# Gradient methods for quadratic-regularized POT

Khoa Nguyen

March 2024

# 1 Quadratic Partial Optimal Transport (QPOT)

## 1.1 Partial Optimal Transport (POT)

Consider two discrete distributions $\mathbf{r}, \mathbf{c} \in \mathbb{R}_+^n$ with possibly different masses. POT seeks a transport plan $\mathbf{X} \in \mathbb{R}_+^{n\times n}$ which maps $\mathbf{r}$ to $\mathbf{c}$ at the lowest cost. Since the masses at two marginals may differ, only a total mass $s$ such that $0 \le s \le \min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\}$ is allowed to be transported [1, 2]. Formally, the POT problem is written as

$$\mathbf{POT}(\mathbf{r}, \mathbf{c}, s) = \min\langle \mathbf{C}, \mathbf{X}\rangle \;\text{ s.t. }\; \mathbf{X} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s), \tag{1}$$

where $\mathcal{U}(\mathbf{r}, \mathbf{c}, s)$ is defined as $\mathcal{U}(\mathbf{r}, \mathbf{c}, s) = \left\{\mathbf{X} \in \mathbb{R}_+^{n\times n} : \mathbf{X}\mathbf{1}_n \le \mathbf{r}, \mathbf{X}^\top \mathbf{1}_n \le \mathbf{c}, \mathbf{1}_n^\top \mathbf{X}\mathbf{1}_n = s\right\}$, i.e. the feasible set for the transport map $\mathbf{X}$ is and $\mathbf{C} \in \mathbb{R}_+^{n\times n}$ is a cost matrix. The goal of this paper is to derive efficient algorithms to find an $\varepsilon$-approximate solution to $\mathbf{POT}(\mathbf{r}, \mathbf{c}, s)$, pursuant to the following definition.

**Definition 1** ($\varepsilon$-approximation)**.** *For $\varepsilon \ge 0$, the matrix $\mathbf{X} \in \mathbb{R}_+^{n\times n}$ is an $\varepsilon$-approximate solution to $\mathbf{POT}(\mathbf{r}, \mathbf{c}, s)$ if $\mathbf{X} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s)$ and*

$$\langle \mathbf{C}, \mathbf{X}\rangle \le \min\langle \mathbf{C}, \mathbf{X}'\rangle + \varepsilon \;\; \text{s.t.} \;\; \mathbf{X}' \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s).$$

## 1.2 Quadratic Partial Optimal Transport (QPOT)

The Quadratic Partial Optimal Transport (QPOT) problem is written as:

$$\begin{aligned} &\mathbf{QPOT}_\eta(\mathbf{r}, \mathbf{c}, s) = \min\left\{ f_\eta(\mathbf{X}) \triangleq \langle \mathbf{C}, \mathbf{X}\rangle + \eta\|\mathbf{X}\|_2^2 \right\} \\ &\text{s.t. } \; \mathbf{X} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s) = \{\mathbf{X} \in \mathbb{R}_+^{n\times n} : \mathbf{X}\mathbf{1}_n \le \mathbf{r}, \mathbf{X}^\top \mathbf{1}_n \le \mathbf{c}, \mathbf{1}_n^\top \mathbf{X}\mathbf{1}_n = s\}. \end{aligned} \tag{2}$$

Let $\mathbf{X}^\eta \in \operatorname{argmin}_{\mathbf{X}\in\mathcal{U}(\mathbf{r},\mathbf{c},s)}\{f_\eta(\mathbf{X})\}$ be the optimal transportation plan of the QPOT problem (2).

# 2 New Iterative Method for QPOT

## 2.1 Penalty method

Skim this paper along the way: [3]. Don't try to understand all the theoretical reasonings. Just briefly understand the interpretation of the claimed results therein.

Among the constraints in $\mathcal{U}(\mathbf{r}, \mathbf{c}, s)$, the $\mathbf{X} \ge 0$ (box constraints) and $\mathbf{1}_n^\top \mathbf{X}\mathbf{1}_n = s$ ($\ell_1$ ball constraints) are easy. In order to handle the remaining inequality constraints, i.e. $\mathbf{X}\mathbf{1}_n \le \mathbf{r}$ and $\mathbf{X}^\top \mathbf{1}_n \le \mathbf{c}$, we would rely on the quadratic exterior penalty method [3]. In particular, consider the following penalty function:

$$P(\mathbf{X}, \alpha) = \alpha \sum_{i=1}^{n} \left[ \min\{0, r_i - (\mathbf{X}\mathbf{1}_n)_i\}^2 + \min\{0, c_i - (\mathbf{X}^\top \mathbf{1}_n)_i\}^2 \right]. \tag{3}$$

**Task 1:** $(\mathbf{X}\mathbf{1}_n)_i$ is the i-th coordinate of $\mathbf{X}\mathbf{1}_n$. Express it in full form. Then break down both $(\mathbf{X}\mathbf{1}_n)_i$ and $(\mathbf{X}^\top\mathbf{1}_n)_i$ (into full forms) in (3).

*Proof.* Given $\mathbf{X} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s) = \{\mathbf{X} \in \mathbb{R}_+^{n\times n} : \mathbf{X}\mathbf{1}_n \leq \mathbf{r}, \mathbf{X}^\top\mathbf{1}_n \leq \mathbf{c}, \mathbf{1}_n^\top\mathbf{X}\mathbf{1}_n = s\}$, we can rewrite the first two constraints as

- $$\begin{aligned}\mathbf{X}\mathbf{1}_n \leq \mathbf{r} \Leftrightarrow \mathbf{X}\mathbf{1}_n = \begin{bmatrix}\sum_{j=1}^n \mathbf{X}_{1j}\\ \sum_{j=1}^n \mathbf{X}_{2j}\\ \vdots \\ \sum_{j=1}^n \mathbf{X}_{nj}\end{bmatrix} &\leq \begin{bmatrix}\mathbf{r}_1\\ \mathbf{r}_2\\ \vdots\\ \mathbf{r}_n\end{bmatrix} = \mathbf{r}\\ \Leftrightarrow \mathbf{r}_i - (\mathbf{X}\mathbf{1}_n)_i &\geq 0, \quad \forall i\end{aligned}$$

- $$\begin{aligned}\mathbf{X}^T\mathbf{1}_n \leq \mathbf{c} \Leftrightarrow \mathbf{X}^T\mathbf{1}_n = \begin{bmatrix}\sum_{i=1}^n \mathbf{X}_{i1}\\ \sum_{i=1}^n \mathbf{X}_{i2}\\ \vdots \\ \sum_{i=1}^n \mathbf{X}_{in}\end{bmatrix} &\leq \begin{bmatrix}\mathbf{c}_1\\ \mathbf{c}_2\\ \vdots\\ \mathbf{c}_n\end{bmatrix} = \mathbf{c}\\ \Leftrightarrow \mathbf{c}_i - (\mathbf{X}^T\mathbf{1}_n)_i &\geq 0, \quad \forall i\end{aligned}$$

□

Next, we consider the penalized objective:

$$F_\eta(\mathbf{X}, \alpha) = f_\eta(\mathbf{X}) + P(\mathbf{X}, \alpha), \tag{4}$$

and the Penalized QPOT (P-QPOT) problem as follows:

$$\textbf{P-QPOT}_{\eta,\alpha}(\mathbf{r}, \mathbf{c}, s) = \min_{\mathbf{X}\in\mathbb{R}_+^{n\times n}:\mathbf{1}_n^\top\mathbf{X}\mathbf{1}_n = s} F_{\eta,\alpha}(\mathbf{X}). \tag{5}$$

Note that in the above, we have removed the inequality constraints from the optimziation problem. Let $F = \{\mathbf{X} \in \mathbb{R}^{n\times n} : \mathbf{X}\mathbf{1}_n \leq \mathbf{r}, \mathbf{X}^\top\mathbf{1}_n \leq \mathbf{c}\}$ be the set of $\mathbf{X}$ satisfying those two constraints. In the following Lemma, we would intuitively establish why the penalty method works.

