École polytechnique fédérale de Lausanne

Mathematics of data: from theory to computation

Laboratory for information and inference systems

Fall 2021

# Homework 3

Bruno Rodriguez Carrillo

Sciper 326180

Professor:
Volkan Cevher

Head TA's:
Fabian Latorre
Ali Kavis

January 7th, 2022

# 1 Crime scene investigation with blind image deconvolution

## 1.1 Computing projections onto $\mathcal{X}$

1. From the lecture notes, we have

$$\delta_{\mathcal{X}}(\mathbf{Z}) = \begin{cases} 0, & \mathbf{Z} \in \mathcal{X} \\ +\infty, & \text{otherwise} \end{cases}$$

and

$$\text{prox}_{\delta_{\mathcal{X}}}(\mathbf{Z}) = \underset{\mathbf{X} \in \mathbb{R}^{p \times m}}{\text{argmin}} \left\{ \delta_{\mathcal{X}}(\mathbf{X}) + \frac{1}{2}\|\mathbf{X} - \mathbf{Z}\|_2^2 \right\} = \underset{\mathbf{X} \in \mathcal{X}}{\text{argmin}} \left\{ \frac{1}{2}\|\mathbf{X} - \mathbf{Z}\|_2^2 \right\}$$

Since the matrix $\mathbf{X}$ is a 1-rank matrix, we have that $\|A\|_2 = \|A\|_F$
Consequently, the result follows; that is

$$\text{proj}_{\mathcal{X}}(\mathbf{Z}) = \text{prox}_{\delta_{\mathcal{X}}}(\mathbf{Z}).$$

2. From the given hint

$$\mathbf{z}^* = \text{proj}_{\mathcal{X}}(\mathbf{x}) \Leftrightarrow \langle \mathbf{x} - \mathbf{z}^*, \mathbf{z} - \mathbf{z}^* \rangle \leq 0, \forall \mathbf{z} \in \mathcal{X}$$

and by definition $\text{proj}_{\mathcal{X}}(\mathbf{y}) \in \mathcal{X}$ for any $\mathbf{y}$; thus we can replace $\mathbf{z}$ by $\text{proj}_{\mathcal{X}}(\mathbf{y})$, which leads to

$$\langle \mathbf{x} - \text{proj}_{\mathcal{X}}(\mathbf{x}), \text{proj}_{\mathcal{X}}(\mathbf{y}) - \text{proj}_{\mathcal{X}}(\mathbf{x}) \rangle \leq 0 \ (*)$$

similarly,

$$\mathbf{w}^* = \text{proj}_{\mathcal{X}}(\mathbf{y}) \Leftrightarrow \langle \mathbf{y} - \mathbf{w}^*, \mathbf{w} - \mathbf{w}^* \rangle \leq 0, \forall \mathbf{w} \in \mathcal{X}$$

Using similar arguments as before, we obtain:

$$\langle \text{proj}_{\mathcal{X}}(\mathbf{y}) - \mathbf{y}, \text{proj}_{\mathcal{X}}(\mathbf{y}) - \text{proj}_{\mathcal{X}}(\mathbf{x}) \rangle \leq 0 \ (**)$$

Adding (*) and (**) and re-arraging terms, we have

$$\langle \text{proj}_{\mathcal{X}}(\mathbf{y}) - \text{proj}_{\mathcal{X}}(\mathbf{x}), \text{proj}_{\mathcal{X}}(\mathbf{y}) - \text{proj}_{\mathcal{X}}(\mathbf{x}) \rangle \leq \langle \mathbf{y} - \mathbf{x}, \text{proj}_{\mathcal{X}}(\mathbf{y}) - \text{proj}_{\mathcal{X}}(\mathbf{x}) \rangle$$

Now we apply the Cauchy-Schwarz inequality and the definition of the standard inner product to the left-hand and right-hand sides of the above expression; we obtain

$$\langle \text{proj}_{\mathcal{X}}(\mathbf{y}) - \text{proj}_{\mathcal{X}}(\mathbf{x}), \text{proj}_{\mathcal{X}}(\mathbf{y}) - \text{proj}_{\mathcal{X}}(\mathbf{x}) \rangle \leq \|\text{proj}_{\mathcal{X}}(\mathbf{y}) - \text{proj}_{\mathcal{X}}(\mathbf{x})\|^2$$

and

$$\langle \mathbf{y} - \mathbf{x}, \text{proj}_{\mathcal{X}}(\mathbf{y}) - \text{proj}_{\mathcal{X}}(\mathbf{x}) \rangle \leq \|\mathbf{y} - \mathbf{x}\| \, \|\text{proj}_{\mathcal{X}}(\mathbf{y}) - \text{proj}_{\mathcal{X}}(\mathbf{x})\|$$

This implies that

$$\|\text{proj}_{\mathcal{X}}(\mathbf{x}) - \text{proj}_{\mathcal{X}}(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$$

as requested to prove.

3. Let us consider the SVD of $\mathbf{Z} \in \mathbb{R}^{p \times m}$ as $\mathbf{Z} = \mathbf{W}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{T}^T$, where $\mathbf{W} \in \mathbb{R}^{p \times p}$ and $\mathbf{T}^T \in \mathbb{R}^{m \times m}$.
We are requested to project $\mathbf{Z}$ onto the nuclear norm ball $\mathcal{X} = \{\mathbf{X} : \mathbf{X} \in \mathbb{R}^{p \times m}, \|\mathbf{X}\| \leq \kappa\}$. Furthermore, we know that such a projection can be computed by:

$$\text{proj}_{\mathcal{X}}(\mathbf{Z}) = \underset{\mathbf{X} \in \mathcal{X}}{\text{argmin}} \left\{ \|\mathbf{X} - \mathbf{Z}\|_F^2 \right\}, \forall \mathbf{Z} \in \mathbb{R}^{p \times m}$$

