EE-556 Mathematics of Data: From Theory to Computation
Homework 3

WEI Jiheng (319872)

January 7, 2022

1 Crime Scene Investigation with Blind Image Deconvolution

1.1 Computing projections onto $\mathcal{X}$

1. With the indicator function

$$\delta_{\mathcal{X}}(\boldsymbol{X}) = \begin{cases} 0, & \text{if } \boldsymbol{X} \in \mathcal{X} \\ +\infty, & \text{otherwise} \end{cases} \tag{1}$$

we have

$$\begin{aligned} \operatorname{prox}_{\delta_{\mathcal{X}}}(\boldsymbol{Z}) &= \arg\min_{\boldsymbol{X}\in\mathbb{R}^{p\times m}} \left\{ \delta_{\mathcal{X}}(\boldsymbol{X}) + \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Z}\|_{\mathrm{F}}^2 \right\} \\ &= \arg\min_{\boldsymbol{X}\in\mathcal{X}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Z}\|_{\mathrm{F}}^2 \\ &= \arg\min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{X}-\boldsymbol{Z}\|_{\mathrm{F}}^2 \\ &= \operatorname{proj}_{\mathcal{X}}(\boldsymbol{Z}). \end{aligned} \tag{2}$$

2. Let $\boldsymbol{x}^* = \operatorname{proj}_{\mathcal{X}}(\boldsymbol{x}), \boldsymbol{y}^* = \operatorname{proj}_{\mathcal{X}}(\boldsymbol{y})$, then

$$\langle \boldsymbol{x}-\boldsymbol{x}^*, \boldsymbol{y}^*-\boldsymbol{x}^* \rangle \leqslant 0, \tag{3}$$

$$\langle \boldsymbol{y}-\boldsymbol{y}^*, \boldsymbol{x}^*-\boldsymbol{y}^* \rangle \leqslant 0, \tag{4}$$

which implies

$$\langle \boldsymbol{x}-\boldsymbol{x}^*, \boldsymbol{y}^*-\boldsymbol{x}^* \rangle + \langle \boldsymbol{y}-\boldsymbol{y}^*, \boldsymbol{x}^*-\boldsymbol{y}^* \rangle \leqslant 0. \tag{5}$$

Arranging the terms of the above inequality,

$$\begin{aligned} \|\boldsymbol{x}^*-\boldsymbol{y}^*\|^2 &\leqslant \langle \boldsymbol{x}, \boldsymbol{x}^* \rangle - \langle \boldsymbol{x}, \boldsymbol{y}^* \rangle + \langle \boldsymbol{y}, \boldsymbol{y}^* \rangle - \langle \boldsymbol{y}, \boldsymbol{x}^* \rangle \\ &= \langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{x}^*-\boldsymbol{y}^* \rangle \\ &\leqslant \|\boldsymbol{x}-\boldsymbol{y}\| \|\boldsymbol{x}^*-\boldsymbol{y}^*\|. \end{aligned} \tag{6}$$

Hence,

$$\|\operatorname{proj}_{\mathcal{X}}(\boldsymbol{x}) - \operatorname{proj}_{\mathcal{X}}(\boldsymbol{y})\| \leqslant \|\boldsymbol{x}-\boldsymbol{y}\|. \tag{7}$$

3. With the non-negativity of sigular values,

$$\|\boldsymbol{X}\|_* = \sum_{i=1}^{s} \sigma_i(\boldsymbol{X}) = \|\boldsymbol{\sigma}_{\boldsymbol{X}}\|_1. \tag{8}$$

So, $\|\boldsymbol{X}\|_* \leqslant \kappa$ is equivalent to $\|\boldsymbol{\sigma}_{\boldsymbol{X}}\|_1 \leqslant \kappa$.

