Tight Nonasymptotic Local Convergence of Sinkhorn-Knopp

Wenzhi Gao*[1], Zhaonan Qu†[2], Yinyu Ye‡[1,3], and Madeleine Udell§[1,3]

[1]ICME, Stanford University
[2]Google
[3]Department of Management Science and Engineering, Stanford University

August 13, 2026

Abstract

We revisit the Sinkhorn-Knopp (SK) algorithm for the matrix scaling problem. Despite extensive literature on the global convergence of SK and its variants, its local linear convergence behavior remains less understood. We address this gap by providing the first nonasymptotic local analysis of SK that matches the rate obtained from existing asymptotic Jacobian-based arguments. We show that under certain connectivity conditions, SK is a polynomial-time algorithm for doubly stochastic matrix scaling. With the developed tools, we showcase the local suboptimality of SK and provide accelerated variants. Finally, for dense matrices, we improve the complexity of existing first-order matrix scaling algorithms from $\mathcal{O}(\frac{n^{7/3}}{\varepsilon^{2/3}})$ to $\mathcal{O}(\frac{n^{9/4}}{\sqrt{\varepsilon}})$.

1 Introduction

Given a nonnegative matrix $A \in \mathbb{R}_+^{m\times n}$ and two positive margin vectors $p \in \mathbb{R}_{++}^m, q \in \mathbb{R}_{++}^n$, the matrix scaling problem (A, p, q) seeks two positive diagonal scaling matrices $D_1^\star, D_2^\star$ such that the scaled matrix $A^\star := D_1^\star A D_2^\star$ (approximately) satisfies the target margin condition $A^\star \mathbf{1}_n = p$ and $(A^\star)^\top \mathbf{1}_m = q$. Matrix scaling arises in a wide range of applications, including computational optimal transport [3, 34], numerical linear algebra [25], choice modeling [35], neural architecture design [39], and many others [19].

Among the algorithms for matrix scaling, Sinkhorn-Knopp (SK) [36] is arguably the simplest and most widely used. It is also known in the literature as the RAS algorithm [4]. Algorithmically, SK alternates between the two scaling matrices: it fixes one scaling matrix and updates the other so that the corresponding marginal condition is satisfied. Theoretically, the global worst-case iteration complexity of SK is now well understood, following a sequence of works [21, 24, 6, 3, 11, 18]. In terms of target accuracy ε, and ignoring the dimension dependence, worst-case $\mathcal{O}(\frac{1}{\varepsilon})$ complexity upper and lower bounds for SK have been established in the literature [11, 18].

Despite this conservative worst-case bound, SK often reaches medium-to-high accuracy in only a few iterations in practice. In particular, it is frequently observed that SK exhibits local linear convergence [34]. On one hand, a line of work has developed nonasymptotic global linear convergence guarantees for SK [13, 35, 18]. These rates are easily computable directly from the problem data (A, p, q), but are generally not tight. On the other hand, by viewing SK as a fixed-point iteration, the literature has obtained a much sharper local linear convergence rate via Jacobian linearization [25]. In particular, the asymptotic convergence rate is

*gwz@stanford.edu
†zhaonan@google.com
‡yyye@stanford.edu
§udell@stanford.edu

$$1 - \sigma_2 \coloneqq \lambda_{m-1}(P^{-1/2} A^\star Q^{-1} (A^\star)^\top P^{-1/2}), \tag{1}$$

where $P \coloneqq \mathrm{diag}(p)$, $Q \coloneqq \mathrm{diag}(q)$, and λ_{m-1} denotes the second-largest eigenvalue. However, since this Jacobian-based argument applies in the limit as the iterates approach the optimum, it only yields an asymptotic guarantee. This gap motivates us to establish a nonasymptotic local linear convergence analysis whose rate matches the sharp asymptotic rate in (1).

Contributions.

- We provide the first tight nonasymptotic local linear convergence analysis of SK, establishing an $\mathcal{O}(c + \frac{1}{\sigma_2}\log(\frac{1}{\varepsilon}))$ iteration complexity. In particular, by explicitly computing the problem-dependent constant c, we show that SK is a polynomial-time algorithm for the doubly-stochastic matrix scaling problem when σ_2 is treated as a fixed constant. Our analysis also provides a general template for proving nonasymptotic local convergence rates of alternating minimization algorithms.
- Building on these tools, we derive accelerated variants of SK that achieve improved nonasymptotic global and local complexity bounds for the matrix scaling problem.

1.1 Related literature

There is a vast literature on the matrix scaling problem. We focus on recent complexity-theoretic advances and refer interested readers to [19] for a comprehensive review.

Matrix scaling and balancing. Matrix scaling is a widely studied problem in numerical linear algebra [25] and optimization [10]. Early works on the complexity of matrix scaling problem treat the problem as a structured convex problem and obtain polynomial-time algorithms that depend on $\log \|A\|_\infty, \log \|A\|_{-\infty}, \log \|p\|_1$ and $\log(\frac{1}{\varepsilon})$ [31, 23]. [29] further removes the dependence on $\log \|A\|_{-\infty}$ and obtains a strongly polynomial-time algorithm. Recent advances on the matrix scaling problem, including [2, 7], develop first- and second-order methods for the matrix scaling problem. The matrix scaling problem is also closely related to matrix balancing and equilibration. See [25, 10] for a more detailed discussion between these problems.

Sublinear convergence of Sinkhorn-Knopp. The global sublinear convergence of SK was first established in [22, 24] and later improved in a sequence of works [3, 11, 6]. The current state-of-the-art complexity of SK is $\mathcal{O}(\frac{D}{\varepsilon})$, where D is an upper bound on the minimum-norm solution pair in log space. Recently, [18] establishes an $\Omega(\frac{\sqrt{n}}{\varepsilon})$ iteration complexity lower bound. Hence, the two bounds match in terms of the order of ε.

Linear convergence of Sinkhorn-Knopp. This line of work aims to establish linear convergence of SK. Some of the results provide global linear convergence rates: for example, [13] establishes a linear convergence rate $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ based on Hilbert's projective metric, where $\kappa = \max_{i,j,k,l} \frac{a_{ik}a_{jl}}{a_{jk}a_{il}}$. Recently, [35] establishes a linear convergence rate based on the second smallest eigenvalue of $\left(\begin{smallmatrix} P & A \\ A^\top & Q \end{smallmatrix}\right)$. Another notable recent result is [18], where the authors define a density parameter based on the normalized version of the matrix and show linear convergence when this parameter is fixed. However, the above results are often conservative upper bounds on the behavior of SK, and some only apply to strictly positive matrices. In contrast, another line of research obtains tight contraction factors based on local arguments [25, 34]. These arguments treat SK as a fixed-point iteration and obtain the tight local linear convergence rate in (1) by analyzing the spectrum of the Jacobian. The local rate matches the true performance of SK, but such arguments are only asymptotic. This paper closes this gap by providing a nonasymptotic local analysis for SK for nonnegative matrices.

2 Sinkhorn-Knopp algorithm for matrix scaling

Notation. Throughout the paper, we use $\langle \cdot, \cdot \rangle$ to denote the Euclidean inner product and $\|\cdot\|$ to denote the Euclidean norm. We denote $\|A\|_1 \coloneqq \sum_{i,j} |a_{ij}|$ and $\|A\|_{-\infty} \coloneqq \min_{|a_{ij}|>0} |a_{ij}|$. Given a vector $d \in \mathbb{R}^n$, we use

$\mathcal{D}(d) = \mathrm{diag}(d)$ to denote a corresponding diagonal matrix with d on its diagonal. Notation $\mathbf{1}_n$ denotes the all-one vector of dimension n. We will frequently use the notation $U = \mathcal{D}(u)$ and $V = \mathcal{D}(v)$. Notation $\exp(\cdot) = \mathrm{e}^{(\cdot)}$ and $\log(\cdot)$ will be applied element-wise to a vector or to the nonzeros of a matrix. i.e., $\exp(U) = \mathcal{D}(\exp(u))$. Given two vectors a, b of the same dimension, we use $a/b = \mathcal{D}(b)^{-1}a$ to denote the element-wise division. Given a symmetric matrix A, $\lambda_k(A)$ denotes its k-th smallest eigenvalue. We often use $(a, b) \coloneqq \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$ to denote the concatenation of column vectors when the context is clear. We denote $P = \mathcal{D}(p), Q = \mathcal{D}(q)$ and $S = \left(\begin{smallmatrix} P & \\ & Q \end{smallmatrix}\right)$.

