A: Proof of Proposition 1

The equivalence in terms of the global optimal solution sets between problem (1) and its penalized reformulation (11) follows directly from Lemma 1.

We prove the equivalence of local optimal solution sets. To begin with, we revisit an existing result stated in Lemma 3, where we use the notation $\operatorname{arglocmin}_{\boldsymbol{x}\in\mathcal{X}}\phi(\boldsymbol{x})$ to denote the set of locally optimal solutions to the problem $\min_{\boldsymbol{x}\in\mathcal{X}}\phi(\boldsymbol{x})$.

Lemma 3 (Proposition 9.1.2 in (Cui and Pang 2021)) *Consider the same setting in Lemma 1. Then:*

- *Given any scalar $\lambda > K$, it holds that*
$$\check{\boldsymbol{x}} \in \operatorname*{arglocmin}_{\boldsymbol{x}\in\mathcal{V}} \phi(\boldsymbol{x}) \Rightarrow \check{\boldsymbol{x}} \in \operatorname*{arglocmin}_{\boldsymbol{x}\in\mathcal{X}} \phi(\boldsymbol{x})+\lambda\psi(\boldsymbol{x}).$$
- *If $\check{\boldsymbol{x}}$ is a point in $\mathcal{V}$, then the following implication is true*
$$\check{\boldsymbol{x}} \in \operatorname*{arglocmin}_{\boldsymbol{x}\in\mathcal{X}} \phi(\boldsymbol{x})+\lambda\psi(\boldsymbol{x}) \Rightarrow \check{\boldsymbol{x}} \in \operatorname*{arglocmin}_{\boldsymbol{x}\in\mathcal{V}} \phi(\boldsymbol{x}).$$

This result needs the assumption $\check{\boldsymbol{x}} \in \mathcal{V}$ to guarantee the equivalence of locally optimal solutions. We show that our proposed penalization formulation eliminates this assumption, leading to equivalence between locally optimal solutions. We begin the proof. Let $\boldsymbol{x} \in [0,1]^n$. By Lemma 2, the distance from $\boldsymbol{x}$ to the set $\mathcal{U}_\kappa^n$ satisfies:

$$\operatorname{dist}(\boldsymbol{x},\mathcal{U}_k^n) \le k + \underbrace{\mathbf{1}^\top\boldsymbol{x} - 2S_k(\boldsymbol{x})}_{=h(\boldsymbol{x})}. \tag{20}$$

Now, suppose $\check{\boldsymbol{x}} \in \operatorname{arglocmin}_{\boldsymbol{x}\in[0,1]^n} F_\lambda(\boldsymbol{x}) := f(\boldsymbol{x}) + \lambda h(\boldsymbol{x})$. Without loss of generality, assume $\check{x}_1 \ge \cdots \ge \check{x}_n$. By the definition of local optimality, there exists a scalar $\epsilon > 0$ such that:

$$F_\lambda(\check{\boldsymbol{x}}) \le F_\lambda(\boldsymbol{x}), \forall \boldsymbol{x} \in [0,1]^n, \|\boldsymbol{x}-\check{\boldsymbol{x}}\|_2 \le \epsilon. \tag{21}$$

Let $\boldsymbol{x}$ be defined as follows:

$$x_i = \begin{cases} \check{x}_i, & \text{if } i = 1,\dots,k,\ k+2,\dots,n \\ \max\{0, \check{x}_{k+1}-\epsilon\}, & \text{if } i = k+1. \end{cases}$$

Note that if $\check{\boldsymbol{x}} \notin \mathcal{U}_k^n$, then we must have $\check{x}_{k+1} > 0$. It follows that $\boldsymbol{x} \ne \check{\boldsymbol{x}}$, $\boldsymbol{x} \in [0,1]^n$, and $\|\boldsymbol{x}-\check{\boldsymbol{x}}\|_2 \le \epsilon$. Furthermore, we have:

$$\begin{aligned} F_\lambda(\check{\boldsymbol{x}}) \ge&\quad f(\boldsymbol{x}) - K\|\boldsymbol{x}-\check{\boldsymbol{x}}\|_2 + \lambda h(\check{\boldsymbol{x}}) \\ =&\quad F_\lambda(\boldsymbol{x}) - K\|\boldsymbol{x}-\check{\boldsymbol{x}}\|_2 + \lambda\big(h(\check{\boldsymbol{x}}) - h(\boldsymbol{x})\big) \\ >&\quad F_\lambda(\boldsymbol{x}), \end{aligned} \tag{22}$$

where for the last inequality, we have $\lambda > K$ and

$$\begin{aligned} h(\check{\boldsymbol{x}}) - h(\boldsymbol{x}) =&\quad \check{x}_{k+1} - x_{k+1} \\ =&\quad |\check{x}_{k+1} - x_{k+1}| \\ =&\quad \|\check{\boldsymbol{x}} - \boldsymbol{x}\|_2. \end{aligned} \tag{23}$$

The result in equation (22) contradicts (21).

Thus, for $\lambda > K$, $\check{\boldsymbol{x}}$ must satisfy $\check{\boldsymbol{x}} \in \mathcal{U}_k^n$. Consequently, we can conclude that:

$$\operatorname*{arglocmin}_{\boldsymbol{x}\in\mathcal{U}_k^n} f(\boldsymbol{x}) = \operatorname*{arglocmin}_{\boldsymbol{x}\in[0,1]^n} f(\boldsymbol{x}) + \lambda h(\boldsymbol{x}). \tag{24}$$

This establishes the equivalence of the local optimality of problem (1) and its penalized reformulation (11). ∎

B: Proof of Proposition 2

The non-convex optimization problem can be rewritten as:

$$\min_{\boldsymbol{x}\in[0,1]^n} u(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 + \mu\left(\sum_{i=k+1}^n x_{[i]} - \sum_{i=1}^k x_{[i]}\right), \tag{25}$$

where $k \le n$ is an integer, and $x_{[i]}$ denotes the i-th largest element of $\boldsymbol{x}$. Without loss of generality, assume $z_1 \ge z_2 \ge \cdots \ge z_n$. The objective can be rewritten as:

$$\begin{aligned} u(\boldsymbol{x}) = &\textstyle\sum_{i=1}^k \left(\frac{1}{2}(x_{[i]} - z_{j_i(\boldsymbol{x})})^2 - \mu x_{[i]}\right) \\ &+ \textstyle\sum_{i=k+1}^n \left(\frac{1}{2}(x_{[i]} - z_{j_i(\boldsymbol{x})})^2 + \mu x_{[i]}\right), \end{aligned} \tag{26}$$

where $j_i(\boldsymbol{x})$ is the index of $x_{[i]}$ in $\boldsymbol{x}$, which depends on $\boldsymbol{x}$. It can be verified that when

$$x_{[i]} = \begin{cases} \left[z_{j_i(\boldsymbol{x})} + \mu\right]_0^1, & \text{if } i \le k, \\ \left[z_{j_i(\boldsymbol{x})} - \mu\right]_0^1, & \text{if } i > k, \end{cases} \tag{27}$$

we have

$$u(\boldsymbol{x}) \ge \tilde{u}(\boldsymbol{x}), \tag{28}$$

where

$$\begin{aligned} &\tilde{u}(\boldsymbol{x}) \\ &= \textstyle\sum_{i=1}^k \left(\frac{1}{2}\left([z_{j_i(\boldsymbol{x})}+\mu]_0^1 - z_{j_i(\boldsymbol{x})}\right)^2 - \mu[z_{j_i(\boldsymbol{x})}+\mu]_0^1\right) \\ &+ \textstyle\sum_{i=k+1}^n \left(\frac{1}{2}\left([z_{j_i(\boldsymbol{x})}-\mu]_0^1 - z_{j_i(\boldsymbol{x})}\right)^2 + \mu[z_{j_i(\boldsymbol{x})}-\mu]_0^1\right). \end{aligned} \tag{29}$$

Note that $\tilde{u}$ is still a function of $\boldsymbol{x}$. To proceed, we need the following lemma.