$$\|\boldsymbol{\Sigma}_{\boldsymbol{X}} - \boldsymbol{\Sigma}_{\boldsymbol{Z}}\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{s} \sigma_i^2(\boldsymbol{\Sigma}_{\boldsymbol{X}} - \boldsymbol{\Sigma}_{\boldsymbol{Z}})} = \sqrt{\sum_{i=1}^{s} [\sigma_i(\boldsymbol{\Sigma}_{\boldsymbol{X}}) - \sigma_i(\boldsymbol{\Sigma}_{\boldsymbol{Z}})]^2} = \|\boldsymbol{\sigma}_{\boldsymbol{X}} - \boldsymbol{\sigma}_{\boldsymbol{Z}}\|_2. \tag{9}$$

With Mirsky's inequality,

$$\min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{X}-\boldsymbol{Z}\|_{\mathrm{F}} \geqslant \min_{\boldsymbol{\Sigma}_{\boldsymbol{X}}:\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{\Sigma}_{\boldsymbol{X}}-\boldsymbol{\Sigma}_{\boldsymbol{Z}}\|_{\mathrm{F}} = \min_{\|\boldsymbol{\sigma}_{\boldsymbol{X}}\|_1\leqslant\kappa} \|\boldsymbol{\sigma}_{\boldsymbol{X}}-\boldsymbol{\sigma}_{\boldsymbol{Z}}\|_2. \tag{10}$$

The right hand side of the inequality is achieved by projecting $\boldsymbol{\sigma}_{\boldsymbol{Z}}$ onto the ℓ_1-norm ball, which implies that

$$\arg\min_{\|\boldsymbol{\sigma}_{\boldsymbol{X}}\|_1\leqslant\kappa} \|\boldsymbol{\sigma}_{\boldsymbol{X}}-\boldsymbol{\sigma}_{\boldsymbol{Z}}\|_2 = \boldsymbol{\sigma}_{\boldsymbol{Z}}^{\ell_1}, \tag{11}$$

$$\arg\min_{\boldsymbol{\Sigma}_{\boldsymbol{X}}:\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{\Sigma}_{\boldsymbol{X}}-\boldsymbol{\Sigma}_{\boldsymbol{Z}}\|_{\mathrm{F}} = \boldsymbol{\Sigma}_{\boldsymbol{Z}}^{\ell_1}. \tag{12}$$

Assume $\boldsymbol{X}=\boldsymbol{U}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{V}^{\mathsf{T}}$, with the rotational invariance of Frobenius norm,

$$\|\boldsymbol{X}-\boldsymbol{Z}\|_{\mathrm{F}} = \|\boldsymbol{\Sigma}_{\boldsymbol{X}}-\boldsymbol{\Sigma}_{\boldsymbol{Z}}\|_{\mathrm{F}}. \tag{13}$$

As a result,

$$\arg\min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{X}-\boldsymbol{Z}\|_{\mathrm{F}} = \boldsymbol{U}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{\ell_1}\boldsymbol{V}^{\mathsf{T}}. \tag{14}$$

Since $\arg\min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{X}-\boldsymbol{Z}\|_{\mathrm{F}} = \arg\min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{X}-\boldsymbol{Z}\|_{\mathrm{F}}^2$,

$$\mathrm{proj}_{\mathcal{X}}(\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{\ell_1}\boldsymbol{V}^{\mathsf{T}}. \tag{15}$$

1.2 Computing the linear minimisation oracle of $\mathcal{X}$

Denoting the largest singular value of $\boldsymbol{Z}$ by σ_1,

$$\langle -\kappa\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}, \boldsymbol{Z}\rangle = -\kappa\,\mathrm{Tr}(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}) = -\kappa\,\mathrm{Tr}(\sigma_1\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}}) = -\kappa\sigma_1, \tag{16}$$

where $\mathrm{Tr}(\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}})=1$ because the $\boldsymbol{v}$ is a column of unitary matrix $\boldsymbol{V}$. For all $\boldsymbol{X}\in\mathcal{X}$, Hölder's inequality gives that $|\langle\boldsymbol{X},\boldsymbol{Z}\rangle| \leqslant \|\boldsymbol{X}\|_*\|\boldsymbol{Z}\|_\infty$. With $\|\boldsymbol{X}\|_*\leqslant\kappa$,