**Lemma 1.** *For $0 < \alpha_1 < \alpha_2$, we have the following:*

$$P(\mathbf{X}, \alpha_1) \leq P(\mathbf{X}, \alpha_2). \tag{6}$$

*Furthermore, for any $\alpha > 0$, we have:*

$$P(\mathbf{X}, \alpha) = 0, \forall \mathbf{X} \in F \tag{7}$$

$$P(\mathbf{X}, \alpha) > 0, \forall \mathbf{X} \notin F, \tag{8}$$

$$\lim_{\alpha\to\infty} P(\mathbf{X}, \alpha) = \infty, \forall \mathbf{X} \notin F. \tag{9}$$

*Proof.* Given $\mathbf{X} \in F = \{\mathbf{X} \in \mathbb{R}^{n\times n} : \mathbf{X}\mathbf{1}_n \leq \mathbf{r}, \mathbf{X}^\top\mathbf{1}_n \leq \mathbf{c}\}$, we can rewrite the first two constraints as:

$$\begin{aligned}\mathbf{X}\mathbf{1}_n \leq \mathbf{r} &\Leftrightarrow \mathbf{r}_i - (\mathbf{X}\mathbf{1}_n)_i \geq 0, \quad \forall i\\ &\Leftrightarrow \min(\mathbf{r}_i - (\mathbf{X}\mathbf{1}_n)_i, 0) = 0 \quad \forall i \quad (1)\\ \mathbf{X}^T\mathbf{1}_n \leq \mathbf{c} &\Leftrightarrow \mathbf{c}_i - (\mathbf{X}^T\mathbf{1}_n)_i \geq 0, \quad \forall i\\ &\Leftrightarrow \min(\mathbf{c}_i - (\mathbf{X}^T\mathbf{1}_n)_i, 0) = 0 \quad \forall i \quad (2)\end{aligned}$$

From (1), (2):

$$
\begin{aligned}
P(\mathbf{X}, \alpha) &= \alpha \sum_{i=1}^{n} \left(\min\{0, \mathbf{r}_i - (\mathbf{X}\mathbf{1}_n)_i\}^2 + \min\{0, \mathbf{c}_i - (\mathbf{X}^T\mathbf{1}_n)_i\}^2\right) \\
&= \alpha \sum_{i=1}^{n} \left(0^2 + 0^2\right) = 0
\end{aligned}
$$

On the other hand, given $\mathbf{X} \notin F$, then there exists $i \in \{1, \cdots, n\}$ such that either

$$
\mathbf{r}_i - (\mathbf{X}\mathbf{1}_n)_i < 0 \quad \text{or} \quad \mathbf{c}_i - (\mathbf{X}^T\mathbf{1}_n)_i < 0
$$

. Then, there exists

$$
\min\{0, \mathbf{r}_i - (\mathbf{X}\mathbf{1}_n)_i\}^2 + \min\{0, \mathbf{c}_i - (\mathbf{X}^T\mathbf{1}_n)_i\}^2 = c > 0
$$

. Hence,

$$
P(\mathbf{X}, \alpha) = \alpha \sum_{i=1}^{n} \left(\min\{0, \mathbf{r}_i - (\mathbf{X}\mathbf{1}_n)_i\}^2 + \min\{0, \mathbf{c}_i - (\mathbf{X}^T\mathbf{1}_n)_i\}^2\right) = \alpha c > 0
$$

Then, given $0 < \alpha_1 < \alpha_2$ and:

- $P(\mathbf{X}, \alpha) = 0$ for $\mathbf{X} \in F$
- $P(\mathbf{X}, \alpha) > 0$ for $\mathbf{X} \notin F$

The following holds

$$
P(\mathbf{X}, \alpha_1) \leq P(\mathbf{X}, \alpha_2)
$$

Ultimately, given $c > 0$ and $P(\mathbf{X}, \alpha) = \alpha c$ when $\mathbf{X} \notin F$, $\lim_{\alpha \to \infty} P(\mathbf{X}, \alpha) = \lim_{\alpha \to \infty} \alpha c = \infty$

□

Using the above Lemma 1, we can derive the equivalence between **QPOT** and **P-QPOT** in the limit sense in the next Theorem.

**Theorem 1.** *We have:*

$$
\textbf{\textit{QPOT}}_{\eta}(\mathbf{r}, \mathbf{c}, s) = \lim_{\alpha \to \infty} \textbf{\textit{P-QPOT}}_{\eta, \alpha}(\mathbf{r}, \mathbf{c}, s) \tag{10}
$$

*Proof.* As proven that:

$$
\begin{aligned}
P(\mathbf{X}, \alpha) = 0, \forall \mathbf{X} \in F \Rightarrow &\lim_{\alpha \to \infty} P(\mathbf{X}, \alpha) = 0, \forall \mathbf{X} \in F \\
\text{and:} \quad &\lim_{\alpha \to \infty} P(\mathbf{X}, \alpha) = \infty, \forall \mathbf{X} \notin F.
\end{aligned}
$$

Let $\mathbf{X} \in \mathcal{U}$ be the solution of $\textbf{P-QPOT}_{\eta,\alpha}(\mathbf{r}, \mathbf{c}, s)$, $\mathbf{X}$ then must satisfy the constraints of $F$. Hence, $\lim_{\alpha \to \infty} \textbf{P-QPOT}_{\eta,\alpha}(\mathbf{r}, \mathbf{c}, s)$ can be re-written as:

$$
\begin{aligned}
\lim_{\alpha \to \infty} \textbf{P-QPOT}_{\eta,\alpha}(\mathbf{r}, \mathbf{c}, s) &= \lim_{\alpha \to \infty} \left(\textbf{QPOT}_{\eta}(\mathbf{r}, \mathbf{c}, s) + P(\mathbf{X}, \alpha)\right) \\
&= \textbf{QPOT}_{\eta}(\mathbf{r}, \mathbf{c}, s) + \lim_{\alpha \to \infty} P(\mathbf{X}, \alpha) \\
&= \textbf{QPOT}_{\eta}(\mathbf{r}, \mathbf{c}, s) + 0 \\
&= \textbf{QPOT}_{\eta}(\mathbf{r}, \mathbf{c}, s)
\end{aligned}
$$

□

However, such equivalence can hold for large enough $\alpha$. By invoking [3], we indeed can derive a bound on how large $\alpha$ is for such equivalence to hold in Theorem 2.

In order to prove Theorem 2, we would need the following supplementary Lemmas.