Lemma 4 *For a given $\mu > 0$, define the function*

$$\begin{aligned} g(z) = &\left(\tfrac{1}{2}(z - [z+\mu]_0^1)^2 - \mu[z+\mu]_0^1\right) \\ &- \left(\tfrac{1}{2}(z - [z-\mu]_0^1)^2 + \mu[z-\mu]_0^1\right). \end{aligned} \tag{30}$$

Then $g(z)$ is monotonically non-increasing.

Proof of Lemma 4: We can evaluate the following functions segment-wise:

$$[z+\mu]_0^1 = \begin{cases} 1, & \text{if } z \ge 1-\mu, \\ z+\mu, & \text{if } -\mu \le z < 1-\mu, \\ 0, & \text{if } z < -\mu, \end{cases} \tag{31}$$

$$[z-\mu]_0^1 = \begin{cases} 1, & \text{if } z \ge 1+\mu, \\ z-\mu, & \text{if } \mu \le z < 1+\mu, \\ 0, & \text{if } z < \mu, \end{cases} \tag{32}$$

$$(z - [z+\mu]_0^1)^2 = \begin{cases} (z-1)^2, & \text{if } z \ge 1-\mu, \\ \mu^2, & \text{if } -\mu \le z < 1-\mu, \\ z^2, & \text{if } z < -\mu, \end{cases} \tag{33}$$

$$(z - [z-\mu]_0^1)^2 = \begin{cases} (z-1)^2, & \text{if } z \ge 1+\mu, \\ \mu^2, & \text{if } \mu \le z < 1+\mu, \\ z^2, & \text{if } z < \mu. \end{cases} \tag{34}$$

Then we are ready to write out the expression for $g(z)$. There are two cases, $\mu \ge \frac{1}{2}$ and $\mu < \frac{1}{2}$. We first consider the case

$\mu \geq \frac{1}{2}$, which means $1-\mu \leq \mu$. The expression of $g(z)$ is given by

$$g(z) = \begin{cases} -2\mu, & \text{if } z \geq 1+\mu, \\ \frac{1}{2}(z-\mu-1)^2 - 2\mu, & \text{if } \mu \leq z < 1+\mu, \\ \frac{1}{2} - z - \mu, & \text{if } 1-\mu \leq z < \mu, \\ -\frac{1}{2}(z+\mu)^2, & \text{if } -\mu \leq z < 1-\mu, \\ 0, & \text{if } z < -\mu. \end{cases} \tag{35}$$

It can be seen that when $\mu \geq 1/2$, $g(z)$ is continuous and segment-wise non-increasing. Hence, $g(z)$ is monotonically non-increasing in this case. We then consider the case $\mu < 1/2$ which means $1-\mu > \mu$. The expression of $g(z)$ is given by

$$g(z) = \begin{cases} -2\mu, & \text{if } z \geq 1+\mu, \\ \frac{1}{2}(z-\mu-1)^2 - 2\mu, & \text{if } 1-\mu \leq z < 1+\mu, \\ -2\mu z, & \text{if } \mu \leq z < 1-\mu, \\ -\frac{1}{2}(\mu+z)^2, & \text{if } -\mu \leq z < \mu, \\ 0, & \text{if } z < -\mu. \end{cases} \tag{36}$$

It can be seen that $g(z)$ is also monotonically non-increasing. ■

Lemma 4 implies that if there exist indices $p \in \{1, ..., k\}$ and $q \in \{k+1, ..., n\}$ such that $z_{j_p} < z_{j_q}$, then we can construct $\hat{\boldsymbol{x}}$ as

$$\hat{x}_i = \begin{cases} x_q, & \text{if } i = p, \\ x_p, & \text{if } i = q, \\ x_i, & \text{if } i \neq p \text{ and } i \neq q. \end{cases} \tag{37}$$