$$-|\langle\boldsymbol{X},\boldsymbol{Z}\rangle| \geqslant -\|\boldsymbol{X}\|_*\|\boldsymbol{Z}\|_\infty = -\sigma_1\|\boldsymbol{X}\|_* \geqslant -\kappa\sigma_1, \tag{17}$$

$$\langle\boldsymbol{X},\boldsymbol{Z}\rangle \geqslant -|\langle\boldsymbol{X},\boldsymbol{Z}\rangle| \geqslant -\kappa\sigma_1 = \langle -\kappa\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}, \boldsymbol{Z}\rangle. \tag{18}$$

Hence,

$$-\kappa\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}} \in \mathrm{lmo}_{\mathcal{X}}(\boldsymbol{Z}). \tag{19}$$

1.3 Comparing the scalability

1.

Table 1: Computation time of the projection operator (in second)

#ratings	No. run					Average
	1	2	3	4	5	
100k	0.4519	0.4489	0.5239	0.4515	0.4858	0.4724
1M	27.21	30.67	27.13	29.37	30.07	28.89

2.

Table 2: Computation time of the lmo (in second)

#ratings	No. run					Average
	1	2	3	4	5	
100k	0.0164	0.0146	0.0136	0.0140	0.0139	0.0145
1M	0.1730	0.1757	0.1695	0.1671	0.1725	0.1716

Compared with the projection operator, lmo does not require full SVD, but only the largest singular value/vectors, so it spends less computation time. Also, implementation with methods from `scipy.sparse` package makes the latter algorithm more efficient.

1.4 Frank-Wolfe for blind image deconvolution

1. The objective function is

$$f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{A}(\boldsymbol{X}) - \boldsymbol{b}\|_2^2, \tag{20}$$

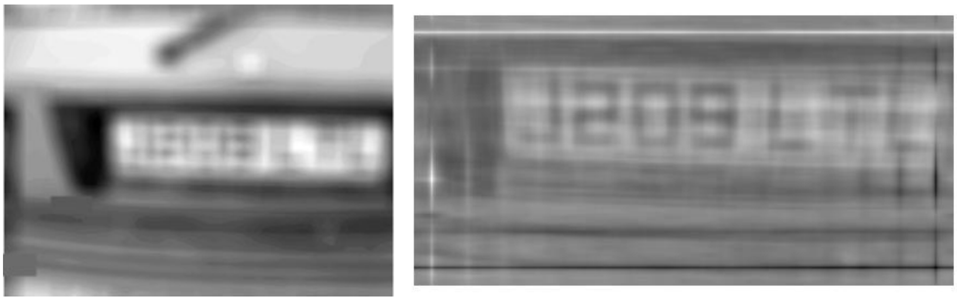
   in which the linear operator $\boldsymbol{A}$ can be treated as a matrix. So, transformation $\boldsymbol{A}(\boldsymbol{X})$ can be written as matrix multiplication $\boldsymbol{AX}$.

$$\nabla f(\boldsymbol{X}) = \boldsymbol{A}^\mathsf{T}(\boldsymbol{AX} - \boldsymbol{b}). \tag{21}$$

$$\begin{aligned} \|\nabla f(\boldsymbol{X}) - \nabla f(\boldsymbol{Y})\| &= \|\boldsymbol{A}^\mathsf{T}(\boldsymbol{AX} - \boldsymbol{b}) - \boldsymbol{A}^\mathsf{T}(\boldsymbol{AY} - \boldsymbol{b})\| \\ &= \|\boldsymbol{A}^\mathsf{T}\boldsymbol{A}(\boldsymbol{X} - \boldsymbol{Y})\| \\ &\leqslant \|\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\|\|\boldsymbol{X} - \boldsymbol{Y}\| \\ &= L\|\boldsymbol{X} - \boldsymbol{Y}\|. \end{aligned} \tag{22}$$

   The gradient is Lipschitz continuous, so the objective function is smooth.

