Supplementary for Accelerated Alternating Minimization, Accelerated Sinkhorn's Algorithm and Accelerated Iterative Bregman Projections

A. Omitted proofs in Section 2: Accelerated Alternating Minimization

Proof of Lemma 2. Let us introduce an auxiliary sequence of functions defined as

$$\psi_0(x) = \frac{1}{2}\|x - x^0\|^2, \quad \psi_{k+1}(x) = \psi_k(x) + a_{k+1}\{f(y^k) + \langle \nabla f(y^k), x - y^k\rangle\}.$$

It is easy to see that $v^k = \operatorname{argmin}_{x\in\mathbb{R}^N} \psi_k(x)$.

Now, we prove inequality (1) by induction over k. For $k = 0$, the inequality holds. Assume that

$$A_k f(x^k) \leqslant \min_{x\in\mathbb{R}^N} \psi_k(x) = \psi_k(v^k).$$

Then

$$\begin{aligned}
\psi_{k+1}(v^{k+1}) &= \min_{x\in\mathbb{R}^N} \left\{\psi_k(x) + a_{k+1}\{f(y^k) + \langle \nabla f(y^k), x - y^k\rangle\}\right\} \geqslant \\
&\geqslant \min_{x\in\mathbb{R}^N} \left\{\psi_k(v^k) + \frac{1}{2}\|x - v^k\|_2^2 + a_{k+1}\{f(y^k) + \langle \nabla f(y^k), x - y^k\rangle\}\right\} \geqslant \\
&\geqslant \psi_k(v^k) + a_{k+1} f(y^k) - \frac{a_{k+1}^2}{2}\|\nabla f(y^k)\|_2^2 + a_{k+1}\langle \nabla f(y^k), v^k - y^k\rangle \geqslant \\
&\geqslant A_k f(x^k) + a_{k+1} f(y^k) - \frac{a_{k+1}^2}{2}\|\nabla f(y^k)\|_2^2 + a_{k+1}\langle \nabla f(y^k), v^k - y^k\rangle \\
&\geqslant A_{k+1} f(y^k) - \frac{a_{k+1}^2}{2}\|\nabla f(y^k)\|_2^2 + a_{k+1}\langle \nabla f(y^k), v^k - y^k\rangle.
\end{aligned}$$

Here we used that ψ_k is a strongly convex function with minimum at v^k and that $f(y^k) \leqslant f(x^k)$. By the optimality conditions for the problem $\min\limits_{\beta\in[0,1]} f\left(x^k + \beta(v^k - x^k)\right)$, there are three possibilities

(1) $\beta_k = 1$, $\langle \nabla f(y^k), x^k - v^k\rangle \geqslant 0$, $y^k = v^k$;
(2) $\beta_k \in (0, 1)$ and $\langle \nabla f(y^k), x^k - v^k\rangle = 0$, $y^k = v^k + \beta_k(x^k - v^k)$;
(3) $\beta_k = 0$ and $\langle \nabla f(y^k), x^k - v^k\rangle \leqslant 0$, $y^k = x^k$.

In all three cases, $\langle \nabla f(y^k), v^k - y^k\rangle \geqslant 0$.

Using the rule for choosing a_{k+1} in the method, we finish the proof of the induction

step:

$$\psi_{k+1}(v^{k+1}) \geqslant A_{k+1} f(x^{k+1}).$$

It remains to show that the equation

$$f(y^k) - \frac{a_{k+1}^2}{2A_{k+1}} \|\nabla f(y^k)\|_2^2 = f(x^{k+1}). \tag{1}$$

has a solution $a_{k+1} > 0$. By the L-smoothness of the objective, we have, for all $i \geqslant 0$,

$$f(y^k) - \frac{1}{2L} \|\nabla_i f(y^k)\|_2^2 \geqslant f(x_i^{k+1}),$$

where $x_i^{k+1} = \operatorname{argmin}_{x \in S_i} f(x)$. Since $A_{k+1} = A_k + a_{k+1}$, we can rewrite (1) as

$$\frac{a_{k+1}^2}{2} \|\nabla f(y^k)\|_2^2 + a_{k+1}(f(x^{k+1}) - f(y^k)) + A_k(f(x^{k+1}) - f(y^k)) = 0.$$

Since $f(x^{k+1}) - f(y^k) < 0$ (otherwise $\|\nabla f(y^k)\| = 0$ and y_k is a solution to the problem), there exists solution $a_{k+1} > 0$.

Let us estimate the rate of the growth for A_k. Since $i_k = \operatorname{argmax}_i \|\nabla_i f(y^k)\|_2^2$,

$$\|\nabla_{i_k} f(y^k)\|_2^2 \geqslant \frac{1}{n} \|\nabla f(y^k)\|_2^2.$$

As a consequence, we have

$$f(y^k) - \frac{1}{2Ln} \|\nabla f(y^k)\|_2^2 \geqslant f(y^k) - \frac{1}{2L} \|\nabla_{i_k} f(y^k)\|_2^2 \geqslant f(x^{k+1}).$$

This in combination with our rule for choosing a_{k+1} implies $\frac{a_{k+1}^2}{2A_{k+1}} \geqslant \frac{1}{2Ln}$. Since $A_1 = a_1 \geqslant \frac{1}{Ln}$, we prove by induction that $a_k \geqslant \frac{k}{2Ln}$ and $A_k \geqslant \frac{(k+1)^2}{4nL} \geqslant \frac{k^2}{4nL}$. Indeed,

$$\begin{aligned} a_{k+1} \geqslant \frac{1 + \sqrt{1 + 4A_k Ln}}{2Ln} &= \frac{1}{2Ln} + \sqrt{\frac{1}{4L^2n^2} + \frac{A_k}{Ln}} \\ &\geqslant \frac{1}{2Ln} + \sqrt{\frac{A_k}{Ln}} \geqslant \frac{1}{2Ln} + \frac{1}{\sqrt{L}} \frac{k+1}{2\sqrt{Ln}} = \frac{k+2}{2Ln}. \end{aligned}$$

Hence,

$$A_{k+1} = A_k + a_{k+1} \geqslant \frac{(k+1)^2}{4Ln} + \frac{k+2}{2Ln} \geqslant \frac{(k+2)^2}{4Ln}.$$

□

B. Omitted proofs in Section 3: Primal-Dual Extension

To prove Theorem 3, we first prove a slightly more general result.

Theorem B.1. *Let the objective ϕ in the problem (P_2) be L-smooth w.r.t. $\|\cdot\|_2$ and the solution of this problem be bounded, i.e. $\|\lambda^*\|_2 \leqslant R$. Then, for the sequences $\hat{x}_{k+1}, \eta_{k+1}$, $k \geqslant 0$, generated by Algorithm 2,*

$$\|\boldsymbol{A}\hat{x}^k - b\|_2 \leqslant \frac{8nLR}{k^2},\ |\phi(\eta^k) + f(\hat{x}^k)| \leqslant \frac{8nLR^2}{k^2},\ \|\hat{x}^k - x^*\|_E \leqslant \frac{4}{k}\sqrt{\frac{2nLR^2}{\gamma}}.$$

Proof. Applying Lemma 2 to problem (P_2), we obtain

$$A_k\phi(\eta^k) \leqslant \min_{\lambda\in\Lambda}\left\{\sum_{j=0}^{k-1}\{a_{j+1}(\phi(\lambda^j) + \langle\nabla\phi(\lambda^j), \lambda - \lambda^j\rangle) + \frac{1}{2}\|\lambda\|_2^2\right\}, \tag{2}$$

Let us introduce the set $\Lambda_R = \{\lambda : \|\lambda\|_2 \leqslant 2R\}$ where R is such that $\|\lambda^*\|_2 \leqslant R$. Then, from (2), we obtain for $h(\lambda) = \sum_{j=0}^{k-1} a_{j+1}\left(\phi(\lambda^j) + \langle\nabla\phi(\lambda^j), \lambda - \lambda^j\rangle\right) + \frac{1}{2}\|\lambda\|_2^2$

$$A_k\phi(\eta^k) \leqslant \min_{\lambda\in\Lambda} h(\lambda) \leqslant \min_{\lambda\in\Lambda_R} h(\lambda) \leqslant 2R^2 + \min_{\lambda\in\Lambda_R}\left\{\sum_{j=0}^{k-1} a_{j+1}(\phi(\lambda^j) + \langle\nabla\phi(\lambda^j), \lambda - \lambda^j\rangle\right\}. \tag{3}$$

On the other hand, from the definition (P_2) of $\phi(\lambda)$, we have

$$\phi(\lambda^i) = \langle\lambda^i, b\rangle + \max_{x\in Q}\left(-f(x) - \langle\boldsymbol{A}^T\lambda^i, x\rangle\right) = \langle\lambda^i, b\rangle - f(x(\lambda^i)) - \langle\boldsymbol{A}^T\lambda^i, x(\lambda^i)\rangle.$$

Combining this equality with (2), we obtain

$$\begin{aligned}&\phi(\lambda^i) - \langle\nabla\phi(\lambda^i), \lambda^i\rangle\\ &\qquad = \langle\lambda^i, b\rangle - f(x(\lambda^i)) - \langle\boldsymbol{A}^T\lambda^i, x(\lambda^i)\rangle - \langle b - \boldsymbol{A}x(\lambda^i), \lambda^i\rangle = -f(x(\lambda^i)).\end{aligned}$$

Summing these equalities from $i = 0$ to $i = k-1$ with the weights $\{a_{i+1}\}_{i=0,\dots k-1}$, we get, using the convexity of f

$$\begin{aligned}&\sum_{i=0}^{k-1} a_{i+1}(\phi(\lambda^i) + \langle\nabla\phi(\lambda^i), \lambda - \lambda^i\rangle) =\\ &\quad = -\sum_{i=0}^{k-1} a_{i+1}f(x(\lambda^i)) + \sum_{i=0}^{k-1} a_{i+1}\langle(b - \boldsymbol{A}x(\lambda^i), \lambda\rangle \leqslant -A_kf(\hat{x}^k) + A_k\langle b - \boldsymbol{A}\hat{x}^k, \lambda\rangle.\end{aligned}$$

Substituting this inequality into (3), we obtain

$$A_k\phi(\eta^k) \leqslant -A_kf(\hat{x}^k) + \min_{\lambda\in\Lambda_R}\left\{A_k\langle b - \boldsymbol{A}\hat{x}^k, \lambda\rangle\right\} + 2R^2$$

Finally, since $\max_{\lambda\in\Lambda_R}\left\{\langle -b + \boldsymbol{A}\hat{x}^k, \lambda\rangle\right\} = 2R\|\boldsymbol{A}\hat{x}^k - b\|_2$, we obtain

$$A_k(\phi(\eta^k) + f(\hat{x}^k)) + 2RA_k\|\boldsymbol{A}\hat{x}^k - b\|_2 \leqslant 2R^2. \tag{4}$$

Since λ^* is an optimal solution of Problem (D_1), we have, for any $x \in Q$

$$Opt[P_1] \leqslant f(x) + \langle \lambda^*, \boldsymbol{A}x - b \rangle.$$

Using the assumption that $\|\lambda^*\|_2 \leqslant R$, we get

$$f(\hat{x}^k) \geqslant Opt[P_1] - R\|\boldsymbol{A}\hat{x}^k - b\|_2. \tag{5}$$