**Lemma 2.** *For any $\mathbf{X} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s)$, we have the bound on problem size:*

$$f_\eta(\mathbf{X}) \leq s\|\mathbf{C}\|_\infty + \eta s^2 \tag{11}$$

*Proof.* Given that: $f_\eta(\mathbf{X}) \triangleq \langle \mathbf{C}, \mathbf{X} \rangle + \eta\|\mathbf{X}\|_2^2$ where:

- $\mathbf{X}, \mathbf{C} \in \mathbb{R}_+^{n\times n}$, $\mathbf{r}, \mathbf{c}, \mathbf{s} \in \mathbb{R}_+^n$
- $\mathbf{X} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s) = \{\mathbf{X} \in \mathbb{R}_+^{n\times n} : \mathbf{X}\mathbf{1}_n \leq \mathbf{r}, \mathbf{X}^\top \mathbf{1}_n \leq \mathbf{c}, \mathbf{1}_n^\top \mathbf{X}\mathbf{1}_n = s\}$

The regularization term can be re-written as:

$$\|\mathbf{X}\|_2^2 = \sum_{i=1}^n \sum_{j=1}^n x_{ij}^2 \leq \sum_{i=1}^n \sum_{j=1}^n x_{ij}^2 + 2 \sum_{\substack{i=1 \\ (i,j)\neq(k,l)}}^n \sum_{j=1}^n x_{ij}x_{kl} = \left(\sum_{i=1}^n \sum_{j=1}^n x_{ij}\right)^2 = \mathbf{s}^2. \quad (*)$$

Meanwhile, consider that: $\|\mathbf{C}\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^n c_{ij}$ and $\mathbf{C} \in \mathbb{R}_+^{n\times n}$. As every element in $\mathbf{C}$ is positive, this inequality holds: $c_{ij} \leq \|\mathbf{C}\|_\infty$. Hence, the inner product $\langle \mathbf{C}, \mathbf{X} \rangle$ can be bounded by:

$$\langle \mathbf{C}, \mathbf{X} \rangle = \sum_{i=1}^n \sum_{j=1}^n c_{ij}x_{ij} \leq \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{C}\|_\infty x_{ij} = \|\mathbf{C}\| \sum_{i=1}^n \sum_{j=1}^n x_{ij} = s\|\mathbf{C}\|_\infty \quad (**)$$

From (*) and (**), $f_\eta(\mathbf{X})$ hence is bounded by:

$$f_\eta(\mathbf{X}) \triangleq \langle \mathbf{C}, \mathbf{X} \rangle + \eta\|\mathbf{X}\|_2^2 \leq s\|\mathbf{C}\|_\infty + \eta s^2 \tag{12}$$

Good Kid □

Next, we would establish the following Slater Lemma 5.

In addition to the feasible domain $\mathcal{U}(\mathbf{r}, \mathbf{c}, s)$ of the POT problem, we now also consider the domain $\Upsilon(\mathbf{a}, \mathbf{b}) = \left\{\mathbf{X} \in \mathbb{R}_+^{n\times n} : \mathbf{X}\mathbf{1}_n = \mathbf{a}, \mathbf{X}^\top \mathbf{1}_n = \mathbf{b}\right\}$ that is relevant to the feasible domain of the OT problem (note that $\Upsilon(\mathbf{a}, \mathbf{b})$ does not contain the constraint $\|\mathbf{X}\|_1 = 1$).

**Lemma 3** (Rim condition for transportation problem)**.** *The necessary and sufficient condition (aka if and only if) for $\Upsilon(\mathbf{a}, \mathbf{b})$ to be feasible (i.e. there exists some $\mathbf{X} \in \Upsilon(\mathbf{a}, \mathbf{b})$) is that $\|\mathbf{a}\|_1 = \|\mathbf{b}\|_1$ given that $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^n$.*

*Proof.* This is a known result. Kid Nguyen, you don't have to prove. But google-search some sources to read and understand it by yourself, and cite some sources here. □

**Lemma 4.** *Show that for any $\mathbf{r}, \mathbf{c} \in \mathbb{R}_+^n, s > 0$, the domain $\mathcal{U}(\mathbf{r}, \mathbf{c}, s)$ is feasible if and only if $s \leq \min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\}$.*

*Proof.* To prove the statement above, we prove it two ways. The first is to prove that if $\exists \mathbf{X} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s)$, then $s \leq \min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\}$.

Let $\mathbf{X} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s)$, knowing 3:

$$\mathbf{X} \in \mathcal{U} \Leftrightarrow \begin{cases} \mathbf{X}\mathbf{1}_n \leq \mathbf{r} \\ \mathbf{X}^T\mathbf{1}_n \leq \mathbf{c} \\ \mathbf{1}_n^T\mathbf{X}\mathbf{1}_n = s \end{cases}$$
$$\Leftrightarrow \begin{cases} \sum_n^i \sum_n^j x_{ij} \leq \|\mathbf{r}\|_1 \\ \sum_n^j \sum_n^i x_{ij} \leq \|\mathbf{c}\|_1 \\ \sum_n^i \sum_n^j x_{ij} = s \end{cases}$$
$$\Rightarrow \begin{cases} s \leq \|\mathbf{r}\|_1 \\ s \leq \|\mathbf{c}\|_1 \end{cases} \quad \Leftrightarrow s \leq \min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\}$$

Now, let $s \leq \min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\}$. Then it is possible to construct a $\mathbf{r}'$ and $\mathbf{c}'$ such that:

- $\mathbf{c}_i' \leq \mathbf{c}_i \ \forall i = 1, ..., n$
- $\mathbf{r}_i' \leq \mathbf{r}_i \ \forall i = 1, ..., n$
- $\|\mathbf{r}'\|_1 \leq \|\mathbf{r}\|_1$, $\|\mathbf{c}'\|_1 \leq \|\mathbf{c}\|_1$
- $\|\mathbf{r}'\|_1 = \|\mathbf{c}'\|_1 = s$

For example, to enforce $\|\mathbf{r}'\|_1 = \|\mathbf{c}'\|_1 = s$ and $\|\mathbf{r}'\|_1 \leq \|\mathbf{r}\|_1$, $\|\mathbf{c}'\|_1 \leq \|\mathbf{c}\|_1$, we can construct $\mathbf{r}'$ and $\mathbf{c}'$ such that:

- $\mathbf{r}_i' = s\mathbf{r}_i/\|\mathbf{r}\|_1 \ i = 1, ..., n$
- $\mathbf{c}_i' = s\mathbf{c}_i/\|\mathbf{c}\|_1 \ i = 1, ..., n$

As constructing such $\|\mathbf{r}'\| = \|\mathbf{c}'\|$ is possible and $\mathbf{1}_n^T\mathbf{X}\mathbf{1} = s$, then there is a feasible domain of $\mathbf{X} \in \Upsilon(\mathbf{r}', \mathbf{c}') = \left\{\mathbf{X} \in \mathbb{R}_+^{n\times n} : \mathbf{X}\mathbf{1}_n = \mathbf{r}', \mathbf{X}^\top\mathbf{1}_n = \mathbf{c}', \mathbf{1}_n^T\mathbf{X}\mathbf{1} = s\right\}$.