2. With $\kappa = 50$ and $K_1 = K_2 = 17$, one can find that the license plate number is J209 LTL.



(a) Blurred (b) Deconvoluted

Figure 1: The license plate

2 k-means Clustering by Semidefinite Programming

2.1 Methods for clustering the fashion-MNIST data

1. With $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{X}$ and $0 \leqslant \lambda \leqslant 1$, define

$$\boldsymbol{Z} = \lambda \boldsymbol{X} + (1-\lambda)\boldsymbol{Y}. \tag{23}$$

   $\lambda, 1-\lambda$ are non-negative and $\boldsymbol{X}, \boldsymbol{Y}$ are positive semidefinite, it is trivial that $\boldsymbol{Z} \succeq 0$.

$$\operatorname{Tr}(\boldsymbol{Z}) = \operatorname{Tr}\{\lambda \boldsymbol{X} + (1-\lambda)\boldsymbol{Y}\} = \lambda \operatorname{Tr}(\boldsymbol{X}) + (1-\lambda)\operatorname{Tr}(\boldsymbol{Y}). \tag{24}$$

   Since $\operatorname{Tr}(\boldsymbol{X}) \leqslant \kappa$ and $\operatorname{Tr}(\boldsymbol{Y}) \leqslant \kappa$,

$$\operatorname{Tr}(\boldsymbol{Z}) \leqslant \max\{\operatorname{Tr}(\boldsymbol{X}), \operatorname{Tr}(\boldsymbol{Y})\} \leqslant \kappa. \tag{25}$$

   The matrix $\boldsymbol{Z}$, which is the affine combination of $\boldsymbol{X}$ and $\boldsymbol{Y}$, also belongs to $\mathcal{X}$. Hence, the set $\mathcal{X}$ is convex.

2. The optimisation problem is

$$\min_{x\in\mathcal{X}} f(x) + \delta_{\{b_1\}}(A_1(x)) + \delta_{\{b_2\}}(A_2(x)) \quad \text{s.t.} \quad B(x) \in \mathcal{K}. \tag{26}$$

   $B(x) \in \mathcal{K}$ is quivalent to $x \succeq 0$, we can deduce that

$$B(x) = x \Longrightarrow B = \boldsymbol{I}. \tag{27}$$

   Then, the quadratic penalty terms are

$$\mathrm{QP}_{\{b_1\}}(x) = \min_{y\in\{b_1\}} \|y - A_1(x)\|^2 = \|A_1(x) - b_1\|^2, \tag{28}$$

$$\mathrm{QP}_{\{b_2\}}(x) = \min_{y\in\{b_2\}} \|y - A_2(x)\|^2 = \|A_2(x) - b_2\|^2, \tag{29}$$

$$\mathrm{QP}_{\mathcal{K}}(x) = \operatorname{dist}^2(B(x), \mathcal{K}) = \operatorname{dist}^2(x, \mathcal{K}). \tag{30}$$

   So, the penalised objective function with parameter $(2\beta)^{-1}$ is

$$f(x) + \frac{1}{2\beta}\|A_1(x) - b_1\|^2 + \frac{1}{2\beta}\|A_2(x) - b_2\|^2 + \frac{1}{2\beta}\operatorname{dist}^2(x, \mathcal{K}). \tag{31}$$

   Danskin's theorem gives

$$\frac{\partial}{\partial x}\operatorname{dist}^2(x, \mathcal{K}) = \frac{\partial}{\partial x}\|k^* - x\|^2 = 2(x - k^*), \tag{32}$$

   where

$$k^* = \arg\min_{k\in\mathcal{K}} \|k - x\|^2 = \operatorname{proj}_{\mathcal{K}}(x). \tag{33}$$

   As before, the linear operators can be regarded as matrices. The gradient of the penalised objective is

$$\nabla f(x) + \frac{1}{\beta}\left[A_1^{\mathsf{T}}(A_1(x) - b_1) + A_2^{\mathsf{T}}(A_2(x) - b_2) + (x - \operatorname{proj}_{\mathcal{K}}(x))\right] = \frac{v}{\beta}. \tag{34}$$

