Math of Data HW4

Hiroki Hayakawa

6th Jan 2020

1. Projection free convex low-rank matrix optimization

1.1: PROJECTION ONTO THE NUCLEAR NORM BALL

(a)

Let Y is a matrix in the nuclear norm ball $\mathcal{X}$ and Σ_Y and Σ_Z are diagonal matrices derived from the singular value decomposition of Y and Z. We also denote the diagonal of Σ_Y and Σ_Z by vectors σ_Y and σ_Z respectively. We have following relations.

$$\begin{aligned}\frac{1}{2}\|Y - Z\|_F^2 &\geq \frac{1}{2}\|\Sigma_Y - \Sigma_Z\|_F^2 \ \text{(Mirsky's inequality)} \\ &= \frac{1}{2}\|\sigma_Y - \sigma_Z\|_2^2 \ \text{(Definition of Frobeniusnorm)}\end{aligned}$$

Since the matrix Y is in the nuclear norm ball $\mathcal{X}$,

$$\frac{1}{2}\|\sigma_Y - \sigma_Z\|_2^2 \geq \frac{1}{2}\|\sigma_{l1} - \sigma_Z\|_2^2$$

is verified, where σ_{l1} is a projection of σ_Y onto l_1 norm ball. We again utilize a definition of Frobenius norm,

$$\begin{aligned}\frac{1}{2}\|\sigma_{l1} - \sigma_Z\|_2^2 &= \frac{1}{2}\|\Sigma_{\sigma_{l1}} - \Sigma_Z\|_F^2 \\ &= \frac{1}{2}U^T\|U(\Sigma_{\sigma_{l1}} - \Sigma_Z)V^T\|_F^2 V\end{aligned}$$

is obtained, where $\Sigma_{\sigma_{l1}}$ is a diagonal matrix whose diagonal entries are entries of σ_{l1}. Thus, equalities and inequalities above shows the projection of Z onto the nuclear ball X can be calculated by projection of σ onto the l_1 norm ball.

(b)

The script was run five times. The average computational time for the 100k and 1M MovieLens datasets are 0.4791 and 44.12 sec respectively.

1.2: LMO of NUCLEAR NORM

(a)

Following the give HINT, we show $< K, Z > \geq < -\kappa uv^T, Z >$ for all $X \in \mathcal{X}$. Utilizing properties of trace operation, we obtain following equation.

$$\begin{aligned} < -\kappa uv^T, Z > &= -\kappa \mathrm{Tr}(Z^T uv^T) \\ &= -\kappa \mathrm{Tr}((U\Sigma V^T)^T uv^T) \\ &= -\kappa \mathrm{Tr}(V\Sigma U^T uv^T) \\ &= -\kappa \mathrm{Tr}(v^T V\Sigma U^T u) \\ &= -\kappa \mathrm{Tr}([1\ 0 \dots\ 0]\Sigma[1\ 0 \dots\ 0]^T) \\ &= -\kappa\sigma_1 \end{aligned}$$

where σ_1 is the largest singular value of Z. In addition, Holder's inequality gives us following relations.

$$\begin{aligned} | < X, Z > | &\leq \|X\|_* \|Z\|_\infty \\ &\leq \kappa\sigma_1 \\ \Leftrightarrow -\kappa\sigma_1 &\leq < X, Z > \leq \kappa\sigma_1 \end{aligned}$$

Thus, we verified $< K, Z > \geq < -\kappa uv^T, Z >$.

(b)

The script was run five times. The average computational time for the 100k and 1M MovieLens datasets are 0.05451 and 0.3934 sec respectively. Comparing the computation times generated from 1.1(b) and 1.2(b), for 100k dataset, method with LMO takes 10 times smaller computation time. looking at the change of the computation time by inputting larger dataset, with the projection operator, the obtained results says that the computation time increases with $\mathcal{O}(\mathrm{size}(dataset)^2)$. By contrast, with LMO, the computation time increases with $\mathcal{O}(\mathrm{size}(dataset))$.