2.1 Matrix scaling and Sinkhorn-Knopp iteration

We formally define the matrix scaling problem: given a nonnegative matrix $A \in \mathbb{R}_+^{m\times n}$ and two target margins $p \in \mathbb{R}_{++}^m, q \in \mathbb{R}_{++}^n$ such that $\langle \mathbf{1}_m, p\rangle = \langle \mathbf{1}_n, q\rangle$, the matrix scaling problem looks for two diagonal scaling matrices D_1, D_2 such that two margin conditions $D_1 A D_2 \mathbf{1}_n = p, D_2 A^\top D_1 \mathbf{1}_m = q$ are approximately satisfied:

$$\max\{\|D_1 A D_2 \mathbf{1}_n - p\|, \|D_2 A^\top D_1 \mathbf{1}_m - q\|\} \le \varepsilon. \tag{2}$$

Remark 1. In the recent literature for matrix scaling, the residual of the margin condition is often measured in ℓ_1-norm when p, q are probability margins. Since our focus is linear convergence, we will adopt ℓ_2-error, and converting linear convergence to ℓ_1 error loses a $\log n$ term.

A pair (D_1, D_2) satisfying (2) is called an ε-approximate scaling. Without loss of generality, we reparametrize D_1, D_2 by $D_1 = \mathrm{e}^U$ and $D_2 = \mathrm{e}^{-V}$ for real diagonal matrices U, V. `SK` alternates between the two scaling matrices by fixing one and forcing the other to satisfy the margin condition:

$$D_1^{k+1} = \mathcal{D}(p/(AD_2^k\mathbf{1}_n)) \quad \text{and} \quad D_2^{k+1} = \mathcal{D}(q/(A^\top D_1^{k+1}\mathbf{1}_m)).$$

Under mild conditions [19], this iteration will converge to an optimal scaling pair $(D_1^\star, D_2^\star)$. Without loss of generality, we assume that such a pair exists and is finite. i.e., the matrix A is scalable with respect to p, q.

A1: The nonnegative matrix A is scalable with respect to margin p, q.

Dual potential. A useful technique for analyzing the Sinkhorn iteration is through the potential function [19]

$$\varphi(u, v) \coloneqq \langle \mathrm{e}^u, A\mathrm{e}^{-v}\rangle - \langle p, u\rangle + \langle q, v\rangle, \tag{3}$$

and `SK` can be written as performing alternating minimization (`AM`) on it:

$$u^{k+1} = \arg\min_u \ \varphi(u, v^k), \quad v^{k+1} = \arg\min_v \ \varphi(u^{k+1}, v),$$

This paper adopts the `AM` perspective and will stick to **Algorithm 1**.

Algorithm 1: Sinkhorn-Knopp (`AM` on the potential function (3))

input (A, p, q), initial v^1 (or u^1)
for $k = 1, 2, \ldots$ **do**
  $u^{k+1} = \arg\min_u \ \varphi(u, v^k)$
  $v^{k+1} = \arg\min_v \ \varphi(u^{k+1}, v)$
end

3 Global and local convergence of Sinkhorn-Knopp

This section presents our main result on the convergence of the Sinkhorn algorithm.

3.1 Algorithm analysis

Global convergence. The local behavior of an algorithm typically follows a phase of global convergence. Define the solution norm diameter constant

$$D \coloneqq \min_{(u^\star,v^\star)\in\arg\min\varphi} \|(u^\star,v^\star)\|_\infty. \tag{4}$$

Under **A1**, there exists a finite scaling pair $(u^\star, v^\star)$, and the constant $D < \infty$ can be explicitly bounded under additional assumptions:

Lemma 3.1 ([28, 24, 2])**.** *Suppose* **A1** *holds. Then*

- *Square doubly-stochastic. If* $p = q = \frac{1}{n}\mathbf{1}_n$*, then* $D \le \frac{1}{2}|\log(\frac{1}{n\|A\|_{-\infty}})| + n\log(\frac{\|A\|_1}{n\|A\|_{-\infty}})$*,*
- *Doubly-stochastic. If* $p = \frac{1}{m}\mathbf{1}_m$*,* $q = \frac{1}{n}\mathbf{1}_n$*, and that* $m \le n$*, then* $D \le \frac{1}{2}|\log(\frac{1}{n\|A\|_{-\infty}})| + n^2\log(\frac{\|A\|_1}{n\|A\|_{-\infty}})$*,*
- *Positive matrix. If* $A > 0$ *and* $\langle \mathbf{1}_m, p\rangle = \langle \mathbf{1}_n, q\rangle = 1$*, then* $D \le \log(\frac{n}{\|A\|_{-\infty}\|(p,q)\|_{-\infty}^2})$*.*

Given **Lemma 3.1**, an explicit convergence rate for the suboptimality of φ is immediate.

Theorem 3.1. *Suppose* **A1** *holds and let* (u^k, v^k) *be generated by* ***Algorithm 1****, then*

$$\varphi(u^{K+1}, v^{K+1}) - \varphi(u^\star, v^\star) \le \frac{2\|p\|_1 D^2}{K},$$

where the diameter D *is defined in* (4)*.*

Local convergence. Given **Theorem 3.1**, the function value gap will vanish and finally enter the local convergence regime. Our local analysis hinges on a simple relation, given by **Proposition 3.1** below.

Proposition 3.1. *Suppose* **A1** *holds. Then we have the following identity:*

$$\varphi(u, v) = \varphi(u^\star, v^\star) + \textstyle\sum_{i=1}^m \sum_{j=1}^n a^\star_{ij}\phi(\Delta_{ij}),$$

where $\phi(\delta) \coloneqq \mathrm{e}^\delta - \delta - 1$*,* $A^\star$ *is the optimally scaled matrix, and* $\Delta_{ij} \coloneqq (u_i - u_i^\star) - (v_j - v_j^\star)$*.*

Proposition 3.1 suggests we could analyze the potential function φ through the scalar function

$$\phi(\delta) = \mathrm{e}^\delta - \delta - 1 = \frac{\delta^2}{2} + \delta^3 \int_0^1 \frac{1-t}{2}\mathrm{e}^{\delta t}\,\mathrm{d}t,$$

whose local behavior is dictated by $\frac{\delta^2}{2}$ and high-order remainder terms. In particular, it is convenient to bound the first and second-order information with zeroth-order information:

Lemma 3.2. *Let* $\phi(\delta) = \mathrm{e}^\delta - \delta - 1$*. Then* $\phi'(\delta) - \delta = \phi(\delta)$ *and* $|\phi'(\delta)| = |\phi''(\delta) - 1| \le \sqrt{2\phi(\delta)} + \phi(\delta)$*.*

As a standard component of local analysis, we introduce the second-order expansion that governs the local behavior. Define $\lambda(u, v) \coloneqq \varphi(u^\star, v^\star) + \frac{1}{2}\|(\Delta u, \Delta v)\|^2_{\nabla^2\varphi(u^\star,v^\star)}$. By definition, we have

$$\lambda(u, v) = \varphi(u^\star, v^\star) + \textstyle\sum_{i,j} \frac{a^\star_{ij}}{2}\Delta_{ij}^2, \quad \nabla\lambda(u, v) = \nabla^2\varphi(u^\star, v^\star)(\Delta u, \Delta v), \quad \text{and} \quad \nabla^2\lambda(u, v) = \nabla^2\varphi(u^\star, v^\star)$$

Since we will mostly focus on single-step progress, from now on, we will resort to the notation

$$(u, v) \;\to\; (u^+, v) \;\to\; (u^+, v^+)$$

to denote one iteration of `SK` (or `AM`). A useful fact is that `AM` on $\lambda(u, v)$ coincides with preconditioned gradient descent with preconditioner P^{-1} and Q^{-1}, whose contraction can be explicitly computed:

Lemma 3.3. *Let $(u, v), (u^+, v)$ and (u^+, v^+) be consecutive iterations obtained by running* `AM` *on λ. Then*

$$\lambda(u^+, v^+) - \varphi(u^\star, v^\star) \le (1 - \sigma_2)[\lambda(u^+, v) - \varphi(u^\star, v^\star)] \le (1 - \sigma_2)^2[\lambda(u, v) - \varphi(u^\star, v^\star)],$$

where $\sigma_2 = \lambda_2(I - P^{-1/2}A^\star Q^{-1}(A^\star)^\top P^{-1/2}) \in (0, 1)$ is the connectivity of the normalized Laplacian.