We have also the Mirsky's inequality:

$$\|\mathbf{X} - \mathbf{Z}\|_F \geq \|\mathbf{\Sigma}_{\mathbf{X}} - \mathbf{\Sigma}_{\mathbf{Z}}\|_F$$

Now, we know that the projection of $\mathbf{Z}$ is such that it minimizes the distance onto the space it is projected; that is

$$\|\text{proj}_{\mathcal{X}}(\mathbf{Z}) - \mathbf{\Sigma}_{\mathbf{Z}}\|_F \leq \|\mathbf{Y} - \mathbf{Z}\|_F, \forall \mathbf{Y} \in \mathcal{X}$$

Thus, we realize that by using Mirsky's inequality, the projection $\text{proj}_{\mathcal{X}}(\mathbf{Z})$ is such that $\|\mathbf{Y} - \mathbf{Z}\|_F = \|\mathbf{\Sigma}_{\mathbf{Y}} - \mathbf{\Sigma}_{\mathbf{Z}}\|_F$. This means that the projection is such that the Mirsky's inequality holds for the equality.

In addition, let us consider the SVD of $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}_{\mathbf{Y}}\mathbf{V}^T$ and recall that $\|\mathbf{\Sigma}_{\mathbf{Y}} - \mathbf{\Sigma}_{\mathbf{Z}}\|_F^2 = \text{tr}\left( \left( \mathbf{\Sigma}_{\mathbf{Y}} - \mathbf{\Sigma}_{\mathbf{Z}} \right)^T \left( \mathbf{\Sigma}_{\mathbf{Y}} - \mathbf{\Sigma}_{\mathbf{Z}} \right) \right)$.

As a result, using the previous SVD's and the Mirsky's equality, we have

$$\begin{aligned} \|\mathbf{Y} - \mathbf{Z}\|_F^2 = \text{tr}\left( \mathbf{V}\mathbf{\Sigma}_{\mathbf{Y}}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}_{\mathbf{Y}}\mathbf{V}^T \right) + \text{tr}\left( \mathbf{T}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{W}^T\mathbf{W}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{T}^T \right) \\ -\text{tr}\left( \mathbf{T}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{W}^T\mathbf{U}\mathbf{\Sigma}_{\mathbf{Y}}\mathbf{V}^T \right) - \text{tr}\left( \mathbf{V}\mathbf{\Sigma}_{\mathbf{Y}}\mathbf{U}^T\mathbf{W}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{T}^T \right) \end{aligned}$$

Using the fact that the matrices $\mathbf{U}, \mathbf{V}$, $\mathbf{W}$ and $\mathbf{T}$ are orthogonal, and $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ for any matrices $\mathbf{A}, \mathbf{B}$ such that the products can be computed.

Thus we have:

$$\|\mathbf{Y} - \mathbf{Z}\|_F^2 = \text{tr}\left( \mathbf{\Sigma}_{\mathbf{Y}}\mathbf{\Sigma}_{\mathbf{Y}} \right) + \text{tr}\left( \mathbf{\Sigma}_{\mathbf{Z}}\mathbf{\Sigma}_{\mathbf{Z}} \right) - 2\text{tr}\left( \mathbf{T}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{W}^T\mathbf{U}\mathbf{\Sigma}_{\mathbf{Y}}\mathbf{V}^T \right)$$

Following a similar procedure, we have

$$\|\mathbf{\Sigma}_{\mathbf{Y}} - \mathbf{\Sigma}_{\mathbf{Z}}\|_F^2 = \text{tr}\left( \mathbf{\Sigma}_{\mathbf{Y}}\mathbf{\Sigma}_{\mathbf{Y}} \right) + \text{tr}\left( \mathbf{\Sigma}_{\mathbf{Z}}\mathbf{\Sigma}_{\mathbf{Z}} \right) - 2\text{tr}\left( \mathbf{\Sigma}_{\mathbf{Y}}\mathbf{\Sigma}_{\mathbf{Z}} \right)$$

Consequently, for the Mirsky's equality to hold we need:

$$-2\text{tr}\left(\mathbf{T}\mathbf{\Sigma_Z}\mathbf{W}^T\mathbf{U}\mathbf{\Sigma_Y}\mathbf{V}^T\right) = -2\text{tr}\left(\mathbf{\Sigma_Y}\mathbf{\Sigma_Z}\right)$$

This, in turn, holds if $\mathbf{W} = \mathbf{U}$ and $\mathbf{V} = \mathbf{T}$. That is, the projected left and right singular vectors are the same as the ones from the original matrix $\mathbf{Z}$.

As a consequence, we have that projecting $\mathbf{Z}$ onto $\mathcal{X}$ is equivalent to projecting its singular values onto the $\ell_1$-norm. Obviously, such a projection is carried out by forming a vector containing the diagonal elements of $\mathbf{\Sigma_Z}$ and then projecting onto the $\ell_1$-norm. This proves the claim.

## 1.2 Computing the linear minimization oracle $\mathcal{X}$

1. Let us consider the SVD of $\mathbf{Z} \in \mathbb{R}^{p\times m}$ as $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma_Z}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{p\times p}$ and $\mathbf{V}^T \in \mathbb{R}^{m\times m}$.