This leads to a non-increasing objective value:

$$\begin{aligned} & \tilde{u}(\boldsymbol{x}) - \tilde{u}(\hat{\boldsymbol{x}}) \\ = & \left(\frac{1}{2}\left(\left[z_{j_p}+\mu\right]_0^1 - z_{j_p}\right)^2 - \mu\left[z_{j_p}+\mu\right]_0^1\right) \\ & + \left(\tfrac{1}{2}\left(\left[z_{j_q}-\mu\right]_0^1 - z_{j_q}\right)^2 + \mu\left[z_{j_q}-\mu\right]_0^1\right) \\ & - \left(\tfrac{1}{2}\left(\left[z_{j_q}+\mu\right]_0^1 - z_{j_q}\right)^2 - \mu\left[z_{j_q}+\mu\right]_0^1\right) \\ & - \left(\tfrac{1}{2}\left(\left[z_{j_p}-\mu\right]_0^1 - z_{j_p}\right)^2 + \mu\left[z_{j_p}-\mu\right]_0^1\right) \\ = & \; g(z_{j_p}) - g(z_{j_q}) \geq 0. \end{aligned} \tag{38}$$

Hence, based on Lemma 4, $\tilde{u}(\boldsymbol{x})$ can be further lower bounded as

$$\begin{aligned} \tilde{u}(\boldsymbol{x}) \geq & \sum_{i=1}^{k}\left(\frac{1}{2}\left([z_i+\mu]_0^1 - z_i\right)^2 - \mu\,[z_i+\mu]_0^1\right) \\ & + \sum_{i=k+1}^{n}\left(\frac{1}{2}\left([z_i-\mu]_0^1 - z_i\right)^2 + \mu\,[z_i-\mu]_0^1\right), \end{aligned} \tag{39}$$

which does not depend on $\boldsymbol{x}$. It can be verified that when

$$x_i = \begin{cases} [z_i+\mu]_0^1, & i \leq k, \\ [z_i-\mu]_0^1, & i > k, \end{cases} \tag{40}$$

the lower bound is obtained. Hence, we can conclude that (40) is an optimal solution to (25).

C: Proof of Proposition 3

Recall that the PGM algorithm iterates as

$$\boldsymbol{z}^\ell = \boldsymbol{x}^\ell + \gamma_\ell(\boldsymbol{x}^\ell - \boldsymbol{x}^{\ell-1}), \tag{41}$$

$$\boldsymbol{x}^{\ell+1} = \mathrm{prox}_{[0,1]^n, \eta_\ell \lambda h}(\boldsymbol{z}^\ell - \eta_\ell \nabla f(\boldsymbol{z}^\ell)), \tag{42}$$

where $\{\gamma_\ell\}_{\ell \geq 1}$ is an extrapolation sequence. We assume that the step sizes satisfy $c_1 L_f \leq 1/\eta_\ell \leq c_2 L_f$ where $1 < c_1 \leq c_2 < \infty$, and $0 \leq \gamma_\ell < \bar{\gamma} < 1$ where $\bar{\gamma} = (c_1-1)/(2+2c_2)$.

To begin with the proof, we define

$$g(\boldsymbol{x}|\tilde{\boldsymbol{x}}, \beta) = f(\tilde{\boldsymbol{x}}) + \langle \nabla f(\tilde{\boldsymbol{x}}), \boldsymbol{x} - \tilde{\boldsymbol{x}} \rangle + \frac{\beta}{2}\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2^2.$$

Since the objective function has L_f-Lipschitz continuous gradient, we have:

$$f(\boldsymbol{x}^{\ell+1}) \leq f(\boldsymbol{x}^\ell) + \nabla f(\boldsymbol{x}^\ell)^\top(\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell) + \frac{L_f}{2}\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2. \tag{43}$$

Then, we can rewrite (42) as

$$\boldsymbol{x}^{\ell+1} \in \arg\min_{\boldsymbol{x} \in [0,1]^n} g(\boldsymbol{x}|\boldsymbol{z}^\ell, 1/\eta_\ell) + \lambda h(\boldsymbol{x}). \tag{44}$$

