}((uv^T)^T (uv^T)}$$

$$= \frac{\mathrm{tr}((uv^T)^T h(uv^T))}{\mathrm{tr}(vu^T uv^T)})$$

$$= \frac{1}{\|v\|_2^2 \|u\|_2^2} \mathrm{tr}(vu^T h(uv^T))$$

$$= \frac{1}{\|v\|_2^2 \|u\|_2^2} \mathrm{tr}(u^T h(uv^T) v)$$

$$= \frac{1}{\|v\|_2^2 \|u\|_2^2} \mathrm{tr}(u^T (\sum_{i=1}^{L} A_i uv^T B_i) v)$$

$$= \frac{1}{\|v\|_2^2 \|u\|_2^2} \mathrm{tr}(\sum_{i=1}^{L} uA_i uv^T B_i v)$$

$$= \frac{1}{\|v\|_2^2 \|u\|_2^2} \sum_{i=1}^{L} (u^T A_i u)(v^T B_i v)$$

3 Gradient Descent Algorithm on a Simplified Neural network

- a: Let $M = R - AWB$

$$\sum_{i=1}^{n} \| y_i - AWB_i \|_2^2$$
$$= \sum_{i=1}^{n} \| (R - AWB) x_i \|_2^2$$
$$= \sum_{i=1}^{N} \langle M x_i, M x_i \rangle$$
$$= \sum_{i=1}^{N} x_i^T M^T M x_i$$
$$= \sum_{i=1}^{N} \text{tr}(x_i^T M^T M x_i)$$
$$= \sum_{i=1}^{N} \text{tr}(x_i x_i^T M^T M)$$
$$= \text{tr}(M^T M)$$

- b:

$$\min_{W \in \mathbf{R}^{p,q}} \frac{1}{2} \sum_{i=1}^{n} \| y_i - AWBX_i \|_2^2$$
$$= \min_{W \in \mathbf{R}^{p,q}} \frac{1}{2} \text{tr}((R - AWB)^T (R - AWB))$$
$$= \frac{1}{2} \Big(\text{tr}(R^T R) - 2\text{tr}(R^T AWB) + \text{tr}((AWB)^T (AWB)) \Big)$$
$$= \frac{1}{2} \Big(\text{tr}(R^T R) - 2\text{tr}(BR^T AW) + \text{tr}(BB^T W^T A^T AW) \Big)$$
$$= \frac{1}{2} \Big(\text{tr}(R^T R) - 2\text{tr}(BR^T AW) + \text{tr}(BB^T W^T A^T AW) \Big)$$
$$\nabla w = \frac{1}{2} (2A^T RB^T + \nabla_w \text{tr}(BB^T W^T A^T AW))$$

Let $g(w) = A^T Aw$, and let $f(w) = BB^T W^T$

$$\nabla_w \text{tr}(fg) = (fg')^T + (gf')^T$$
$$= A^T AWBB^T + A^T AWBB^T$$
$$= 2(A^T AWBB^T)$$
$$\nabla_w = -A^T RB^T + A^T AWBB^T$$

Therefore, our update rule is as follows:

$$W_{k+1} = W_k + \eta (A^T RB^T - A^T AW_k BB^T)$$

- c:

  By the Real Spectral Theorem, we know that there exist distinct and orthonormal eigenvectors $u_1, ..., u_n$, and $v_1, ..., v_m$ for the respective transformations $A^T A$ and BB^T.

  $A^T A$ is PSD
  Proof:

  $$x^T A^T A x = \|Ax\|_2^2 \geq 0$$

  BB^T is PSD
  Proof:

  $$x^T BB^T x = \|B^T x\|_2^2 \geq 0$$

  Let λ_i be the eigenvalue corresponding to the ith eigenvector of $A^T A$ and let γ_i be eigenvalue corresponding to ith eigenvector of BB^T

  $$h(u_i v_j^T) = u_i v_j^T + \eta(A^T A u_i v_j^T BB^T)$$

  $$= u_i v_j^T + \eta(\lambda_i u_i v_j^T \gamma_j)$$

  $$= u_i v_j^T + \eta(\lambda_i \gamma_i u_i v_j^T)$$

  $$= u_i v_j^T (1 - \eta \lambda_i \gamma_i)$$

  $$h(u_i v_j^T) = (1 - \eta \lambda_i \gamma_j) u_i v_j^T$$

  $$\lambda_{\min}(h) = 1 - \eta(\lambda_{\max}(A^T A)\gamma_{\max}(BB^T)) > -1$$

  $$\eta < \frac{2}{\lambda_{\max}(A^T A)\gamma_{\max}(BB^T)}$$

  $$\lambda_{\max}(h) = 1 - \eta(\lambda_{\min}(A^T A)\gamma_{\min}(BB^T)) < 1$$

  $$\eta > 0$$

  Therefore:

  $$0 < \eta < \frac{2}{\lambda_{\max}(A^T A)\gamma_{\max}(BB^T)}$$

- d:

  $$\nabla_w = -A^T R B^T + A^T A W B B^T = 0$$

  $$A^T A W B B^T = A^T R B^T$$

  $$W B B^T = (A^T A)^{-1} A^T R B^T$$

  $$W^* = (A^T A)^{-1} A^T R B^T (BB^T)^{-1}$$

4 Sensitivity and Dual Variables

- a:
  $p^*(\lambda, \nu)$ is convex:
  Proof:
  We know that if we define
  $$g(y) = \inf_{x \in C} F(x, y)$$
  Where F is jointly convex, and C is a convex set, then the function g is also convex.

  By definition:
  $$p^*(\lambda, \nu) = \{\inf f_0(x) | f_i(x) \leq \lambda_i, i = 1, .., n : h_i(x) = \nu_i, i = 1, ...m\}$$
  Let $F(\lambda, \nu, x) = f_0(x)$, and by taking note that f_0 is convex, over x, we can see that reformulation of the objective function is jointly convex over $x, \lambda,$ and ν

  We know define the set C as follows:
  $$C = \{(x, \lambda, \nu) | f_(i) \leq \lambda_i i = 1, ..., m : h_i(x) \leq \nu_i\}$$
  Proof: C is convex
  Lemma: $c_1 = \{(x, \lambda, \nu) | f_i(x) \leq \nu_i, i = 1, ...m\}$ is convex Consider $f_i(x) - \lambda_i \leq 0$ for a fixed i
  The hessian of $g(i) = f_i(x) - \lambda_i$ is as follows:
  $$\begin{bmatrix} \frac{\partial^2 g}{\partial x^2} & \frac{\partial^2 g}{\partial x \partial \lambda_i} \\ \frac{\partial^2 g}{\partial \lambda_i x} & \frac{\partial^2 g}{\partial_x^2} \end{bmatrix}$$
  $$= \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & 0 \\ 0 & 0 \end{bmatrix}$$
  Which is PSD. Furthermore, because the set we define is an intersection of m convex sets, then c_1 is a convex set. If we now define c_2 as follows:
  $$c_2 = \{(x, \lambda, \nu) | h_i(x) \leq \nu_i, i = 1, ...m\}$$
  and if we compute the hessian $g(i) = h_i(x) - \nu_i$, we can similarily see c_2 is a convex set. Because the intersection of c_1 and c_2 is also a convex set and take note that the intersection is equal to C, then we can conclude the C is a convex set.

  Credit to the Boyd book:

- b:
  $$p^*(\lambda, \nu) \geq p^*(0, 0) - u^T \lambda - v^T \nu$$
  Proof: Let $x \in C$: Because we are told that strong duality holds, then for all u, v:
  $$p^*(0, 0) = g(\lambda^*, \nu^*) \leq f_0(x) + \sum_{i=1}^{n} \lambda_i^* f_i(x) + \nu_i^* h_i(x)$$

$$\leq f_0(x) + \lambda_*^T v + \nu_*^T u$$

By construction of the Lagrangian. Working with the inequalities we can thus see:

$$f_0(x) \geq p^*(0,0) - \lambda^{*T} u - \nu^{*T} v$$

- c:

  More credit to the boyd book:

  - If u^* is large and we pick $\lambda < 0$ Consider the case where u, λ are both scalar, and by part b, we know the objective function is always greater than $p^*(0,0) - u^*\lambda$, then we can see that the value of the optimum will be increased as a the $-u^*\lambda$ term will increase.
  - Similarily, if we know allow $\lambda > 0$, then we know the value of the right-hand side will decrease as the $-u^*\lambda$ term will decrease.
  - Finally, if we choose v^* as large and positive and pick $\nu < 0$, Then we can also guarantee that the optimal value $p^*(u, v)$ will increase as the $-v^{*T}\nu$ term will increase.