   Computing the matrix product of $\sigma_1\mathbf{u}_1\mathbf{v}_1^T$ and $\mathbf{Z}$, where $\mathbf{u}_1$ and $\mathbf{v}_1^T$ are the left and singular vectors of the largest singular value of $\mathbf{Z}$, $\sigma_1$.

$$\left\langle \mathbf{u}_1\mathbf{v}_1^T, \mathbf{Z}\right\rangle = \text{tr}\left(\mathbf{Z}^T\sigma_1\mathbf{u}_1\mathbf{v}_1^T\right) = \text{tr}\left(\mathbf{V}\mathbf{\Sigma_Z}\mathbf{U}^T\sigma_1\mathbf{u}_1\mathbf{v}_1^T\right) = \sigma_1\text{tr}\left(\sum_{i=1}^{\min\{p,m\}}\mathbf{v}_i\mathbf{u}_i^T\mathbf{u}_1\mathbf{v}_1^T\right)$$

$$\rightarrow -\kappa\left\langle \mathbf{u}_1\mathbf{v}_1^T, \mathbf{Z}\right\rangle = -\kappa\sigma_1$$

   since $\mathbf{u}_i^T\mathbf{u}_j = 0$ for $i \neq j$ and $\text{tr}(\mathbf{v}_1\mathbf{v}_1^T) = \mathbf{v}_1^T\mathbf{v}_1 = 1$.

   From [1], we have that for Schatten $p$-norms given matrix $\mathbf{A}$, $\|\mathbf{A}\|_\infty = \sigma_{\max}(\mathbf{A})$, where $\sigma_{\max}(\mathbf{A})$ is the largest singular value of such a matrix. From the same reference, we have the following inequality

$$\left|\text{tr}\left(\mathbf{Z}^T\mathbf{X}\right)\right| \leq \|\mathbf{Z}\|_p\|\mathbf{X}\|_q$$

   where $1/p + 1/q = 1$. Using $q = 1$ (this corresponds to the nuclear norm), $p \rightarrow \infty$, $\|\mathbf{X}\|_1 \leq \kappa$ and $\|\mathbf{Z}\|_\infty = \sigma_1$, we obtain the following

$$-\kappa\left\langle \mathbf{u}_1\mathbf{v}_1^T, \mathbf{Z}\right\rangle = -\kappa\sigma_1 = -\|\mathbf{Z}\|_\infty\|\mathbf{X}\|_1 \leq \text{tr}\left(\mathbf{Z}^T\mathbf{X}\right).$$

   Thus, we have that $\langle \mathbf{X}, \mathbf{Z}\rangle \geq -\kappa\left\langle \mathbf{u}_1\mathbf{v}_1^T, \mathbf{Z}\right\rangle$, which proves the claim.

## 1.3 Comparing the scalability of $\text{proj}_{\mathcal{X}}(\mathbf{Z})$ and $\text{lmo}_{\mathcal{X}}(\mathbf{Z})$

1. See next point.

2. We run the codes five time for both $\text{proj}_{\mathcal{X}}(\mathbf{Z})$ and $\text{lmo}_{\mathcal{X}}(\mathbf{Z})$. Computing the average timing, we have the following chart

| method | 100k data | 1M data |
|---|---|---|
| proj_nuc | 1.5318 | 119.1945 |
| lmo_nuc | 0.01137 | 0.1037 |

Table 1: Average time for 5 runs

As observed, computation time of the linear minimization oracle lmo is much faster than its counterpart projector. Such difference is more evident when we run on the 1M data set. This may happen since the lmo is computed using a single singular value. We should also keep in mind that lmo, we implement the SVD using a Python method for sparse matrices, which is faster than the standard thin SVD.

## 1.4 Frank-Wolfe for blind image deconvolution

1. The objective function is $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{A}(\mathbf{X}) - \mathbf{b}\|_2^2$, where $\|\mathbf{X}\|_* \leq \kappa, \mathbf{X} \in \mathbb{R}^{pxm}$, then computing the gradient and taking norms we have

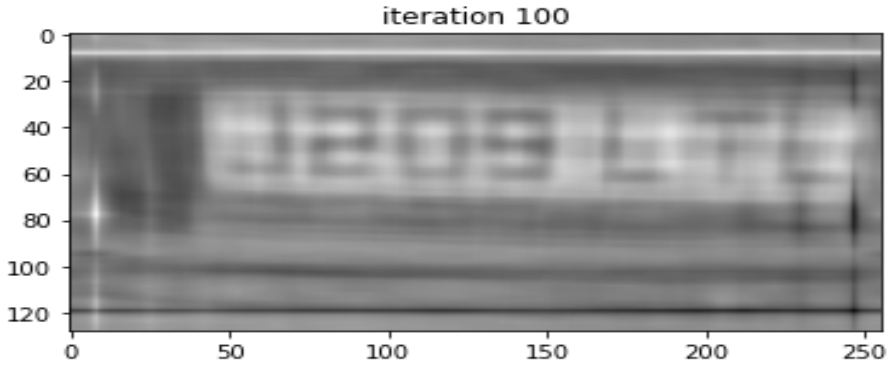
$$\|\nabla f(\mathbf{Y}) - \nabla f(\mathbf{X})\|_2 = \left\|\mathbf{A}^\top(\mathbf{A}(\mathbf{Y}) - \mathbf{b}) - \mathbf{A}^\top(\mathbf{A}(\mathbf{X}))\right\|_2 = \left\|\mathbf{A}^\top\mathbf{A}(\mathbf{Y} - \mathbf{X})\right\|_2$$

where we have applied the linearity of the operator $\mathbf{A}(\alpha\mathbf{X} + \beta\mathbf{Y}) = \alpha\mathbf{A}(\mathbf{X}) + \beta\mathbf{A}(\mathbf{Y})$, thus we have