The quantity σ_2 already appeared in the `SK` literature [34, 25, 35], and it is obtained by treating `SK` as a fixed-point iteration and linearizing its Jacobian. This quantity tightly characterizes the asymptotic behavior of the algorithm and is therefore a desirable measure of local convergence. Given that φ behaves like λ in the local regime, and that `AM` applied to quadractic function λ produces a contraction of $1 - \sigma_2$, it remains to connect φ to λ. With the help of **Lemma 3.2**, **Lemma 3.4** below makes this connection explicit.

Lemma 3.4. *Suppose* **A1** *holds and let $\varepsilon = \varphi(u, v) - \varphi(u^\star, v^\star) > 0$. Then we have*

1. $\|\nabla\varphi(u, v)\|_{S^{-1}} \le 2\sqrt{\varepsilon} + \frac{2}{\sqrt{s}}\varepsilon$ *and that* $\|\nabla\varphi(u, v) - \nabla\lambda(u, v)\|_{S^{-1}} \le \frac{2}{\sqrt{s}}\varepsilon$, *where* $s \coloneqq \|(p, q)\|_{-\infty}$;
2. $\|S^{-1/2}[\nabla^2\varphi(u, v) - \nabla^2\lambda(u, v)]S^{-1/2}\| \le 4\sqrt{\frac{\varepsilon}{s}} + \frac{2\varepsilon}{s}$, *and recall that* $S = \begin{pmatrix} P & \\ & Q \end{pmatrix}$;
3. *if* $\varepsilon \le \frac{s\sigma_2^2}{1296}$, *then* $|\varphi(u, v) - \lambda(u, v)| \le \frac{168}{\sigma_2}(\frac{1}{\sqrt{s}}\varepsilon^{3/2} + \frac{1}{s}\varepsilon^2)$.

Using **Lemma 3.4**, we obtain a local recursion on the potential function gap.

Lemma 3.5. *Suppose* **A1** *holds. If $\varphi(u, v) - \varphi(u^\star, v^\star) \le \frac{s\sigma_2^2}{1296}$ with $s \coloneqq \|(p, q)\|_{-\infty}$, then*

$$\varphi(u^+, v) - \varphi(u^\star, v^\star) \le (1 - \sigma_2)[\varphi(u, v) - \varphi(u^\star, v^\star)] + \frac{336}{\sqrt{s}\sigma_2}[\varphi(u, v) - \varphi(u^\star, v^\star)]^{3/2} + \frac{338}{s\sigma_2}[\varphi(u, v) - \varphi(u^\star, v^\star)]^2,$$

and the same result holds for $\varphi(u^+, v^+) - \varphi(u^\star, v^\star)$ with respect to $\varphi(u^+, v) - \varphi(u^\star, v^\star)$.

We are ready to present the nonasymptotic complexity of Sinkhorn that only depends on (A, p, q) and σ_2.

Theorem 3.2. *Suppose* **A1** *holds and $\varepsilon \in (0, 16\|p\|_1]$. Then* `SK` *finds an ε-approximate scaling in*

$$K_\varepsilon = \left\lceil \frac{2\cdot 1344^2\|p\|_1 D^2}{\sigma_2^4} + \frac{1}{\sigma_2}\log(128[\|A\|_1\|p\|_1 + \|p\|_1^2]D) + \frac{2}{\sigma_2}\log(\tfrac{1}{\varepsilon}) \right\rceil = \mathcal{O}\big(\tfrac{\|p\|_1 D^2}{\min\{\sigma_2^4, s\sigma_2^2\}} + \tfrac{1}{\sigma_2}\log(\tfrac{1}{\varepsilon})\big)$$

iterations, where solution norm bound D is defined in (4), *$\|A\|_1 = \sum_{i,j} a_{ij}$, and $s = \|(p, q)\|_{-\infty}$.*

3.2 Implications and extensions

Tightness. Since σ_2 has been shown to match the rate obtained by linearizing the Jacobian, it is easy to find instances where our analysis matches the asymptotic behavior of the algorithm.

Proposition 3.2 (Informal)**.** *There exists (a family of) matrix scaling instances $\{(A, p, q)\}$ such that $\varphi(u^k, v^k) - \varphi(u^\star, v^\star) \ge \varepsilon$ for $k = \mathcal{O}(c + \frac{1}{\sigma_2}\log(\frac{1}{\varepsilon}))$, where $c \ge 0$ does not depend on ε.*

Polynomiality of `SK`. For doubly stochastic matrix scaling, our analysis shows that `SK` is a polynomial-time algorithm if $\sigma_2 > 0$ is considered a fixed constant.

Corollary 3.1. *Let $p = q = \frac{1}{n}\mathbf{1}_n$ (case 1 in* ***Lemma 3.1****) and suppose $\sigma_2 > 0$ is a fixed constant. Then* `SK` *finds an ε-approximate scaling in $\mathcal{O}(n^3 + \log(\frac{n}{\varepsilon}))$ iterations. If $A > 0$ is positive (case 3 in* ***Lemma 3.1****), then the complexity is further improved to $\mathcal{O}(n + \log(\frac{n}{\varepsilon}))$.*

Remark 2. The result exhibits an undesired sharp complexity transition between $\mathcal{O}(n)$ for positive matrices and $\mathcal{O}(n^3)$ for general nonnegative matrices. This transition comes from the norm bound D in **Lemma 3.1** and is likely an artifact of analysis. A smoothed measure of sparsity may overcome this weakness (e.g. [17, 18]).

Matrix balancing and equilibration. Given a real matrix $A \in \mathbb{R}^{m\times n}$, the matrix ℓ_ω matrix equilibration problem finds positive diagonal matrices D_1, D_2 such that

$$|D_1 A D_2|^\omega \mathbf{1}_m \approx p \quad \text{and} \quad |D_1 A^\top D_2|^\omega \mathbf{1}_n \approx q,$$

where $|A|$ denotes element-wise absolute value and $\omega \in (0,\infty)$. This problem [12, 10] can be reduced to matrix scaling with data $(|A|^\omega, p, q)$, and our analysis follows immediately.

Corollary 3.2. *For $\omega \in (0,1)$, **Algorithm 1** applied to $(|A|^\omega, p, q)$ finds an ε-approximate ℓ_ω-equilibration of A in $\mathcal{O}(\frac{D^2}{\min\{\sigma_2^4, s\sigma_2^2\}} + \frac{1}{\sigma_2}\log(\frac{1}{\varepsilon}))$ iterations, where $\sigma_2 = \lambda_2(I - P^{-1/2}|A^\star|^\omega Q^{-1}(|A^\star|^\omega)^\top P^{-1/2})$.*

Finally, since `SK` can also be used for matrix balancing [25], our result are similarly applicable.

4 Nonasymptotic acceleration of matrix scaling

This section develops accelerated variants of `SK`. Given the tightness result **Proposition 3.2**, acceleration is generally unachievable without modifying the algorithm. Hence, we resort to the semi-dual formulation [8].

4.1 Semi-dual function

Since `SK` performs `AM` on $\varphi(u,v)$, it is natural to consider the partially minimized objective

$$\zeta(u) \coloneqq \min_v \varphi(u,v) = \varphi(u, -\log(q/A^\top \mathrm{e}^u)) = \textstyle\sum_{j=1}^n q_j \log\langle a_{[:,j]}, \mathrm{e}^u\rangle - \langle p, u\rangle - \sum_{j=1}^n q_j \log q_j + \langle \mathbf{1}_n, q\rangle,$$

where $a_{[:,j]}$ is the j-th column of A. Function ζ is known as the semi-dual in the literature [8]. Its gradient can be computed by a half `SK` iteration: define nonlinear map $v(u) \coloneqq -\log(q/A^\top \mathrm{e}^u)$ and its Jacobian $\mathcal{J}_v$. We have

$$\nabla\zeta(u) = \nabla_u\varphi(u, v(u)) + \nabla_v\varphi(u, v(u))\mathcal{J}_v(u) = \nabla_u\varphi(u, v(u)),$$

since optimality condition implies $\nabla_v\varphi(u, v(u)) = 0$. Therefore, minimizing residual $\|\nabla\varphi(u,v)\|$ reduces to reducing the gradient norm $\|\nabla\zeta(u)\|$. **Lemma 4.1** shows that ζ has desirable properties.