Moreover, since $\mathbf{r}_i' \leq \mathbf{r}_i$ and $\mathbf{c}_i' \leq \mathbf{c}_i \ \forall i = 1, ..., n$, implying that $\mathbf{X}\mathbf{1}_n = \mathbf{r}' \leq \mathbf{r}, \mathbf{X}^\top\mathbf{1}_n = \mathbf{c}' \leq \mathbf{c}$ making $\Upsilon(\mathbf{r}', \mathbf{c}') \subseteq \mathcal{U}(\mathbf{r}, \mathbf{c}, s)$ and also making $\mathcal{U}(\mathbf{r}, \mathbf{c}, s)$ feasible. Hence, $\mathbf{X} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s)$ [4]

$\square$

**Lemma 5** (Slater's condition)**.** *Assume that* $\min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\} - s > 0$*. Let* $r_{min} = \min_i\{r_i\}$, $c_{min} = \min_i\{c_i\}$ *and:*

$$\zeta = \min\left\{r_{min}, c_{min}, \frac{1}{n}\big(\min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\} - s\big)\right\}. \tag{13}$$

*Then there exists some* $\bar{\mathbf{X}} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s)$ *such that:*

$$\bar{\mathbf{X}}\mathbf{1}_n + \zeta\mathbf{1}_n \leq \mathbf{r} \tag{14}$$
$$\bar{\mathbf{X}}^\top\mathbf{1}_n + \zeta\mathbf{1}_n \leq \mathbf{c}. \tag{15}$$

*Proof.* Try to leverage the above Lemma 4 to prove this. In particular, ask yourself whether the domain $\mathcal{U}(\mathbf{r} - \zeta\mathbf{1}_n, \mathbf{c} - \zeta\mathbf{1}_n, s)$ is feasible.

Knowing

$$\zeta = \min\left\{r_{min}, c_{min}, \frac{1}{n}\big(\min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\} - s\big)\right\}.$$

Then:

$$\begin{aligned}
&\zeta \leq \frac{1}{n}\big(\min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\} - s\big)\\
\Leftrightarrow & s + n\zeta \leq \min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\}\\
\Leftrightarrow & s + \|\zeta \mathbf{1}_n\|_1 \leq \min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\}\\
\Leftrightarrow & s \leq \min\{\|\mathbf{r}\|_1 - \|\zeta \mathbf{1}_n\|_1, \|\mathbf{c}\|_1 - \|\zeta \mathbf{1}_n\|_1\}\\
\Leftrightarrow & s \leq \min\{\|\mathbf{r} - \zeta \mathbf{1}_n\|_1, \|\mathbf{c} - \zeta \mathbf{1}_n\|_1\}
\end{aligned}$$

Note: $\|\mathbf{r}\|_1 = \|(\mathbf{r} - \zeta \mathbf{1}_n) + \zeta \mathbf{1}_n\|_1 \leq \|\mathbf{r} - \zeta \mathbf{1}_n\|_1 + \|\zeta \mathbf{1}_n\|_1 \implies \|\mathbf{r}\|_1 - \|\zeta \mathbf{1}_n\|_1 \leq \|\mathbf{r} - \zeta \mathbf{1}_n\|_1$

which is necessary and sufficient to prove that there exists a feasible domain $\mathcal{U}(\mathbf{r} - \zeta \mathbf{1}_n, \mathbf{c} - \zeta \mathbf{1}_n, s)$

On the other hand

In other words, there exists some $\bar{\mathbf{X}} \in \mathcal{U}(\mathbf{r}, \mathbf{c}, s)$ such that:

$$\begin{aligned}
\bar{\mathbf{X}}\mathbf{1}_n + \zeta \mathbf{1}_n &\leq \mathbf{r}\\
\bar{\mathbf{X}}^\top \mathbf{1}_n + \zeta \mathbf{1}_n &\leq \mathbf{c}.
\end{aligned}$$

□

**Theorem 2.** *For* $\alpha > \frac{(s\|\mathbf{C}\|_\infty + \eta s^2)(\sqrt{2n}+1)}{2\zeta\varepsilon} = O\big(\frac{n^{1.5}}{\varepsilon}\big)$, *we have:*

$$\boldsymbol{QPOT}_\eta(\mathbf{r}, \mathbf{c}, s) = \boldsymbol{P\text{-}QPOT}_{\eta,\alpha}(\mathbf{r}, \mathbf{c}, s) \pm \varepsilon. \tag{16}$$

*Furthermore, if* $\mathbf{X}^{\eta,\alpha}$ *is a solution to* $\boldsymbol{P\text{-}QPOT}_{\eta,\alpha}(\mathbf{r}, \mathbf{c}, s)$, *then we have:*

$$\mathbf{X}^{\eta,\alpha}\boldsymbol{1}_n - \varepsilon \boldsymbol{1}_n \leq \mathbf{r} \tag{17}$$

$$\mathbf{X}^{\eta,\alpha\top}\boldsymbol{1}_n - \varepsilon \boldsymbol{1}_n \leq \mathbf{c}. \tag{18}$$

*Proof.* This is direct yet non-trivial application of Theorem 1 from [3]. Note that:

- (17) and (18) are different beasts from (14) and (15). The latter concerns Slater's condition, while the former concerns $\varepsilon$ convergence to the feasibility domain; see Definition 4 in [3].
- [3, Theorem 1] would involve their quantity $r_0$ in the paper. $r_0$ is hard to estimate, yet we can instead use the RHS of the bound (8), say $r'_0 > r_0$ in their paper to replace $r_0$. The goal is that we then would choose some $\alpha > r'_0$ (or anything that is $> r_0$) to guarantee $\varepsilon$ convergence to both the objective (16) and the feasibility domain in the sense of (17) and (18). ask if u need; this can be confusing
- Such $r'_0$, i.e. RHS of the bound (8), would essentially the ratio between the problem size bound in Lemma 2 and the Slater coefficient $\zeta$ in Lemma 5. Convince yourself this.ask if u need; this can be confusing

Now, note that Theorem 1 of [3] involves the following parameter $r'_0$ (as the bound/estimate for their $r_0$) aka their equation (8), $h = (m^{1/2} + 1)/2$. Read and understand the meaning of these parameters in the paper, and write down explicitly here what $r'_0$, $m$ and thus $h$ are in our context.

Let's do it, Kid Nguyen.

Info:

- [3] Theorem 1: $\varepsilon$-converge occurs when $r_\varepsilon = \frac{r_0 h}{\varepsilon}$

- [3] Theorem 1: Knowing $h = \frac{\sqrt{m}+1}{2}$ and $P(\mathbf{X}, \alpha)$ is penalty function of two vectors in $\mathbb{R}^n$, then $h = \frac{\sqrt{2n}+1}{2}$
- [3] inequation (7): finite convergence occurs when $(x^0, u^0)$ is a saddle point of the Lagrangian function and $r_0 > max_i u_i^0$ be the appropriate penalty weight of the penalty function.
- [3] inequation (8): $r_0$ of (7) is hard to estimate. However, if $\exists \overline{x}$ where $g_i(\overline{x}) > 0 \forall i$ and an upper bound z of $f(x^0)$ where $x^0$ is the optimal value of the original optimization problem, then $r_0$ can be estimated by: $\overline{r} > \frac{z - f(\overline{x})}{min_i g_i(\overline{x})}$.
- Our problem (equivalently) is to $\max(-f_\eta(\mathbf{X}))$. An upperbound for $-f_\eta(\mathbf{X}_{optimal})$ is $z = 0$. So the term in the paper can be written as $\frac{z - f(\overline{x})}{min_i g_i(\overline{x})} = \frac{0 - (-f_\eta(\mathbf{X}))}{min_i g_i(\overline{x})} = \frac{f_\eta(\mathbf{X})}{min_i g_i(\overline{x})} \leq \frac{s\|\mathbf{C}\|_\infty + \eta s^2}{\zeta}$
- From Lemma (2), the upper bound of $f_\eta(\mathbf{X})$ is $(s\|\mathbf{C}\|_\infty + \eta s^2)$
- From (13), $\zeta$ is a Slater coefficient $\zeta = \min\left\{ r_{min}, c_{min}, \frac{1}{n}\big(\min\{\|\mathbf{r}\|_1, \|\mathbf{c}\|_1\} - s\big) \right\}$. while $g_i(x)$ be the constraints, we can somewhat imply $\zeta \leq min_i g_i(\overline{x})$