Hence,

$$\begin{aligned}\phi(\eta^k) + f(\hat{x}^k) &= \phi(\eta^k) - Opt[P_2] + Opt[P_2] + Opt[P_1] - Opt[P_1] + f(\hat{x}^k) = \\ &= \phi(\eta^k) - Opt[P_2] - Opt[P_1] + f(\hat{x}^k) \geqslant -Opt[P_1] + f(\hat{x}^k) \overset{(5)}{\geqslant} -R\|\boldsymbol{A}\hat{x}^k - b\|_2. \end{aligned} \tag{6}$$

This and (4) give $R\|A_k(\boldsymbol{A}\hat{x}^k - b)\|_2 \leqslant 2R^2$. Hence, from (6) we obtain $A_k(\phi(\eta^k) + f(\hat{x}^k)) \geqslant -2R^2$. On the other hand, from (4) we have $A_k(\phi(\eta^k) + f(\hat{x}^k)) \leqslant 2R^2$. Combining all of these results, we conclude

$$A_k\|\boldsymbol{A}\hat{x}^k - b\|_2 \leqslant 2R, \quad A_k|\phi(\eta^k) + f(\hat{x}^k)| \leqslant 2R^2. \tag{7}$$

From 2, for any $k \geqslant 0$, $A_k \geqslant \frac{k^2}{4Ln}$. Combining this and (7), we obtain the first two inequalities of the statement:

$$\|\boldsymbol{A}\hat{x}^k - b\|_2 \leqslant \frac{8nLR}{k^2}, \ |\phi(\eta^k) + f(\hat{x}^k)| \leqslant \frac{8nLR^2}{k^2}.$$

It remains to prove the third inequality. By the optimality condition for Problem (P_1), we have

$$\langle \nabla f(x^*) + \boldsymbol{A}^T\lambda^*, \hat{x}_k - x^* \rangle \geqslant 0, \quad \boldsymbol{A}x^* = b.$$

Then

$$\begin{aligned}\langle \nabla f(x^*), \hat{x}_k - x^* \rangle &\geqslant -\langle \boldsymbol{A}^T\lambda^*, \hat{x}_k - x^* \rangle = -\langle \lambda^*, \boldsymbol{A}\hat{x}_k - b \rangle \\ &\geqslant -R\|\boldsymbol{A}\hat{x}_k - b\|_2 \geqslant -\frac{8nLR^2}{k^2}, \end{aligned} \tag{8}$$

where we used the same reasoning as while deriving (5). Using this inequality and the γ-strong convexity of f, we obtain

$$\begin{aligned}\frac{\gamma}{2}\|\hat{x}_k - x^*\|_E^2 &\leqslant f(\hat{x}_k) - Opt[P_1] - \langle \nabla f(x^*), \hat{x}_k - x^* \rangle \\ &\leqslant f(\hat{x}_k) + \phi(\eta^k) + \langle \nabla f(x^*), \hat{x}_k - x^* \rangle \leqslant \frac{8nLR^2}{k^2} + \frac{8nLR^2}{k^2} = \frac{16nLR^2}{k^2}, \end{aligned}$$

or

$$\|\hat{x}_k - x^*\|_E \leqslant \frac{4}{k}\sqrt{\frac{2nLR^2}{\gamma}}.$$

□

Proof of Theorem 3

Proof. The result follows from the previous theorem and the bound $L \leqslant \frac{\|\boldsymbol{A}\|^2_{E\to H}}{\gamma}$, which is shown in [7]. □

C. Fixed-Step Accelerated Alternating Minimization

In this section we introduce another variant of accelerated alternating minimization method. Algorithm 2 in the main text uses full relaxation on a segment to find the next iterate y^k. On the contrary, the method which we introduce in this section tries to adaptively find an approximation for the constant L – Lipschitz constant of the gradient. Based on this approximation, a fixed stepsize is used to find y^k. Thus, compared to the AAM algorithm presented in Section 2 of the main paper, this algorithm does not require solving any one-dimensional minimization problems during each iteration, but instead requires adapting to the smoothness parameter of the problem. This typically results in repeating each iteration twice. In our experience, which of the two method turns out to be more efficient significantly depends on the problem being solved (generally, the more difficult the function is to compute, the more taxing the line-search becomes) and the implementation of the line-search procedure. We also point out that we can not guarantee the convergence of Algorithm 1 to a stationary point for non-convex objectives. In the experiments for the OT problem we use this algorithm and the result is denoted by AAM-A.

Algorithm 1 Fixed-Step Accelerated Alternating Minimization

Input: starting point x_0, initial estimate of the Lipschitz constant L_0.
Output: x^k

$x^0 = y^0 = v^0$.
for $k \geqslant 0$ **do**
  Set $L_{k+1} = L_k/2$
  while True **do**
    Set $a_{k+1} = \frac{1}{2L_{k+1}} + \sqrt{\frac{1}{4L^2_{k+1}} + a_k^2 \frac{L_k}{L_{k+1}}}$ Find a_{k+1} s.t. $A_{k+1} := a^2_{k+1}L_{k+1} = a_k^2 L_k + a_{k+1}$.
    Set $\tau_k = \frac{1}{a_{k+1}L_{k+1}}$
    Set $y^k = \tau_k v^k + (1-\tau_k)x^k$ {Extrapolation step}
    Choose $i_k = \underset{i\in\{1,\dots,n\}}{\operatorname{argmax}} \|\nabla_i f(y^k)\|_2^2$
    Set $x^{k+1} = \underset{x\in S_{i_k}(y^k)}{\operatorname{argmin}} f(x)$
    Set $v^{k+1} = v^k - a_{k+1}\nabla f(y^k)$
    if $f(x^{k+1}) \leqslant f(y^k) - \frac{\|\nabla f(y^k)\|_2^2}{2L_{k+1}}$ **then**
      break
    end if
    Set $L_{k+1} = 2L_{k+1}$.
  end while
  $k = k+1$
end for

The convergence rate of Algorithm 1 is given by the following theorem

Theorem C.1. *Let the objective f be convex and L-smooth. If $L_0 \leqslant 4nL$, then after k steps of Algorithm 1 it holds that*

$$f(x^k) - f(x^*) \leqslant \frac{4nL\|x^0 - x^*\|_2^2}{k^2}. \tag{9}$$

Unlike the AM algorithm, this method requires computing the whole gradient of the objective, which makes the iterations of this algorithm considerably more expensive. Also, even when the number of blocks is 2, the convergence rate of Algorithm 1 depends on the smoothness parameter L of the whole objective, and not on the Lipschitz constants of each block on its own, which is the case for the AM algorithm [2]. On the other hand, if we compare the Algorithm 1 algorithm to an adaptive accelerated gradient method, we will see that the theoretical worst-case time complexity of Algorithm 1 method is only $\sqrt{n}$ times worse, while in practice block-wise minimization steps may perform much better than gradient descent steps simply because they directly use some specific structure of the objective.

This convergence rate is n times worse than that of an adaptive accelerated gradient method [4], or, equivalently, this means that in the worst case it may take $\sqrt{n}$ times more iterations to guarantee accuracy ε compared to an adaptive accelerated gradient method. To prove the convergence rate of the method, we will need a technical result.

Lemma C.2. *For any $u \in \mathbb{R}^N$*

$$a_{k+1}\langle \nabla f(y^k), v^k - u\rangle \leqslant a_{k+1}^2 L_{k+1}\left(f(y^k) - f(x^{k+1})\right) + \frac{1}{2}\|v^k - u\|_2^2 - \frac{1}{2}\|v^{k+1} - u\|_2^2.$$

Proof.

$$\begin{aligned}
a_{k+1}\langle \nabla f(y^k), v^k - u\rangle &= a_{k+1}\langle \nabla f(y^k), v^k - v^{k+1}\rangle + a_{k+1}\langle \nabla f(y^k), v^{k+1} - u\rangle \\
&= a_{k+1}^2\|\nabla f(y^k)\|_2^2 + \langle v^k - v^{k+1}, v^{k+1} - u\rangle \\
&= a_{k+1}^2\|\nabla f(y^k)\|_2^2 + \frac{1}{2}\|v^k - u\|_2^2 - \frac{1}{2}\|v^{k+1} - u\|_2^2 - \frac{1}{2}\|v^{k+1} - v^k\|_2^2 \\
&\leqslant a_{k+1}^2 L_{k+1}\Big(f(y^k) - f(x^{k+1})\Big) + \frac{1}{2}\|v^k - u\|_2^2 - \frac{1}{2}\|v^{k+1} - u\|_2^2.
\end{aligned}$$

Here the last inequality follows from line 11 of Algorithm 1. □

Lemma C.3. *For any $u \in \mathbb{R}^N$ and any $k \geqslant 0$*

$$\begin{aligned}
a_{k+1}^2 L_{k+1} f(x^{k+1}) - \left(a_{k+1}^2 L_{k+1} - a_{k+1}\right) f(x^k) \\
+ \frac{1}{2}\|v^k - u\|_2^2 - \frac{1}{2}\|v^{k+1} - u\|_2^2 \leqslant a_{k+1} f(u).
\end{aligned}$$

Proof.

$$\begin{aligned}
a_{k+1}(f(y^k)-f(u)) &\leqslant a_{k+1}\langle \nabla f(y^k), y^k-u\rangle \\
&= a_{k+1}\langle \nabla f(y^k), y^k-v^k\rangle + a_{k+1}\langle \nabla f(y^k), v^k-u\rangle \\
&\overset{①}{=} \frac{(1-\tau_k)a_{k+1}}{\tau_k}\langle \nabla f(y^k), x^k-y^k\rangle + a_{k+1}\langle \nabla f(y^k), v^k-u\rangle \\
&\overset{②}{\leqslant} \frac{(1-\tau_k)a_{k+1}}{\tau_k}\left(f(x^k)-f(y^k)\right) + a_{k+1}^2 L_{k+1}\left(f(y^k)-f(x^{k+1})\right) \\
&\qquad + \frac{1}{2}\|v^k-u\|_2^2 - \frac{1}{2}\|v^{k+1}-u\|_2^2 \\
&\overset{③}{=} (a_{k+1}^2 L_{k+1}-a_{k+1})f(x^k) - a_{k+1}^2 L_{k+1} f(x^{k+1}) + a_{k+1}f(y^k) \\
&\qquad + \frac{1}{2}\|v^k-u\|_2^2 - \frac{1}{2}\|v^{k+1}-u\|_2^2. \quad (10)
\end{aligned}$$

Here, ① uses the fact that our choice of y^k satisfies $\tau_k(y^k-v^k)=(1-\tau_k)(x^k-y^k)$. ② is by convexity of $f(\cdot)$ and Lemma C.2 , while ③ uses the choice of $\tau_k=\frac{1}{a_{k+1}L_{k+1}}$. □

Proof of Theorem C.1. Note that

$$a_{k+1}=\frac{1}{2L_{k+1}}+\sqrt{\frac{1}{4L_{k+1}^2}+a_k^2\frac{L_k}{L_{k+1}}}$$

satisfies the equation $a_{k+1}^2L_{k+1}=a_k^2L_k+a_{k+1}$. We also have $a_1=\frac{1}{L_{k+1}}$. With that in mind, we sum up the inequality in the statement of Lemma C.3 for $k=0,\ldots,T-1$ and set $u=x^*$:

$$L_T a_T^2 f(x^T)+\frac{1}{2}\|v^0-x^*\|_2^2-\frac{1}{2}\|v^T-x^*\|_2^2 \leqslant \sum_{k=0}^{T-1} a_k f(x^*) = L_T a_T^2 f(x^*).$$

Denote $A_k=a_k^2L_k$. Since $v^0=x^0$, we now have that for any $T\geqslant 1$

$$f(x^T)-f(x^*)\leqslant \frac{\|x^0-x^*\|_2^2}{2A_T}.$$

It remains to estimate A_T from below. We will now show by induction that $A_k\geqslant \frac{nk^2}{8L}$. From the L-smoothness of the objective we have

$$f(x^{k+1})=\operatorname*{argmin}_{x\in S_{i_k}(y^k)} f(x)\leqslant f(y^k-\frac{1}{L}\nabla_{i_k}f(y^k))\leqslant f(y^k)-\frac{1}{2L}\|\nabla_{i_k}f(y^k)\|_2^2.$$

Also, since i_k is chosen by the Gauss–Southwell rule, it is true that

$$\|\nabla_{i_k}f(y^k)\|_2^2\geqslant\frac{1}{n}\|\nabla f(y^k)\|_2^2.$$

As a result,

$$f(x^{k+1}) \leqslant f(y^k) - \frac{1}{2nL}\|\nabla f(y^k)\|_2^2.$$

This implies that the condition in line 11 of Algorithm 1 is automatically satisfied if $L_{k+1} \geqslant nL$. Combined with the fact that we multiply L_{k+1} by 2 if this condition is not met, this means that if $L_{k+1} \leqslant 2Ln$ at the beginning of the while loop during iteration k, then it is sure to hold at the end of the iteration too. This is guaranteed by our assumption that $L_0 \leqslant 4Ln$.

