Mathematics of Data **Fall 2019**

Homework #4 - Due date: 5th January 2020

Student: Oriol Barbany Mayor

Part I - Projection-free convex low-rank matrix optimization

Problem I.1 - Projection onto the nuclear norm ball

(a) The projection of Z onto $\mathcal{X}$ is given by

$$\Pi_{\mathcal{X}}(Z) = \underset{X \in \mathcal{X}}{\arg\min} \|X - Z\|_F \tag{1}$$

Using Mirsky's inequality and the definition of Frobenius norm, it follows that

$$\|X - Z\|_F \geq \|\Sigma_X - \Sigma_Z\|_F := \sqrt{\sum_{i=1}^{s}\sum_{j=1}^{s} |\Sigma_X(i,j) - \Sigma_Z(i,j)|^2} \tag{2}$$

$$= \sqrt{\sum_{i=1}^{s} |\sigma_X(i) - \sigma_Z(i)|^2} =: \|\sigma_X - \sigma_Z\|_2 \tag{3}$$

where $\Sigma_X, \Sigma_Z \in \mathbb{R}^{s\times s}$ are the diagonal matrices of the singular values of X, Z respectively.

Using the latter, we have that

$$\min_{X\in\mathcal{X}} \|X - Z\|_F \geq \min_{\Sigma_X \in \mathcal{X}} \|\Sigma_X - \Sigma_Z\|_F = \min_{\|\sigma_X\|_1 \leq \kappa} \|\sigma_X - \sigma_Z\|_2 \tag{4}$$

so we can equivalently minimize the left hand side and obtain a solution for (1). This latter has a minimum attained at $\sigma_X = \sigma_Z^{\ell_1}$, the projection of σ_Z onto the ℓ_1-norm ball of radius κ. This means that $\Sigma_Z^{\ell_1} := \text{diag}(\sigma_Z^{\ell_1})$ minimizes the equivalent matrix version. Finally, using (4) we have that

$$\Pi_{\mathcal{X}}(Z) = U\Sigma_Z^{\ell_1}V^T \tag{5}$$

(b) After performing 10 runs, the projection took 1.452±0.129 and 77.376±3.188 seconds for the 1K and 1M MovieLens datasets respectively.

Problem I.2 - LMO of nuclear norm

(a) Let $Z = U\Sigma V^T$ be the singular value decomposition of Z with $\sigma_{\max}$ its singular value and $\mathbf{u}$ and $\mathbf{v}$ the associated left and right singular vectors respectively. Since the matrices U and V are unitary,

$$\left\langle -\kappa\mathbf{u}\mathbf{v}^T, Z \right\rangle = Tr(-\kappa Z^T\mathbf{u}\mathbf{v}^T) = -\kappa Tr(V\Sigma U^T\mathbf{u}\mathbf{v}^T) = -\kappa\sigma_{\max}Tr(\mathbf{v}\mathbf{v}^T) = -\kappa\sigma_{\max}Tr(\mathbf{v}^T\mathbf{v}) \tag{6}$$

$$= -\kappa\sigma_{\max} = -\kappa\|Z\| \leq -\|X\|_*\|Z\| \leq -|\langle X, Z\rangle| \leq \langle X, Z\rangle \tag{7}$$

where the penultimate step holds by Hölder's inequality since the spectral norm is the dual of the nuclear norm. Given that $-\kappa\mathbf{u}\mathbf{v}^T \in \mathcal{X}$, it follows that $-\kappa\mathbf{u}\mathbf{v}^T \in \text{lmo}_{\mathcal{X}}(Z)$.

(b) After performing 10 runs, the projection took 0.029±0.005 and 0.294±0.008 seconds for the 1K and 1M MovieLens datasets respectively.

Part II - Crime Scene Investigation with Blind Deconvolution

(a)

Lemma 1. *For every linear operator $A : V \to W$, where V and W are finite-dimensional vector spaces, A can be expressed as a matrix.*

Proof. Let $\mathcal{B}_V = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ and $\mathcal{B}_W = \{\mathbf{f}_1, \dots, \mathbf{f}_m\}$ be a basis of V and W respectively. By linearity of A, we know that

$$A(c_1\mathbf{e}_1 + \cdots, c_n\mathbf{e}_n) = c_1A(\mathbf{e}_1) + \cdots + c_nA(\mathbf{e}_n) \tag{8}$$

and thus the output of any vector $\mathbf{x} \in V$ is fully determined by its decomposition in $\mathcal{B}_V$ and $\{A(\mathbf{e}_i)\}_{i=1}^n$. Moreover, since $A(\mathbf{e}_i) \in W \; \forall i \in \{1, \dots, n\}$, it can be represented as a linear combination of basis vectors in $\mathcal{B}_W$, i.e.