Putting the pieces together, according to [3], to ensure $\varepsilon$-*convergence*, a penalty weight of $r_\varepsilon = \frac{r_0 h}{\varepsilon}$ is sufficient to achieve convergence. Then, we can establish a lower bound of:

$$r_\varepsilon = \frac{r_0 h}{\varepsilon} > \frac{(s\|\mathbf{C}\|_\infty + \eta s^2)(\sqrt{2n}+1)}{2\varepsilon min_i g_i(\overline{x})} \geq \frac{(s\|\mathbf{C}\|_\infty + \eta s^2)(\sqrt{2n}+1)}{2\zeta\varepsilon} = \mathcal{O}\left(\frac{1n^{0.5}}{n^{-1}\varepsilon}\right) = \mathcal{O}\left(\frac{n^{1.5}}{\varepsilon}\right)$$

$\square$

After this, would need to bound objective gap. And $\varepsilon$ violation of constraints would mean we need the rounding algorithm from [5]

Next step: Use projected accelerated gradient methods to

## 2.2 Algorithmic development

From Theorem 2, we now know that we can solve $\textbf{P-QPOT}_{\eta,\alpha}(\mathbf{r}, \mathbf{c}, s)$ in (5) instead of $\textbf{QPOT}_\eta(\mathbf{r}, \mathbf{c}, s)$. The problem (5) corresponds to strongly-convex and smooth optimization with simple constraints: the box constraint $\mathbf{X} \in \mathbb{R}_+^{n\times n}$ and the $\ell_1$ ball constraint $\mathbf{1}_n^\top \mathbf{X} \mathbf{1}_n = s$. To this end, let $\mathcal{S} = \{\mathbf{X} \in \mathbb{R}^{n\times n} : \mathbf{X} \geq 0, \mathbf{1}_n^\top \mathbf{X} \mathbf{1}_n = s\}$ be the domain of such simple constraint. Then the problem (5) reads:

$$\textbf{P-QPOT}_{\eta,\alpha}(\mathbf{r}, \mathbf{c}, s) = \min_{\mathbf{X}\in\mathcal{S}} F_{\eta,\alpha}(\mathbf{X}). \tag{19}$$

First, familiarize yourself with Nesterov's Accelerated Gradient Descent for Smooth and Strongly Convex Optimization by reading this:

https://web.archive.org/web/20210121055037/https://blogs.princeton.edu/imabandit/2014/03/06/nesterovs-accelerated-gradient-descent-for-smooth-and-strongly-convex-optimization/

Note that from the above:

- The algorithm is for constraint-free optimization. In fact, we would use a more generalized version that can handle simple constraints, where the final complexity of the overall algorithm will be added with the cost for projecting onto the domain $\mathcal{S}$ of simple constraints. In particular, we would use Algorithm 20 and Corollary 4.23 in [6].

- The complexity of the algorithm depends con the condition numbers, comprised of the strong-convexity number and smoothness number. Next, the following Lemmas establish such condition numbers for $F_{\eta,\alpha}(\mathbf{X})$ in $\mathcal{S}$.

You can review some stuffs on strong-convexity (for vectors) in: https://xingyuzhou.org/blog/notes/strong-convexity and smoothness (for vectors) in: https://xingyuzhou.org/blog/notes/Lipschitz-gradient (Note that people also use "Lipschitz continuous gradient" mean smoothness.)

**Lemma 6.** *$F_{\eta,\alpha}(\mathbf{X})$ is $\eta$-strongly convex.*

*Proof.* Hint: $\eta\|\mathbf{X}\|_2^2$ is the name of the game. □

**Lemma 7.** *$F_{\eta,\alpha}(\mathbf{X})$ is $\beta$-smooth in $\mathcal{S}$ with $\beta = blabla = O(n\alpha)$.*

*Proof.* Kid Nguyen, prove that $\|\nabla F_{\eta,\alpha}(\mathbf{X})\|_2 \leq \beta, \forall \mathbf{X} \in \mathcal{S}$ and figure out such value for $\beta$ along the way. I will explain why to you later. For a hint, you can see proof of [7, Lemma 11 ] on how to compute the smoothness number.

We have:

$$\frac{\partial F_\eta(\mathbf{X},\alpha)}{\partial X_{ij}} = C_{ij} + 2\eta X_{ij} - 2\alpha\big(r_i - \sum_{k=1}^n X_{ik}\big)\mathbb{I}\Big(r_i - \sum_{k=1}^n X_{ik} < 0\Big) - 2\alpha\big(c_j - \sum_{k=1}^n X_{kj}\big)\mathbb{I}\Big(c_j - \sum_{k=1}^n X_{kj} < 0\Big) \tag{20}$$

Thus, by Cauchy-Swchartz and using $\|.\|_2 \leq \|.\|_1$, for $\mathbf{X} \in \mathcal{S}$, we have:

$$\begin{aligned}
\|\nabla F_{\eta,\alpha}(\mathbf{X})\|_2^2 &\leq 4\sum_{i,j}\Big[C_{ij}^2 + 4\eta^2 X_{ij}^2 + 8\alpha^2 r_i^2 + 8\alpha^2(\sum_{k=1}^n X_{ik})^2 + +8\alpha^2 c_j^2 + 8\alpha^2(\sum_{k=1}^n X_{kj})^2\Big] &(21)\\
&\leq 4\|\mathbf{C}\|_2^2 + 16\eta^2\|\mathbf{X}\|_1^2 + 64\alpha^2 n^2\|\mathbf{X}\|_1^2 + 32\alpha^2\|\mathbf{r}\|_1^2 + 32\alpha^2\|\mathbf{c}\|_1^2 &(22)\\
&= 4\|\mathbf{C}\|_2^2 + 16\eta^2 s^2 + +64n^2\alpha^2 s^2 + 32\alpha^2\|\mathbf{r}\|_1^2 + 32\alpha^2\|\mathbf{c}\|_1^2 &(23)\\
&= O(n^2\alpha^2) &(24)\\
\therefore \|\nabla F_{\eta,\alpha}(\mathbf{X})\|_2 &\leq n\alpha. &(25)
\end{aligned}$$

Besides, we would also have:

$$\|\nabla F_{\eta,\alpha}(\mathbf{X})\|_\infty \leq O\Big(\|C\|_\infty + \eta + \alpha\Big) \tag{26}$$

By Lemma 8, we have $\forall \mathbf{X}, \mathbf{X}'$:

$$\begin{aligned}
&\Big|2\alpha\big(r_i - \sum_{k=1}^n X_{ik}\big)\mathbb{I}\Big(r_i - \sum_{k=1}^n X_{ik} < 0\Big) - 2\alpha\big(r_i - \sum_{k=1}^n X'_{ik}\big)\mathbb{I}\Big(r_i - \sum_{k=1}^n X'_{ik} < 0\Big)\Big|^2 &(27)\\
&\leq 4\alpha^2|\sum_{k=1}^n X_{ik} - \sum_{k=1}^n X'_{ik}|^2 \leq 4\alpha^2 n\sum_{k=1}^n (X_{ik} - X'_{ik})^2 &(28)
\end{aligned}$$