We have just shown that $L_k \leqslant 2Ln$ for $k \geqslant 1$.The base case $k = 0$ is trivial. Now assume that $A_k \geqslant \frac{k^2}{8nL}$ for some k. Note that $A_{k+1} = L_k a_k^2 + a_{k+1} = A_k + a_{k+1}$ and $L_{k+1} = \frac{A_{k+1}}{a_{k+1}^2}$.

$$a_{k+1} = \frac{1}{2L_{k+1}} + \sqrt{\frac{1}{4L_{k+1}^2} + a_k^2 \frac{L_k}{L_{k+1}}} \geqslant \frac{1}{4nL} + \sqrt{\frac{1}{16n^2L^2} + a_k^2 \frac{L_k}{2nL}} \geqslant$$
$$\geqslant \frac{1}{4nL}\left(1 + \sqrt{1 + 8A_k nL}\right) \geqslant \frac{k+1}{4nL}.$$

Finally,

$$A_{k+1} = A_k + a_{k+1} \geqslant \frac{k^2 + 2(k+1)}{8nL} \geqslant \frac{(k+1)^2}{8nL}.$$

By induction, we have $\forall k \geqslant 1$

$$A_k \geqslant \frac{k^2}{8nL} \tag{11}$$

and

$$f(x^k) - f(x^*) \leqslant \frac{4nL\|x^0 - x^*\|_2^2}{k^2}.$$

□

We also note that the assumption $L_0 \leqslant 4nL$ is not really crucial. In fact, if $L_0 > 4nL$, then after $O(\log_2 \frac{L_0}{4L})$ iterations L_k is surely lesser than $4L$, so overestimating L only results in a logarithmic in $\frac{L_0}{L}$ amount of additional iterations needed to converge.

C.1. Primal-Dual Extension for Fixed Step Accelerated Alternating Minimization

Our primal-dual algorithm based on Algorithm 1 for Problem (P_1) is listed below as Algorithm 2.

Algorithm 2 Primal-Dual Accelerated Alternating Minimization

Input: initial estimate of the Lipschitz constant L_0.

$A_0 = a_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.
for $k \geqslant 0$ **do**
  Set $L_{k+1} = L_k/2$
  while True **do**
    Set $a_{k+1} = \frac{1}{2L_{k+1}} + \sqrt{\frac{1}{4L_{k+1}^2} + a_k^2 \frac{L_k}{L_{k+1}}}$
    Set $\tau_k = \frac{1}{a_{k+1}L_{k+1}}$
    Set $\lambda^k = \tau_k \zeta^k + (1-\tau_k)\eta^k$
    Choose $i_k = \underset{i \in \{1,\dots,n\}}{\operatorname{argmax}} \|\nabla_i \varphi(\lambda^k)\|_2^2$
    Set $\eta^{k+1} = \underset{\eta \in S_{i_k}(\lambda^k)}{\operatorname{argmin}} \varphi(\eta)$
    Set $\zeta^{k+1} = \zeta^k - a_{k+1}\nabla f(\lambda^k)$
    if $\varphi(\eta^{k+1}) \leqslant \varphi(\lambda^k) - \frac{\|\nabla \varphi(\lambda^k)\|_2^2}{2L_{k+1}}$ **then**
      $\hat{x}^{k+1} = \frac{a_{k+1}x(\lambda^k) + L_k a_k^2 \hat{x}^k}{L_{k+1}a_{k+1}^2}$.
      break
    end if
    Set $L_{k+1} = 2L_{k+1}$.
  end while
end for

Output: The points $\hat{x}^{k+1}$, η^{k+1}.

The key result for this method is that it guarantees convergence in terms of the constraints and the duality gap for the primal problem, provided that it is strongly convex.

Theorem C.4. *Let the objective φ in the problem (P_2) be L-smooth and the solution of this problem be bounded, i.e. $\|\lambda^*\|_2 \leqslant R$. Then, for the sequences $\hat{x}_{k+1}, \eta_{k+1}$, $k \geqslant 0$, generated by Algorithm 2,*

$$\|\boldsymbol{A}\hat{x}^k - b\|_2 \leqslant \frac{16nLR}{k^2}, |\varphi(\eta^k) + f(\hat{x}^k)| \leqslant \frac{16nLR^2}{k^2}.$$

Proof. Once again, denote $A_k = a_k^2 L_k$ and note that $A_{k+1} = A_k + a_{k+1}$. From the proof of Lemma C.3 we have for all $\lambda \in H$

$$a_{j+1}\langle \nabla\varphi(\lambda^j), \lambda^j - \lambda\rangle \\ \leqslant A_j\varphi(\eta^j) - A_{j+1}\varphi(\eta^{j+1}) + a_{j+1}\varphi(\lambda^j) + \frac{1}{2}\|\zeta^j - \lambda\|_2^2 - \frac{1}{2}\|\zeta^{j+1} - \lambda\|_2^2.$$

We take a sum of these inequalities for $j = 0, \dots, k-1$ and rearrange the terms:

$$A_k\varphi(\eta^k) \leqslant \sum_{j=0}^{k-1} \left\{a_{j+1}\left(\varphi(\lambda^j) + \langle \nabla\varphi(\lambda^j), \lambda - \lambda^j\rangle\right)\right\} + \frac{1}{2}\|\zeta^0 - \lambda\|_2^2 - \frac{1}{2}\|\zeta^k - \lambda\|_2^2.$$

If we drop the last negative term and notice that this inequality holds for all $\lambda \in H$,

we arrive at

$$A_k\varphi(\eta^k) \leqslant \min_{\lambda\in\Lambda}\left\{\sum_{j=0}^{k-1}\{a_{j+1}(\varphi(\lambda^j) + \langle\nabla\varphi(\lambda^j), \lambda - \lambda^j\rangle) + \frac{1}{2}\|\lambda\|_2^2\right\},$$

From this point onwards, the proof mimics the proof of Theorem B.1 word-for-word. The only difference is the different bound on A_k, which is $A_k \geqslant \frac{k^2}{8Ln}$ as in Theorem C.1. □

D. Details for Section 5: Application to Optimal Transport and Wasserstein Barycenter

D.1. Derivation of the dual entropy-regularized OT problem

The dual problem is constructed as follows.

$$\begin{aligned}
&\min_{X\in Q\cap\mathcal{U}(r,c)}\langle C, X\rangle + \gamma\langle X, \ln X\rangle \\
&\quad= \min_{X\in Q}\max_{y,z\in\mathbb{R}^N}\left\{\langle C, X\rangle + \gamma\langle X, \ln X\rangle + \langle y, X\mathbf{1} - r\rangle + \left\langle z, X^T\mathbf{1} - c\right\rangle\right\} \\
&\quad= \max_{y,z\in\mathbb{R}^N}\left\{-\langle y, r\rangle - \langle z, c\rangle + \min_{X\in Q}\sum_{i,j=1}^{N} X^{ij}\left(C^{ij} + \gamma\ln X^{ij} + y^i + z^j\right)\right\}
\end{aligned}$$

Since the derivative of the entropy grows exponentially as $X^{ij} \to 0$, the objective under $\min_{X\in Q}$ grows as $X^{ij} \to 0$. This means that at the minimum point all the components $X^{ij} > 0$. Our next goal is to find $\min_{X\in Q}$. Using Lagrange multipliers for the constraint $\mathbf{1}^T X\mathbf{1} = 1$, we obtain the problem

$$\min_{X^{ij}>0}\max_{\nu}\left\{\sum_{i,j=1}^{N}\left[X^{ij}\left(C^{ij} + \gamma\ln X^{ij} + y^i + z^j\right)\right] - \nu\left[\sum_{i,j=1}^{N} X^{ij} - 1\right]\right\},$$

we obtain that the solution to this problem is

$$X^{ij} = \frac{\exp\left(-\frac{1}{\gamma}\left(y^i + z^j + C^{ij}\right)\right)}{\sum_{i,j=1}^{n}\exp\left(-\frac{1}{\gamma}\left(y^i + z^j + C^{ij}\right)\right)}$$

This allows us to write the dual problem as

$$\min_{y,z\in\mathbb{R}^N}\phi(y,z) = \gamma\ln\left(\sum_{i,j=1}^{N}\exp\left(-(y^i + z^j + C^{ij})/\gamma\right)\right) + \langle y, r\rangle + \langle z, c\rangle. \tag{12}$$

By performing a change of variables $u = -y/\gamma, v = -z/\gamma$ in (10) we arrive at an equivalent, but possibly more well-known formulation

$$\min_{u,v\in\mathbb{R}^N}\varphi(u,v) = \gamma(\ln\left(\mathbf{1}^T B(u,v)\mathbf{1}\right) - \langle u, r\rangle - \langle v, c\rangle), \tag{13}$$

$$[B(u,v)]^{ij} = \exp\left(u^i + v^j - \frac{C^{ij}}{\gamma}\right). \tag{14}$$

Note that to distinguish between the dual problem in terms of variables (y, z) and its reformulation in terms of variables (u, v) we use $\phi(y, z)$ in the first case and $\varphi(u, v)$ in the second. This also means that $\phi(y, z) = \varphi(-y/\gamma, -z/\gamma)$ by definition.

D.2. Deriving Sinkhorn's algorithm as AM for the dual problem

Lemma D.1. *The iterations*

$$u^{k+1} \in \underset{u\in\mathbb{R}^N}{\operatorname{argmin}}\, \varphi(u, v^k),\ v^{k+1} \in \underset{v\in\mathbb{R}^N}{\operatorname{argmin}}\, \varphi(u^{k+1}, v),$$

can be written explicitly as

$$u^{k+1} = u^k + \ln r - \ln\left(B\left(u^k, v^k\right)\mathbf{1}\right),$$

$$v^{k+1} = v^k + \ln c - \ln\left(B\left(u^{k+1}, v^k\right)^T\mathbf{1}\right).$$

Proof. From optimality conditions, for u to be optimal, it is sufficient to have $\nabla_u\varphi(u, v) = 0$, or

$$r - (\mathbf{1}^T B(u, v^k)\mathbf{1})^{-1} B(u, v^k)\mathbf{1} = 0. \tag{15}$$

Now we check that it is, indeed, the case for $u = u^{k+1}$ from the statement of this lemma. We manually check that

$$B(u^{k+1}, v^k)\mathbf{1} = \operatorname{diag}(e^{(u^{k+1}-u^k)})B(u^k, v^k)\mathbf{1} = \operatorname{diag}(e^{\ln r - \ln(B(u^k,v^k)\mathbf{1})})B(u^k, v^k)\mathbf{1} =$$
$$= \operatorname{diag}(r)\operatorname{diag}(B(u^k, v^k)\mathbf{1})^{-1}B(u^k, v^k)\mathbf{1} = \operatorname{diag}(r)\mathbf{1} = r$$

and the conclusion then follows from the fact that

$$\mathbf{1}^T B(u^{k+1}, v^k)\mathbf{1} = \mathbf{1}^T r = 1.$$

The optimality of v^{k+1} can be proven in the same way. □

D.3. Complexity bound for the non-regularized optimal transport

Next we describe how to apply our Algorithm 2 and Theorem 3 to find the *non-regularized* OT distance with accuracy ε, i.e. find $\widehat{X} \in \mathcal{U}(r, c)$ s.t. $\langle C, \widehat{X}\rangle - \langle C, X^*\rangle \leqslant \varepsilon$. Algorithm 3 is the pseudocode of our new algorithm for approximating the *non-regularized* OT distance.