Lemma 4.1. *The semi-dual function ζ satisfies the following properties*

1. *It is convex, with $L = \frac{1}{2}\|p\|_1$-Lipschitz gradient and $H = \frac{3}{2}\|p\|_1$-Lipschitz Hessian.*
2. *If $\zeta(u) - \zeta(u^\star) \le \frac{s\sigma_2^2}{1296}$, then $\zeta(u) - \zeta(u^\star) \ge \frac{\sigma_2}{4}\|\Pi(u - u^\star)\|_P^2$ and*

$$\tfrac{4}{\sigma_2}\|\nabla\zeta(u)\|_{P^{-1}}^2 \ge \zeta(u) - \zeta(u^\star) \ge \tfrac{1}{16}\|\nabla\zeta(u)\|_{P^{-1}}^2.$$

3. *If $\zeta(u) - \zeta(u^\star) \le \min\{\frac{s\sigma_2^2}{1296}, \frac{\sigma_2^3 s^2}{24\|p\|_1}\}$, then $\frac{\sigma_2}{2} \le \lambda_2(P^{-1/2}\nabla^2\zeta(u)P^{-1/2})$ and $\lambda_m(P^{-1/2}\nabla^2\zeta(u)P^{-1/2}) \le 2$*

where $\Pi = I - \frac{1}{\|p\|^2}pp^\top$ is the orthogonal projection onto $p^\perp$.

Remark 3. To our knowledge, the global and local smoothness (case 1 of **Lemma 4.1**) of the semi-dual function has been established and leveraged in the literature for optimal transport [8, 16, 40]. However, the explicit local PL constants (case 2 and 3 of **Lemma 4.1**) are not yet available.

Given that ζ is a smooth, convex function with a finite-sum structure, there are several techniques from optimization theory for making its gradient small.

4.2 Global acceleration

Finding an ε-approximate scaling corresponds to making $\|\nabla\zeta(u)\| \le \varepsilon$. For an L-smooth convex function, Nesterov's regularization technique achieves this goal with an iteration complexity of $\mathcal{O}\big(\sqrt{\frac{L\|u^\star\|}{\varepsilon}}\log(\frac{L\|u^\star\|}{\varepsilon})\big)$ [32], where the log factor can be removed using the recently developed performance estimation techniques [26].

Theorem 4.1. *Under the same assumptions as **Theorem 3.2**, there exists an algorithm $\mathcal{A}_1$ that outputs an ε-approximate scaling in*

$$K_\varepsilon = \left\lceil \frac{8\sqrt{\|p\|_1\|u^\star\|}}{\sqrt{\varepsilon}} \right\rceil \le \left\lceil \frac{8n^{1/4}\|p\|_1^{1/2}\sqrt{D}}{\sqrt{\varepsilon}} \right\rceil$$

iterations, where each iteration has the same cost as `SK`.

Following [2], assuming $D = \mathcal{O}(1)$, the arithmetic complexity of **Theorem 4.1** is $\mathcal{O}(\frac{\mathrm{nnz}(A)}{\sqrt{\varepsilon}}n^{1/4}\|p\|_1^{1/2}\sqrt{D})$, which improves on the $\mathcal{O}(\frac{\mathrm{nnz}(A)}{\varepsilon^{2/3}}\|p\|_1^{1/3}D^{2/3})$ result from [2] in terms of ε.
Finally, given the finite-sum structure of ζ, applying stochastic gradient descent with variance reduction (e.g., `Katyusha` [1]) further reduces the complexity to $\tilde{\mathcal{O}}(\frac{\mathrm{nnz}(A)}{\sqrt{\varepsilon}}n^{1/4}\|p\|_\infty^{1/2}\sqrt{D})$ in achieving $\mathbb{E}[\|\nabla\zeta(u)\|] \le \varepsilon$.

Theorem 4.2. *Under the same assumptions as **Theorem 3.2**, there exists an algorithm $\mathcal{A}_2$ that outputs $\hat{u}$ such that $\mathbb{E}[\|\nabla\zeta(\hat{u})\|] \le \varepsilon$ with arithmetic complexity $\tilde{\mathcal{O}}(\frac{\mathrm{nnz}(A)}{\sqrt{\varepsilon}}n^{1/4}\|p\|_\infty^{1/2}\sqrt{D})$ in expectation.*

Remark 4. For $p = q = \mathbf{1}_n$, $\tilde{\mathcal{O}}(\frac{\mathrm{nnz}(A)}{\sqrt{\varepsilon}}n^{1/4}\sqrt{D})$ is better than $\mathcal{O}(\frac{\mathrm{nnz}(A)}{\varepsilon^{2/3}}n^{1/3}D^{2/3})$ in expectation.

Remark 5. Specializing to entropically-regularized optimal transport with smoothing parameter η and target accuracy $\hat{\varepsilon}$, we have [28] $\mathrm{nnz}(A) = n^2, D = \mathcal{O}(\frac{\|C\|_\infty}{\eta}), \eta = \Theta(\varepsilon)$ and $\|p\|_\infty = n^{-1}$. Taking `SK` accuracy $\varepsilon \leftarrow n^{-1/2}\hat{\varepsilon}$ ensures $\|\nabla\zeta(\hat{u})\|_1 \le \sqrt{n}\|\nabla\zeta(\hat{u})\| \le \hat{\varepsilon}$. It results in an

$$\tilde{\mathcal{O}}\Big(\frac{n^2\cdot n^{1/4}}{\sqrt{\hat{\varepsilon}}n^{1/4}}\sqrt{\frac{\|C\|_\infty}{\hat{\varepsilon}}}\Big) = \tilde{\mathcal{O}}\big(\frac{n^2\sqrt{\|C\|_\infty}}{\hat{\varepsilon}}\big)$$

complexity for entropy-regularized optimal transport, with a better dependence on $\|C\|_\infty$ than the previous $\mathcal{O}(\frac{n^2\|C\|_\infty}{\hat{\varepsilon}})$ result [20, 5]. This improvement comes at the cost of randomness in the algorithm.

While global acceleration is interesting from a worst-case perspective, it is local behavior that determines how quickly the algorithm reaches a high-accuracy solution. The rest of this section explores the local behavior of `SK` and introduces two acceleration mechanisms based on the semi-dual formulation: one adopts Nesterov's acceleration and improves the complexity from $\mathcal{O}(\frac{1}{\sigma_2}\log(\frac{1}{\varepsilon}))$ to $\mathcal{O}(\frac{1}{\sqrt{\sigma_2}}\log(\frac{1}{\varepsilon}))$; the other notices that locally, `SK` corresponds to a *suboptimal* diagonal preconditioner. A better preconditioner also improves the local contraction.

4.3 Local suboptimality of Sinkhorn-Knopp

Consider a step of `SK`: $(u, v) \to (u^+, v)$ with $\nabla_v\varphi(u, v) = 0$. By our previous analysis, locally, we have

$$u^+ = u - P^{-1}\nabla_u\varphi(u, v) = u - P^{-1}\nabla\zeta(u). \tag{5}$$

Therefore, the local convergence rate of `SK` can be reproduced by running preconditioned gradient descent on ζ with preconditioner P^{-1} (or Q^{-1} if v is kept in the semi-dual formulation). Since a preconditioner can be viewed as a matrix stepsize, a natural question is whether this stepsize is optimal. To simplify the notation, for now we assume $p = q = \mathbf{1}_m$ and $P = I$, and (5) simplifies to vanilla gradient descent with stepsize 1. With this simplification, the local linear rate of preconditioned gradient descent is dictated by the ratio between smoothness and strong convexity over $\mathbf{1}_m^\perp$, the orthogonal complement of $\mathrm{span}\{\mathbf{1}_m\}$:

$$\sigma_2 I \ \preceq_\Pi \ \nabla^2\zeta(u^\star) = I - RR^\top \ \preceq_\Pi \ \sigma_m I,$$

where $\preceq_\Pi$ denotes the semidefinite order over $\mathbf{1}_m^\perp$ and $\sigma_m \coloneqq \lambda_m(I - RR^\top) \leq 1$. The effective local condition number is $\frac{\sigma_m}{\sigma_2} \leq \frac{1}{\sigma_2}$, and the $(1-\sigma_2)^2$ contraction of SK (two half-iterations) matches this bound. However, this perspective also shows that SK is locally suboptimal: for an L-smooth μ-strongly convex problem, the minimax optimal stepsize [37] $\frac{2}{L+\mu} > \frac{1}{L}$ produces a better contraction

$$(\tfrac{L-\mu}{L+\mu})^2 = (1 - \tfrac{2}{\kappa+1})^2 < (1 - \tfrac{1}{\kappa})^2.$$

