Improved Complexity Bounds in Wasserstein Barycenter Problem (Supplementary Materials)

Darina Dvinskikh, Daniil Tiapkin

1 MISSING PROOFS

1.1 Notation

First, we redefine spaces $\mathcal{X} \triangleq \prod^m \Delta_{n^2} \times \Delta_n$ and $\mathcal{Y} \triangleq [-1,1]^{2mn}$, where $\prod^m \Delta_{n^2} \times \Delta_n$ is the short form of $\underbrace{\Delta_{n^2} \times \ldots \times \Delta_{n^2}}_{m} \times \Delta_n$. Then we rewrite the WB problem for column vectors $\mathbf{x} = (x_1^\top, \ldots, x_m^\top, p^\top)^\top \in \mathcal{X}$ and $\mathbf{y} = (y_1^\top, \ldots, y_m^\top)^\top \in \mathcal{Y}$ in a saddle-point formulation

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{m}\left\{\boldsymbol{d}^\top\mathbf{x} + 2\|d\|_\infty\left(\mathbf{y}^\top\boldsymbol{A}\mathbf{x} - \boldsymbol{c}^\top\mathbf{y}\right)\right\}, \tag{1}$$

where $\boldsymbol{d} = (d^\top, \ldots, d^\top, \mathbf{0}_n^\top)^\top$, $\boldsymbol{c} = (\mathbf{0}_n^\top, q_1^\top, \ldots, \mathbf{0}_n^\top, q_m^\top)^\top$ and $\boldsymbol{A} = \begin{pmatrix}\hat{A} & \mathcal{E}\end{pmatrix} \in \{-1,0,1\}^{2mn\times(mn^2+n)}$ with block-diagonal matrix $\hat{A}$ of m blocks

$$\hat{A} = \begin{pmatrix} A & 0_{2n\times n^2} & \cdots & 0_{2n\times n^2} \\ 0_{2n\times n^2} & A & \cdots & 0_{2n\times n^2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{2n\times n^2} & 0_{2n\times n^2} & \cdots & A \end{pmatrix}$$

and matrix

$$\mathcal{E}^\top = \left(\underbrace{\begin{pmatrix}-I_n & 0_{n\times n}\end{pmatrix}}_{-B_{\mathcal{E}}^\top}\underbrace{\begin{pmatrix}-I_n & 0_{n\times n}\end{pmatrix}}_{-B_{\mathcal{E}}^\top}\cdots\underbrace{\begin{pmatrix}-I_n & 0_{n\times n}\end{pmatrix}}_{-B_{\mathcal{E}}^\top}\right).$$

We define the following regularizer

$$r(\mathbf{x},\mathbf{y}) = \frac{2\|d\|_\infty}{m}\left(10\sum_{i=1}^m\langle x_i, \log x_i\rangle + 5m\langle p, \log p\rangle + \hat{x}^\top\hat{A}^\top(\mathbf{y})^2 - p^\top\mathcal{E}^\top(\mathbf{y})^2\right), \tag{2}$$

where $\log x$ and $(x)^2$ are entry-wise, and $\hat{x} = (x_1^\top, \ldots, x_m^\top)^\top$. Also we define $\bar{\mathbf{z}} = (\bar{\mathbf{x}}, \bar{\mathbf{y}})$ as a minimizer of r.

Also we recall the gradient operator for the problem (1):

$$G(\mathbf{x},\mathbf{y}) = \frac{1}{m}\begin{pmatrix}\boldsymbol{d} + 2\|d\|_\infty\boldsymbol{A}^\top\mathbf{y} \\ 2\|d\|_\infty(\boldsymbol{c} - \boldsymbol{A}\mathbf{x})\end{pmatrix}. \tag{3}$$

1.2 Proof of Theorem 4.3

Theorem (Theorem 4.3)**.** *r is 3-area-convex with respect to the gradient operator G.*

Proof. Firstly, we define some notation connected to block-diagonal matrices. Assume that D is a block diagonal matrix of size $ak \times bk$

$$D = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_k \end{pmatrix},$$

where matrices B_i of size $a \times b$. We refer to i-th block of D as $D_{(i)} = B_i$. Also we define $D_{[i]}$ as a matrix D with all blocks zeroes except the i-th one. Equivalent, we can write $D_{[i]} = \delta_{ii}^{(k)} \otimes D_{(i)}$, where $\delta_{ij}^{(k)}$ is a matrix of size $k \times k$ with 1 on the position i, j position and 0 in any other, and $\otimes$ is a Kronecker product of matrices.