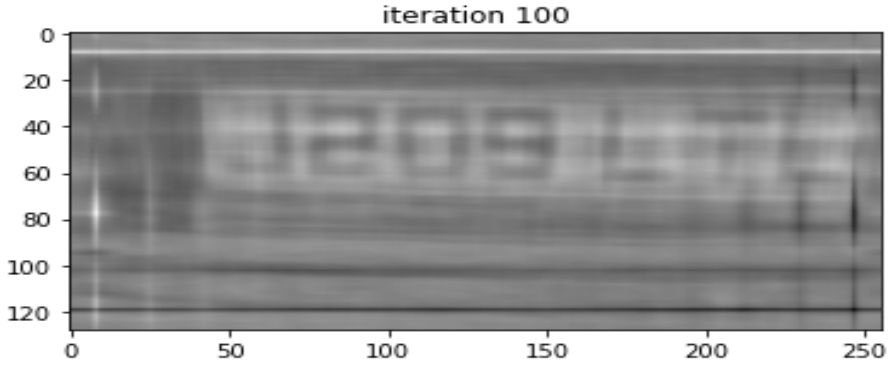
$$\|\nabla f(\mathbf{Y}) - \nabla f(\mathbf{X})\|_2 \leq \left\|\mathbf{A}^\top\mathbf{A}\right\|_2 \|\mathbf{Y} - \mathbf{X}\|_2$$

Then the Lipschitz $L = \left\|\mathbf{A}^\top\mathbf{A}\right\|_2$.

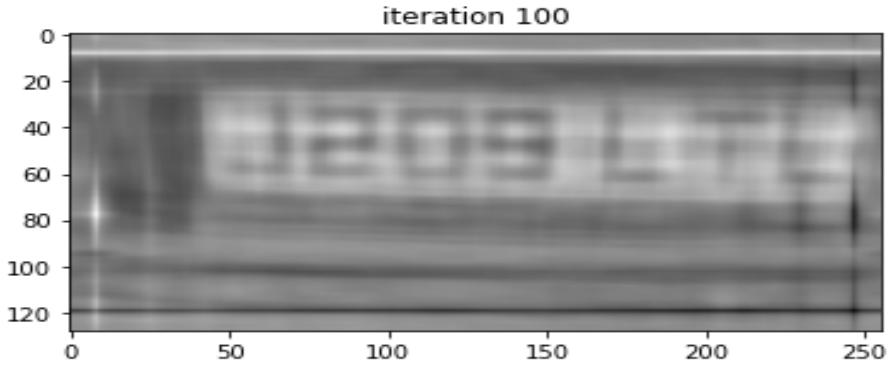
2. After completing the missing lines in the provided codes, we have the following images for two different values of $\kappa$ for 100 iterations and kernel size (17, 17) each




(a) $\kappa = 500$.




(b) $\kappa = 1000$.




(c) $\kappa = 2000$.

Figure 1: Plate number deblurring.

As can be seen, the plate number is J209LTL.

# 2 Hands-on experiment 2: $k$-means clustering by semidefinite programming(SDP)

## 2.1 Methods for clustering the fashion-MNIST data

1. We have the domain $\mathcal{X} = \{\mathbf{X} : \text{tr}(\mathbf{X}) \leq \kappa, \mathbf{X} \in \mathbb{C}^{p\times p}, \mathbf{X} \geq 0\}$

   Moreover, we have the definition of set convexity: the set $\mathcal{X}$ is convex if for all $\mathbf{X}, \mathbf{Y} \in \mathcal{X}$ and $0 \leq \mu \leq 1$ with $\mu \in \mathbb{R}$, $\mu\mathbf{Y} + (1-\mu)\mathbf{X} \in \mathcal{X}$ holds.

   Then, if we assume $\mathbf{X}, \mathbf{Y} \in \mathcal{X}$, then $\text{tr}\,(\mathbf{X}) \leq \kappa$ and $\text{tr}\,(\mathbf{Y}) \leq \kappa$; we obtain

   $$\text{tr}\,(\mu\mathbf{Y} + (1-\mu)\mathbf{X}) = (\mu)\text{tr}\,(\mathbf{Y}) + (1-\mu)\text{tr}\,(\mathbf{X}) \leq \kappa$$

   where we have used the fact that for any matrices $\mathbf{A}$, $\mathbf{B}$, $\text{tr}\,(\mathbf{A}+\mathbf{B}) = \text{tr}\,(\mathbf{A}) + \text{tr}\,(\mathbf{A})$.

   Consequently, the set $\mathcal{X}$ is convex.

2. We are given liner inclusion constraint $Tx \in \mathcal{Y}$ and the corresponding quadratic penalty function is given by

   $$QP_{\mathcal{Y}}(x) = \text{dist}^2(Tx, \mathcal{Y}) = \min_{y\in\mathcal{Y}} \|y - Tx\|^2 \ .$$

   Similarly, from the problem description, we have

   $$g_1 = \delta_{b_1}(A_1(x)),\ g_2 = \delta_{b_2}(A_2(x)) \text{ and } \mathcal{K} \text{ a positive orthant}$$

   and it can been clearly seen from the problem description that the constraints of the problem are

   $$A_1(\mathbf{X}) = b_1, A_2(\mathbf{X}) = b_2 \text{ and } B(\mathbf{X}) \in \mathcal{K}$$

   Here we do not consider $\text{tr}(\mathbf{X}) \leq \kappa$ as a constraint since it is given in the description of the set $\mathcal{X}$ along with the positive-definite characterization of $\mathbf{X}$.

   From such constraints and based on the examples shown in the lecture notes, their quadratic form are:

   $$QP_{\mathcal{X},1}(x) = \min_{x\in\mathcal{X}} \|A_1(x) - b_1\|^2,\ QP_{\mathcal{X},2}(x) = \min_{x\in\mathcal{X}} \|A_2(x) - b_2\|^2,\ QP_{\mathcal{X},3}(x) = \text{dist}^2(x, \mathcal{K})$$

   the penalized objective function $f_{\text{penalized}}(x)$ consequently is

   $$f_{\text{penalized}}(x) = f(x) + \frac{1}{2\beta}\|A_1(x) - b_1\|^2 + \frac{1}{2\beta}\|A_2(x) - b_2\|^2 + \frac{1}{2\beta}\,\text{dist}^2\,(x, \mathcal{K})$$

   where $\beta$ is the penalty parameter.