From (20), we have:

$$\|\nabla F_\eta(\mathbf{X},\alpha) - \nabla F_\eta(\mathbf{X}',\alpha)\|_2^2 \tag{29}$$

The Hessian of $P(\mathbf{X}, \alpha)$ with respect to $X_{ij}$ is given by:

$$\frac{\partial^2 P(\mathbf{X}, \alpha)}{\partial X_{ij} \partial X_{kl}} = -2\alpha \Big[ \mathbb{I}\Big( r_i - \sum_{m=1}^{n} X_{im} < 0 \Big) \mathbb{I}(i = k) + \mathbb{I}\Big( c_j - \sum_{m=1}^{n} X_{mj} < 0 \Big) \mathbb{I}(j = l) \Big].$$

□

**Lemma 8.** *The function $g(x) = 2\alpha(r_i - x)\mathbb{I}_{\{r_i - x < 0\}}$ is Lipschitz continuous with Lipschitz constant $L = 2\alpha$.*

*Proof.* The function $g(x)$ is defined as:

$$g(x) = \begin{cases} 0, & \text{if } x \le r_i, \\ 2\alpha(r_i - x), & \text{if } x > r_i. \end{cases}$$

To verify that $g(x)$ is Lipschitz continuous, we need to show that there exists a constant $L \ge 0$ such that for all $x_1, x_2 \in \mathbb{R}$,

$$|g(x_1) - g(x_2)| \le L|x_1 - x_2|.$$

We consider the following cases:

**Case 1:** $x_1, x_2 \le r_i$.
In this case, $g(x_1) = g(x_2) = 0$. Therefore,

$$|g(x_1) - g(x_2)| = 0 \le L|x_1 - x_2| \quad \text{for any } L.$$

**Case 2:** $x_1, x_2 > r_i$.
Here, $g(x) = 2\alpha(r_i - x)$. Thus,

$$|g(x_1) - g(x_2)| = |2\alpha(r_i - x_1) - 2\alpha(r_i - x_2)| = 2\alpha|x_2 - x_1|.$$

This inequality holds with $L = 2\alpha$.

**Case 3:** One of $x_1 \le r_i$ and $x_2 > r_i$.
Without loss of generality, assume $x_1 \le r_i$ and $x_2 > r_i$. Then $g(x_1) = 0$ and $g(x_2) = 2\alpha(r_i - x_2)$. Therefore,

$$|g(x_1) - g(x_2)| = |0 - 2\alpha(r_i - x_2)| = |2\alpha(r_i - x_2)|.$$

Since $r_i - x_2 < 0$, we have $|g(x_1) - g(x_2)| = 2\alpha|x_2 - r_i|$. Moreover, $|x_2 - r_i| \le |x_2 - x_1|$. Thus,

$$|g(x_1) - g(x_2)| \le 2\alpha|x_2 - x_1|.$$

**Conclusion:** In all cases, $|g(x_1) - g(x_2)| \le 2\alpha|x_1 - x_2|$. Hence, $g(x)$ is Lipschitz continuous with Lipschitz constant $L = 2\alpha$. □

### 2.2.1 Projection onto the probability simplex

**Probability simplex:** a mathematical construct used to represent the space of probability distributions over a finite set of discrete outcomes. A subset of a higher-dimensional space that satisfies:

- Non-negativity
- Sum to 1

For a set with $n$ possible outcomes, the probability simplex is defined as:

$$\Delta^n = \{\mathbf{p} \in \mathbb{R}^n \mid p_i \geq 0 \; \forall i, \; \sum_{i=1}^{n} p_i = 1\}$$

where:

- $\mathbf{p} = (p_1, p_2, ..., p_n)$ is a vector of probabilities
- $\Delta^n$ denotes the simplex $n$-dimensional space

**Projection onto the probability simplex:** Consider the problem of computing the Euclidean projection of a point $\mathbf{y} = [y_1, ..., y_D]^T \in \mathbb{R}^D$ onto the probability simplex. Denote the solution by $\mathbf{x} = [x_1, ..., x_D]^T$, the problem is defined by:

$$\min_{x \in \mathbb{R}^D} \frac{1}{2}\|x - y\|^2 \tag{30}$$

$$\text{s.t. } x^T 1 = 1 \tag{31}$$

$$x \geq 0 \tag{32}$$

which is a quadratic programming problem with a strictly convex objective function

**Algorithm** The following $\mathcal{O}(DlogD)$ algorithm finds the optimal solution $\mathbf{x}$

**Algorithm 1** Euclidean projection of a vector onto the probability simplex

**Require:** $\mathbf{y} \in \mathbb{R}^D$

1: Sort $\mathbf{y}$ into $\mathbf{u}$ such that $u_1 \geq u_2 \geq \cdots \geq u_D$
2: Find $\rho = \max\left\{1 \leq j \leq D : u_j + \frac{1}{j}\left(1 - \sum_{i=1}^{j} u_i\right) > 0\right\}$ (finding the max number of parameter $\rho$ such that $y_1 \geq ... \geq y_\rho$ correspond to the components of the optimal solution $\mathbf{x}$ that are non-zero
3: Define $\lambda = \frac{1}{\rho}\left(1 - \sum_{i=1}^{\rho} u_i\right)$
4: Output $\mathbf{x}$ such that $x_i = \max\{y_i + \lambda, 0\}, \quad i = 1, \ldots, D$ =0

### 2.2.2 Convex Optimization Over a Probability Simplex

**Optimization over the probability simplex** involves minimizing an assumed-convex-function $f(\mathbf{w})$ with $\mathbf{w} \in \mathbf{R}^n$ with in the probability simplex

$$\min_{\mathbf{w} \in \Delta^n} f(\mathbf{w}), \quad \text{where: } \Delta^n = \{\mathbf{p} \in \mathbb{R}^n \mid p_i \geq 0 \; \forall i, \; \sum_{i=1}^{n} p_i = 1\}$$

The paper provides a new algorithm to solve this problem for general convex function $f$ named *Cauchy-Simplex* (CS): Over the iteration t:

$$\begin{aligned} & w^{t+1} = w^t - \eta_t d^t, \\ \text{where } & d^t = w^t(\nabla f - w^t \cdot \nabla f) \\ & 0 < \eta_t \leq \eta_{t,max} \text{ and } \eta_{t,max}^{-1} = \max_i(\nabla_i f - w^t \cdot \nabla f) \end{aligned}$$

With the upper bound of the learning rate $\eta_t$ ensures that $w_i^{t+1}$ is positive for all i. Summing over the indices of $d^t$:

$$\sum_i w_i^t(\nabla_i f - w^t \cdot \nabla f = (w^t \cdot \nabla f)\left(1 - \sum_i w_i^t\right)$$

Thus, if $\sum_i w_i^t = 1$ then $d^t$ lies in the null space of $\sum_i w_i^t$ and $w^{t+1}$ satisfies the unit-sum constraint, giving a scheme where each iteration remains explicitly within the probability simplex

---

**Algorithm 2** Cauchy-Simplex

---

**Require:** $\epsilon \leftarrow 10^{-10}$ (Tolerance for the zero set)

$\mathbf{w} \leftarrow (1/n, \dots, 1/n)$
**while** termination conditions not met **do**
  $S \leftarrow \{i = 1, \dots, n \mid w_i > \epsilon\}$
  $Q \leftarrow \{i = 1, \dots, n \mid w_i \leq \epsilon\}$