We will use a second-order criteria proposed by Jambulapati et al. (2019). We will show that

$$\begin{pmatrix} \nabla^2 r(\mathbf{z}) & -J \\ J & \nabla^2 r(\mathbf{z}) \end{pmatrix} \succeq 0,$$

where

$$J = \frac{2\|d\|_\infty}{m} \begin{pmatrix} 0 & \mathbf{A}^T \\ -\mathbf{A} & 0 \end{pmatrix} = \frac{2\|d\|_\infty}{n} \begin{pmatrix} 0 & 0 & \hat{A}^\top \\ 0 & 0 & \mathcal{E}^\top \\ -\hat{A} & -\mathcal{E} & 0 \end{pmatrix}$$

is the Jacobian matrix for $F(\mathbf{x}, \mathbf{y})$.

A good idea to remove a positive multiplicative term $2\|d\|_\infty m^{-1}$ to simplify the statement. Define $r'(\mathbf{z}) = 1/(2\|d\|_\infty m^{-1}) r(\mathbf{z})$ and $J' = 1/(2\|d\|_\infty m^{-1}) J$. Hence we only should show that

$$P = \begin{pmatrix} \nabla^2 r'(\mathbf{z}) & -J' \\ J' & \nabla^2 r'(\mathbf{z}) \end{pmatrix} = \frac{m}{2\|d\|_\infty} \begin{pmatrix} \nabla^2 r(\mathbf{z}) & -J \\ J & \nabla^2 r(\mathbf{z}) \end{pmatrix} \succeq 0.$$

Then we can rewrite r' in the following manner

$$\begin{aligned} r'(\mathbf{x}, \mathbf{y}) &= \sum_{i=1}^m \left[10\langle x_i, \log x_i \rangle + \langle Ax_i, (y_i)^2 \rangle \right] + 5m\langle p, \log p \rangle - p^T \mathcal{E}^T (\mathbf{y}^2) = \\ &= \sum_{i=1}^m \left[10\langle x_i, \log x_i \rangle + \langle Ax_i, (y_i)^2 \rangle \right] + \sum_{i=1}^m \left[5\langle p, \log p \rangle + \langle B_{\mathcal{E}} p, (y_i)^2 \rangle \right]. \end{aligned}$$

In this case, we can easily calculate the hessian of r', divide it into blocks:

$$\begin{aligned} \nabla^2 r'(\mathbf{z}) &= \begin{pmatrix} \nabla^2_{\hat{x},\hat{x}} r'(\mathbf{z}) & \nabla^2_{\hat{x},p} r'(\mathbf{z}) & \nabla^2_{\hat{x},\mathbf{y}} r'(\mathbf{z}) \\ \nabla^2_{p,\hat{x}} r'(\mathbf{z}) & \nabla^2_{p,p} r'(\mathbf{z}) & \nabla^2_{p,\mathbf{y}} r'(\mathbf{z}) \\ \nabla^2_{\mathbf{y},\hat{x}} r'(\mathbf{z}) & \nabla^2_{\mathbf{y},p} r'(\mathbf{z}) & \nabla^2_{\mathbf{y},\mathbf{y}} r'(\mathbf{z}) \end{pmatrix} \\ &= \begin{pmatrix} 10\operatorname{diag}((\hat{x})^{-1}) & 0_{mn^2 \times n} & 2\hat{A}^\top \operatorname{diag}(\mathbf{y}) \\ 0_{n \times mn^2} & 5m \operatorname{diag}((p)^{-1}) & -2\mathcal{E}^\top \operatorname{diag}(\mathbf{y}) \\ 2\operatorname{diag}(\mathbf{y})\hat{A} & -2\operatorname{diag}(y)\mathcal{E} & 2\operatorname{diag}(\hat{A}\hat{x}) - 2\operatorname{diag}(\mathcal{E}p) \end{pmatrix}, \end{aligned}$$

where $\operatorname{diag}(v)$ for a vector $v \in \mathbb{R}^n$ produces a diagonal matrix with v on diagonal and v^{-1} is a entry-wise operation on vector.