3.

$$\nabla f(x_k) = \langle \boldsymbol{C}, x_k\rangle = \operatorname{Tr}(\boldsymbol{C}^{\mathsf{T}} x_k) = \boldsymbol{C}. \tag{35}$$

   The projection of an arbitrary matrix onto the positive orthant is the element-wise maximum function, i.e.

$$\operatorname{proj}_{\mathcal{K}}(x_k) = \max(0, x_k). \tag{36}$$

For a matrix $\boldsymbol{X}$ and its projection $\boldsymbol{K}$, this function ensures that

$$|\boldsymbol{X}_{ij} - \boldsymbol{K}_{ij}|_{\boldsymbol{X}_{ij} \geqslant 0} = 0 \tag{37}$$

and

$$\begin{aligned} |\boldsymbol{X}_{ij} - \boldsymbol{K}_{ij}|_{\boldsymbol{X}_{ij}<0} &= -\boldsymbol{X}_{ij} \\ &= \arg\min_{\boldsymbol{K}\in\mathcal{K}} -\boldsymbol{X}_{ij} + \boldsymbol{K}'_{ij} \\ &= \arg\min_{\boldsymbol{K}'\in\mathcal{K}} |\boldsymbol{X}_{ij} - \boldsymbol{K}'_{ij}|_{\boldsymbol{X}_{ij}<0}. \end{aligned} \tag{38}$$

Hence,

$$v_k = \beta_k \boldsymbol{C} + A_1^\top(A_1(x_k) - b_1) + A_2^\top(A_2(x_k) - b_2) + \max(0, x_k). \tag{39}$$

4. Denote $\sigma^{-1}y^k + A(\tilde{x}^{k+1})$ by z,

$$\operatorname{prox}_{\sigma^{-1}g}(z) = \arg\min_{w} \left\{ g(w) + \frac{\sigma}{2}\|w - z\|^2 \right\}. \tag{40}$$

g_1 and g_2 are indicator functions of singletons $\{b_1\}$ and $\{b_2\}$, respectively. Hence,

$$\operatorname{prox}_{\sigma^{-1}g_1}(z_1) = b_1, \quad \operatorname{prox}_{\sigma^{-1}g_2}(z_2) = b_2, \tag{41}$$

$$\operatorname{prox}_{\sigma^{-1}g_3}(z_3) = \arg\min_{w_3\in\mathcal{K}} \left\{ \delta_{\mathcal{K}}(w_3) + \frac{\sigma}{2}\|w_3 - z_3\|^2 \right\}. \tag{42}$$

As before, the point in $\mathcal{K}$ nearest to z_3 is the latter's projection onto the set, i.e.

$$\operatorname{prox}_{\sigma^{-1}g_3}(z_3) = \operatorname{proj}_{\mathcal{K}}(z_3). \tag{43}$$

With

$$A = \begin{bmatrix} A_1 \\ A_2 \\ B \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \\ \boldsymbol{I} \end{bmatrix}, \quad A(\tilde{x}^{k+1}) = \begin{bmatrix} A_1(\tilde{x}^{k+1}) \\ A_2(\tilde{x}^{k+1}) \\ \tilde{x}^{k+1} \end{bmatrix} \tag{44}$$