Plugging in $L = 1$ and $\mu = \sigma_2$, we see that the stepsize $\frac{2}{\sigma_2+1}$ provides a local contraction of $(\frac{1-\sigma_2}{1+\sigma_2})^2 < (1-\sigma_2)^2$. If $\sigma_2 \to 1$, this stepsize improves on the contraction of SK by nearly a factor of 4. Furthermore, we have been using $L = 1$, an upper bound of σ_m; in contrast, the true condition number $\frac{\sigma_m}{\sigma_2}$ can be smaller than $\frac{1}{\sigma_2}$, making this change even more impactful. Indeed, when R has full row rank, $RR^\top$ is positive definite, and

$$\sigma_m = \lambda_m(I - RR^\top) < 1.$$

Together with the previous analysis, the stepsize $\frac{2}{\sigma_2+\sigma_m}$ further improves the contraction to $(\frac{\sigma_m-\sigma_2}{\sigma_m+\sigma_2})^2 < (1-\sigma_2)^2$.

Remark 6. Our analysis reveals an asymmetry between u and v blocks: to ensure R is full row rank, it is necessary that $m \leq n$. In other words, $\sigma_m < 1$ happens only if one uses the semi-dual objective on the block with a smaller dimension. We are not aware of this viewpoint being exploited in existing SK analyses.

As a summary, the local perspective reveals the intrinsic suboptimality of SK. This observation also carries over to the local behavior of other two-block alternating minimization algorithms. Using a slightly larger stepsize provably improves the local contraction. However, there is one challenge left: the above stepsizes rely on unknown information such as σ_2 and σ_m. It is computationally expensive to obtain accurate estimates of these quantities. To use this strategy, we need an algorithm that automatically finds the locally best stepsize. This goal is achieved by the online scaled gradient method (OSGM) [15].

4.4 Local online acceleration

Consider optimizing ζ with gradient descent preconditioned (scaled) by a diagonal matrix $\mathcal{D}(w)$ (vector w)

$$u^+ = u - \mathcal{D}(w)\nabla\zeta(u). \tag{6}$$

The descent lemma in P-norm motivates the definition of relative progress of the preconditioner vector w with respect to the scaled gradient norm $\|\nabla\zeta(u)\|_{P^{-1}}^2$: $h_u(w) \coloneqq \frac{\zeta(u-\mathcal{D}(w)\nabla\zeta(u))-\zeta(u)}{\|\nabla\zeta(u)\|_{P^{-1}}^2}$. The function h_u inherits properties like convexity and smoothness from ζ.

Lemma 4.2. *The function h_u has the following properties*

1. *It is convex and $L = \frac{1}{2}\|p\|_1$-smooth in P^{-1}-norm,*
2. *$h_u(P^{-1}\mathbf{1}_m) \leq -\frac{1}{4}$ for all u such that $\zeta(u) - \zeta(u^\star) \leq \frac{\sigma_2^2 s}{1296}$.*

Algorithm 2: Matrix scaling with online scaling (OSMS)

input (A, p, q) with $L = \frac{1}{2}\|p\|_1, H = \frac{3}{2}\|p\|_1$ (as defined in **Lemma 4.1**), initial u^1 and w^1
for $k = 1, 2, \ldots$ **do**
  $w^{k+1} = w^k - \frac{1}{L}P^{-1}\nabla h_{u^k}(w^k)$
  $u^{k+1/2} = u^k - \mathcal{D}(w^{k+1})\nabla\zeta(u^k)$
  Choose u^{k+1} that yields smaller potential value ζ between $u^{k+1/2}$ and u^k.
end

Online acceleration (**Algorithm 2**) alternates between **1)** (hyper-)gradient descent on w (with gradient $\nabla h_u(w)$) to learn $\hat{w}$ and **2)** preconditioned gradient update (6) on u (with gradient $\nabla\zeta(u)$). Given any preconditioner

vector $\hat{w}$, define the learning potential

$$\Omega_{\hat{w}}(u,w) \coloneqq \log(\zeta(u)-\zeta(u^\star)) + \tfrac{L\sigma_2}{8}\|w^+ - \hat{w}\|_P^2.$$

The following result characterizes the local behavior of **Algorithm 2**.

Lemma 4.3. *Under the same assumptions as **Theorem 3.2**, for any fixed scaling vector $\hat{w} \in \mathbb{R}^m$, we have $\Omega_{\hat{w}}(u^+,w^+) \le \Omega_{\hat{w}}(u,w) + \frac{\sigma_2}{4}h_u(\hat{w})$ for all u such that $\zeta(u)-\zeta(u^\star) \le \frac{s\sigma_2^2}{1296}$.*

Telescoping **Lemma 4.3** implies the relation $\Omega_{\hat{w}}(u^{K+1},w^{K+1}) \le \Omega_{\hat{w}}(u^1,w^1) + \frac{\sigma_2}{4}\sum_{k=1}^K h_{u^k}(\hat{w})$ holds for any $\hat{w}$. Given the freedom to choose $\hat{w}$, taking $\hat{w} = P^{-1}\mathbf{1}_m$ gives $h_u(P^{-1}\mathbf{1}_m) \le -\frac{1}{4}$ for all local u (by **Lemma 4.2**), which gives $\Omega_{\hat{w}}(u^{K+1},w^{K+1}) \le \Omega(u^1,w^1) - \frac{\sigma_2}{16}K$ and a final complexity of

$$K_\varepsilon \coloneqq \left\lceil \tfrac{16}{\sigma_2}\tfrac{L\sigma_2}{8}\|w^1 - P^{-1}\mathbf{1}_m\|_P^2 + \tfrac{16}{\sigma_2}\log(\tfrac{1}{\varepsilon})\right\rceil = \left\lceil 2L\|w^1 - P^{-1}\mathbf{1}_m\|_P^2 + \tfrac{16}{\sigma_2}\log(\tfrac{1}{\varepsilon})\right\rceil.$$

Alternatively, we can take $\hat{w} = w^\star$ that minimizes $\sum_{k=1}^K h_{u^k}(\hat{w})$. Define $\sigma_2^\star \coloneqq -\frac{\sigma_2}{4}\max_u h_u(w^\star)$. Then $\Omega_{w^\star}(u^{K+1},w^{K+1}) \le \Omega_{w^\star}(u^1,w^1) - \sigma_2^\star K$ implies a complexity of

$$K_\varepsilon \coloneqq \lceil \tfrac{L\sigma_2}{8\sigma_2^\star}\|w^1 - w^\star\|_P^2 + \tfrac{1}{\sigma_2^\star}\log(\tfrac{1}{\varepsilon})\rceil,$$

which can be faster than $\mathcal{O}(\frac{1}{\sigma_2}\log(\frac{1}{\varepsilon}))$ of `SK` if $\sigma_2 < \sigma_2^\star$. Finally, the above local arguments are automatically globalized since `OSGM` has an $\mathcal{O}(\frac{D^2}{K})$ global convergence rate on smooth convex functions [15].

Theorem 4.3 (Informal)**.** ***Algorithm 2*** *outputs an ε-approximate scaling in $\mathcal{O}(\frac{1}{\sigma_2^\star}\log(\frac{1}{\varepsilon}))$ iterations.*

We see that while online acceleration can speed up matrix scaling, its effect is problem-dependent. On the other hand, extrapolation-based schemes often yield a strict acceleration effect. The next section explores Nesterov's accelerated gradient method to accelerate matrix scaling locally.

4.5 Local Nesterov's acceleration

A few attempts have been made in the literature to locally accelerate `SK`, mostly by interpreting it as fixed-point iteration and employing overrelaxation [27, 38, 33]. However, existing arguments based on fixed-point iteration are asymptotic. In this section, we instead make use of the preconditioned gradient descent perspective, whose accelerated variant is preconditioned accelerated gradient descent (**Algorithm 3**, `PAGD`).