From the previous point, we compute the gradient of the penalized objective function $f_{\text{penalized}}(x)$; computing $\nabla_{x_k}$ separately for each term, we have the following computations. Since $f(x_k) = \langle \mathbf{C}, x_k \rangle$, we have

$$\nabla_{x_k} f(x_k) = \text{tr}\left(\mathbf{C}^T x_k\right) = \mathbf{C}$$

Regarding the term $\frac{1}{2\beta} \|A_1(x) - b_1\|^2$, we have:

$$\nabla_{x_k} \left( \frac{1}{2\beta} \|A_1(x_k) - b_1\|^2 \right) = \frac{1}{\beta} A_1^T \left( A_1(x_k) - b_1 \right)$$

We can proceed similarly for $\frac{1}{2\beta} \|A_2(x) - b_2\|^2$ and from the lecture notes

$$\nabla_{x_k} \left( \frac{1}{2\beta} \text{dist}^2 \left( x_k, \mathcal{K} \right) \right) = \frac{1}{\beta} \left( x_k - \text{proj}_{\mathcal{K}}(x_k) \right)$$

then putting all gradients together we have

$$\nabla_{x_k} f_{\text{penalized}}(x_k) = \mathbf{C} + \frac{1}{\beta} A_1^T \left( A_1(x_k) - b_1 \right) + \frac{1}{\beta} A_2^T \left( A_2(x_k) - b_2 \right) + \frac{1}{\beta} \left( x_k - \text{proj}_{\mathcal{K}}(x_k) \right),$$

which corresponds to $v_k/\beta_k$, taking $\beta = \beta_k$ given in the Vu-Condat algorithm provided in the homework sheet

$$v_k = \beta_k \nabla f(x_k) + A_1^T \left( A_1(x_k) - b_1 \right) + A_2^T \left( A_2(x_k) - b_2 \right) + \left( x_k - \text{proj}_{\mathcal{K}}(x_k) \right)$$

Then $\nabla_{x_k} f_{\text{penalized}}(x_k) = v_k/\beta$.

3. In the homework sheet, we are given the function $f(x_k) = \langle \mathbf{C}, x_k \rangle = \text{tr}\left(\mathbf{C}^T x_k\right)$. From [1], we know that the gradient of the trace is given by

$$\nabla_{x_k} f(x_k) = \mathbf{C}$$

Regarding the term $(x_k - \text{proj}_{\mathcal{K}}(x_k))$, we have that $\mathcal{K}$ is a positive orthant, then when we project onto $\mathcal{K}$, we are solving the following minimization problem

$$\text{proj}_{\mathcal{K}}(x_k) = \arg \min_{z \in \mathcal{K}, z \geq 0} \|z - x_k\|,$$

where $\mathcal{K}$ represents a positive orthant; that is

$$\mathcal{K} = \{ z : z_i \geq 0, \forall i \}$$

such a minimization problem has a trivial solution $\hat{z}$ given by:

$$\hat{z} = [x_k]_+ \text{ where } [\cdot]_+ = \max\{\cdot, 0\}$$

and $[x_k]_+$ is computed element-wise, which leads to $\hat{z}^i = [x_k^i]_+ = \max\left\{x_k^i, 0\right\}, \forall i$. The super index $i$ is not an exponent but the $i$-th element if $x_k$. Furthermore, we have that the projection term is now:

$$x_k - \text{proj}_{\mathcal{K}}(x_k) = x_k - \max\{x_k, 0\} = \min\{x_k, 0\}$$

Finally, the term $v_k$ specific to problem (4) is

$$\beta_k \mathbf{C} + A_1^T\left(A_1(x_k) - b_1\right) + A_2^T\left(A_2(x_k) - b_2\right) + \min\{x_k, 0\},$$

where as before, $\min\{x_k, 0\}$ is computed element-wise.

4. We have the following definition for $A$

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} A_1 x \\ A_2 x \\ Bx \end{bmatrix}, \; A = \begin{bmatrix} A_1 \\ A_2 \\ B \end{bmatrix} \rightarrow z = A(x)$$

Using the Moreau's decomposition, we have

$$y^{k+1} = \text{prox}_{\sigma g^x}\left(y^k + \sigma A\left(\tilde{x}^{k+1}\right)\right) = y^k + \sigma A\left(\tilde{x}^{k+1}\right) - \sigma \text{prox}_{\sigma^{-1} g}\left(\sigma^{-1} y^k + A\left(\tilde{x}^{k+1}\right)\right)$$

Now, we observe that the function $\delta_g(z) = g(z) = \delta_{b_1}(z_1) + \delta_{b_2}(z_2) + \delta_{\mathcal{K}}(z_3)$ is an indicator function where

$$\delta_g(z) = \begin{cases} 0, & z_1 \in b_1, z_2 \in b_2, z_3 \in \mathcal{K} \\ +\infty, & \text{otherwise} \end{cases}.$$

It is important to keep in mind that the function $\delta_g(z)$ is zero if all the conditions in its definition hold; that is, if at least $z_1, z_2$ or $z_3$ do not fulfill the constraints, $\delta_g(z)$ goes to infinity.