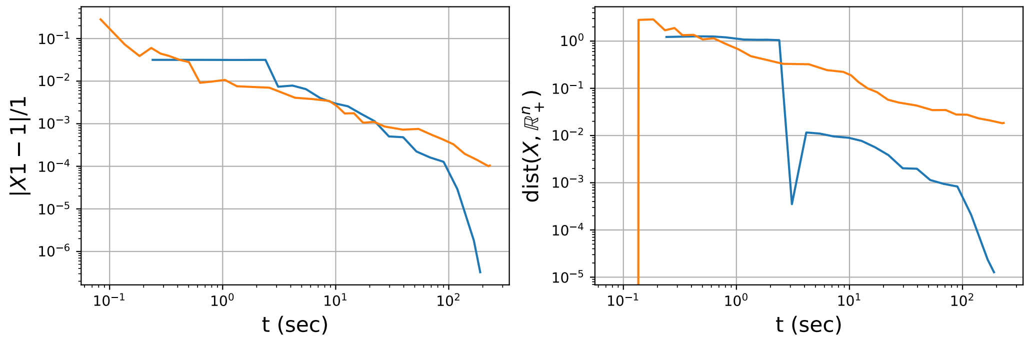
we have

$$\begin{aligned} y^{k+1} &= y^k + \sigma A(\tilde{x}^{k+1}) - \sigma \operatorname{prox}_{\sigma^{-1}g}(z) \\ &= \begin{bmatrix} y_1^k \\ y_2^k \\ y_3^k \end{bmatrix} + \sigma \begin{bmatrix} A_1(\tilde{x}^{k+1}) - b_1 \\ A_2(\tilde{x}^{k+1}) - b_2 \\ \tilde{x}^{k+1} - \operatorname{proj}_{\mathcal{K}}(\sigma^{-1}y_3^k + \tilde{x}^{k+1}) \end{bmatrix}. \end{aligned} \tag{45}$$

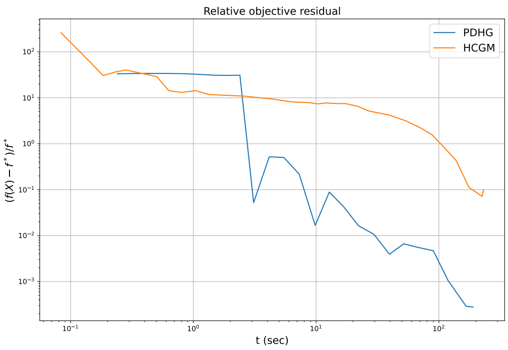
Then, it is obvious that

$$A^\top y^{k+1} = A^\top y^k + \sigma \left[A_1^\top(A_1(\tilde{x}^{k+1}) - b_1) + A_2^\top(A_2(\tilde{x}^{k+1}) - b_2) + \tilde{x}^{k+1} - \operatorname{proj}_{\mathcal{K}}(\sigma^{-1}y_3^k + \tilde{x}^{k+1}) \right]. \tag{46}$$

5.



(a)



(b)

Figure 2: Convergence results

Table 3: Final objective value

Optimal	HCGM	PDHG
57.053	51.384	57.037

Both final objective values are below the given optimal one. The reason is that the constraint relaxation of SDP may introduce infeasible points to the original optimisation problem.

Table 4: k-means value

Initial	HCGM	PDHG
150.9680	28.7269	28.7269

Table 5: k-means value of `kmeans` function

No. run	1	2	3	4	5
Value	182.5713	110.0894	189.9882	178.6771	168.4946

Whilst `HCGM` and `PDHG` converge and give the same k-means value, `kmeans` returns each time a different result. The latter's performance depends on random initialisation, and its solutions are generally worse.

3 Computing a geometric embedding for the Sparsest Cut Problem via Semidefinite Programming

1. For problem (7), there are $O(1)$ constraint for $A(\boldsymbol{X})$; $O(p^3)$ constraints for $B_{i,j,k}(\boldsymbol{X})$ due to the number of (i,j,k) triplets; and $O(p)$ constraints for $\boldsymbol{X} \in \mathcal{X}$ (i.e. $\sigma_i(\boldsymbol{X}) \geqslant 0$, $\forall i$). So, the problem has $O(p^3)$ constraints in total.

   For problem (3), there are $O(p)$ constraints for row and column sums, $O(p^2)$ constraints for $\boldsymbol{X} \geqslant 0$ and $O(p)$ constraints for $\boldsymbol{X} \in \mathcal{X}$. These gives $O(p^2)$ constraints in total, which is $O(p)$ times fewer than that of problem (7).