Taking the bounds in (6) instead of bounds in [4][Theorem 3] and repeating the proof steps in [4][Theorem 4] together with [4][Theorem 2], we obtain the final bound of the complexity to find an ε-approximation for the non-regularized OT problem to

Algorithm 3 Accelerated Sinkhorn for OT

Input: Accuracy ε.

1: Set $\gamma = \frac{\varepsilon}{3\ln N}$, $\varepsilon' = \frac{\varepsilon}{8\|C\|_\infty}$.
2: Set $(\tilde{r}, \tilde{c}) = \left(1 - \frac{\varepsilon'}{8}\right)\left((r, c) + \frac{\varepsilon'}{8N}(\mathbf{1}, \mathbf{1})\right)$
3: **for** $k = 1, 2, \ldots$ **do**
4: Perform an iteration of Algorithm 2 for the OT problem with marginals $\tilde{r}, \tilde{c}$ and calculate $\widehat{X}_k$ and η_k.
5: Find $\widehat{X}$ as the projection of $\widehat{X}_k$ on $\mathcal{U}(r, c)$ by Algorithm 2 of [1].
6: **if** $\langle C, \widehat{X} - \widehat{X}_k\rangle \leqslant \frac{\varepsilon}{6}$ and $f(\hat{x}_k) + \phi(\eta_k) \leqslant \frac{\varepsilon}{6}$
7: **then** Return $\widehat{X}$.
8: **end for**

be $O\left(\frac{N^{5/2}\sqrt{\ln N}\|C\|_\infty}{\varepsilon}\right)$. To show this, we equip the primal space E with 1-norm and the dual space H with 2-norm. We define $\boldsymbol{A} : \mathbb{R}^{N\times N} \to \mathbb{R}^{2N}$ as the linear operator defining the linear constraints of the problem (8), which is in this case defined as $\boldsymbol{A}\operatorname{vec} X = ((X\mathbf{1})^T, (X^T\mathbf{1}))^T$. Then, $\|\boldsymbol{A}\|^2_{1\to 2} = 2$. Besides the Lipschitz constant, we need to bound the norm of the solution to the dual problem (10) since that norm enters the convergence rate in Theorem 3. To obtain the bound we need two following lemmas.

Lemma D.2. *Denote* $\nu = \min\limits_{i,j} K^{ij} = e^{\frac{-\|C\|_\infty}{\gamma}}$. *Any solution* (u^*, v^*) *of the dual problem* (13) *satisfies*

$$\max u_i^* - \min u_i^* \leqslant -\ln\nu \min_i r_i, \quad \max v_i^* - \min v_i^* \leqslant -\ln\nu \min_i c_i.$$

Proof. Taking the derivative of the dual objective with respect to u and denoting $\Sigma = \mathbf{1}^T B(u^*, v^*)\mathbf{1}$, we obtain that

$$\nabla_u \varphi(u^*, v^*) = r - \Sigma^{-1} B(u^*, v^*)\mathbf{1}.$$

From the first order optimality conditions we have $\nabla_u\varphi(u^*, v^*) = 0$. Then we have

$$1 \geqslant r_i = \Sigma^{-1}[B(u^*, v^*)\mathbf{1}]_i \geqslant \Sigma^{-1} e^{u_i^*}\nu\langle\mathbf{1}, e^{v^*}\rangle.$$

From this for all i we get an upper bound

$$u_i^* \leqslant \ln\Sigma - \ln\nu\langle\mathbf{1}, e^{v^*}\rangle.$$

On the other hand, since $C^{ij} > 0$, we have $K^{ij} \leqslant 1$ and

$$r_i = \Sigma^{-1}[B(u^*, v^*)\mathbf{1}]_i \leqslant \Sigma^{-1} e^{u_i^*}\langle\mathbf{1}, e^{v^*}\rangle, \quad u_i^* \geqslant \ln\Sigma + \ln r_i - \ln\langle\mathbf{1}, e^{v^*}\rangle.$$

Combining the two above results, we obtain

$$\max u_i^* - \min u_i^* \leqslant -\ln\nu \min_i r_i.$$

The result for v_i^* holds by the same exact argument. □

Lemma D.3. *There exists a solution* (y^*, z^*) *of* (10) *such that*

$$\|(y^*, z^*)\|_2 \leqslant R := \sqrt{N/2}\left(\|C\|_\infty - \frac{\gamma}{2}\ln\min_{i,j}\{r_i, c_j\}\right).$$

Proof. We begin by deriving an upper bound on $\|(u^*, v^*)\|_2$. Using the results of the previous lemma, it remains to notice that the objective $\varphi(u, v)$ is invariant under transformations $u \to u + t_u\mathbf{1}$, $u \to u + t_v\mathbf{1}$, with $t_u, t_v \in \mathbb{R}$, so there must exist some solution with $\max_i u_i^* = -\min_i u_i^* = \|u^*\|_\infty$, $\max_i v_i = -\min_i v_i = \|v^*\|_\infty$, so

$$\|u^*\|_\infty \leqslant -\frac{1}{2}\ln\nu\min_i r_i,\ \|v^*\|_\infty \leqslant -\frac{1}{2}\ln\nu\min_i c_i.$$

As a consequence,

$$\begin{aligned}\|(u^*, v^*)\|_2 \leqslant \sqrt{2N}\|(u^*, v^*)\|_\infty &\leqslant -\sqrt{N/2}\ln\nu\min_{i,j}\{r_i, c_j\} \\ &\leqslant \sqrt{N/2}\left(\frac{\|C\|_\infty}{\gamma} - \frac{1}{2}\ln\min_{i,j}\{r_i, c_j\}\right).\end{aligned}$$

By definition, $u = -\frac{1}{\gamma}y - \frac{1}{2}\mathbf{1}$, $v = -\frac{1}{\gamma}z - \frac{1}{2}\mathbf{1}$, so we have the inverse transformation $y = -\gamma u - \frac{\gamma}{2}\mathbf{1}$, $z = -\gamma v - \frac{\gamma}{2}\mathbf{1}$. Finally,

$$\begin{aligned}R = \|(y^*, z^*) - (y^0, z^0)\|_2 &= \left\|(-\gamma u^* - \frac{\gamma}{2}\mathbf{1}, -\gamma v^* - \frac{\gamma}{2}\mathbf{1}) - (-\frac{\gamma}{2}\mathbf{1}, -\frac{\gamma}{2}\mathbf{1})\right\|_2 = \\ &= \|-\gamma(u^*, v^*)\|_2 = \gamma\|(u^*, v^*)\|_2 \leqslant \sqrt{N/2}\left(\|C\|_\infty - \frac{\gamma}{2}\ln\min_{i,j}\{r_i, c_j\}\right)\end{aligned}$$

□

Next, consider the non-regularized OT problem

$$\min_{X \in Q \cap \mathcal{U}(r,c)} \langle C, X\rangle. \tag{16}$$

Let X^* be the solution of the problem (16) and X_γ^* be the solution of the regularized problem

$$\min_{X \in Q \cap \mathcal{U}(r,c)} \langle C, X\rangle + \gamma\langle X, \ln X\rangle. \tag{17}$$

Then, we have

$$\langle C, \widehat{X}\rangle = \langle C, X^*\rangle + \langle C, X_\gamma^* - X^*\rangle + \langle C, \widehat{X}_k - X_\gamma^*\rangle + \langle C, \widehat{X} - \widehat{X}_k\rangle. \tag{18}$$

Now we estimate the second and third term in the r.h.s.

$$\begin{aligned}\langle C, X_\gamma^* - X^*\rangle &= \langle C, X_\gamma^*\rangle - \gamma H(X_\gamma^*) + \gamma H(X_\gamma^*) - \min_{X\in\mathcal{U}(r,c)}\langle C, X\rangle \\ &= \min_{X\in\mathcal{U}(r,c)}\{\langle C, X\rangle - \gamma H(X)\} + \gamma H(X_\gamma^*) - \min_{X\in\mathcal{U}(r,c)}\langle C, X\rangle \quad (19)\end{aligned}$$

Furthermore, since our algorithm solves problem (P_1) with $f(x) = \langle C, X\rangle - \gamma H(X)$ and X_γ^* is the solution, we have

$$\begin{aligned}\langle C, \widehat{X}_k - X_\gamma^*\rangle &= \langle C, \widehat{X}_k\rangle - \gamma H(\widehat{X}_k)) - (\langle C, X_\gamma^*\rangle - \gamma H(X_\gamma^*)) + \gamma(H(\widehat{X}_k) - H(X_\gamma^*) \\ &\overset{①}{\leqslant} f(\hat{x}_k) + \varphi(\eta_k) + \gamma(H(\widehat{X}_k) - H(X_\gamma^*)), \quad (20)\end{aligned}$$

where ① follows from the duality gap bound $f(\hat{x}_k) - f^* \leqslant f(\hat{x}_k) + \varphi(\eta_k)$.

Then by (20) and (19) we have

$$\begin{aligned}&\langle C, X_\gamma^* - X^*\rangle + \langle C, \widehat{X}_k - X_\gamma^*\rangle \\ &\quad\leqslant \min_{X\in\mathcal{U}(r,c)}\{\langle C, X\rangle - \gamma H(X)\} + \gamma H(X_\gamma^*) - \min_{X\in\mathcal{U}(r,c)}\langle C, X\rangle \\ &\qquad + f(\hat{x}_k) + \varphi(\eta_k) + \gamma(H(\widehat{X}_k) - H(X_\gamma^*)).\end{aligned}$$

Next we use that $-H(X) \in [-2\ln n, 0]$ for any $X \in \mathcal{U}(r,c)$, which implies

$$\min_{X\in\mathcal{U}(r,c)}\{\langle C, X\rangle - \gamma H(X)\} - \min_{X\in\mathcal{U}(r,c)}\langle C, X\rangle \leqslant 0. \quad (21)$$

and finally implies

$$\langle C, X_\gamma^* - X^*\rangle + \langle C, \widehat{X}_k - X_\gamma^*\rangle \leqslant f(\hat{x}_k) + \varphi(\eta_k) + 2\gamma\ln n. \quad (22)$$

Combining (18) and (22), we obtain

$$\langle C, \widehat{X}\rangle \leqslant \langle C, X^*\rangle + \langle C, \widehat{X} - \widehat{X}_k\rangle + f(\hat{x}_k) + \varphi(\eta_k) + 2\gamma\ln n. \quad (23)$$

We immediately see that, when the stopping criterion in step 6 of Algorithm 3 is fulfilled, the output $\widehat{X} \in \mathcal{U}(r,c)$ satisfies $\langle C, \widehat{X}\rangle - \langle C, X^*\rangle \leqslant \varepsilon$.