2. Hands-on experiment 1: Crime Scene Investigation with Blind Deconvolution

2.1: FRANK-WOLF FOR BLIND DECONVOLUTION

(a)

Let $L \times K$ matrix L and $N \times 1$ matrix U be representative matrices of the linear operator $A(\cdot)$. The objective function $f(X)$ can be expressed as

$$\begin{aligned} f(X) &= \frac{1}{2}\|A(X) - b\|_2^2 \\ &= \frac{1}{2}\|LXU - b\|_2^2. \end{aligned}$$

The gradient of the objective function is

$$\begin{aligned} \nabla f(X) &= \nabla \frac{1}{2}(LXU - b)^T(LXU - b) \\ &= \nabla \frac{1}{2}(U^T X^T L^T - b^T)(LXU - b) \\ &= \nabla \frac{1}{2}(U^T X^T L^T LXU - U^T X^T L^T b - b^T LXU + b^T b) \\ &= (L^T LXUU^T - L^T bU^T). \end{aligned}$$

For points X, Y in the nuclear norm ball, we obtain following relations thanks to cauchy schwarz inequality.

$$\begin{aligned}\|\nabla f(X) - \nabla f(Y)\| &= \|L^T LXUU^T - L^T LYUU^T\| \\ &= \|L^T L(X-Y)VV^T\| \\ &\leq \|L^T L\|\|UU^T\|\|X-Y\|\end{aligned}$$

, which indicates the Lipschitz continuity of the gradient of the objective function and the smoothness of the objective function.

(b)

The license plate number is 'J209 LTL'. The parameter κ of 5,000 generate the readable licence plate number with given support of the blue kernel.

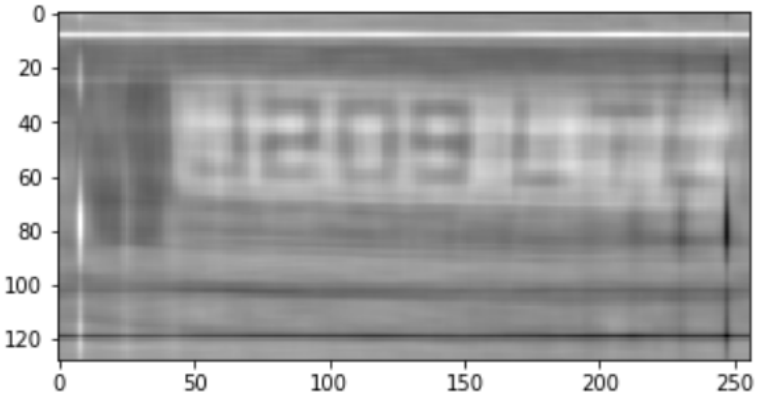

Figure 1: Deblurred Image

3. Hands-on experiment 2: k-means Clustering by Semidefinite Programming

3.1: CONDITIONAL GRADIENT METHOD FOR CLUSTERING FASHION-MNIST DATA

(a)

Let $\theta \in [0, 1]$ and matrices X, Y be in the domain $\mathcal{X}$. We show $\theta X + (1-\theta)Y$ is also in the domain $\mathcal{X}$. From the property of the trace operation, we obtain

$$\begin{aligned}\mathrm{Tr}(\theta X + (1-\theta)Y) &= \theta \mathrm{Tr}(X) + (1-\theta)\mathrm{Tr}(Y) \\ &\leq \theta\kappa + (1-\theta)l\kappa \\ &= \kappa.\end{aligned}$$

Since both X, Y are positive semidefinite, for all vector $x \in \mathbb{R}^p$ $x^T Xx \geq 0$ and $x^T Yx \geq 0$ hold. Thus,

$$\begin{aligned}&\theta x^T Xx + (1-\theta)x^T Yx \geq 0 \\ \Leftrightarrow\ &x^T(\theta X + (1-\theta)Y)x \\ \Leftrightarrow\ &\theta X + (1-\theta)Y \succeq 0\end{aligned}$$

also holds. Now we verified $\theta X + (1-\theta)Y$ is also in the domain $\mathcal{X}$ and the set $\mathcal{X}$ is convex.

(b)

The constraint $B(x) \in (K)$ is expressed as

$$\text{dist}^2(B(x), \mathcal{K}) = \min_{k \in \mathcal{K}} \|k - B(x)\|.$$

The semidefinite program has three constraints $x1 = 1(A_1(x) = b_1), x^T 1 = 1(A_2(x) = b_2)$,and $x \geq 0(B(X) \in \mathcal{K})$. Quadratic penalty functions corresponding to $x1 = 1(A_1(x) = b_1)$ and $x^T 1 = 1(A_2(x) = b_2)$ are $\|A_1(x) - b_1\|^2$ and $\|A_2(x) - b_2\|^2$ respectively. By describing $B(x) = Tx$, the penalized of objective of the objective function $f(x)$ can be expressed as $f(x) + \frac{1}{2\beta}\|A_1(x) - b_1\|^2 + \|A_2(x) - b_2\|^2 + \frac{1}{2\beta}\text{dist}^2(Tx, \mathcal{K})$. Let L and U be representative matrices of the linear operator $A_1(\cdot)$. The gradient of $\frac{1}{2\beta}\|A_1(x) - b_1\|^2$ is