As a result and from point 1.a, the proximal operator is equivalent to the projection operator. We can use this fact to view that

$$\text{prox}_{\sigma^{-1} g}\left(\sigma^{-1} y^k + A\left(\tilde{x}^{k+1}\right)\right) = \text{proj}_{\sigma^{-1} g}\left(\sigma^{-1} y^k + A\left(\tilde{x}^{k+1}\right)\right)$$

If we project $\text{proj}_{\sigma^{-1} g}\left(\sigma^{-1} y^k + A\left(\tilde{x}^{k+1}\right)\right)$ onto $b_1$, we obtain $b_1$ itself. A similar argument can be used for $b_2$. On the other hand, when we project onto $\mathcal{K}$, we actually compute the projection explicitly for the component $y_3^k$. Then we have that the proximal operator is

$$\text{prox}_{\sigma^{-1} g}\left(\sigma^{-1} y^k + A\left(\tilde{x}^{k+1}\right)\right) = \begin{bmatrix} b_1 \\ b_2 \\ \text{proj}_{\mathcal{K}}\left(\sigma^{-1} y_3^k + \tilde{x}^{k+1}\right) \end{bmatrix}$$

Combining this with the Moreau's decomposition and using the definition $A(x)$, we have

$$y^{k+1} := \begin{bmatrix} y_1^{k+1} \\ y_2^{k+1} \\ y_3^{k+1} \end{bmatrix} = \begin{bmatrix} y_1^{k} \\ y_2^{k} \\ y_3^{k} \end{bmatrix} + \sigma \begin{bmatrix} A_1\widetilde{x}^{k+1} - b_1 \\ A_2\widetilde{x}^{k+1} - b_2 \\ \widetilde{x}^{k+1} - \mathrm{proj}_k\left(\sigma^{-1}y_3^k + \tilde{x}^{k+1}\right) \end{bmatrix}$$

We now observe that we can use the definition of $A(x)$ to compute $A^T(x)$ and if we apply $A^T(x)$ to the decomposition above, we obtain the requested result

$$A^\top y^{k+1} = A^\top y^k + \sigma\left(A_1^T\left(A_1\left(\tilde{x}^{k+1}\right) - b_1\right) + A_2^T\left(A_2\left(\tilde{x}^{k+1}\right) - b_2\right) + \tilde{x}^{k+1} - \mathrm{proj}_k\left(\sigma^{-1}y_3^k + \tilde{x}^{k+1}\right)\right))$$

We point out that the we keep in mind that the operator $B(x)$ just ensures that element-wise $x \geq 0$ and then when computing $B^T$, there is no change in its 'effect'.

5. After running both the algorithms, we have the following final objective values

| | |
|---|---|
| k-means value initial | 150.9680 |
| k-means value for HCGM | 28.7269 |
| k-means value for Vu-Condat | 28.7269 |

Table 2: k-means values for different runs

| run | run 1 | run 2 | run 3 | run 4 | run 5 | run 6 | run 7 | run 8 |
|---|---|---|---|---|---|---|---|---|
| k-means value | 28.732 | 294.404 | 129.519 | 107.818 | 181.481 | 178.488 | 114.810 | 210.431 |

Table 3: k-means values for different runs

and the misclassification rate for HCGM and Vu-Condat is 0.1250.

# 3 Hands-on experiment 3: Computing a geometric embedding for sparsest cut problem via SDP

1. We have the following constraints where $p$ is the number of total nodes $V$ in the graph $G = (V, E)$

$$\text{Constraints 1: } p\,\mathrm{tr}(\mathbf{X}) - \mathrm{tr}\left(\mathbf{1}_{p\times p}\mathbf{X}\right) = \frac{p^2}{2} \rightarrow\equiv A(\mathbf{X}) = \frac{p^2}{2}.$$

$$\text{Constraints 2: } \mathbf{X}_{i,j} + \mathbf{X}_{j,k} - \mathbf{X}_{i,k} - \mathbf{X}_{j,j} \leq 0, \forall i \neq j \neq k \neq i \in V \rightarrow\equiv B_{i,j,k}(\mathbf{X}) \in \mathbf{X} = (-\infty, 0].$$

$$\text{Constraints 3: } \mathrm{tr}(\mathbf{X}) \leq p, \mathbf{X} \in \mathbb{R}^{p\times p}, \mathbf{X} \succcurlyeq 0 \rightarrow\equiv \mathbf{X} \in \mathcal{X}.$$

We observe that *Constraints 1* has $p$ constraints since the trace operator acts on the diagonal elements. The matrix multiplication $\text{tr}(\mathbf{1}_{p\times p}\mathbf{X}) = \text{tr}(\mathbf{X})$ which has $p$ constraints. Then *Constraints 2* is of order $O(p)$.

In case of *Constraints 2*, we compute the order in big-O notation for the these constraints. That is, we are interested in the upper bound for the number of constraints. By observing *Constraints 2*, we notice that in the worst case, $i, j, k$ will run a number of times that is proportional to $p^3$ since the graph G has $p$ nodes. As a consequence, *Constraints 2* is of order $O(p^3)$.

Similarly, for *Constraints 3*, we observe that as before the constraints due to the trace are of order $O(p)$. On the other hand, since we also require that each element in $\mathbf{X} \geq 0$, this number of constraints is of order $O(p^2)$.

Finally, the total number of constraints is upper bound by a number proportional to $p^3+p^2+p$; that is, the total number of constraints is of order $O(p^3)$.

Even if we do not consider *Constraints 3* as a set of constraints since it is the domain where we look for $\mathbf{X}$, the upper bound still holds asymptotically for sufficiently large $p$.

In contrast and by following a similar analysis, problem (3) is of order $O(p^2)$.