We notice that matrices $\operatorname{diag}((\hat{x})^{-1}), \hat{A}^\top \operatorname{diag}(\mathbf{y}), \operatorname{diag}(\hat{A}\hat{x})$ have a block-diagonal structure with m blocks. Define the following matrices

$$B_i(\mathbf{z}) = \begin{pmatrix} 10\operatorname{diag}((\hat{x})^{-1})_{[i]} & 0_{mn^2 \times n} & 2(\hat{A}^\top \operatorname{diag}(\mathbf{y}))_{[i]} \\ 0_{n \times mn^2} & 0_{n \times n} & 0_{n \times 2mn} \\ 2(\operatorname{diag}(\mathbf{y})\hat{A})_{[i]} & 0_{2mn \times n} & 2\operatorname{diag}(\hat{A}\hat{x})_{[i]} \end{pmatrix}$$

and

$$R(\mathbf{z}) = \begin{pmatrix} 0_{mn^2 \times mn^2} & 0_{mn^2 \times n} & 0_{mn^2 \times 2mn} \\ 0_{n \times mn^2} & 5m \operatorname{diag}((p)^{-1}) & -2\mathcal{E}^\top \operatorname{diag}(\mathbf{y}) \\ 0_{2mn \times mn^2} & -2\operatorname{diag}(y)\mathcal{E} & -2\operatorname{diag}(\mathcal{E}p) \end{pmatrix}.$$

Using these matrices, the decomposition of Hessian can be observed: $\nabla^2 r'(\mathbf{z}) = \sum_{i=1}^m B_i(\mathbf{z}) + R(\mathbf{z})$.

We notice that the matrix J' has the same block decomposition:

$$C_i = \begin{pmatrix} 0 & 0 & (\hat{A}^\top)_{[i]} \\ 0 & 0 & 0 \\ -(\hat{A})_{[i]} & 0 & 0 \end{pmatrix}, \quad S = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \mathcal{E}^\top \\ 0 & -\mathcal{E} & 0 \end{pmatrix}.$$

Clearly we have $J' = \sum_{i=1}^m C_i + S$. Using these two decompositions, we get the following:

$$P = \sum_{i=1}^m \underbrace{\begin{pmatrix} B_i(\mathbf{z}) & -C_i \\ C_i & B_i(\mathbf{z}) \end{pmatrix}}_{P_i} + \begin{pmatrix} R(\mathbf{z}) & -S \\ S & R(\mathbf{z}) \end{pmatrix}.$$

It can be observed that each matrix P_i is almost a corresponding matrix for the area-convex regularizer for the optimal transportation problem with variables x_i, y_i in (Jambulapati et al., 2019), except the rows and columns of zeros. Moreover, it was proven that these matrices are positive semi-definite. Hence, only the remaining term is need to be examined.

Firstly, we write the action of non-zero corner of $R(\mathbf{z})$, called $\hat{R}(\mathbf{z})$, as a quadratic form:

$$Q_{\hat{R}(\mathbf{z})}(u, v) = (u\top, v\top)\hat{R}(\mathbf{z}) \begin{pmatrix} u \\ v \end{pmatrix} = (u\top, v\top) \begin{pmatrix} 5m \operatorname{diag}((p)^{-1}) & -2\mathcal{E}^\top \operatorname{diag}(\mathbf{y}) \\ -2 \operatorname{diag}(y)\mathcal{E} & -2 \operatorname{diag}(\mathcal{E}p) \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}.$$

The we can use the trick induced by the structure of the matrix $\mathcal{E}$ to compute the quadratic form. The trick is about to rewrite m in the following way: $m = \|\mathcal{E}_{:,j}\|_1 = -\sum_{i=1}^{2mn} \mathcal{E}_{ij}, \forall j \in [n]$.

Then, we can calculate the quadratic form:

$$Q_{\hat{R}(\mathbf{z})}(u, v) = \sum_{i,j} (-\mathcal{E}_{ij}) \left(\frac{5u_j^2}{p_j} + 4u_j v_i y_i + 2v_i^2 p_j \right).$$