It remains to obtain the complexity bound. First, we estimate the number of iterations in Algorithm 3 to guarantee $\langle C, \widehat{X} - \widehat{X}_k\rangle \leqslant \frac{\varepsilon}{6}$ and, after that, estimate the number of iterations to guarantee $f(\hat{x}_k) + \varphi(\eta_k) \leqslant \frac{\varepsilon}{6}$. By Hölder's inequality, we have $\langle C, \widehat{X} - \widehat{X}_k\rangle \leqslant \|C\|_\infty \|\widehat{X} - \widehat{X}_k\|_1$. By Lemma 7 in [1],

$$\|\widehat{X} - \widehat{X}_k\|_1 \leqslant 2\left(\|\widehat{X}_k\mathbf{1} - r\|_1 + \|\widehat{X}_k^T\mathbf{1} - c\|_1\right). \quad (24)$$

Next, we obtain two estimates for the r.h.s of this inequality. First, by the definition of

the operator A and the vector b,

$$\|\widehat{X}_k\mathbf{1} - r\|_1 + \|\widehat{X}_k^T\mathbf{1} - c\|_1 \leqslant \sqrt{2N}\|\boldsymbol{A}\mathrm{vec}(\widehat{X}_k) - b\|_2 \\ \leqslant \frac{16R\|\boldsymbol{A}\|_{E\to H}^2\sqrt{2N}}{\gamma k^2} \leqslant \frac{32R\sqrt{2N}}{\gamma k^2}. \quad (25)$$

Where we used Theorem 3 and the bound for R defined in Lemma D.3. Note that the statement of Theorem 3 involves n, the number of blocks, which in this case is simply equal to 2. Here we used the choice of the norm $\|\cdot\|_1$ in $E = \mathbb{R}^{n^2}$ and the norm $\|\cdot\|_2$ in $H = \mathbb{R}^{2n}$. Indeed, in this setting $\|\boldsymbol{A}\|_{E\to H}$ is equal to the maximum Euclidean norm of a column of A. By definition, each column of A contains only two non-zero elements, which are equal to one. Hence, $\|A\|_{E\to H} = \sqrt{2}$.

Combining (24) and (25) we obtain

$$\langle C, \widehat{X} - \widehat{X}_k\rangle \leqslant 2\|C\|_\infty \frac{32R\sqrt{2N}}{\gamma k^2}.$$

Setting $\gamma = \frac{\varepsilon}{3\ln N}$, we have that, to obtain $\langle C, \widehat{X} - \widehat{X}_k\rangle \leqslant \frac{\varepsilon}{6}$, it is sufficient to choose

$$k = O\left(\frac{N^{1/4}\sqrt{R\|C\|_\infty \ln N}}{\varepsilon}\right). \quad (26)$$

At the same time, since $\|\boldsymbol{A}\|_{E\to H} = \sqrt{2}$, by Theorem 3,

$$f(\hat{x}_k) + \varphi(\eta_k) \leqslant \frac{32R^2}{\gamma k^2}.$$

Since we set $\gamma = \frac{\varepsilon}{3\ln N}$, we conclude that in order to obtain $f(\hat{x}_k) + \varphi(\eta_k) \leqslant \frac{\varepsilon}{6}$, it is sufficient to choose

$$k = O\left(\frac{R\sqrt{\ln N}}{\varepsilon}\right). \quad (27)$$

To estimate the number of iterations required to reach the desired accuracy, we should take maximum of (26) and (27). We return to the bound established in Lemma D.3:

$$R \leqslant \sqrt{N/2}\left(\|C\|_\infty - \frac{\gamma}{2}\ln\min_{i,j}\{r_i, c_j\}\right).$$

In Algorithm 3 of the main part of the paper we modify the marginals $r,\ c$ to have $\min\limits_{i,j}\{r_i, c_j\} \geqslant \frac{\varepsilon}{64N\|C\|_\infty}$. As it was shown in the proof of Theorem 1 of [1], the optimal value of this problem differs from the optimal value of the original problem by no more than $2\ln N\gamma + \frac{\varepsilon}{2} = \frac{7}{6}\varepsilon$. For the modified problem we hence have the bound

$$R \leqslant \sqrt{N/2}\left(\|C\|_\infty - \frac{\varepsilon}{2\ln N}\ln\frac{\varepsilon}{64N\|C\|_\infty}\right) = O\left(\sqrt{N}\|C\|_\infty\right).$$

The ratio of the bounds (26) and (27) is equal to $\frac{\sqrt{R}}{N^{1/4}\sqrt{\|C\|_\infty}}$, so from our estimate of R we can see that these bounds are of the same order. Hence, we finally obtain the estimate on the number of iterations

$$O\left(\frac{N^{1/2}\sqrt{\ln N}\|C\|_\infty}{\varepsilon}\right).$$

Since each iteration requires $O(N^2)$ arithmetic operations, which is the same as in the Sinkhorn's algorithm, we get the total complexity

$$O\left(\frac{N^{5/2}\sqrt{\ln N}\|C\|_\infty}{\varepsilon}\right).$$

We would also like to note that the additional factor $N^{1/2}$ compared to the complexity of the Sinkhorn's algorithm seems to be the result of the very rough estimate of $\|\boldsymbol{A}\text{vec}(\widehat{X}_k) - b\|_2$ in (25), and in our experiments our method scales approximately in the same way as the Sinkhorn's algorithm when increasing the size of the problem N. Figure 1 should illustrate it.

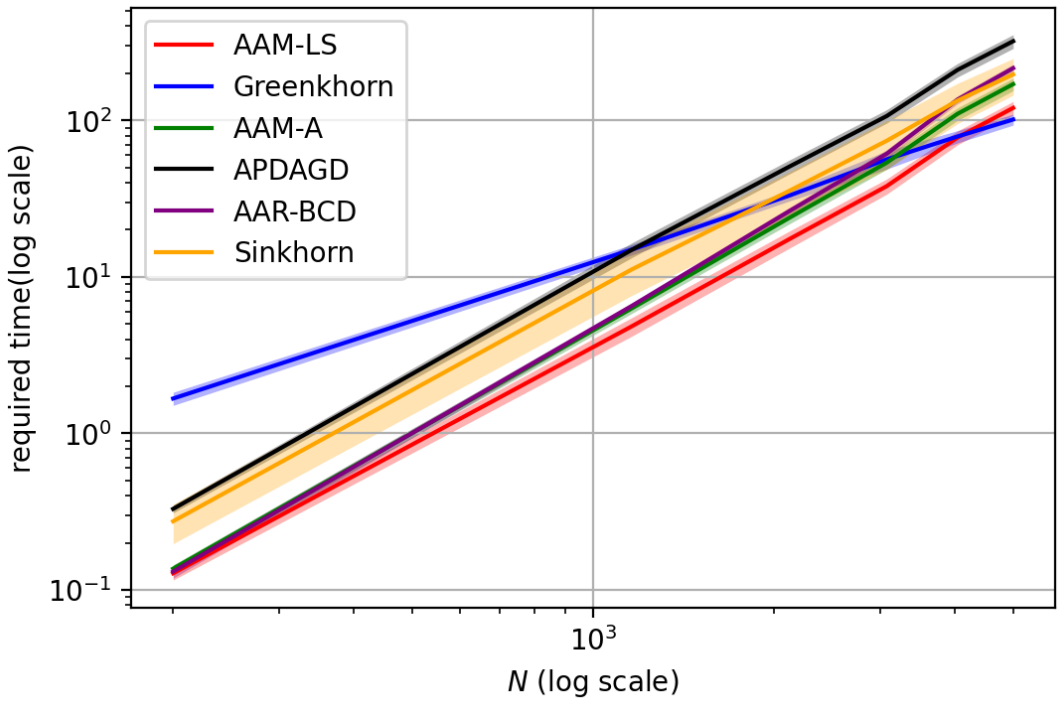

Figure 1. Experiments for OT with $\varepsilon = 0.04$ and varying dimension N

We also add to comparison the rate of decay of the dual objective in Figure 2.

Numerical experiments in [5] were performed with an instance of Mirror-prox algorithm. Authors shared their code, and now the python implementation of the method

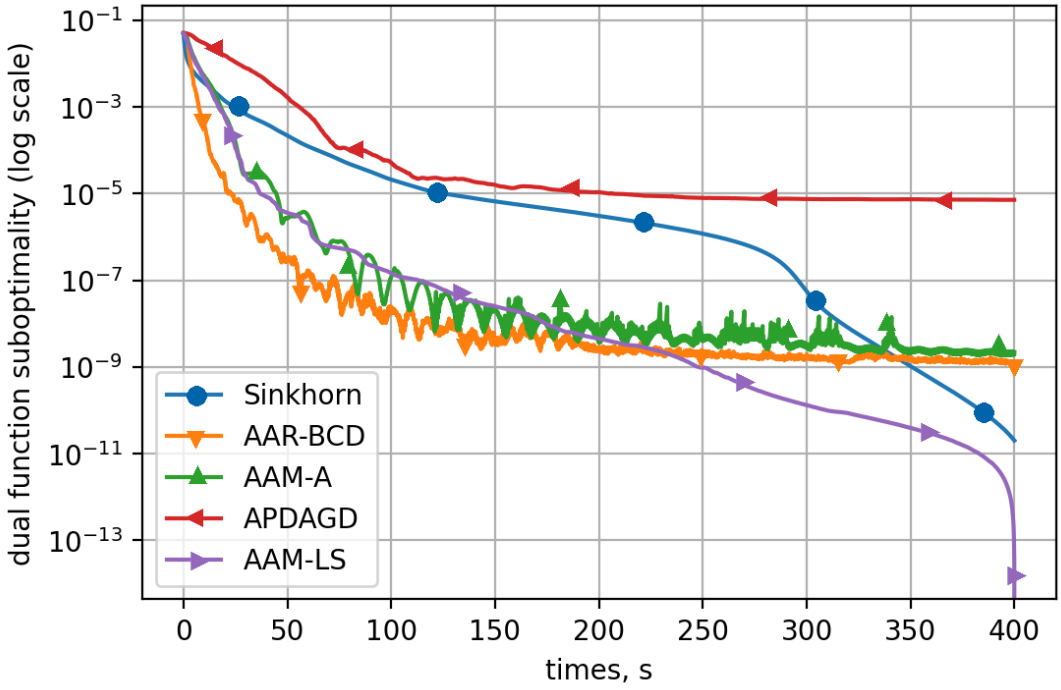

Figure 2. Decrease of the dual objective for $\varepsilon = 0.004$, $N = 1568$

is available at `https://github.com/kumarak93/numpy_ot`. We compared the rate of decay of primal non-regularized function from a transportation plan, which is projected on the feasible set with Algorithm 2 from [1]. The results is presented in Figure 3. For AAM-LS algorithm $\varepsilon = 4e - 4$.