2. As we did previously, the constraints in the penalty form are

$$QP_{\mathcal{X},1}(x) = \min_{\mathbf{X}\in\mathcal{X}} \|A(\mathbf{X}) - p^2/2\|^2,\ QP_{\mathcal{X},2}(\mathbf{X}) = \text{dist}^2(\mathbf{X}, \mathcal{K}),$$

where $A(\mathbf{X})$ is given above. Then the penalized objective function is

$$f(\mathbf{X}) + \frac{1}{2\beta}\|A(\mathbf{X}) - p^2/2\|^2 + \frac{1}{2\beta}\text{dist}^2(\mathbf{X}, \mathcal{K}),$$

where $\mathcal{K} = (-\infty, 0]$, $\mathcal{X} = \{\mathbf{X} : \text{tr}(\mathbf{X}) \leq p, \mathbf{X} \in \mathbb{R}^{p\times p}, \mathbf{X} \succcurlyeq 0\}$ and $\beta$ the penalty term.

3. After filling out the missing parts in the code and run it for each of the provided data sets, we obtain the following plots.

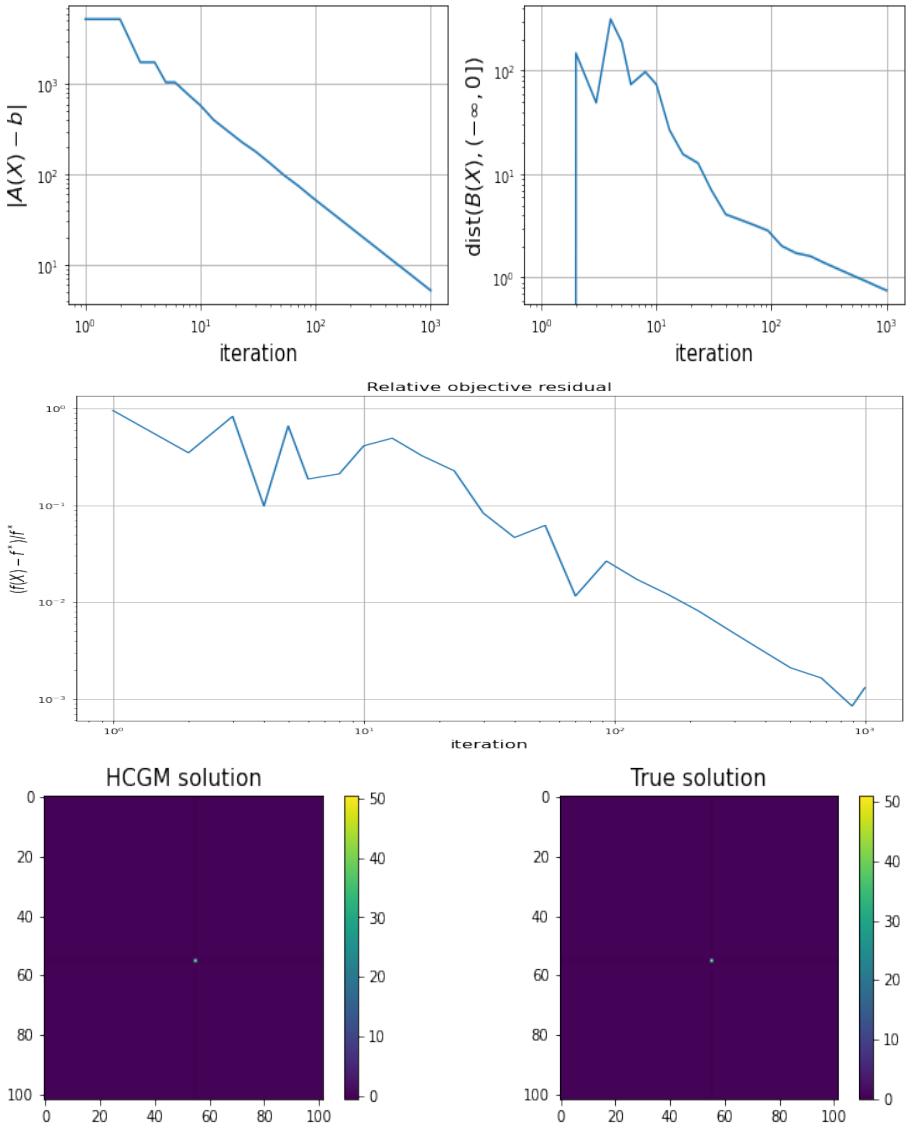



Figure 2: Results for data set *102n-insecta-ant-colony4-day10*.