Since $\boldsymbol{x}^{\ell+1}$ is a critical point of the problem in (44), $\boldsymbol{x}^{\ell+1}$ satisfies

$$\begin{aligned} \mathbf{0} & \in \nabla g(\boldsymbol{x}^{\ell+1}|\boldsymbol{z}^\ell, 1/\eta_\ell) + \lambda \partial h(\boldsymbol{x}^{\ell+1}), \\ & = \nabla f(\boldsymbol{z}^\ell) + (\boldsymbol{x}^{\ell+1} - \boldsymbol{z}^\ell)/\eta_\ell + \lambda \partial h(\boldsymbol{x}^{\ell+1}). \end{aligned} \tag{45}$$

It follows that

$$\begin{aligned} & \mathrm{dist}(\mathbf{0}, \partial F_\lambda(\boldsymbol{x}^{\ell+1}))^2 \\ = & \mathrm{dist}(\mathbf{0}, \nabla f(\boldsymbol{x}^{\ell+1}) + \lambda \partial h(\boldsymbol{x}^{\ell+1}))^2 \\ = & \inf_{\boldsymbol{v} \in \partial h(\boldsymbol{x}^{\ell+1})} \|\nabla f(\boldsymbol{x}^{\ell+1}) + \boldsymbol{v}\|_2^2 \\ \leq & \left\|\nabla f(\boldsymbol{x}^{\ell+1}) - \left(\nabla f(\boldsymbol{z}^\ell) + (\boldsymbol{x}^{\ell+1} - \boldsymbol{z}^\ell)/\eta_\ell\right)\right\|_2^2 \\ \leq & 2(L_f^2 + \eta_\ell^{-2})\|\boldsymbol{x}^{\ell+1} - \boldsymbol{z}^\ell\|_2^2, \end{aligned} \tag{46}$$

where we recall that the distance is defined as

$$\mathrm{dist}(\boldsymbol{x}, \mathcal{A}) := \inf_{\boldsymbol{y} \in \mathcal{A}} \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Hence, we have the result

$$\begin{aligned} & \mathrm{dist}(\mathbf{0}, \partial F_\lambda(\boldsymbol{x}^{\ell+1})) \\ \leq & \sqrt{2(L_f^2 + \eta_\ell^{-2})}\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell - \gamma_\ell(\boldsymbol{x}^\ell - \boldsymbol{x}^{\ell-1})\|_2, \\ \leq & C_1\left(\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2 + \|\boldsymbol{x}^\ell - \boldsymbol{x}^{\ell-1}\|_2\right), \end{aligned} \tag{47}$$

where $C_1 = \sqrt{2(1+c_2^2)}L_f$.

On the other hand, we see from (44) that

$$\begin{aligned} & g(\boldsymbol{x}^{\ell+1} \mid \boldsymbol{z}^\ell, 1/\eta_\ell) + \lambda h(\boldsymbol{x}^{\ell+1}) \\ \leq & g(\boldsymbol{x}^\ell \mid \boldsymbol{z}^\ell, 1/\eta_\ell) + \lambda h(\boldsymbol{x}^\ell). \end{aligned} \tag{48}$$

Adding (43) and (48) we get

$$\begin{aligned}
&F_\lambda(\boldsymbol{x}^\ell) - F_\lambda(\boldsymbol{x}^{\ell+1})\\
\geq& \left(\nabla f(\boldsymbol{z}^\ell) - \nabla f(\boldsymbol{x}^\ell)\right)^\top (\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell) + \frac{1}{2\eta_\ell}\|\boldsymbol{x}^{\ell+1} - \boldsymbol{z}^\ell\|_2^2\\
&- \frac{L_f}{2}\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2 - \frac{1}{2\eta_\ell}\|\boldsymbol{x}^\ell - \boldsymbol{z}^\ell\|_2^2\\
=& \left(\nabla f(\boldsymbol{z}^\ell) - \nabla f(\boldsymbol{x}^\ell)\right)^\top (\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell)\\
&+ \frac{1-\eta_\ell L_f}{2\eta_\ell}\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2 + \frac{1}{\eta_\ell}(\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell)^\top(\boldsymbol{x}^\ell - \boldsymbol{z}^\ell)\\
\geq& -\left(L_f + \frac{1}{\eta_\ell}\right)\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2\|\boldsymbol{x}^\ell - \boldsymbol{z}^\ell\|_2\\
&+ \frac{1-\eta_\ell L_f}{2\eta_\ell}\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2\\
=& -\gamma_l\left(L_f + \frac{1}{\eta_\ell}\right)\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2\|\boldsymbol{x}^\ell - \boldsymbol{x}^{\ell-1}\|_2\\
&+ \frac{1-\eta_\ell L_f}{2\eta_\ell}\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2\\
\geq& d_\ell\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2 - e_\ell\|\boldsymbol{x}^\ell - \boldsymbol{x}^{\ell-1}\|_2^2,
\end{aligned} \tag{49}$$