E. Accelerating IBP

E.1. Derivation of the dual entropy-regularized WB problem

The Iterative Bregman Projections algorithm for solving the regularized Wasserstein Barycenter problem is also an instance of an alternating minimizations procedure [3, 6]. Hence, our accelerated alternating minimizations method may also be used for this problem. Denote by Δ^N the N-dimensional probability simplex. Given two probability measures p, q and a cost matrix $C \in \mathbb{R}_+^{N\times N}$ we define optimal transportation distance between them as

$$W_C(p,q) = \min_{\pi\in\Pi(p,q)} \langle \pi, C\rangle.$$

For a given set of probability measures p_i and cost matrices C_i we define their weighted barycenter with weights $w \in \Delta^m$ as a solution of the following convex optimization

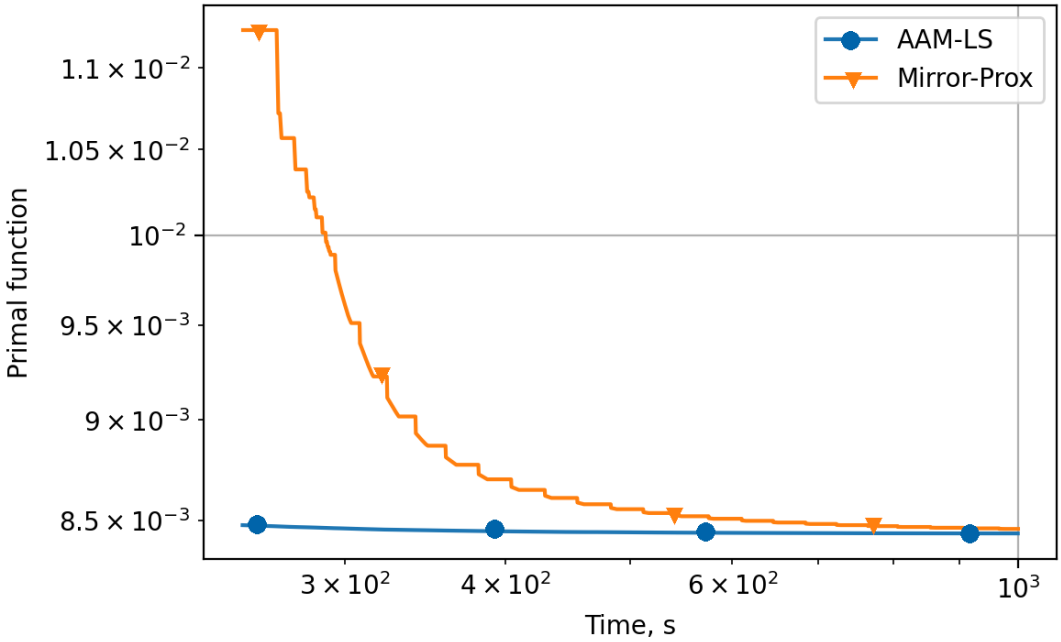

Figure 3. Decrease of the primal non-regularized objective for $\varepsilon = 0.00004$, $N = 1568$

problem:

$$\min_{q \in \Delta^N} \sum_{i=1}^{m} w_i W_{C_i}(p_i, q).$$

We use c to denote $\max\limits_{i=1,\dots,m} \|C_i\|_\infty$. We will also be using the notation $p = [p_1, \dots, p_m]$. Using the entropic regularization we define the regularized OT-distance for $\gamma > 0$:

$$W_{C,\gamma}(p, q) = \min_{\pi \in \Pi(p,q)} \langle \pi, C \rangle + \gamma H(\pi),$$

where $H(\pi) := \sum_{i,j=1}^{N} \pi_{ij} \ln \pi_{ij} = \langle \pi, \ln \pi \rangle$. One may also consider the regularized barycenter which is the solution to the following problem:

$$\min_{q \in \Delta^N} \sum_{l=1}^{m} w_l \mathcal{W}_{C_l,\gamma}(p_l, q) \tag{28}$$

The following lemma is referring to Lemma 1 from [6].

Lemma E.1. *The dual (minimization) problem of* (28) *is*

$$\min_{\sum_l w_l v_l = 0} \varphi(\mathrm{u}, \mathrm{v}), \tag{29}$$

where

$$\min_{\substack{u,v \\ \sum_{l=1}^m w_l v_l = 0}} \gamma \sum_{l=1}^m w_l \left\{ \ln \left(\mathbf{1}^T B_l \left(u_l, v_l \right) \mathbf{1} \right) - \langle u_l, p_l \rangle \right\} \tag{30}$$

$u = [u_1, \ldots, u_m], v = [v_1, \ldots, v_m], u_l, v_l \in \mathbb{R}^N$, *and*

$$B_l(u_l, v_l) := \operatorname{diag}(e^{u_l}) K_l \operatorname{diag}(e^{v_l})$$

$$K_l = \exp\left(-\frac{C_l}{\gamma} \right)$$

Moreover, the solution π_γ^* *to* (28) *is given by the formula*

$$\left[\boldsymbol{\pi}_\gamma^* \right]_l = B_l \left(u_l^*, v_l^* \right) / \left(\mathbf{1}^T B_l(u_l^*, v_l^*) \mathbf{1} \right),$$

where (u^*, v^*) *is a solution to the problem* (29).

Proof. Set $Q = \{ X \in \mathbb{R}_+^{N \times N} : \mathbf{1}^T X \mathbf{1} = 1 \}$. In its expanded form, the primal problem takes the following form:

$$\min_{\substack{\pi_l \in Q \\ \pi_l \mathbf{1} = p_l \\ \mathbf{1}^T \pi_1 = \ldots = \mathbf{1}^T \pi_m = q}} \sum_{l=1}^m w_l \left\{ \langle \pi_l, C_l \rangle + \gamma \langle \pi_l, \ln \pi_l \rangle \right\} \tag{31}$$

The above problem is equivalent to the problem

$$\min_{\pi_l \in Q} \max_{\lambda_l, \mu_l} \sum_{l=1}^m \left(w_l \left\{ \langle \pi_l, C_l \rangle + \gamma \langle \pi_l, \ln \pi_l \rangle \right\} + \langle \lambda_l, \pi_l \mathbf{1} - p_l \rangle \right) + \sum_{l=1}^{m-1} \langle \mu_l, \mathbf{1}^T \pi_l - \mathbf{1}^T \pi_m \rangle, \tag{32}$$

$$\min_{\pi_l \in Q} \max_{\lambda_l, \mu_l} \sum_{l=1}^m w_l \left\{ \langle \pi_l, C_l \rangle + \gamma \langle \pi_l, \ln \pi_l \rangle \right\} + \langle \lambda_l, \pi_l \mathbf{1} - p_l \rangle + \langle \mu_l, \mathbf{1}^T \pi_l \rangle$$

where $\mu_m = -\sum_{l=1}^{m-1} \mu_l$.

We introduce new variables $u_l = -\frac{\lambda_l}{\gamma w_l}$, $v_l = -\frac{\mu_l}{\gamma w_l}$, $l = 1, ..., m$. We can now manipulate each term in the sum above exactly as we did for the optimal transportation problem. This way we arrive at the following problem.

$$\min_{\substack{u,v \\ v_m = -\frac{1}{w_m} \sum_{l=1}^{m-1} w_l v_l}} \gamma \sum_{l=1}^m w_l \left\{ \ln \left(\mathbf{1}^T B_l \left(u_l, v_l \right) \mathbf{1} \right) - \langle u_l, p_l \rangle \right\}. \tag{33}$$

The constraints $v_m = -\frac{1}{w_m} \sum_{l=1}^{m-1} w_l v_l$ is equivalent to $\sum_{l=1}^m w_l v_l = 0$, that leads to final dual minimization problem:

$$\min_{\substack{u,v \\ \sum_{l=1}^m w_l v_l = 0}} \gamma \sum_{l=1}^{m} w_l \left\{ \ln \left(\mathbf{1}^T B_l \left(u_l, v_l \right) \mathbf{1} \right) - \langle u_l, p_l \rangle \right\}. \tag{34}$$

□

E.2. Deriving IBP algorithm as AM for the dual problem

The next result is well-known, but we include its proof in here for the sake of completeness: the objective can also be minimized exactly over the variables u, v.

Lemma E.2. *Iterations*

$$u^{k+1} = \operatorname*{argmin}_u \varphi(u, v^k),\ v^{k+1} = \operatorname*{argmin}_v \varphi(u^k, v),$$

may be written explicitly as

$$u_l^{k+1} = u_l^k + \ln p_l - \ln \left(B_l \left(u_l, v_l \right) \mathbf{1} \right),$$

$$v_l^{k+1} = v_l^k + \sum_{j=1}^{m} w_j \ln(B_j(u_j^k, v_j^k)^T 1) - \ln B_l(u_l, v_l)^T \mathbf{1}.$$

Proof. Since each term in the sum in the objective only depends on one pair of vectors (u_l, v_l), minimizing over u equivalent to minimizing over each u_l. We now have to find a solution of

$$\min_{u_l} \ln \left(\mathbf{1} B_l(u_l, v_l^k) \mathbf{1} \right) - \langle u_l, p_l \rangle.$$

This is the same problem as in Lemma D.1 with p_l instead of r, so the solution has the same form.

To minimize over v we will use Lagrange multipliers:

$$\begin{aligned} L(u, v, \tau) &= \gamma \sum_{l=1}^{m} w_l \left\{ \ln \left(\mathbf{1}^T B_l \left(u_l, v_l \right) \mathbf{1} \right) - \langle u_l, p_l \rangle \right\} + \langle \tau, \sum_{l=1}^{m} w_l v_l \rangle \\ &= \gamma \sum_{l=1}^{m} w_l \left\{ \ln \left(\mathbf{1}^T B_l \left(u_l, v_l \right) \mathbf{1} \right) - \langle u_l, p_l \rangle - \langle v_l, \frac{1}{\gamma} \tau \rangle \right\}. \end{aligned}$$

Again, we can minimize this Lagrangian independently over each v_l. By the results from Lemma D.1, we have

$$v_l^{k+1} = v_l^k + \ln \frac{1}{\gamma} \tau - \ln B_l(u_l, v_l)^T \mathbf{1}.$$

This iterate needs to satisfy the constraint $\sum_{l=1}^{m} w_l v_l^{k+1} = 0$. Assuming that the previous

iterate satisfies this constraint, we have an equation for τ:

$$\sum_{l=1}^{m} w_l \ln \frac{1}{\gamma}\tau = \sum_{l=1}^{m} w_l \ln B_l(u_l, v_l)^T \mathbf{1}.$$

Since $\sum_{l=1}^{m} w_l = 1$, we have

$$\ln \frac{1}{\gamma}\tau = \sum_{l=1}^{m} w_l \ln B_l(u_l, v_l)^T \mathbf{1}.$$

By plugging this into the formula for v_l^{k+1} we obtain the explicit form of the alternating minimization iteration from the statement of the lemma. □

This result allows us to immediately apply our acceleration scheme to this problem. The resulting method is presented as Algorithm 4. We also adopt problem-specific notation: here $\varphi(\cdot)$ denotes the dual objective (30), the first mN coordinates of the dual points $\eta^k, \zeta^k, \lambda^k$ correspond to the coordinate block u, the other coordinates – to the block v. For example, η_1^k denotes the vector of variables u_1 corresponding to the point η^k, η_{m+2}^k denotes the vector of variables v_2 corresponding to the point η^k. The map $x(\lambda)$ defined previously also takes the explicit form $x_l(u, v) = (\mathbf{1}^T B_l(u, v)\mathbf{1})^{-1} B_l(u, v)$ for $l = 1, \ldots, m$.