$$T(\mathbf{e}_i) = a_{i,1}\mathbf{f}_1 + \cdots + a_{i,m}\mathbf{f}_m \tag{9}$$

which means that the operator A can be implemented as a $M \times N$ matrix. □

Using Lemma 1 and with a slight abuse of notation naming $A(X) = AX$ as the linear operator expressed as a matrix multiplication, we have that

$$\nabla f(X) = A^T(AX - b) \tag{10}$$

Let L be the Lipschitz constant of ∇f.

$$\|\nabla f(X) - \nabla f(X)\| = \left\|A^T(AX - b) - A^T(AY - b)\right\| = \left\|A^TA(X - Y)\right\| \tag{11}$$

$$\leq \left\|A^TA\right\| \|X - Y\| =: L \|X - Y\| \tag{12}$$

where the inequality follows from the definition of the spectral norm.

(b) The result with $\kappa = 100$ and kernel support $K_1 = K_2 = 17$ is depicted in Figure 1, where one can easily read the plate with number J209LTL.

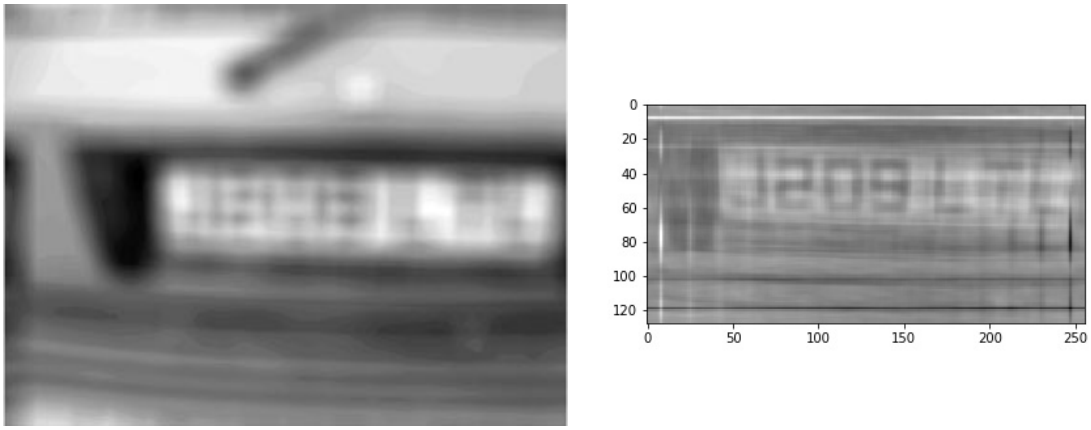

Figure 1: Blurred license plate (left) and result of the blind deconvolution (right).

Part III - K-Means Clustering by Semidefinite Programming

Problem III.1 - Conditional gradient method for clustering

(a) Let $X, Y \in \mathcal{X}$ any two matrices and $\alpha \in [0,1]$. Then, we have that $\alpha X + (1-\alpha)Y \in \mathcal{X}$ and thus the set $\mathcal{X}$ is convex iff $Tr(\alpha X + (1-\alpha)Y) \leq \kappa$ and $\alpha X + (1-\alpha)Y \succeq 0$.

Given that the trace is a linear operator,

$$Tr(\alpha X + (1-\alpha)Y) = \alpha Tr(X) + (1-\alpha)Tr(Y) \leq \alpha\kappa + (1-\alpha)\kappa = \kappa \tag{13}$$

where the inequality holds since $Tr(X), Tr(Y) \leq \kappa$.

By definition of positive semidefiniteness,

$$\alpha X + (1-\alpha)Y \succeq 0 \iff \mathbf{x}^T(\alpha X + (1-\alpha)Y)\mathbf{x} \geq 0 \qquad \forall \mathbf{x} \in \mathbb{R}^p \tag{14}$$

Since $X, Y \succeq 0$,

$$\mathbf{x}^T(\alpha X + (1-\alpha)Y)\mathbf{x} = \alpha \mathbf{x}^T X \mathbf{x} + (1-\alpha)\mathbf{x}^T Y \mathbf{x} \geq 0 \tag{15}$$

which concludes the proof.