where

$$d_\ell = \frac{1/\eta_\ell - L_f}{4} > 0,\ e_\ell = \frac{\gamma_\ell^2(L_f + 1/\eta_\ell)^2}{1/\eta_\ell - L_f} > 0;$$

we used the Cauchy–Schwarz inequality and the Lipschitz continuity of $\nabla f(\boldsymbol{x})$ in the second inequality; in the last inequality, we used Young's inequality $ab \leq \frac{a^2}{2\varepsilon} + \frac{b^2\varepsilon}{2}, \forall a, b \in \mathbb{R},\ \forall \varepsilon \in \mathbb{R}_+$ with $a = (L_f + 1/\eta_\ell)\gamma_\ell\|\boldsymbol{x}^\ell - \boldsymbol{x}^{\ell-1}\|_2$, $b = \|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2$, and $\varepsilon = (1/\eta_\ell - L_f)/2$. Note that the formula (49) is an adaptation of Lemma 1 in (Xu and Yin 2017). As a result, we have

$$\begin{aligned}
&F_\lambda(\boldsymbol{x}^0) - F_\lambda^\star\\
\geq& F_\lambda(\boldsymbol{x}^0) - F_\lambda(\boldsymbol{x}^{J+1})\\
=& \sum_{\ell=0}^{J} F_\lambda(\boldsymbol{x}^\ell) - F_\lambda(\boldsymbol{x}^{\ell+1})\\
\geq& \sum_{\ell=0}^{J} d_l\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2 - e_l\|\boldsymbol{x}^\ell - \boldsymbol{x}^{\ell-1}\|_2^2\\
=& d_\ell\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2 + \sum_{\ell=0}^{J-1}(d_l - e_{\ell+1})\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2\\
\geq& \sum_{\ell=0}^{J}(d_l - e_{\ell+1})\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2\\
\geq& \frac{C_2(J+1)}{2}\min_{\ell=0,\ldots,J}\|\boldsymbol{x}^{\ell+1} - \boldsymbol{x}^\ell\|_2^2 + \|\boldsymbol{x}^\ell - \boldsymbol{x}^{\ell-1}\|_2^2.
\end{aligned} \tag{50}$$

where

$$\begin{aligned}
C_2 &= \min_{l=0,\ldots,\ell} d_l - e_{l+1}\\
&= \left(\frac{c_1 - 1}{4} - \frac{\bar{\gamma}^2(c_2+1)^2}{c_1 - 1}\right)L_f > 0.
\end{aligned}$$

By using $a + b \leq \sqrt{2a^2 + b^2}$, we get

$$\begin{aligned}
&\min_{l=0,\ldots,J}\|\boldsymbol{x}^{l+1} - \boldsymbol{x}^l\|_2 + \|\boldsymbol{x}^l - \boldsymbol{x}^{l-1}\|_2\\
\leq& \sqrt{\frac{4(F_\lambda(\boldsymbol{x}^0) - F_\lambda^\star)}{C_2(J+1)}}.
\end{aligned} \tag{51}$$

Combining (51) and (47), we finally obtain the sublinear convergence rate:

$$\min_{l=0,\ldots,J}\operatorname{dist}(\mathbf{0}, \partial F_\lambda(\boldsymbol{x}^{\ell+1})) \leq C_1\sqrt{\frac{4(F_\lambda(\boldsymbol{x}^0) - F_\lambda^\star)}{C_2(J+1)}}. \tag{52}$$

D: Lipschitz Constants

Consider the function $f(\boldsymbol{x}) = -\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$. The following holds:

i) $\nabla f(\boldsymbol{x})$ is $2\|\boldsymbol{A}\|_2$-Lipschitz continuous on $\mathbb{R}^n$;

ii) $f(\boldsymbol{x})$ is $2\sqrt{k}\|\boldsymbol{A}\|_2$-Lipschitz continuous on $\operatorname{conv}(\mathcal{U}_k^n)$;

iii) $f(\boldsymbol{x})$ is $2\sqrt{n}\|\boldsymbol{A}\|_2$-Lipschitz continuous on $[0,1]^n$.