Algorithm 4 Accelerated Iterative Bregman Projection (Line Search)

$A_0 = \alpha_0 = 0$, $\eta_0 = \zeta_0 = \lambda_0 = 0$.
for $k \geqslant 0$ **do**
  Set $\beta_k = \underset{\beta \in [0,1]}{\operatorname{argmin}} \varphi\left(\eta^k + \beta(\zeta^k - \eta^k)\right)$
  Set $\lambda^k = \beta_k \zeta^k + (1 - \beta_k)\eta^k$
  Choose $i_k = \underset{i \in \{1,2\}}{\operatorname{argmax}} \|\nabla_i \varphi(\lambda^k)\|^2$
  if $i_k = 1$ **then**
    for $l = 1, \ldots, m$ **do**
      $\eta_l^{k+1} = \lambda_l^k + \ln p_l - \ln\left(B_l\left(\lambda_1^k, \lambda_2^k\right)\mathbf{1}\right)$
      $\eta_{m+l}^{k+1} = \lambda_{m+l}^k$
    end for
  else
    for $l = 1, \ldots, m$ **do**
      $\eta_l^{k+1} = \lambda_l^k$
      $\eta_{m+l}^{k+1} = \lambda_{m+l}^k + \sum_{j=1}^m w_j \ln(B_j(u_j^k, v_j^k)^T 1) - \ln B_l(u_l, v_l)^T \mathbf{1}$
    end for
  end if
  Find a_{k+1}, $A_{k+1} = A_k + a_{k+1}$ from

$$\varphi(\lambda^k) - \frac{a_{k+1}^2}{2(A_k + a_{k+1})} \|\nabla \varphi(\lambda^k)\|_2^2 = \varphi(\eta^{k+1})$$

  Set $\zeta^{k+1} = \zeta^k - a_{k+1} \nabla \varphi(\lambda^k)$
  Set $\hat{x}^{k+1} = \frac{a_{k+1} x(\lambda^k) + A_k \hat{x}^k}{A_{k+1}}$.
end for
Output: Transportation matrices x_l^{k+1}, dual point η^{k+1}.

Note that on each iteration of this method we take a block-wise minimization step over mN variables out of the whole $2mN$ variables, i.e. we are applying our accelerated Alternating Minimization scheme with the number of blocks $n = 2$. Since in this case our method has the exact same primal-dual properties as the accelerated method used in [6], while the complexity of our method only differs by a value dependent only on n, which in this case is simply equal to 2, the same complexity analysis applies and our method has the same complexity $O\left(\frac{mN^{5/2}\sqrt{\ln N} \max_l \|C_l\|_\infty}{\varepsilon}\right)$ as the PDAGD method in [6].

E.3. Complexity bound for the non-regularized WB problem

Next we describe how to apply our Algorithm 2 and Theorem 3 to find the *non-regularized* WB distance with accuracy ε, i.e. find $\widehat{X} \in \mathcal{U}(r, c)$ s.t. $\langle C, \widehat{X} \rangle - \langle C, X^* \rangle \leqslant \varepsilon$. Algorithm 5 is the pseudocode of our new algorithm for approximating the *non-regularized* WB distance.

Taking the bounds in (6) instead of bounds in [4][Theorem 3] and repeating the proof steps in [4][Theorem 4] together with [4][Theorem 2], we obtain the final bound of the complexity to find an ε-approximation for the non-regularized WB problem to be $O\left(\frac{N^{5/2}\sqrt{\ln N}\|C\|_\infty}{\varepsilon}\right)$. We need to bound the norm of the solution to the dual problem

Algorithm 5 Accelerated IBP

Input: Accuracy ε.

1: Set $\gamma = \frac{\varepsilon}{2\ln N}$, $\varepsilon' = \frac{\varepsilon}{8\max_l \|C_l\|_\infty}$.
2: Set $\tilde{p}_l = \left(1 - \frac{\varepsilon'}{4}\right)\left(p_l + \frac{\varepsilon'}{4N}\mathbf{1}\right)$
3: **for** $k = 1, 2, ...$ **do**
4: Perform an iteration of Algorithm 2 for the WB problem with marginals $\tilde{p}$ and calculate $\widehat{X}_l^k$, $l = 1, \cdots, m$ and η^k.
5: Find $\bar{q} = \sum_{l=1}^m w_l (\widehat{X}_l^k)^T \mathbf{1}$
6: Calculate $\widehat{X}_l$ as the projection of $\widehat{X}_l^k$ on $\mathcal{U}(\tilde{p}, \bar{q})$ by Algorithm 2 of [1].
7: **if** $\sum_{l=1}^m w_l \left\{\langle C, \widehat{X}_l\rangle - \langle C, X_l^*\rangle\right\} \leqslant \frac{\varepsilon}{4}$ and $f(\hat{x}_k) + \phi(\eta_k) \leqslant \frac{\varepsilon}{4}$
8: **then** Return $\widehat{X}$.
9: **end for**

(32) since that norm enters the convergence rate in Theorem 3. The bound is given by the two following lemmas.

Lemma E.3. *Any solution* (u^*, v^*) *of the problem* (30) *satisfies*

$$\max[u_l^*]_i - \min[u_l^*]_i \leqslant \frac{\|C_l\|\infty}{\gamma} - \ln \min_i [p_l]_i,$$

$$\max[v_l^*]_i - \min[v_l^*]_i \leqslant \frac{\|C_l\|\infty}{\gamma} + \sum_{k=1}^m w_k \frac{\|C_k\|_\infty}{\gamma}.$$

Proof. The proof of the first inequality is the same as in Lemma D.2, since the derivatives of the objective in the problem (30) with respect to u_l have the same for as in the problem (10).

For the dual iterates v^{k+1} we have the formula

$$\begin{aligned} v_l^{k+1} = v_l^k + \sum_{j=1}^m w_j \ln(B_j(u_j^k, v_j^k)^T 1) - \ln B_l(u_l, v_l)^T \mathbf{1} = \\ = v_l^k + \sum_{j=1}^m w_j \ln e^{v_j^k} + \sum_{j=1}^m w_j \ln(K_j^T e^{u^k}) - \ln e^{v_l^k} - \ln K_l^T e^{u_l^k} = \\ = \sum_{j=1}^m w_j \ln(K_j^T e^{u_j^k}) - \ln K_l^T e^{u_l^k}. \end{aligned}$$

Since this was derived from the equality of the gradient to zero and holds for any u^k, which from now on we will denote as simply u, it must also hold for v_l^*. Denote $\nu_j = e^{-\frac{\|C_j\|_\infty}{\gamma}}$. We then have

$$\ln \nu_j \langle \mathbf{1}, e^{u_j}\rangle \leqslant [\ln(K_j^T e^{u_j})]_i \leqslant \ln\langle \mathbf{1}, e^{u_j}\rangle.$$

Then

$$\sum_{j=1}^{m} w_j \ln \nu_j \langle \mathbf{1}, e^{u_j} \rangle - \ln \langle \mathbf{1}, e^{u_l} \rangle \leqslant [v_l^*]_i \leqslant \sum_{j=1}^{m} w_j \ln \langle \mathbf{1}, e^{u_j} \rangle - \ln \nu_l \langle \mathbf{1}, e^{u_l} \rangle.$$

Finally,

$$\max[v_l^*]_i - \min[v_l^*]_i \leqslant -\sum_{j=1}^{m} w_j \ln \nu_j - \ln \nu_l = \frac{\|C_l\|_\infty}{\gamma} + \sum_{j=1}^{m} w_j \frac{\|C_j\|_\infty}{\gamma}.$$

□

Set (u^0, v^0). Once again, we know the exact value of the smoothness parameter of the dual problem in terms of variables λ_i, μ_l, where $i \in \{1, \ldots, m\}, l \in \{1, \ldots, m-1\}$. Using the above Lemma we will now derive the bound on the distance to the dual solution in these variables.

Lemma E.4. *With $(\lambda^0, \mu^0) = (0, 0)$ there exists a solution of the dual problem (32) in the coordinate space (λ, μ) such that*

$$R^2 = \|(\lambda^*, \mu^*)\|_2^2 \leqslant N \left(\left(\max_l \|C_l\|_\infty - \frac{\gamma}{2} \min_{l,i} [p_l]_i \right)^2 + \max_l \|C_l\|_\infty^2 \right).$$

Proof. The coordinates (λ, μ) and (u, v) are connected by the transformation $\lambda_l = -\gamma w_l u_l$, $l \leqslant m$, $\mu_i = -\gamma w_i v_i$, $i < m$.

As a function of (u, v) the dual objective $\phi(u, v)$ is invariant under transformations of the form $u_l \to u_l + t_l \mathbf{1}$ with arbitrary $t_l \in \mathbb{R}$, and $v_l \to v_l + s_l \mathbf{1}$ with s_l such that $\sum_{l=1}^{m} w_l s_l = 0$. Hence, there exists a solution (u^*, v^*) such that for $l \in 1, \ldots, m$

$$\max[u_l^*]_i = -\min[u_l^*]_i = \|u_l^*\|_\infty,$$

and for $j \in 1, \ldots, m-1$

$$\max[v_j^*]_i = -\min[v_j^*]_i = \|v_j^*\|_\infty.$$

Using the result of the previous Lemma, we have now guaranteed the existence of a solution (u^*, v^*) such that

$$\|u_l^*\|_\infty \leqslant \frac{\|C_l\|_\infty}{2\gamma} - \frac{1}{2} \ln \min_i [p_l]_i,$$

$$\|v_l^*\|_\infty \leqslant \frac{\|C_l\|_\infty}{2\gamma} + \sum_{k=1}^{m} w_k \frac{\|C_k\|_\infty}{2\gamma}.$$

$$\|\lambda_l^*\|_\infty = \gamma w_l \|u_l^*\|_\infty \leqslant w_l \left(\frac{\|C_l\|_\infty}{2} - \frac{\gamma}{2} \ln \min_i [p_l]_i \right) \leqslant$$
$$\leqslant w_l \left(\max_l \|C_l\|_\infty - \frac{\gamma}{2} \min_{l,i} [p_l]_i \right),$$

$$\|\mu_l^*\|_\infty = \gamma w_l \|v_l^*\|_\infty \leqslant w_l \max_l \|C_l\|_\infty,\ l \in \{1, \ldots, m-1\}$$

Finally,

$$\|(\lambda^*, \mu^*)\|_2^2 = \sum_{l=1}^{m} \|\lambda_l\|_2^2 + \sum_{j=1}^{m-1} \|\mu_j^*\|_2^2 \leqslant N \left(\sum_{l=1}^{m} \|\lambda_l\|_\infty^2 + \sum_{j=1}^{m-1} \|\mu_j^*\|_\infty^2 \right)$$
$$\leqslant N \left(\left(\max_l \|C_l\|_\infty - \frac{\gamma}{2} \min_{l,i} [p_l]_i \right)^2 + \max_l \|C_l\|_\infty^2 \right)$$

□

Next, consider the non-regularized WB problem

$$\min_{\substack{X \in Q \\ \boldsymbol{A}\mathrm{vec}(X) = b}} \sum_{l=1}^{m} w_l \langle C_l, X_l \rangle, \tag{35}$$

where $\boldsymbol{A}\,\mathrm{vec}(X) = (X_1\mathbf{1}, \cdots, X_m\mathbf{1}, (X_1^T\mathbf{1} - X_m^T\mathbf{1}), (X_2^T\mathbf{1} - X_m^T\mathbf{1}), \cdots, (X_{m-1}^T\mathbf{1} - X_m^T\mathbf{1}))^T$ and $b = (p_1, \cdots, p_m, 0, \cdots, 0)^T$

Let X^* be the solution of the problem (35) and X_γ^* be the solution of the regularized problem

$$\min_{\substack{X \in Q \\ \boldsymbol{A}\mathrm{vec}(X) = b}} \sum_{l=1}^{m} w_l \langle C_l, X_l \rangle + \gamma \langle X_l, \ln X_l \rangle. \tag{36}$$