Secondly, we wrtie the action of non-zero corner of S, called $\hat{S}$, as a bilinear form

$$B_{\hat{S}}((a, b), (u, v)) = (x\top, y\top) \begin{pmatrix} 0 & \mathcal{E}^\top \\ -\mathcal{E} & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \sum_{i,j} \mathcal{E}_{ij} \left(a_j v_i - u_j b_i \right),$$

and, as a result, we have the complete analytic expression for the quadratic form induced by the remaining term of P:

$$\begin{aligned}
&((a^\top, b^\top), (u^\top, v^\top)) \begin{pmatrix} \hat{R}(\mathbf{z}) & -\hat{S} \\ \hat{S} & \hat{R}(\mathbf{z}) \end{pmatrix} \begin{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \\ \begin{pmatrix} u \\ v \end{pmatrix} \end{pmatrix} \\
&\qquad = \sum_{i,j} (-\mathcal{E}_{ij}) \left(\frac{5a_j^2}{p_j} + 4a_j b_i y_i + 2b_i^2 p_j + 2a_j v_i - 2u_j b_i + \frac{5u_j^2}{p_j} + 4u_j v_i y_i + 2v_i^2 p_j \right) \\
&\qquad = \sum_{i,j} (-\mathcal{E}_{ij}) \frac{1}{p_j} \Big((2a_j y_i + b_i p_j)^2 + (2u_j y_i + v_i p_j)^2 \\
&\qquad + (a_j + v_i p_j)^2 + (u_j + b_i p_j)^2 + (1 - (y_i)^2)(a_j^2 + u_j^2)) \Big) \geq 0.
\end{aligned}$$

The final inequality follows from the range of $y_i \in [-1, 1]$ and finishes the proof. □

1.3 Proof of Theorem 4.4

Theorem (Theorem 4.4). *Let at each iteration, Dual Extrapolation algorithm calls Alternating minimization (AM) scheme to make the proximal steps. Then for $N = \lceil \frac{4\kappa\Theta}{\varepsilon} \rceil$ iterations of Dual Extrapolation algorithm running with regularizer (2) and $\kappa = 3$, AM scheme accumulates additive error $\varepsilon/2$ running with*

$$M = 24 \log \left(\left(\frac{88\|d\|_\infty}{\varepsilon^2} + \frac{4}{\varepsilon} \right) \Theta + \frac{36\|d\|_\infty}{\varepsilon} \right)$$

iterations in $O(mn^2 \log \gamma)$ time, where $\gamma = \varepsilon^{-1}\|d\|_\infty \log n$.

The target function for this procedure can be written in the general form as following:

$$H(\mathbf{x}, \mathbf{y}) = \langle \boldsymbol{v}, \mathbf{x} \rangle + \langle \boldsymbol{u}, \mathbf{y} \rangle + r(\mathbf{x}, \mathbf{y}). \tag{4}$$

To prove this theorem we will use the results from (Jambulapati et al., 2019) about their Alternating minimization scheme. Firstly, we need to obtain a linear convergence and we can do it by adapting an argument of Jambulapati et al. (2019, Lemma 6) to our setup.

Lemma 1.1. *For some $\mathbf{x}^{k+1}, \mathbf{y}_k$, let $\mathcal{X}_{k+1} = \{\mathbf{x} \mid \mathbf{x} \geq \frac{1}{2}\mathbf{x}^{k+1}\}$ where inequality is entrywise, and let $\mathcal{Y}_k$ be the entire domain of $\mathbf{y}$ (i.e. $\mathcal{Y}$). Then for any $\mathbf{x}' \in \mathcal{X}_{k+1}, \mathbf{y}', \mathbf{y}'' \in \mathcal{Y}_k$,*

$$\nabla^2 r(\mathbf{x}', \mathbf{y}') \succeq \frac{1}{12} \nabla^2_{\mathbf{yy}} r(\mathbf{x}^{k+1}, \mathbf{y}'').$$

Proof. The only thing that differs in the analysis is a diagonal approximation then does not depends on $\mathbf{y}$. Hence, we only need to show that for any $\mathbf{y}$

$$D(\mathbf{x}) \preceq \nabla^2 r(\mathbf{x}, \mathbf{y}) \preceq 6D(\mathbf{x}),$$

where $D(\mathbf{x})$ is the diagonal approximation

$$D(\mathbf{x}) = \begin{pmatrix} 2\operatorname{diag}((\hat{x})^{-1}) & 0_{mn^2 \times n} & 0_{mn^2 \times 2mn} \\ 0_{n \times mn^2} & m\operatorname{diag}((p)^{-1}) & 0_{n \times 2mn} \\ 0_{2mn \times mn^2} & 0_{2mn \times n} & \operatorname{diag}(\hat{A}\hat{x}) - \operatorname{diag}(\mathcal{E}p) \end{pmatrix}.$$