2. $$g(A(\boldsymbol{X})): \quad \mathrm{QP}_{\{p^2/2\}}(\boldsymbol{X}) = \left\| A(\boldsymbol{X}) - \frac{p^2}{2} \right\|^2. \tag{47}$$

   $$B_{i,j,k}(\boldsymbol{X}): \quad \mathrm{QP}_{\mathcal{K}}(\boldsymbol{X}) = \mathrm{dist}^2(B_{i,j,k}(\boldsymbol{X}), \mathcal{K}) = \|B_{i,j,k}(\boldsymbol{X}) - \mathrm{proj}_{\mathcal{K}}(B_{i,j,k}(\boldsymbol{X}))\|^2. \tag{48}$$

   Thus, the penalised objective function is

   $$f(x) + \alpha \left\| A(\boldsymbol{X}) - \frac{p^2}{2} \right\|^2 + \beta \sum_{i \neq j \neq k \neq i \in V} \|B_{i,j,k}(\boldsymbol{X}) - \mathrm{proj}_{\mathcal{K}}(B_{i,j,k}(\boldsymbol{X}))\|^2, \tag{49}$$

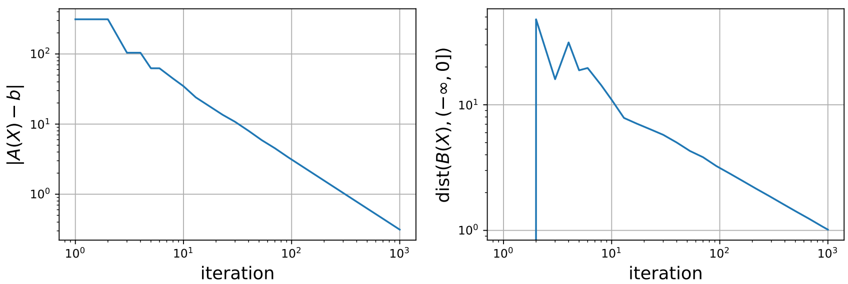
   where $\alpha \geqslant 0$ and $\beta \geqslant 0$ are penalty parameters.

3. There are $O(10^4)$, $O(10^5)$ and $O(10^6)$ constraints for G1, G2 and G3, respectively.

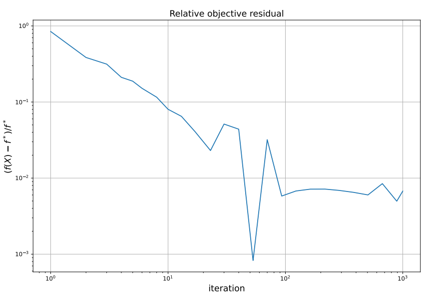
Table 6: Running time of HCGM

#nodes	25	55	105
Time (s)	76.5	782.4	5032.8

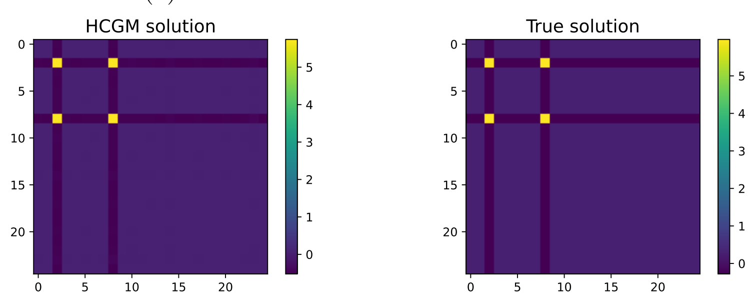
HCGM does not scale well for this problem, as one needs to wait for more than an hour on a graph of 102 nodes. Consequently, with a huge amount of constraints, the algorithm is infeasible on large graphs.