Then, we have

$$\sum_{l=1}^{m} w_l \langle C_l, \widehat{X}_l \rangle$$
$$= \sum_{l=1}^{m} w_l \left\{ \langle C_l, X_l^* \rangle + \langle C_l, X_{l\gamma}^* - X_l^* \rangle + \langle C_l, \widehat{X}_l^k - X_{l\gamma}^* \rangle + \langle C_l, \widehat{X}_l - \widehat{X}_l^k \rangle \right\}. \tag{37}$$

Now we estimate the second and third term in the r.h.s.

$$\sum_{l=1}^{m} w_l \langle C_l, X_{l\gamma}^* - X_l^* \rangle$$
$$= \sum_{l=1}^{m} w_l \left\{ \langle C_l, X_{l\gamma}^* \rangle - \gamma H(X_{l\gamma}^*) + \gamma H(X_{l\gamma}^*) \right\} - \min_{\substack{X \in Q \\ \boldsymbol{A}\mathrm{vec}(X)=b}} \sum_{l=1}^{m} w_l \langle C_l, X_l \rangle$$
$$= \min_{\substack{X \in Q \\ \boldsymbol{A}\mathrm{vec}(X)=b}} \sum_{l=1}^{m} w_l \left\{ \langle C_l, X_l \rangle - \gamma H(X_l) \right\} - \min_{\substack{X \in Q \\ \boldsymbol{A}\mathrm{vec}(X)=b}} \sum_{l=1}^{m} w_l \langle C_l, X_l \rangle + \gamma \sum_{l=1}^{m} w_l H(X_{l\gamma}^*) \tag{38}$$

Furthermore, since our algorithm solves problem (P_1) with $f(x) = \sum_{l=1}^{m} w_l \left\{ \langle C_l, X_l \rangle - \gamma H(X_l) \right\}$ and $X_{l\gamma}^*$ is the solution, we have

$$\sum_{l=1}^{m} w_l \langle C_l, \widehat{X}_l^k - X_{l\gamma}^* \rangle = \sum_{l=1}^{m} w_l \left\{ \langle C_l, \widehat{X}_l^k \rangle - \gamma H(\widehat{X}_l^k) \right\}$$
$$- \sum_{l=1}^{m} w_l \left\{ \langle C_l, X_{l\gamma}^* \rangle - \gamma H(X_{l\gamma}^*) \right\} + \gamma \sum_{l=1}^{m} w_l \left\{ H(\widehat{X}_l^k) - H(X_{l\gamma}^*) \right\}$$
$$\overset{①}{\leqslant} f(\hat{x}_k) + \varphi(\eta_k) + \gamma \sum_{l=1}^{m} w_l \left\{ H(\widehat{X}_l^k) - H(X_{l\gamma}^*) \right\}, \tag{39}$$

where ① follows from the duality gap bound $f(\hat{x}_k) - f^* \leqslant f(\hat{x}_k) + \varphi(\eta_k)$.

Then by (39) and (38) we have

$$\sum_{l=1}^{m} w_l \left\{ \langle C_l, X_{l\gamma}^* - X_l^* \rangle + \langle C_l, \widehat{X}_l^k - X_{l\gamma}^* \rangle \right\}$$
$$\leqslant \min_{\substack{X \in Q \\ \boldsymbol{A}\mathrm{vec}(X)=b}} \sum_{l=1}^{m} w_l \left\{ \langle C_l, X_l \rangle - \gamma H(X_l) \right\} + \gamma \sum_{l=1}^{m} w_l H(X_{l\gamma}^*) - \min_{\substack{X \in Q \\ \boldsymbol{A}\mathrm{vec}(X)=b}} \sum_{l=1}^{m} w_l \langle C_l, X_l \rangle$$
$$+ f(\hat{x}_k) + \varphi(\eta_k) + \gamma \sum_{l=1}^{m} w_l \left\{ H(\widehat{X}_l^k) - H(X_{l\gamma}^*) \right\}.$$

Next we use that $-H(X_l) \in [-2 \ln n, 0]$ for any $X_l \in Q$, which implies

$$\min_{\substack{X \in Q \\ \boldsymbol{A}\mathrm{vec}(X)=b}} \sum_{l=1}^{m} w_l \left\{ \langle C_l, X_l \rangle - \gamma H(X_l) \right\} - \min_{\substack{X \in Q \\ \boldsymbol{A}\mathrm{vec}(X)=b}} \sum_{l=1}^{m} w_l \langle C_l, X_l \rangle \leqslant 0. \tag{40}$$

and finally implies

$$\sum_{l=1}^{m} w_l \left\{ \langle C_l, X_{l\gamma}^* - X_l^* \rangle + \langle C_l, \widehat{X}_l^k - X_{l\gamma}^* \rangle \right\} \leqslant f(\hat{x}_k) + \varphi(\eta_k) + 2\gamma \ln n. \tag{41}$$

Combining (37) and (41), we obtain

$$\sum_{l=1}^{m} w_l \langle C_l, \widehat{X}_l \rangle \leqslant \sum_{l=1}^{m} w_l \langle C_l, X_l^* \rangle + \sum_{l=1}^{m} w_l \langle C_l, \widehat{X}_l - \widehat{X}_l^k \rangle + f(\hat{x}_k) + \varphi(\eta_k) + 2\gamma \ln n. \tag{42}$$

We immediately see that, when the stopping criterion in step 6 of Algorithm 5 is fulfilled, the output $\widehat{X}_l \in \{X \in Q | \boldsymbol{A}\mathrm{vec}(X) = b\}$ satisfies $\sum_{l=1}^{m} w_l \langle C, \widehat{X}_l \rangle - \sum_{l=1}^{m} w_l \langle C, X_l^* \rangle \leqslant \varepsilon$.

It remains to obtain the complexity bound. First, we estimate the number of iterations in Algorithm 5 to guarantee $\sum_{l=1}^{m} w_l \langle C_l, \widehat{X}_l - \widehat{X}_l^k \rangle \leqslant \frac{\varepsilon}{4}$ and, after that, estimate the number of iterations to guarantee $f(\hat{x}_k) + \varphi(\eta_k) \leqslant \frac{\varepsilon}{4}$.

Denote $q_l = (\widehat{X}_l^k)^T \mathbf{1}$. From the scheme of [6] and since $\|\boldsymbol{A}\,\mathrm{vec}(X) - b\|_1 = \sum_{l=1}^{m} \|q_l - q_{l+1}\|_1$ after an update of u variables we have

$$\begin{aligned}
\sum_{l=1}^{m} w_l \langle C_l, \widehat{X}_l - \widehat{X}_l^k \rangle &\leqslant \max_l \|C_l\|_\infty \sum_{l=1}^{m} w_l \|\widehat{X}_l - \widehat{X}_l^k\|_1 \\
&\leqslant 2 \max_l \|C_l\|_\infty \sum_{l=1}^{m} w_l \left(\|(\tilde{p}_l - p_l\|_1 + \|(\widehat{X}_l^k)^T \mathbf{1} - \bar{q}\|_1 \right) \\
&\leqslant 2 \max_l \|C_l\|_\infty \varepsilon' + 2 \max_l \|C_l\|_\infty \max_l w_l \|\boldsymbol{A}\,\mathrm{vec}(X) - b\|_1.
\end{aligned} \tag{43}$$

It remains to show that $2 \max_l \|C_l\|_\infty \max_l w_l \|\boldsymbol{A}\,\mathrm{vec}(X) - b\|_1 \leqslant \varepsilon/4$.

By Theorem 3

$$\|\boldsymbol{A}\,\mathrm{vec}(X) - b\|_1 \leqslant \frac{16R\|\boldsymbol{A}\|_{E\to H}^2 \sqrt{2N}}{\gamma k^2}.$$

Setting

$$\frac{16RL\sqrt{2N}}{k^2} = \frac{16R\|\boldsymbol{A}\|_{E\to H}^2 \sqrt{2N}}{\gamma k^2} \leqslant \frac{\varepsilon}{8 \max_l \|C_l\|_\infty \max_l w_l}, \tag{44}$$

together with the choice of $\gamma = \frac{\varepsilon}{2 \ln N}$ and since $\|\boldsymbol{A}\|_{E\to H} = \sqrt{2}$, we have that, to obtain $\langle C, \widehat{X}_l - \widehat{X}_l^k \rangle \leqslant \frac{\varepsilon}{4}$, it is sufficient to choose

$$k = O\left(\frac{N^{1/4} \sqrt{\|C_l\|_\infty \max_l w_l R \|C\|_\infty \ln N}}{\varepsilon} \right). \tag{45}$$

At the same time, by Theorem 3,

$$f(\hat{x}_k) + \varphi(\eta_k) \leqslant \frac{32R^2}{\gamma k^2}.$$

Since we set $\gamma = \frac{\varepsilon}{2 \ln N}$, we conclude that in order to obtain $f(\hat{x}_k) + \varphi(\eta_k) \leqslant \frac{\varepsilon}{4}$, it is

sufficient to choose

$$k = O\left(\frac{R\sqrt{\ln N}}{\varepsilon}\right). \tag{46}$$

To estimate the number of iterations required to reach the desired accuracy, we should take maximum of (45) and (46). We return to the bound established in Lemma D.3:

$$R^2 = \|(\lambda^*, \mu^*)\|_2^2 \leqslant N\left(\left(\max_l \|C_l\|_\infty - \frac{\gamma}{2}\min_{l,i}[\tilde{p}_l]_i\right)^2 + \max_l \|C_l\|_\infty^2\right)$$

or one can write

$$R = O\left(\sqrt{N}\|C\|_\infty\right).$$

The ratio of the bounds (45) and (46) is equal to $\frac{\sqrt{R}}{N^{1/4}\sqrt{\max_l w_l \|C\|_\infty}}$, so from our estimate of R we can see that these bounds are of the same order. Hence, we finally obtain the estimate on the number of iterations

$$O\left(\frac{N^{1/2}\sqrt{\ln N}\|C\|_\infty}{\varepsilon}\right).$$

Since each iteration requires $O(mN^2)$ arithmetic operations, which is the same as in the IBP algorithm, we get the total complexity

$$O\left(\frac{mN^{5/2}\sqrt{\ln N}\|C\|_\infty}{\varepsilon}\right).$$

F. Implementation Details

Looking through the proof of convergence for Algorithm 1 one can notice that line search subroutine need to fulfill two conditions: $\langle \nabla f(y^k), v^k - y^k\rangle \geqslant 0$ and $f(y^k) \leqslant f(x^k)$. We got significant increase of performance, when were using these condition as a stopping criteria for line search subroutine. Another increase of performance came from the observation that the value of β satisfying the condition is often close to $\frac{k-1}{k+2}$, the value appearing in Nesterov's type accelerated methods [8]. The other observation is that the value of β satisfying the conditions frequently does not change from iteration to iteration with the same parity. So we use the value β_{t-2} as a starting point for the line search subroutine to find β_t on t-th iteration. These and other implementation details are available on `https://github.com/nazya/AAM`

References

[1] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approxfimation algorithms for optimal transport via sinkhorn iteration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1961–1971. Curran Associates, Inc., 2017. arXiv:1705.09634.

[2] A. Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25(1):185–209, 2015.

[3] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[4] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376, 2018. arXiv:1802.04367.

[5] Arun Jambulapati, Aaron Sidford, and Kevin Tian. A direct tilde$o(1/\varepsilon)$ iteration parallel algorithm for optimal transport. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11359–11370. Curran Associates, Inc., 2019.

[6] Alexey Kroshnin, Nazarii Tupitsa, Darina Dvinskikh, Pavel E. Dvurechensky, Alexander Gasnikov, and Cesar A. Uribe. On the complexity of approximating wasserstein barycenters. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 3530–3540, 2019.

[7] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[8] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.