It is easy to see that $D(\mathbf{x})$ has the same block structure as $\nabla^2 r(\mathbf{x}, \mathbf{y})$ and we can prove our inequalities for each block separately. But all blocks connected to $\hat{x}$ is blocks that appears in optimal transport problem and the required inequalities were proven in (Jambulapati et al., 2019). Hence, we only need to show that

$$\hat{D}_p(\mathbf{x}) \preceq \hat{R}(\mathbf{x}, \mathbf{y}) \preceq 6\hat{D}_p(\mathbf{x}),$$

where

$$\hat{D}_p(\mathbf{x}) = \begin{pmatrix} m\operatorname{diag}((p)^{-1}) & 0_{n \times 2mn} \\ 0_{2mn \times n} & -\operatorname{diag}(\mathcal{E}p) \end{pmatrix}.$$

and $\hat{R}$ was defined in the proof of Theorem 1.2.

Also, in the proof of Theorem 1.2 we show that

$$Q_{\hat{R}(\mathbf{z})}(u, v) = \sum_{i,j} (-\mathcal{E}_{ij}) \left(\frac{5u_j^2}{p_j} + 4u_j v_i y_i + 2v_i^2 p_j \right).$$

Using the same idea, we can write the action of quadratic form induced by $\hat{D}_p$:

$$Q_{\hat{D}_p(\mathbf{x})}(u, v) = \sum_{i,j} (-\mathcal{E}_{ij}) \left(\frac{u_j^2}{p_j} + v_i^2 p_j \right).$$

Using the fact that $y_i \in [-1, 1]$, we can obtain the required by the following inequalities and finish the proof:

$$\frac{u_j^2}{p_j} + v_i^2 p_j \leq \frac{5u_j^2}{p_j} + 4u_j v_i y_i + 2v_i^2 p_j \leq \frac{6u_j^2}{p_j} + 6v_i^2 p_j.$$

□

By the exactly same arguments, we obtain the linear rate of converge for our Alternating Minimization (AM) scheme. We need to show last two points

- Bound the complexity of each iteration
- Bound the initial range

Lemma 1.2. *For $H(\mathbf{x}, \mathbf{y})$, defined in (4), we can implement the steps*

1. $\mathbf{x}^{k+1} \triangleq \arg\min_{\mathbf{x}\in\mathcal{X}} H(\mathbf{x}, \mathbf{y}^k)$,
2. $\mathbf{y}^{k+1} \triangleq \arg\min_{\mathbf{y}\in\mathcal{Y}} H(\mathbf{x}^{k+1}, \mathbf{y})$,

in time $O(mn^2)$.

Proof. First of all, divide a vector $\boldsymbol{v}$ from the definition of function (4) into $m+1$ part and vector $\boldsymbol{u}$ into m parts. We have the following function to optimize by some regrouping and rewriting a regularizer in homogeneous manner

$$H(\mathbf{x}, \mathbf{y}) = \frac{2\|d\|_\infty}{m} \sum_{i=1}^{m} \Bigg(\frac{m}{2\|d\|_\infty} \langle v_i, x_i \rangle + \langle (y_i)^2, Ax_i \rangle + 10 \langle x_i, \log x_i \rangle + \frac{m}{2\|d\|_\infty} \langle u_i, y_i \rangle + \langle B_\varepsilon p, (y_i)^2 \rangle \Bigg) + 10\|d\|_\infty \langle p, \log p \rangle + \langle v_{m+1}, p \rangle.$$

We notice that each x_i is independent from others and we can compute $x_i^{(k+1)}$ apart as a solutions of the following optimization problems:

$$x_i^{k+1} = \arg\min_{x\in\Delta^{n^2}} \left\langle \underbrace{\frac{m}{20\|d\|_\infty} v_i + \frac{1}{10} A^\top (y_i^k)^2}_{\gamma_i}, x \right\rangle + \langle x, \log x \rangle,$$

and the solution of this type of problems is well-known and proportional to $\exp(-\gamma_i)$. The multiplication on the matrix A and $A^\top$ can be computed in $O(n^2)$ time, because these matrices consists of $O(n^2)$ non-zero entries, and all these steps can be performed in $O(mn^2)$.

Also we need to compute an optimal p by the same idea

$$p^{k+1} = \arg\min_{p\in\Delta^n} \left\langle \underbrace{\frac{1}{10\|d\|_\infty} v_{m+1} - \frac{1}{5m} \mathcal{E}^\top (\mathbf{y}^k)^2}_{\gamma_{m+1}}, p \right\rangle + \langle p, \log p \rangle.$$

As in the previous case, an optimal p^{k+1} is proportional to $\exp(-\gamma_{m+1})$ and it can be computed in $O(mn^2)$ time.