Algorithm 3: Matrix scaling with accelerated preconditioned gradient descent (`PAGD`)

input (A,p,q), initial point $u^1 = w^1$
for $k = 1,2,\dots$ **do**
  $y^k = u^k + \frac{\sqrt{\sigma_2}}{\sqrt{\sigma_2}+2}(w^k - u^k)$
  $u^{k+1/2} = y^k - \frac{1}{2}P^{-1}\nabla\zeta(y^k)$
  $w^{k+1} = (1-\frac{1}{2}\sqrt{\sigma_2})w^k + \frac{1}{2}\sqrt{\sigma_2}(y^k - \frac{4}{\sigma_2}P^{-1}\nabla\zeta(y^k))$
  Let u^{k+1} be the better between $u^{k+1/2}$ and u^k.
end

Our analysis adopts the following potential function [9]

$$f(u,w) \coloneqq \zeta(u) + \tfrac{\sigma_2}{4}\|\Pi(w-u^\star)\|_P^2,$$

where recall that $\Pi = I - \frac{1}{\|p\|^2}pp^\top$. By **Lemma 4.1**, ζ is 2-smooth and $\frac{\sigma_2}{2}$-strongly convex over $\mathbf{1}_m^\perp$ in P-norm. Then, a standard potential function-based analysis [9] establishes the desired accelerated rate.

Lemma 4.4. *Let $(u, w), (u^+, w^+)$ be consecutive iterates generated by **Algorithm 3**. If $\zeta(u^1) - \zeta(u^\star) \leq \min\{\frac{\sigma_2^2 s}{1296}, \frac{\sigma_2^3 s^2}{24\|p\|_1}\} \cdot \min\{4\|p\|_1^{-1}, 1\}$, then $f(u^+, w^+) - \zeta(u^\star) \leq (1 - \frac{1}{2}\sqrt{\sigma_2})[f(u, w) - \zeta(u^\star)]$.*

Theorem 4.4. *Suppose $u^1 = w^1$ satisfies the condition from **Lemma 4.4**, then **Algorithm 3** outputs an ε-approximate scaling in $\mathcal{O}(\frac{2}{\sqrt{\sigma_2}} \log(\frac{1}{\varepsilon}))$ iterations.*

5 Numerical experiments

This section conducts numerical experiments to validate our findings.

Experiment setup. We compare `SK`, gradient descent (`GD`) on ζ, `OSMS` (**Algorithm 2**), and `PAGD` (**Algorithm 3**). Both `OSMS` and `PAGD` warm-start with `SK` until $\|\mathrm{e}^U A \mathrm{e}^{-v} - p\| \leq 10^{-3}$; `OSMS` initializes its preconditioner at $P_1 = P^{-1}$ (so its first step is a `SK` step).

Dataset. We use entropy-regularized optimal transport instances with Gibbs kernel $A = \mathrm{e}^{-C/\eta}$, $\eta = 2 \times 10^{-3}$, where C is obtained by 1). `random`: support points uniform in $[0, 1]^2$, C the squared Euclidean cost, and $p, q \sim \mathcal{U}[0, 1]$ 2). `MNIST`: p, q are normalized digit images with $m = n = 784$.

5.1 Local suboptimality of Sinkhorn-Knopp

This section validates our finding that `SK` is locally suboptimal (using fixed stepsize 1.0). In **Figure 1**, we compare the performance of `SK` to `GD` with two fixed stepsizes $\frac{2}{1+\sigma_2}$ and $\frac{2}{\sigma_2+\sigma_m}$ from **Section 4.3**.

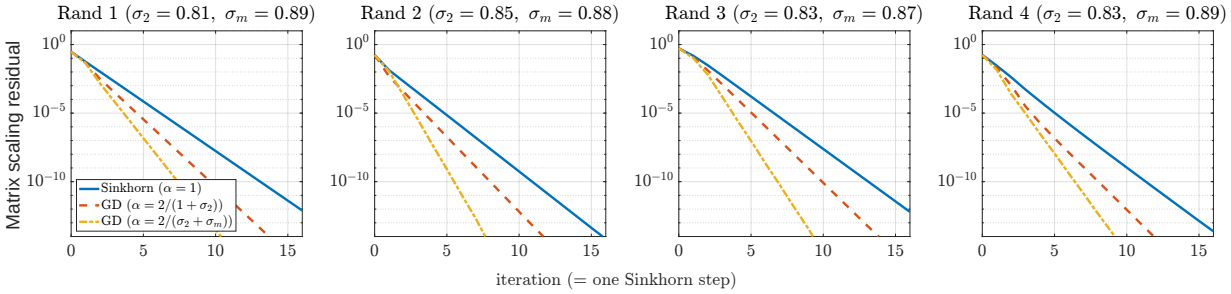

Figure 1: Local suboptimality of `SK` ($m = 3$, $n = 300$): the larger minimax ($\alpha = 2/(1 + \sigma_2)$) and optimal ($\alpha = 2/(\sigma_2 + \sigma_m)$) stepsizes contract faster than `SK` ($\alpha = 1$)

As our theory predicts, using a larger step size yields better local contraction than `SK`.

5.2 Accelerated variants of Sinkhorn-Knopp

Figures 2 and **3** compare the accelerated variants on eight random and eight `MNIST` instances. Both `PAGD` and `OSMS` beat `SK` by several orders of magnitude within the budget.

6 Conclusion

This paper investigates the nonasymptotic local convergence of `SK` and provides new insights into its behavior. We also develop nonasymptotic accelerated variants through the semi-dual formulation. Our techniques extend to related problems, such as unbalanced optimal transport [16], and our analysis template applies to general two-block alternating minimization algorithms.

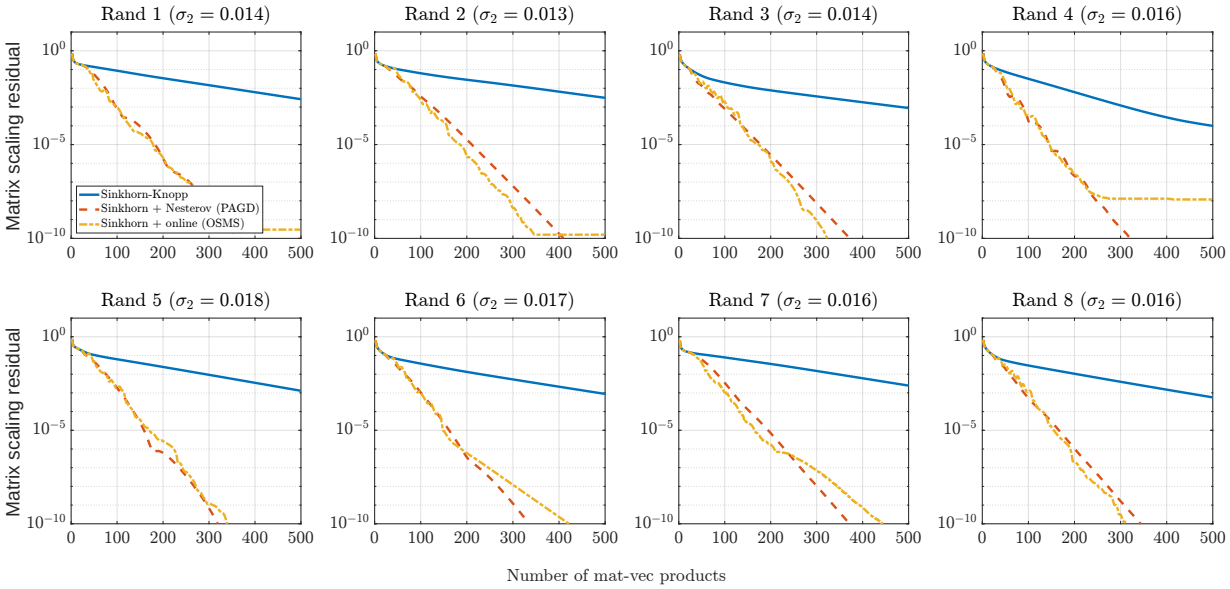

Figure 2: Accelerated variants of SK on eight random entropic optimal transport instances ($m = n = 200$): residual versus number of matrix-vector products.