Proof of i): Note that $\nabla f(\boldsymbol{x}) = -2\boldsymbol{A}\boldsymbol{x}$. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have

$$\begin{aligned}
\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 &= \|-2\boldsymbol{A}\boldsymbol{x} + 2\boldsymbol{A}\boldsymbol{y}\|_2\\
&= \|2\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y})\|_2\\
&\leq 2\|\boldsymbol{A}\|_2\|\boldsymbol{x} - \boldsymbol{y}\|_2.
\end{aligned} \tag{53}$$

Therefore, the Lipschitz constant is $2\|\boldsymbol{A}\|_2$. ■

Proof of ii): For any $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{conv}(\mathcal{U}_k^n)$, we have:

$$\begin{aligned}
|f(\boldsymbol{x}) - f(\boldsymbol{y})| &= |\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}^\top\boldsymbol{A}\boldsymbol{y}|\\
&= |(\boldsymbol{x}+\boldsymbol{y})^\top\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y})|\\
&\leq \|\boldsymbol{x}+\boldsymbol{y}\|_2\|\boldsymbol{A}\|_2\|\boldsymbol{x} - \boldsymbol{y}\|_2\\
&\leq 2\sqrt{k}\|\boldsymbol{A}\|_2\|\boldsymbol{x} - \boldsymbol{y}\|_2,
\end{aligned} \tag{54}$$

where we used the fact that $\|\boldsymbol{x}+\boldsymbol{y}\|_2 \leq 2\sqrt{k}$. The Lipschitz constant is $2\sqrt{k}\|\boldsymbol{A}\|_2$. ■

Proof of iii): For any $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^n$, we have:

$$\begin{aligned}
|f(\boldsymbol{x}) - f(\boldsymbol{y})| &\leq \|\boldsymbol{x}+\boldsymbol{y}\|_2\|\boldsymbol{A}\|_2\|\boldsymbol{x} - \boldsymbol{y}\|_2\\
&\leq 2\sqrt{n}\|\boldsymbol{A}\|_2\|\boldsymbol{x} - \boldsymbol{y}\|_2,
\end{aligned} \tag{55}$$

where we used the fact that $\|\boldsymbol{x}+\boldsymbol{y}\|_2 \leq 2\sqrt{n}$. The Lipschitz constant is $2\sqrt{n}\|\boldsymbol{A}\|_2$. ■

E: Formulation of the Dk_1k_2BS Problem (17)

Consider the Dk_1k_2BS Problem (17). Let $\boldsymbol{a} = [\boldsymbol{x}^\top\ \boldsymbol{y}^\top]^\top \in \mathbb{R}^n$ where $n = n_1 + n_2$, and define the symmetric adjacency matrix:

$$\boldsymbol{A} = \begin{bmatrix} \mathbf{0} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{n\times n}. \tag{56}$$

It follows that $\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{y} = \frac{1}{2}\boldsymbol{a}^\top\boldsymbol{A}\boldsymbol{a}$. The feasible set can be rewritten as:

$$\mathcal{U}_{k_1,k_2}^{n_1,n_2} := \left\{\boldsymbol{a} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \in \{0,1\}^n \mid \boldsymbol{x} \in \mathcal{U}_{k_1}^{n_1},\ \boldsymbol{y} \in \mathcal{U}_{k_2}^{n_2}\right\}. \tag{57}$$