For the computation of $\mathbf{y}^{(k+1)}$ we notice that each $[y_i^{(k+1)}]_j$ can be computed separately as a solution of the following 1-D optimization problem:

$$[y_i^{k+1}]_j = \arg\min_{y\in[-1,1]} \frac{m}{2\|d\|_\infty} [u_i]_j \cdot y + ([Ax_i^{k+1}]_j + [B_\varepsilon p^{k+1}]_j) \cdot y^2.$$

It could be easily solved in constant time if we know Ax_i^{k+1} and $B_\varepsilon p^{k+1} = (p^\top, 0_n)^\top$

$$[y_i^{k+1}]_j = \begin{cases} -1, & \alpha \le -1 \\ 1, & \alpha \ge 1 \\ \alpha, & \alpha \in [-1, 1] \end{cases}, \quad \text{where } \alpha = \frac{-m[u_i]_j}{4\|d\|_\infty ([Ax_i]_j + [B_\varepsilon p]_j)}.$$

Hence, we can make all calculations in $O(mn^2)$. □

Now we are ready to write the final proof.

Proof of Theorem 4.4. To proof the final result, we need to remind the proximal operator for r:

$$\text{prox}_{\mathbf{z}}^r(v) = \arg\min_{\mathbf{z}\in\mathcal{Z}} \langle v, \mathbf{z} \rangle + B_r(\bar{z}, \mathbf{z}) = \arg\min_{\mathbf{z}\in\mathcal{Z}} \langle v - \nabla r(\bar{\mathbf{z}}), \mathbf{z} \rangle + r(\mathbf{z}).$$

We notice, that it is equivalent to the next view, separate over $\mathbf{x}$ and $\mathbf{y}$:

$$\text{prox}^r_{\bar{\mathbf{x}},\bar{\mathbf{y}}}(v) = \arg\min_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \langle v_x - \nabla_{\mathbf{x}} r(\bar{\mathbf{x}},\bar{\mathbf{y}}), \mathbf{x}\rangle + \langle v_y - \nabla_{\mathbf{y}} r(\bar{\mathbf{x}},\bar{\mathbf{y}}), \mathbf{y}\rangle + r(\mathbf{x},\mathbf{y}). \tag{5}$$

We have precisely the type of problems that can be solved using AM scheme described above in linear time, moreover, each step reduces error by 1/24 factor (similar as (Jambulapati et al., 2019)).

The only thing we need to bound is an initial error. For this goal we should bound the norm of the gradient and the argument of the proximal function in all calls during the algorithm.

Firstly, divide gradient operator $G(\mathbf{z}) = (G_{\mathbf{x}}(\mathbf{z})^\top, G_{\mathbf{y}}(\mathbf{z})^\top)^\top$, defined in (3), into two parts and bound uniformly ℓ_∞ and ℓ_1 norms of each part respectively

$$\begin{aligned}
\|G_{\mathbf{x}}(\mathbf{z})\|_\infty &= \frac{1}{m}\|\boldsymbol{d} + 2\|d\|_\infty \boldsymbol{A}^\top \mathbf{y}\|_\infty \le \frac{\|d\|_\infty}{m} + \frac{2\|d\|_\infty}{m}\|\boldsymbol{A}^\top \mathbf{y}\|_\infty \le 3\|d\|_\infty,\\
\|G_{\mathbf{x}}(\mathbf{z})\|_1 &= \frac{1}{m}\|2\|d\|_\infty(\boldsymbol{c} - \boldsymbol{A}\mathbf{x})\|_1 \le \frac{2\|d\|_\infty}{m}(\|\boldsymbol{c}\|_1 + \|\boldsymbol{A}\mathbf{x}\|_1) \le 8\|d\|_\infty.
\end{aligned}$$

In the inequality in the first row we used the fact $m \ge 1$ for simplicity and in the second one we use the fact that matrix A and vector x_i are non-negative, hence, $\|Ax_i\|_1 = \langle \mathbf{1}_n, Ax_i\rangle = 2\langle \mathbf{1}_n, x_i\rangle = 2$, where $\mathbf{1}_n$ is a vector consists of ones.