References

[1] Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018. (cited on 7, 27)

[2] Zeyuan Allen-Zhu, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Much faster algorithms for matrix scaling. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 890–901. IEEE, 2017. (cited on 2, 4, 7)

[3] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in Neural Information Processing Systems*, 30, 2017. (cited on 1, 2)

[4] Michael Bacharach. Estimating nonnegative matrices from marginal data. *International Economic Review*, 6(3):294–310, 1965. (cited on 1)

[5] Jose Blanchet, Arun Jambulapati, Carson Kent, and Aaron Sidford. Towards optimal running times for optimal transport. *arXiv preprint arXiv:1810.07717*, 2018. (cited on 7)

[6] Deeparnab Chakrabarty and Sanjeev Khanna. Better and simpler error analysis of the sinkhorn–knopp algorithm for matrix scaling. *Mathematical Programming*, 188(1):395–407, 2021. (cited on 1, 2)

[7] Michael B Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. Matrix scaling and balancing via box constrained newton's method and interior point methods. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 902–913. IEEE, 2017. (cited on 2)

[8] Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Review*, 60(4):941–965, 2018. (cited on 6)

[9] Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021. (cited on 9, 31)

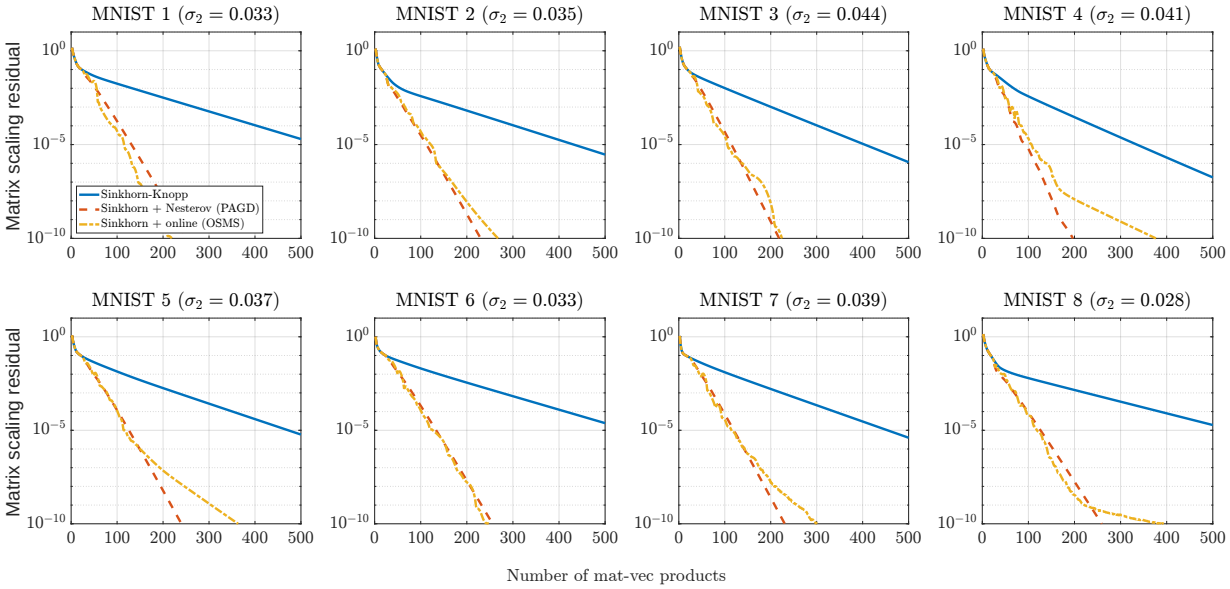

Figure 3: Accelerated variants of `SK` on eight `MNIST` entropic optimal transport instances ($n = 784$, $\eta = 2\times10^{-3}$).

[10] Steven Diamond and Stephen Boyd. Stochastic matrix-free equilibration. *Journal of Optimization Theory and Applications*, 172(2):436–454, 2017. (cited on 2, 6)

[11] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In *International Conference on Machine Learning*, pages 1367–1376. PMLR, 2018. (cited on 1, 2, 17)

[12] Christopher Fougner and Stephen Boyd. Parameter selection and preconditioning for a graph form solver. In *Emerging Applications of Control and Systems Theory: A Festschrift in Honor of Mathukumalli Vidyasagar*, pages 41–61. Springer, 2018. (cited on 6)

[13] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989. (cited on 1, 2)

[14] Wenzhi Gao. Non-asymptotic local convergence analysis of alternating minimization, April 2026. Blog Post #4.

[15] Wenzhi Gao, Ya-Chi Chu, Yinyu Ye, and Madeleine Udell. Gradient methods with online scaling part I. theoretical foundations. *arXiv preprint arXiv:2505.23081*, 2025. (cited on 8, 9)

[16] Ferdinand Genans. Fast and large-scale unbalanced optimal transport via its semi-dual and adaptive gradient methods. *arXiv preprint arXiv:2602.10697*, 2026. (cited on 6, 10)

[17] Kun He. On the efficiency of sinkhorn-knopp for entropically regularized optimal transport. *arXiv preprint arXiv:2604.03787*, 2026. (cited on 5)

[18] Kun He. Phase transition of the sinkhorn-knopp algorithm. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4238–4284. SIAM, 2026. (cited on 1, 2, 5)

[19] Martin Idel. A review of matrix scaling and sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016. (cited on 1, 2, 3)

[20] Arun Jambulapati, Aaron Sidford, and Kevin Tian. A direct tilde {O}(1/epsilon) iteration parallel algorithm for optimal transport. *Advances in Neural Information Processing Systems*, 32, 2019. (cited on 7)

[21] Bahman Kalantari. A theorem of the alternative for multihomogeneous functions and its relationship to diagonal scaling of matrices. *Linear Algebra and its Applications*, 236:1–24, 1996. (cited on 1)

[22] Bahman Kalantari and Leonid Khachiyan. On the rate of convergence of deterministic and randomized ras matrix scaling algorithms. *Operations research letters*, 14(5):237–244, 1993. (cited on 2)

[23] Bahman Kalantari and Leonid Khachiyan. On the complexity of nonnegative-matrix scaling. *Linear Algebra and its applications*, 240:87–103, 1996. (cited on 2)

[24] Bahman Kalantari, Isabella Lari, Federica Ricca, and Bruno Simeone. On the complexity of general matrix scaling and entropy minimization via the ras algorithm. *Mathematical Programming*, 112(2):371–401, 2008. (cited on 1, 2, 4, 16)

[25] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. (cited on 1, 2, 5, 6)

[26] Jongmin Lee, Chanwoo Park, and Ernest Ryu. A geometric structure of acceleration and its role in making gradients small fast. *Advances in Neural Information Processing Systems*, 34:11999–12012, 2021. (cited on 7, 26)

[27] Tobias Lehmann, Max-K Von Renesse, Alexander Sambale, and André Uschmajew. A note on overrelaxation in the sinkhorn algorithm. *Optimization Letters*, 16(8):2209–2220, 2022. (cited on 9)

[28] Tianyi Lin, Nhat Ho, and Michael Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning*, pages 3982–3991. PMLR, 2019. (cited on 4, 7, 16)

[29] Nathan Linial, Alex Samorodnitsky, and Avi Wigderson. A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 644–652, 1998. (cited on 2)

[30] Pravin Nair. Softmax is $1/2$-lipschitz: A tight bound across all ℓ_p norms. *Transactions on Machine Learning Research*, 2026. (cited on 27)

[31] Arkadi Nemirovski and Uriel Rothblum. On complexity of matrix scaling. *Linear Algebra and its Applications*, 302:435–460, 1999. (cited on 2)

[32] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013. (cited on 7, 28)

[33] Gabriel Peyré, Lenaic Chizat, François-Xavier Vialard, and Justin Solomon. Quantum entropic regularization of matrix-valued optimal transport. *European Journal of Applied Mathematics*, 30(6):1079–1102, 2019. (cited on 9)

[34] Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019. (cited on 1, 2, 5)

[35] Zhaonan Qu, Alfred Galichon, Wenzhi Gao, and Johan Ugander. On sinkhorn's algorithm and choice modeling. *Operations Research*, 2025. (cited on 1, 2, 5, 16)

[36] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. (cited on 1)

[37] Adrien B Taylor, Julien M Hendrickx, and François Glineur. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *Journal of Optimization Theory and Applications*, 178(2):455–476, 2018. (cited on 8)

[38] Alexis Thibault, Lénaïc Chizat, Charles Dossal, and Nicolas Papadakis. Overrelaxed sinkhorn–knopp algorithm for regularized optimal transport. *Algorithms*, 14(5):143, 2021. (cited on 9)

[39] Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Kuai Yu, et al. mhc: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025. (cited on 1)

[40] Zeyi Xu and Long Chen. Accelerating sinkhorn for entropy-regularized optimal transport. *arXiv preprint arXiv:2605.30267*, 2026. (cited on 6)

Appendix

Table of Contents

A Proof of results in Section 3

A.1 Auxiliary result

Lemma A.1. *Suppose* **A1** *holds and that* $p = \frac{1}{m}\mathbf{1}_m$, $q = \frac{1}{n}\mathbf{1}_n$. *Then there exists a solution* $(u^\star, v^\star)$ *such that*

$$\|(u^\star, v^\star)\|_\infty \leq \tfrac{1}{2}\big|\log\big(\tfrac{\|(p,q)\|_\infty}{\|A\|_{-\infty}}\big)\big| + \mathrm{lcm}(m,n)\log\big(\tfrac{\|(p,q)\|_\infty \|A\|_1}{\|p\|_1\|A\|_{-\infty}}\big),$$

where $\mathrm{lcm}(a,b)$ *denotes the least common multiple of* m *and* n.