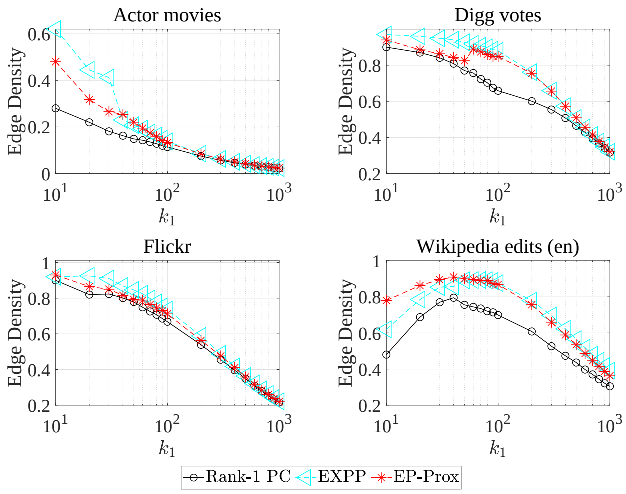

Figure 5: Edge density under different $k_1 = k_2$ for Dk_1k_2BS.

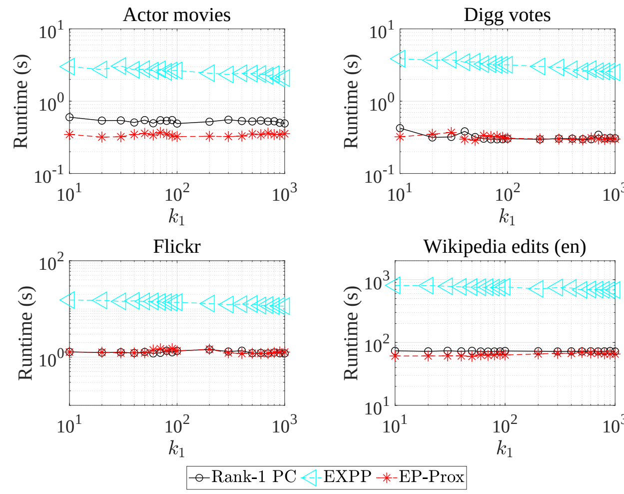

Figure 6: Runtime under different $k_1 = k_2$ for Dk_1k_2BS.

Then, the Dk_1k_2BS problem (17) is equivalent to the following quadratic minimization problem:

$$\min_{\boldsymbol{a}\in\mathcal{U}^{n_1,n_2}_{k_1,k_2}} \quad f(\boldsymbol{a}) := -\boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{a}. \tag{58}$$

The difference to the DkS problem is that the above problem has two cardinality constraints on two disjoint segments of the variable $\boldsymbol{a}$. There is no need to treat the two vector variables $\boldsymbol{x}$ and $\boldsymbol{y}$ separately; our error bound function and the proposed PGM algorithm can be directly applied with the variable $\boldsymbol{a}$.

F: More Descriptions and Results for the Dk_1k_2BS Problem

We preprocess all the datasets by converting each graph into a simple, undirected form by removing edge directions (if present), eliminating multi-edges, and discarding isolated nodes. Furthermore, all nonzero edge weights are set to 1. We follow the implementation in Algorithm 1. All the methods are initialized with the vector $\boldsymbol{a}^0 = \mathbf{1}/(k_1 + k_2)$. The initial penalty parameter is set to $\lambda_0 = 10^{-10}$. The whole algorithm terminates when either $\|\boldsymbol{a}^{\ell+1} - \boldsymbol{a}^\ell\|_2^2 \le 10^{-15}$, or the number of iterations exceeds 100. The penalty parameter λ is updated by $\lambda_{\ell+1} = 10\lambda_\ell$ when either $\|\boldsymbol{a}^{\ell+1} - \boldsymbol{a}^\ell\|_2/\|\boldsymbol{a}^{\ell+1}\|_2 < 0.5$ or 10 iterations have passed since the last update.

For the experimental evaluation, we also consider varying the value of k_1 from 10 to 1000, and set $k_2 = k_1$. The corresponding results of edge density and runtime are shown in Figure 5 and 6. The proposed method, EP-Prox, consistently outperforms Rank-1 PC and demonstrates competitive performance compared to EXPP, while being faster than EXPP and a little slower than Rank-1 PC.