Then we can use the fact that the argument of the first prox-operator $\mathbf{s}^k = (\mathbf{s}^k_{\mathbf{x}}, \mathbf{s}^k_{\mathbf{y}})$ is a sum of k gradients multiplied by $1/2\kappa$, computed in different points. In the second operator we also add gradient operator, multiplied by $1/\kappa$. Since $k \le 4\kappa\Theta\cdot\varepsilon^{-1}$, we have by triangle inequality

$$\begin{aligned}
\|\mathbf{s}^k_{\mathbf{x}}\|_\infty &\le \frac{k}{2\kappa}\cdot 3\|d\|_\infty \le \frac{6\Theta\|d\|_\infty}{\varepsilon},\\
\|\mathbf{s}^k_{\mathbf{y}}\|_1 &\le \frac{k}{2\kappa} 8\|d\|_\infty \le \frac{16\Theta\|d\|_\infty}{\varepsilon}.
\end{aligned}$$

Then, all our arguments of the proximal operator during the running time can be bounded in the following way (for $\kappa = 3$)

$$\begin{aligned}
\|v_{\mathbf{x}}\|_\infty &\le \frac{6\Theta\|d\|_\infty}{\varepsilon} + \|d\|_\infty,\\
\|v_{\mathbf{y}}\|_1 &\le \frac{16\Theta\|d\|_\infty}{\varepsilon} + \frac{8}{3}\|d\|_\infty.
\end{aligned}$$

Then fix $\mathbf{x}^*$ and $\mathbf{y}^*$ as minimizers for the proximal operator (5) and remind the bound for $\Theta \le 40\log n\|d\|_\infty + 6\|d\|_\infty$. Also we can compute $\|\nabla_{\mathbf{x}} r(\bar{\mathbf{x}},\bar{\mathbf{y}})\|_\infty \le 20\|d\|_\infty(2\log n + 1)$ and $\|\nabla_{\mathbf{y}} r(\bar{x},\bar{y})\|_1 = 0$.

Then we can write a suboptimality gap δ_0 for our algorithm for any initial $\mathbf{x}^0$ and $\mathbf{y}^0$:

$$\begin{aligned}
\delta_0 &= \langle v_{\mathbf{x}} - \nabla_{\mathbf{x}} r(\bar{\mathbf{x}},\bar{\mathbf{y}}), \mathbf{x}^0 - \mathbf{x}^*\rangle + \langle v_{\mathbf{y}} - \nabla_{\mathbf{y}} r(\bar{\mathbf{x}},\bar{\mathbf{y}}), \mathbf{y}^0 - \mathbf{y}^*\rangle + r(\mathbf{x}^0,\mathbf{y}^0) - r(\mathbf{x}^*,\mathbf{y}^*)\\
&\le \|v_{\mathbf{x}} - \nabla_{\mathbf{x}} r(\bar{\mathbf{x}},\bar{\mathbf{y}})\|_\infty \|\mathbf{x}^0 - \mathbf{x}^*\|_1 + \|v_{\mathbf{y}} - \nabla_{\mathbf{y}} r(\bar{\mathbf{x}},\bar{\mathbf{y}})\|_1 \|\mathbf{y}^0 - \mathbf{y}^*\|_\infty + \Theta\\
&\le 2\|d\|_\infty \cdot \left(\frac{6\Theta}{\varepsilon} + 20\log n + 10\right) + \|d\|_\infty + 2\cdot\frac{16\Theta\|d\|_\infty}{\varepsilon} + \frac{8}{3}\|d\|_\infty + \Theta\\
&\le \left(\frac{44\|d\|_\infty}{\varepsilon} + 2\right)\Theta + 18\|d\|_\infty.
\end{aligned}$$

Then we can compute the total number of iterations to obtain $\varepsilon/2$ desired accuracy:

$$N = \log_{24/23}\frac{2\delta_0}{\varepsilon} \le 24\log\left(\left(\frac{88\|d\|_\infty}{\varepsilon^2} + \frac{4}{\varepsilon}\right)\Theta + \frac{36\|d\|_\infty}{\varepsilon}\right) = O(\log\gamma),$$

where $\gamma = \|d\|\varepsilon^{-1}\log n$, as desired. Each iteration can be done in $O(mn^2)$ time and we obtain the required complexity. □

References

Jambulapati, A., Sidford, A., and Tian, K. (2019). A direct $\tilde{O}(1/\varepsilon)$ iteration parallel algorithm for optimal transport. In *Advances in Neural Information Processing Systems*, pages 11359–11370.