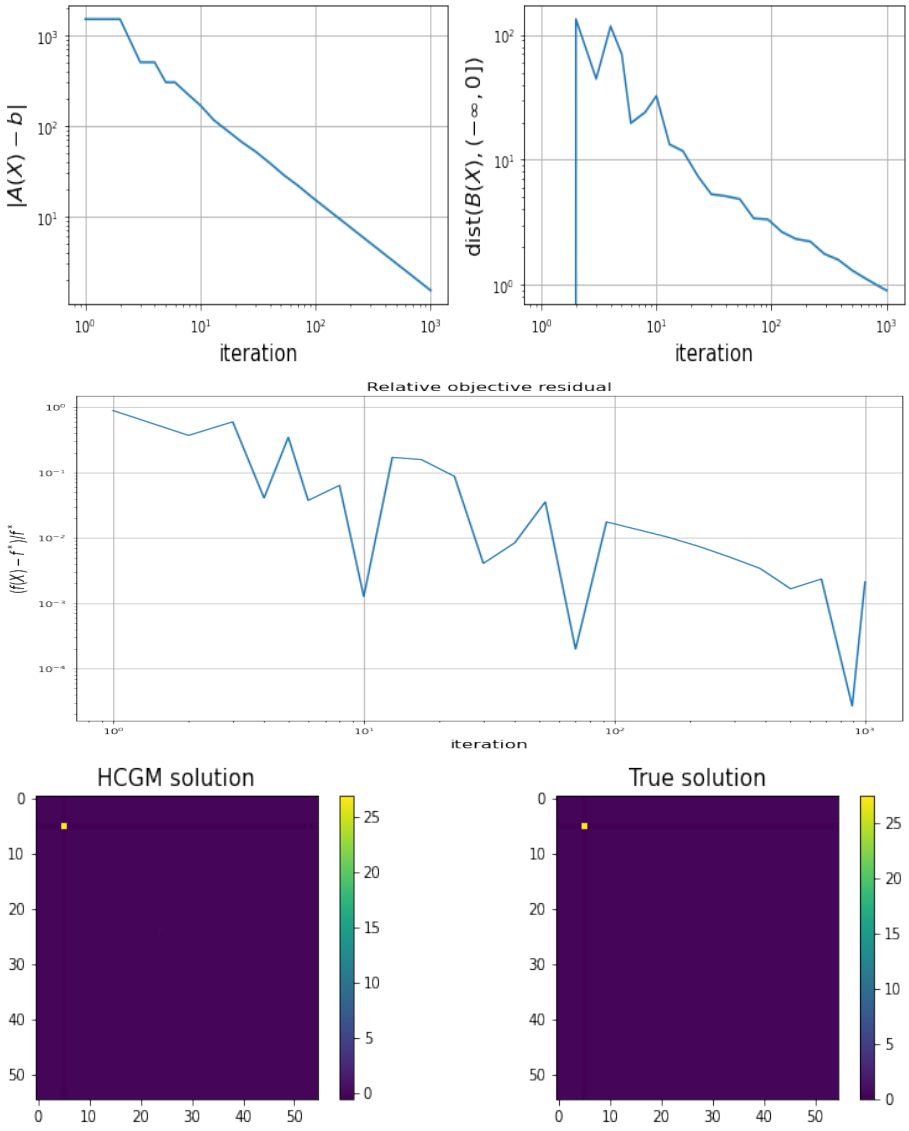



Figure 3: Results for data set *55n-insecta-ant-colony1-day37*.

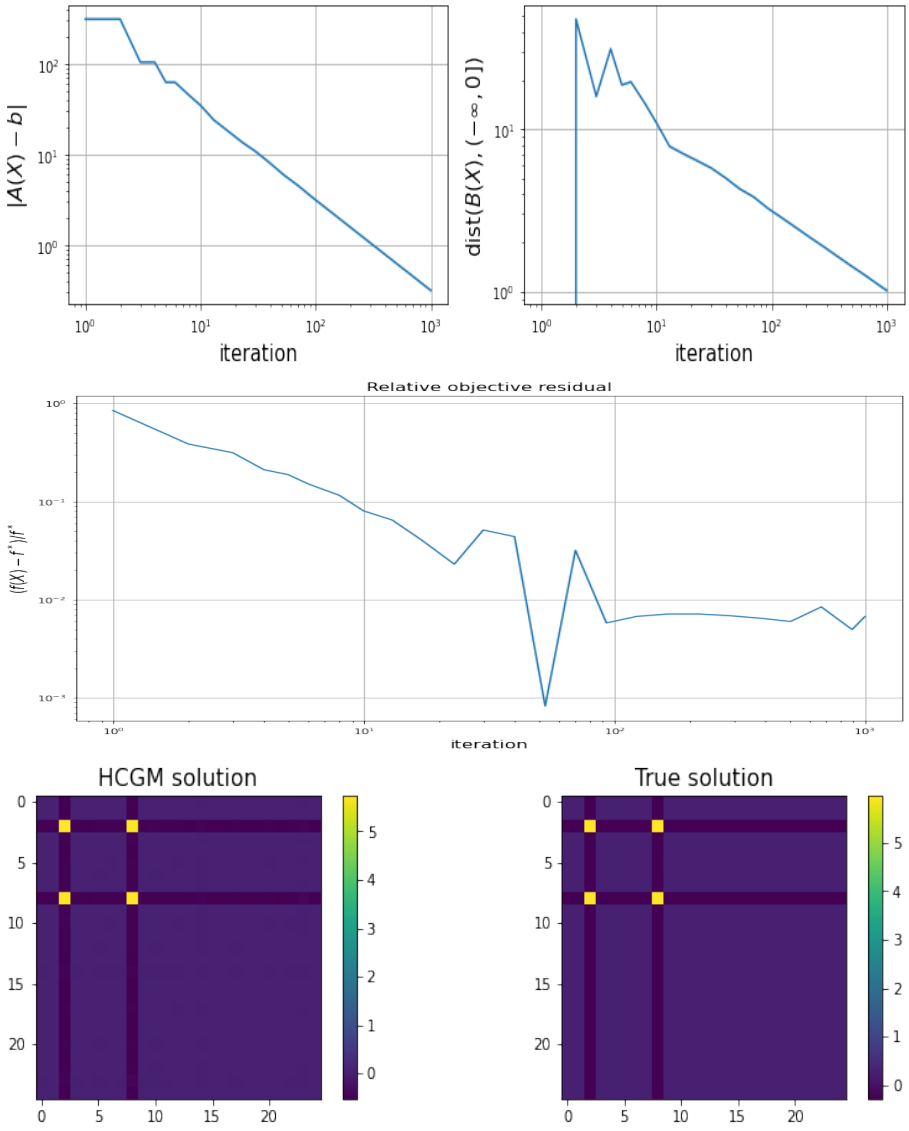



Figure 4: Results for data set *25mammalia-primate-association-13*.

And running times:

| Data set | Running time in sec | number constraints |
|---|---|---|
| 102n-insecta-ant-colony4-day10 | 4535.8765 | 10404 |
| 55n-insecta-ant-colony1-day37 | 792.8296 | 3025 |
| 25mammalia-primate-association-13 | 76.1102 | 625 |

Table 4: Running time for HCGM and estimated number of constraints over three data sets

# References

[1] Rajendra Bhatia. *Matrix analysis*. Vol. 169. Springer Science & Business Media, 2013.