  Choose $\eta_t \geq 0$
  $\eta_{\max} \leftarrow \frac{1}{\max_{i \in S}(\nabla f_i) - \mathbf{w} \cdot \nabla f}$
  $\eta_t \leftarrow \min(\eta_t, \eta_{\max})$

  $\hat{\mathbf{w}}^{t+1} \leftarrow \mathbf{w}^t - \eta_t \mathbf{w}^t(\nabla f - \mathbf{w}^t \cdot \nabla f)$

  $\hat{w}_i^{t+1} \leftarrow 0, \quad \forall i \in Q$
  $\mathbf{w}_i^{t+1} \leftarrow \frac{\hat{w}_i^{t+1}}{\sum_j \hat{w}_j^{t+1}}, \quad \forall i$ (Normalizing for numerical stability)
**end while**=0

---

**Proximal Gradient Descent:** Start with the problem:

$$\begin{aligned} &\min f(x) = g(x) + h(x) \\ \text{or:} \quad &\min F_{\eta,\alpha}(\mathbf{X}) = f_\eta(\mathbf{X}) + P(\mathbf{X}, \alpha) \\ &= \langle \mathbf{C}, \mathbf{X} \rangle + \eta \|\mathbf{X}\|_2^2 + \alpha \sum_i^n \left[\min(0, r_i - (\mathbf{X}1_n)_i)^2 + \min(0, c_i - (\mathbf{X}^T 1_n)_i)^2\right] \end{aligned}$$

where:

- $P(\mathbf{X}, \alpha)$ is closed and convex
- $f_\eta(\mathbf{X})$ is differentiable
- there exist constants $m \geq 0$ and $L > 0$ such that

$$f_\eta(\mathbf{X}) - \frac{m}{2}\mathbf{X}^T\mathbf{X}, \quad \frac{L}{2}\mathbf{X}^T\mathbf{X} - f_\eta(\mathbf{X})$$

- optimal value $F^*_{\eta,\alpha}$ is finite and attained at $\mathbf{X}^*$

---

**Algorithm 3** Proximal Gradient Descent

---

$\mathbf{X}_0$, step size $t > 0$
**for** $k = 1, 2, \dots$ **do**
  Compute gradient step: $y^{(k)} = x^{(k-1)} - t\nabla g(x^{(k-1)})$
  Apply proximal operator: $x^{(k)} = \text{prox}_t(y^{(k)})$ which is the projection of $y^{(k)}$ on the probability simplex
**end for**
**return** $x^{(k)}$ =0

---

**Accelerated Proximal Nesterov's Method:**

Start with the problem:

$$\begin{aligned} &\min f(x) = g(x) + h(x) \\ \text{or:}\quad &\min F_{\eta,\alpha}(\mathbf{X}) = f_\eta(\mathbf{X}) + P(\mathbf{X},\alpha) \\ &\qquad = \langle \mathbf{C}, \mathbf{X}\rangle + \eta\|\mathbf{X}\|_2^2 + \alpha \sum_i^n \left[\min(0, r_i - (\mathbf{X}1_n)_i)^2 + \min(0, c_i - (\mathbf{X}^T 1_n)_i)^2\right] \end{aligned}$$

where:

- $P(\mathbf{X},\alpha)$ is closed and convex
- $f_\eta(\mathbf{X})$ is differentiable
- there exist constants $m \geq 0$ and $L > 0$ such that

$$f_\eta(\mathbf{X}) - \frac{m}{2}\mathbf{X}^T\mathbf{X}, \quad \frac{L}{2}\mathbf{X}^T\mathbf{X} - f_\eta(\mathbf{X})$$

- optimal value $F^*_{\eta,\alpha}$ is finite and attained at $\mathbf{X}^*$

---

**Algorithm 4** Accelerated Proximal Nesterov's Method

---

**Require:** $\theta \in (0,1]$

1: Initialize $\mathbf{X}_0 \in \mathbb{R}^{n\times n},\ \mathbf{V}_0 = \mathbf{X}_0,\ \theta \in (0,1]$
2: Set stepsize $t_k$ fixed ($t_k = 1/L$) or obtained from line search
3: **while** Not converge **do**
4: Set $\gamma_k = \frac{\theta_{k-1}^2}{t_{k-1}}$
5: Calculate $\theta_k$ as the positive solution of equation

$$\frac{\theta_k^2}{t_k} = (1-\theta_k)\gamma_k + m\theta_k$$

6: Set $\begin{cases} \mathbf{Y} = \mathbf{X}_0 & \text{if } k = 0 \\ \mathbf{Y} = \mathbf{X}_k + \frac{\theta_k\gamma_k}{\gamma_k + m\theta_k}(\mathbf{V}_k - \mathbf{X}_k) & \text{if } k > 0 \end{cases}$
7: Update $\mathbf{X}_{k+1} = prox_{t_k h}(\mathbf{Y} - t_k\nabla f(\mathbf{Y}))$
8: Update $\mathbf{V}_{k+1} = \mathbf{X}_k + \frac{\mathbf{X}_{k+1} - \mathbf{X}_k}{\theta_k}$
9: **end while**=0

---

The *prox* function is obtained as the projection of $\mathbf{Y} - t_k\nabla f(\mathbf{Y}))$ onto the probability simplex with Algorithm Algorithm 1 and read about the proximal gradient descent (see lecture), say $\mathbf{Y}'$? Keypoint:

This week (and maybe next week) task:

Task 1: Implement the projection step $prox_{t_k h}(.)$ where $h$ is the indicator function.

- Re-implement the projection algorithm that projects a vector onto probability simplex in Python.
- Make the above algorithm works for 2D array being treated as an 1D vector. For now, just convert 2D array to 1D to run projection ($\mathcal{O}(n^2)$), then convert back to 2D. Future (DO NOT do it now): customize logic in the projection algorithm to directly deal with 2D.
- Finally, given $\mathbf{Y} = \mathbf{Y} - t_k\nabla f(\mathbf{Y})$, we convert $\mathbf{Y}$ to a vector, feed $\frac{1}{s}\mathbf{Y}$ into the above algorithm, which return some $\mathbf{X}'$ such that $1^T\mathbf{X}'1 = 1$. Then we would output the vector $\mathbf{X} = s\mathbf{X}'$ and turn this $\mathbf{X}'$ into a matrix with the original dimension.

Task 2: Implement the proximal gradient descent algorithm (see Keypoints) and Round-POT to round the output of the proximal gradient descent.

Task 3: Implement its Nesterov's version using the same proximal function

Proof of $r_\varepsilon$

*Proof.* Given the general optimization problem:

$$\min_{x \in \mathcal{X}} f(x)$$

with the inequality constraints:

$$g_i(x) \leq k, \quad \forall i = 1, \ldots, m,$$

Assume $x^*$ is an approximate solution for the optimization problem with the exterior penalty function:

$$P(x, r) = f(x) + r \cdot \sum_{i=1}^{m} \max(0, g_i(x) - k)^2,$$

where $r > 0$ is the penalty parameter and $k$ is the bound for $g_i(x) \leq k$.

To prove that $P(x, r)$ satisfies $\varepsilon$-convergence when $r_\varepsilon$ achieves the stated value, we essentially show that $P(x^*, r)$, or more precisely $f(x^*) \geq f^*$, holds.