$$\begin{aligned}
\nabla\|A_1(x) - b_1\|^2 &= \|LxU - b_1\|^2 \\
&= \nabla\frac{1}{2\beta}(LxU - b_1)^T(LxU - b_1) \\
&= \nabla\frac{1}{2\beta}(U^T x^T L^T - b_1^T)(LxU - b_1) \\
&= \nabla\frac{1}{2\beta}(U^T x^T L^T LxU - U^T x^T L^T b_1 - b_1^T LxU + b_1^T b_1) \\
&= \frac{1}{\beta}(L^T LxUU^T - L^T b_1 U^T) \\
&= \frac{1}{\beta}A_1^T(A_1(x) - b_1).
\end{aligned}$$

With similar discussion, The gradient of $\frac{1}{2\beta}\|A_2(x) - b_2\|^2$ is expressed as $\frac{1}{\beta}A_2^T(A_2(x) - b_2)$. Following the given HINT, we write $\text{dist}^2(Tx, \mathcal{K}) = \|k^* - Tx\|^2$, where $y^* = \text{argmin}_{y \in \mathcal{K}}\|k - Tx\|^2$, which means y^* is the projection of Tx on the positive orthant. The gradient of $\text{dist}^2(Tx, \mathcal{K})$ is

$$\begin{aligned}
\nabla\frac{1}{2\beta}\|k^* - Tx\|^2 &= \nabla\frac{1}{2\beta}(k^* - Tx)^T(k^* - Tx) \\
&= \nabla\frac{1}{2\beta}(k^{*T} - x^T T^T)(k^* - Tx) \\
&= \nabla\frac{1}{2\beta}(k^{*T} y^* - 2T^T k^{*T} Tx + x^T T^T Tx) \\
&= \frac{1}{\beta}(Tx - k^*) \\
&= \frac{1}{\beta}(T^T Tx - \text{proj}_{\mathcal{K}}(Tx)).
\end{aligned}$$

Thus, we verified that the gradient of the penalized objective is equal to v_k/β.

(c)

We calculate the gradient of the following objective function and quadratic penalty functions, $f(x) =< C, x >$, $\frac{1}{2\beta}\|A_1(x) - b_1\|^2 = \frac{1}{2\beta}\|X1 - 1\|^2$, $\frac{1}{2\beta}\|A_2(x) - b_2\|^2 = \frac{1}{2\beta}\|X^T 1 - 1\|^2$, and $\frac{1}{2\beta}\text{dist}^2(Tx, \mathcal{K}) = \frac{1}{2\beta}\|k^* - x\|^2$.

$$\nabla < C, x >= \nabla\text{Tr}(x^T C) = C$$

$$\begin{aligned}
\nabla \frac{1}{2\beta}\|x1-1\|^2 &= \frac{1}{2\beta}\nabla(x1-1)^T(x1-1) \\
&= \frac{1}{2\beta}\nabla(1^Tx^T-1^T)(x1-1) \\
&= \frac{1}{2\beta}\nabla(1^Tx^Tx1-1^Tx^T1-1^Tx1-1^T1) \\
&= \frac{1}{\beta}(x11^T-11^T)
\end{aligned}$$

$$\begin{aligned}
\nabla \frac{1}{2\beta}\|x^T1-1\|^2 &= \frac{1}{2\beta}\nabla(x^T1-1)^T(x^T1-1) \\
&= \frac{1}{2\beta}\nabla(1^Tx-1^T)(x^T1-1) \\
&= \frac{1}{2\beta}\nabla(1^Txx^T1-1^Tx1-1^Tx^T1-1^T1) \\
&= \frac{1}{\beta}(x^T11^T-11^T)
\end{aligned}$$

$$\begin{aligned}
\frac{1}{2\beta}\nabla\|k^*-x\|^2 &= \nabla\frac{1}{2\beta}(k^*-x)^T(k^*-x) \\
&= \frac{1}{2\beta}\nabla(k^{*T}-x^T)(k^*-x) \\
&= \frac{1}{2\beta}\nabla(k^{*T}k^*-2k^{*T}x+x^Tx) \\
&= \frac{1}{\beta}(x-k^*) \\
&= \frac{1}{\beta}(x-\text{proj}_{\mathcal{K}}(x)) \\
&= \frac{1}{\beta}(x-\max(x,0)) \\
&= \frac{1}{\beta}\min(x,0)
\end{aligned}$$