(b) SDP relaxation can be formulated as

$$\min_{X \in \mathcal{X}} f(X) + g_1(A_1(X)) + g_2(A_2(X)) \quad \text{subject to } X \in \mathcal{K} \tag{16}$$

where g_1 and g_2 are the indicator functions of singletons $\{b_1\}$ and $\{b_2\}$ respectively. Writing this constraints in the quadratic penalty form yields:

$$g_1(A_1(X)) \longrightarrow \mathrm{QP}_{\{b_1\}}(X) = \min_{Y \in \{b_1\}} \|Y - A_1(X)\|^2 = \|b_1 - A_1(X)\|^2 \tag{17}$$

$$g_2(A_2(X)) \longrightarrow \mathrm{QP}_{\{b_2\}}(X) = \min_{Y \in \{b_2\}} \|Y - A_2(X)\|^2 = \|b_2 - A_2(X)\|^2 \tag{18}$$

$$X \in \mathcal{K} \longrightarrow \mathrm{QP}_{\mathcal{K}}(X) = \mathrm{dist}^2(X, \mathcal{K}) = \|\Pi_{\mathcal{K}}(X) - X\|^2 \tag{19}$$

where $\Pi_{\mathcal{K}}(X) = \arg\min_{Y \in \mathcal{K}} \|Y - X\|$ is the projection of X onto $\mathcal{K}$.

The penalized objective with penalization parameter $\frac{1}{2\beta}$ takes the form

$$f(X) + \frac{1}{2\beta}\|b_1 - A_1(X)\|^2 + \frac{1}{2\beta}\|b_2 - A_2(X)\|^2 + \frac{1}{2\beta}\mathrm{dist}^2(X, \mathcal{K}) \tag{20}$$

which has a gradient of

$$\nabla f(X) + \frac{1}{\beta}A_1^T(A_1X - b_1) + \frac{1}{\beta}A_2^T(A_2X - b_2) + \frac{1}{\beta}(X - \Pi_{\mathcal{K}}(X)) \tag{21}$$

where I used Lemma 1 to express the linear operators A_1, A_2 as a matrix and Danskin's theorem to derivative as if $\Pi_{\mathcal{K}}(X)$ was not a function of X.

(c) Following the proposed notation,

$$v_k := \beta\nabla f(X_k) + A_1^T(A_1X_k - b_1) + A_2^T(A_2X_k - b_2) + (X_k - \Pi_{\mathcal{K}}(X_k)) \tag{22}$$

so the gradient found in (21) can be expressed as $\frac{v_k}{\beta}$.

(d) The initial k-means value is 150.9680, and after running the algorithm it drops to 28.7269. The solution is depicted in Figure 3. The final objective value is below the optimal value (51.63 and 57.05 respectively), which is due to the problem relaxation. This latter includes all the feasible solutions of the original problem and thus its optimal solution, but also includes others that are not feasible on the original problem. This explains why the relative error slightly increases in the final iterations, even though it's hard to see in Figure 2.

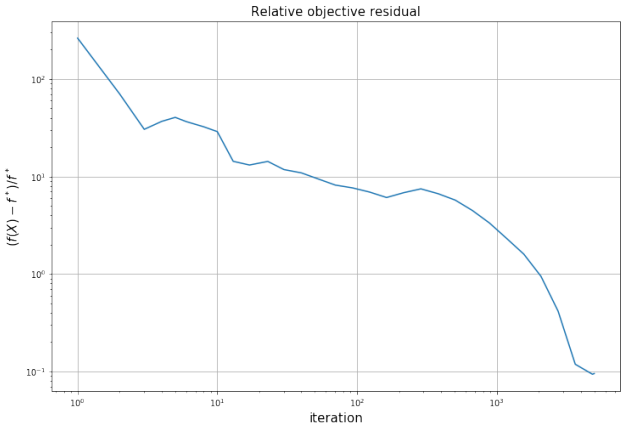

Figure 2: Evolution of relative objective residual with f^* the optimal value of the original problem.

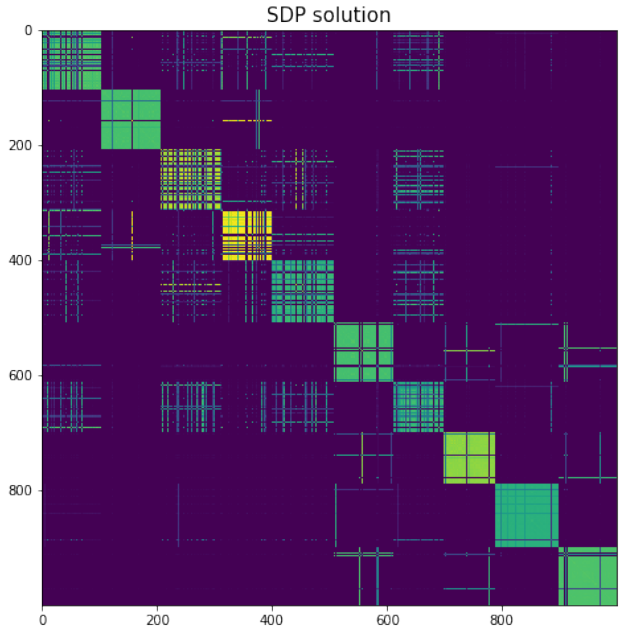

Figure 3: Final SDP solution to the relaxation of $k-$means clustering.