Assume $x^*$ minimizes $P(x^*, r_\varepsilon)$, but $\varepsilon$-convergence is not satisfied. Let the sets of constraints that violate $\varepsilon$-convergence be defined as $I$ while those that satisfies it is $J$:

$$I = \{i \mid g_i(x^*) > k + \varepsilon\}$$
$$J = \{i \mid g_i(x^*) \leq k + \varepsilon\}$$

Under this assumption, there exists at least one violated constraint, i.e., $I \neq \emptyset$, and the maximum number of elements $J$ can have is $J = m - 1$.

- Given the two sets $I$ and $J$, we rewrite $P(x^*, r_\varepsilon)$ as:

$$\begin{aligned} P(x^*, r_\varepsilon) &= f(x^*) + r_\varepsilon \sum_{i=1}^{m} [\min(0, k - g_i(x^*)]^2 \quad (*) \\ &= f(x^*) + r_\varepsilon \sum_{i \in I} [g_i(x^*) - k]^2 + r_\varepsilon \sum_{i \in J} [g_i(x^*) - k]^2 \quad (**) \end{aligned}$$

- Here, we know that $g_i(x^*)$ for $i \in I$ and $i \in J$ exceeds the constraints by a certain margin (since $\varepsilon$-convergence is not satisfied). Therefore, we decompose the penalty function in (*) into the residuals of $g_i(x^*)$ for $i \in I$ and $i \in J$ in (**).
- Using the value of $r_\varepsilon$ above, we analyze the case where $i \in I$:
  - Substitute $r_\varepsilon$ with $r_0$ and add/subtract $\sum_{i \in I} |g_i(x^*) - k|$:

$$r_\varepsilon \sum_{i \in I} [g_i(x^*) - k]^2 = r_0 \sum_{i \in I} |g_i(x^*) - k| + r_0 \sum_{i \in I} \left[ \frac{(g_i(x^*) - k)^2}{\varepsilon/h} - |g_i(x^*) - k| \right]$$

  - When $g_i(x^*) - k$ for $i \in I$, $|g_i(x^*) - k| > \varepsilon \leq \varepsilon/h$, let $y = |g_i(x^*) - k|$:

$$d(y) = \sum_{i \in I} \left[ \frac{(g_i(x^*) - k)^2}{\varepsilon/h} - |g_i(x^*) - k| \right] = \frac{y^2}{\varepsilon/h} - y > 0, \quad \forall y > \varepsilon$$

- Since $d(y)$ is monotonically increasing for $\forall y > \varepsilon$ (as $d(y)$ is a parabola):

$$d(y) > \frac{\varepsilon^2}{\varepsilon/h} - \varepsilon$$

- Substituting this into the original equality, we get:

$$r_\varepsilon \sum_{i \in I} [g_i(x^*) - k]^2 > r_0 \sum_{i \in I} |g_i(x^*) - k| + r_0 \frac{\varepsilon^2}{\varepsilon/h} - \varepsilon = r_0 \sum_{i \in I} |g_i(x^*) - k| + r_0 \varepsilon (h-1) \quad (1)$$

- Using the value of $r_\varepsilon$ above, we analyze the case where $i \in J$:
  - Substitute $r_\varepsilon$ with $r_0$ and add/subtract $\sum_{i \in J} |g_i(x^*) - k|$:

$$r_\varepsilon \sum_{i \in J} [g_i(x^*) - k]^2 = r_0 \sum_{i \in J} |g_i(x^*) - k| + r_0 \sum_{i \in J} \left[ \frac{(g_i(x^*) - k)^2}{\varepsilon/h} - |g_i(x^*) - k| \right]$$

  - Let $y = |g_i(x^*) - k|$. When $g_i(x^*) - k$ for $i \in J$, $0 \leq |g_i(x^*) - k| \leq \varepsilon$:

$$d(y) = \sum_{i \in J} \left[ \frac{(g_i(x^*) - k)^2}{\varepsilon/h} - |g_i(x^*) - k| \right] = \frac{y^2}{\varepsilon/h} - y$$

  - Note that when $0 \leq y \leq \varepsilon$, $d(y)$, which is parabolic, achieves its minimum value at $y = \frac{\varepsilon}{2}$. Substituting this $y$:

$$d(y) \geq -\frac{\varepsilon}{4h}$$

  - Recall that the maximum number of elements $J$ can have is $J = m - 1$:

$$r_\varepsilon \sum_{i \in J} [g_i(x^*) - k]^2 \geq r_0 \sum_{i \in J} |g_i(x^*) - k| - r_0 (m-1) \frac{\varepsilon}{4h} \quad (2)$$

- Combining (1) and (2) into (*), using $h = \frac{\sqrt{m}+1}{2}$:

$$\begin{aligned}
P(x^*, r_\varepsilon) &= f(x^*) + r_\varepsilon \sum_{i \in I} [g_i(x^*) - k]^2 + r_\varepsilon \sum_{i \in J} [g_i(x^*) - k]^2 \\
&> f(x^*) + r_0 \sum_{i \in I} |g_i(x^*) - k| + r_0 \sum_{i \in J} |g_i(x^*) - k| + r_0 \varepsilon (h-1) - r_0 (m-1) \frac{\varepsilon}{4h} \\
&> f(x^*) + r_0 \sum_{i \in I} |g_i(x^*) - k| + r_0 \sum_{i \in J} |g_i(x^*) - k| \\
&> f^*
\end{aligned}$$

This shows that when $h$ and $r_\varepsilon$ achieve the values stated, $P(x^*, r_\varepsilon)$ achieves its optimal value, and the assumption $I \neq \emptyset$ is false.

Thus, proving $f(x^*) \geq f^*$ is sufficient to establish $\varepsilon$-convergence of $P(x, r)$. This ensures that for sufficiently large $r_\varepsilon$, the penalty function $P(x^*, r_\varepsilon)$ leads to a feasible and optimal solution for the original constrained problem, where all constraints $g_i(x^*) \leq k$ are satisfied, and the objective value is near-optimal.

□

# References


[1] L. Chapel, M. Z. Alaya, and G. Gasso, "Partial optimal tranport with applications on positive-unlabeled learning," in *Advances in Neural Information Processing Systems 33*, 2020. 1

[2] K. Le, H. Nguyen, T. Pham, and N. Ho, "On multimarginal partial optimal transport: Equivalent forms and computational complexity," 2021. [Online]. Available: https://arxiv.org/abs/2108.07992 1

[3] K. Truemper, "Note on finite convergence of exterior penalty functions," *Management Science*, vol. 21, no. 5, pp. 600–606, 1975. [Online]. Available: http://www.jstor.org/stable/2630043 1, 4, 6, 7

[4] *Transportation Problem and Variations.* New York, NY: Springer New York, 2003, pp. 207–229. [Online]. Available: https://doi.org/10.1007/0-387-21569-7_7 5

[5] A. D. Nguyen, T. D. Nguyen, Q. M. Nguyen, H. H. Nguyen, L. M. Nguyen, and K.-C. Toh, "On partial optimal transport: Revising the infeasibility of sinkhorn and efficient gradient methods," 2023. [Online]. Available: https://arxiv.org/abs/2312.13970 7

[6] A. d'Aspremont, D. Scieur, and A. Taylor, "Acceleration methods," *Foundations and Trends® in Optimization*, vol. 5, no. 1–2, p. 1–245, 2021. [Online]. Available: http://dx.doi.org/10.1561/2400000036 7

[7] Q. M. Nguyen, H. H. Nguyen, Y. Zhou, and L. M. Nguyen, "On unbalanced optimal transport: Gradient methods, sparsity and approximation error," 2022. [Online]. Available: https://arxiv.org/abs/2202.03618 8