(a)

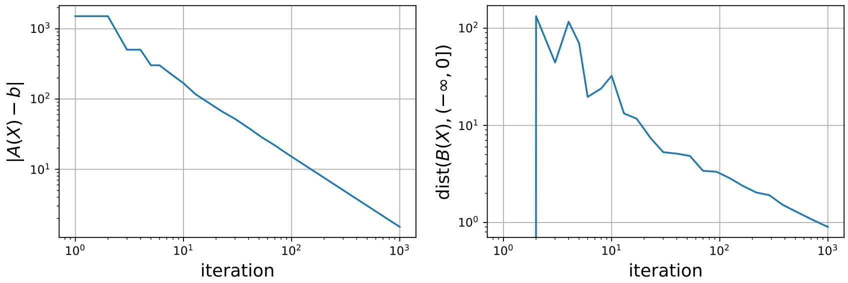


(b)

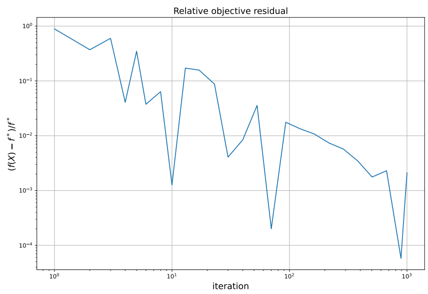


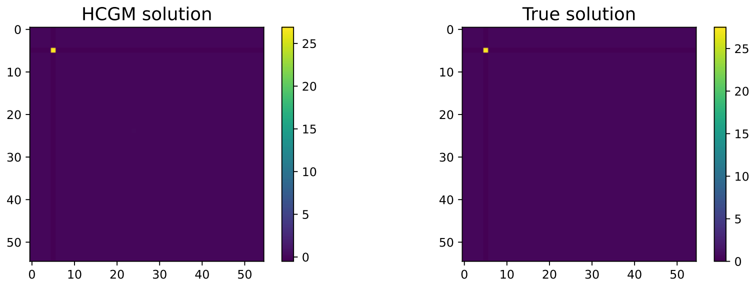
(c)

Figure 3: Results of G1



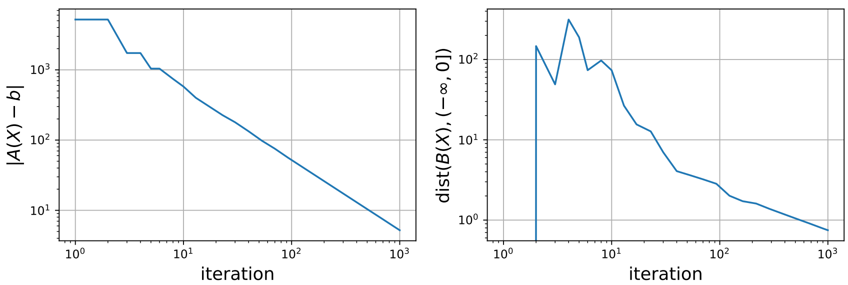
(a)



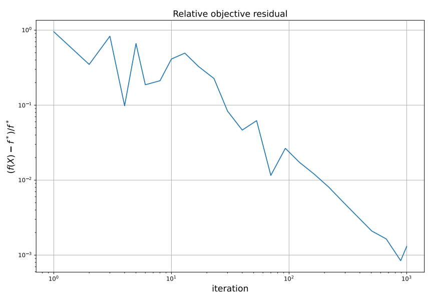


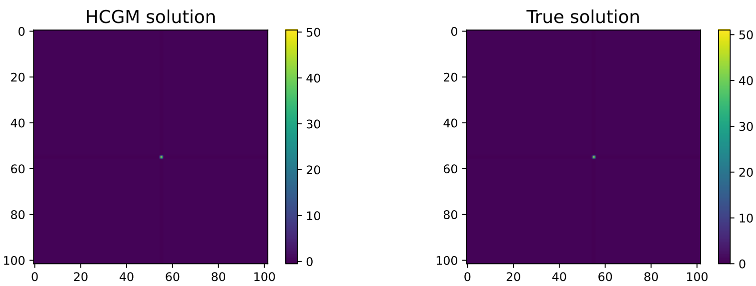
(b) (c)

Figure 4: Results of G2



(a)





(b) (c)

Figure 5: Results of G3