Thus, the explicit form of v_k is as follows.

$$v_k = \beta C + x11^T - 11^T + x^T11^T - 11^T + \min(x,0)$$

(d)

The k-means values before and after running the algorithm are 150.9680 and 28.7269 respectively. Obtained results and SDP solution are shown in the following figures. The final objective value is 54.54 and it is smaller than the given optimal value of 57.05. Due to the existence of the penalized objectives, the objective $f(x)$ is not converge to its optimal. If the penalized objectives have large weights, the optimal algorithm prioritize their convergence rather than $f(x)$.

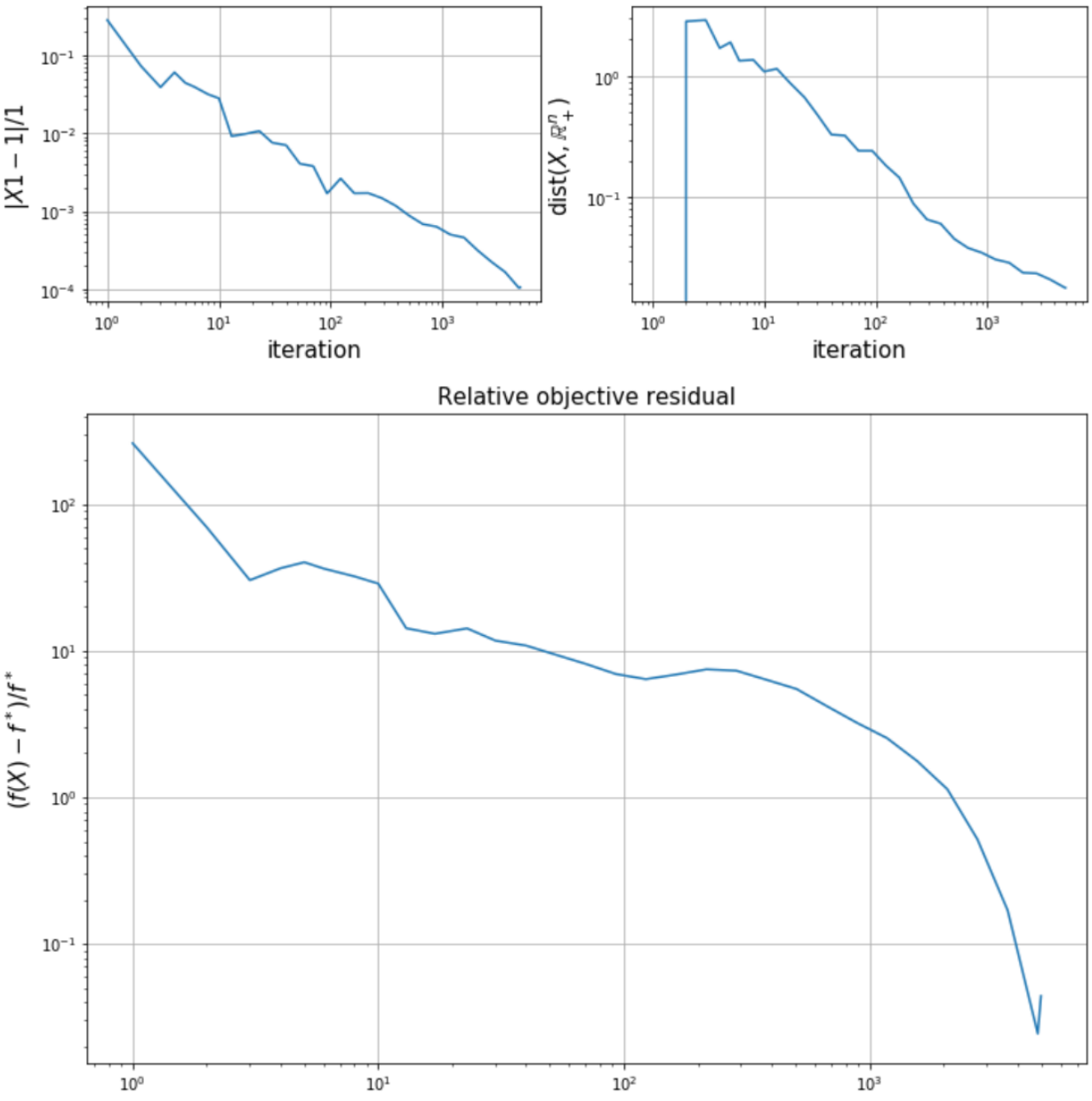

Figure 2: Results

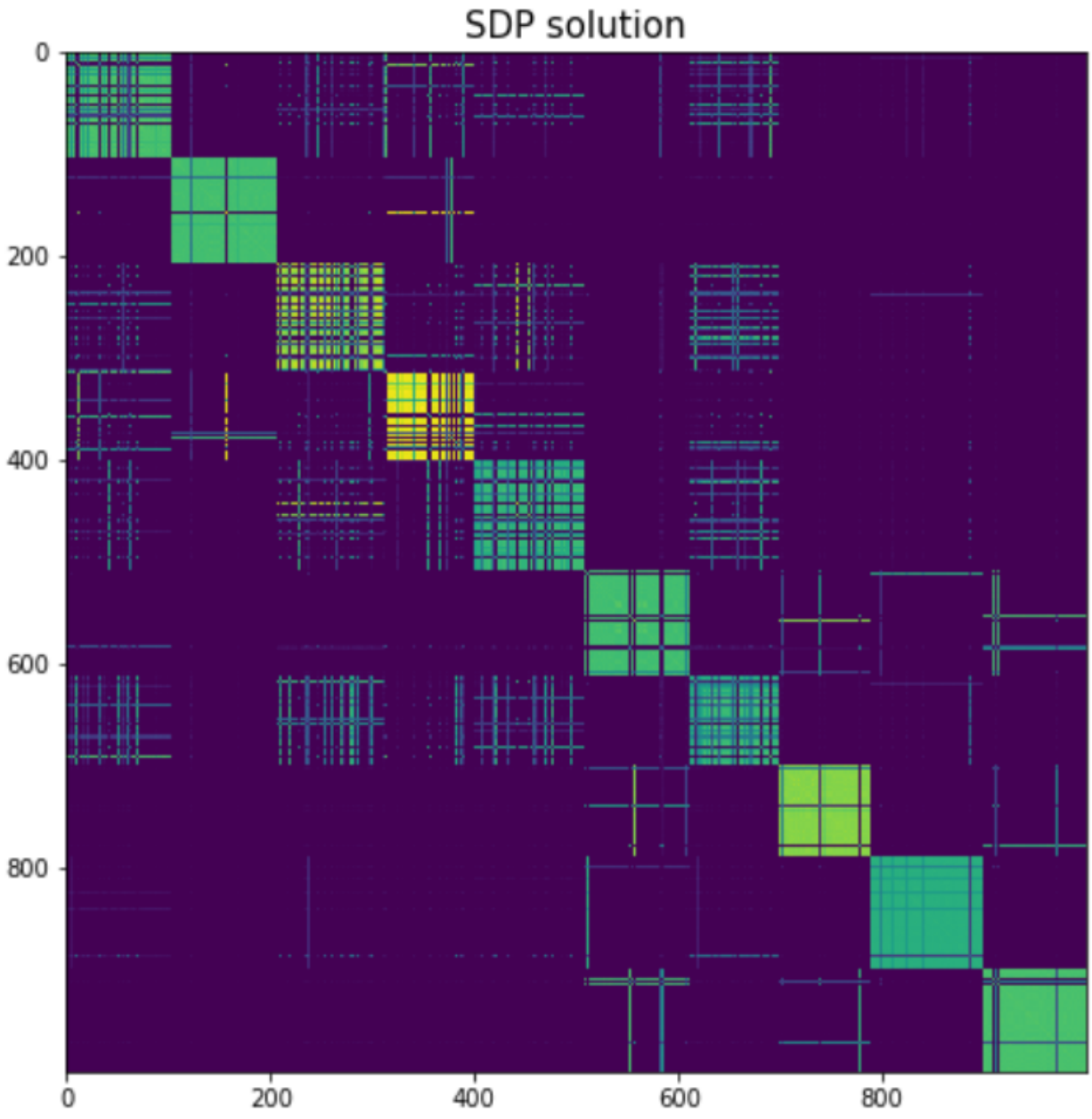

Figure 3: SDP Solution