Proof. According to Theorem 5.1 of [24], there exist $(u^\star, v^\star)$ such that

$$u_i^\star = \tfrac{1}{2}\log\big(\tfrac{\|(p,q)\|_\infty}{\|A\|_{-\infty}}\big) + t \quad \text{and} \quad v_j^\star = \tfrac{1}{2}\log\big(\tfrac{\|(p,q)\|_\infty}{\|A\|_{-\infty}}\big) - t$$

where

$$|t| \leq \tfrac{1}{\delta_{m,n}}\big| - \|p\|_1 \log\big(\tfrac{\|A\|_1}{\|p\|_1}\big) - \|p\|_1 \log\big(\tfrac{\|(p,q)\|_\infty}{\|A\|_{-\infty}}\big)\big| = \tfrac{1}{\delta_{m,n}}\|p\|_1 \log\big(\tfrac{\|(p,q)\|_\infty \|A\|_1}{\|p\|_1\|A\|_{-\infty}}\big),$$

and $\delta_{mn} = \frac{1}{\mathrm{lcm}(m,n)}$, where $\mathrm{lcm}(m,n)$ denotes the least common multiple between m and n. Therefore, with $\|p\|_1 = 1$, we have

$$|u_i^\star| \leq \tfrac{1}{2}\big|\log\big(\tfrac{\|(p,q)\|_\infty}{\|A\|_{-\infty}}\big)\big| + \mathrm{lcm}(m,n)\log\big(\tfrac{\|(p,q)\|_\infty \|A\|_1}{\|A\|_{-\infty}}\big),$$

and the same result holds for $v^\star$. □

Lemma A.2. *Suppose* **A1** *holds and that* $m = n, A > 0$ *and* $\|p\|_1 = 1$. *Then there exists a solution* $(u^\star, v^\star)$ *such that* $\|(u^\star, v^\star)\|_\infty \leq \log\big(\frac{n}{\|A\|_{-\infty}\|(p,q)\|_{-\infty}^2}\big)$.

Proof. The proof is immediate by taking $C_{ij} = \mathrm{e}^{-a_{ij}}$ and adopting $v \leftarrow -v$ in [28]. □

A.2 Proof of Lemma 3.1

The first two bullet points follow from **Lemma A.1** with $\mathrm{lcm}(m,n) \leq mn$ and that $\mathrm{lcm}(n,n) = n$. The last bullet point follows from **Lemma A.2**.

A.3 Proof of Theorem 3.1

We start by showing that the Sinkhorn iterates satisfy $\|u^{k+1} - u^\star\|_\infty \leq \|u^1 - u^\star\|_\infty$. According to Lemma 2 of [35], consider any $i \in [m]$ and γ, η such that

$$\gamma \mathrm{e}^{u_i^\star} \leq \mathrm{e}^{u_i^k} \leq \eta \mathrm{e}^{u_i^\star},$$

we always have $\gamma \mathrm{e}^{u_i^\star} \leq \mathrm{e}^{u_i^{k+1}} \leq \eta \mathrm{e}^{u_i^\star}$. Taking $\gamma = \mathrm{e}^{u_i^\star - u_i^1}$ and $\eta = \mathrm{e}^{u_i^1 - u_i^\star}$, we have, inductively, that

$$\mathrm{e}^{u_i^\star - u_i^1}\mathrm{e}^{u_j^\star} \leq \mathrm{e}^{u_i^k} \leq \mathrm{e}^{u_i^1 - u_i^\star}\mathrm{e}^{u_i^\star},$$

which implies $\mathrm{e}^{u_i^\star - u_i^1} \leq \mathrm{e}^{u_i^k - u_i^\star} \leq \mathrm{e}^{u_i^1 - u_i^\star}$ and that $|u_i^k - u_j^\star| \leq |u_i^1 - u_i^\star|$, giving $\|u^k - u^\star\|_\infty \leq \|u^1 - u^\star\|_\infty$. The same argument holds for v^k. Next, we deduce that

$$\begin{aligned}
\varphi(u^k, v^k) - \varphi(u^\star, v^\star) &\leq \langle \nabla\varphi(u^k, v^k), (u^k - u^\star, v^k - v^\star)\rangle && (7)\\
&\leq \|\nabla\varphi(u^k, v^k)\|_1 \|(u^k - u^\star, v^k - v^\star)\|_\infty && (8)\\
&\leq \|\nabla\varphi(u^k, v^k)\|_1 \|(u^1 - u^\star, v^1 - v^\star)\|_\infty && (9)\\
&= \|\nabla\varphi(u^k, v^k)\|_1 \|(u^\star, v^\star)\|_\infty && (10)\\
&\leq D\|\nabla\varphi(u^k, v^k)\|_1,
\end{aligned}$$

where (7) uses convexity of φ; (8) uses Hölder's inequality $\langle a, b\rangle \le \|a\|_1\|b\|_\infty$; (9) uses the previously derived relation $\|u^k - u^\star\|_\infty \le \|u^1 - u^\star\|_\infty$ and $\|v^k - v^\star\|_\infty \le \|v^1 - v^\star\|_\infty$; the relation (10) uses the initialization $(u^1, v^1) = (\mathbf{0}_m, \mathbf{0}_n)$. Finally, the proof of sublinear convergence is similar to that in [11]. Denote $\delta_k \coloneqq \varphi(u^k, v^k) - \varphi(u^\star, v^\star)$. According to the Theorem 1 of [11], given $\varphi(u^k, v^k) - \varphi(u^\star, v^\star) \le D\|\nabla\varphi(u^k, v^k)\|_1$, we have

$$\delta_{k+1} \le \delta_k - \tfrac{\delta_k^2}{2\|p\|_1 D^2},$$

which implies $\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} = \frac{\delta_k - \delta_{k+1}}{\delta_k\delta_{k+1}} \ge \frac{1}{2\|p\|_1 D^2}\frac{\delta_k}{\delta_{k+1}} \ge \frac{1}{2\|p\|_1 D^2}$ since `AM` enforces $\delta_{k+1} \le \delta_k$. Hence

$$\tfrac{1}{\varphi(u^{K+1}, v^{K+1}) - \varphi(u^\star, v^\star)} = \tfrac{1}{\delta_{K+1}} \ge \tfrac{1}{\delta_1} + \tfrac{K}{2\|p\|_1 D^2}$$

and we arrive at $\varphi(u^{K+1}, v^{K+1}) - \varphi(u^\star, v^\star) \le \frac{2\|p\|_1 D^2}{K}$. This completes the proof.

A.4 Proof of Proposition 3.1

This equation can be directly verified. Let $\Delta u \coloneqq u - u^\star$ and $\Delta v \coloneqq v - v^\star$ and consider

$$\begin{aligned}\varphi(u, v) - \varphi(u^\star, v^\star) &= \langle \mathrm{e}^u, A\mathrm{e}^{-v}\rangle - \langle p, u\rangle + \langle q, v\rangle - (\langle \mathrm{e}^{u^\star}, A\mathrm{e}^{-v^\star}\rangle - \langle p, u^\star\rangle + \langle q, v^\star\rangle)\\ &= \langle \mathrm{e}^u, A\mathrm{e}^{-v}\rangle - \langle \mathrm{e}^{u^\star}, A\mathrm{e}^{-v^\star}\rangle - \langle p, u - u^\star\rangle + \langle q, v - v^\star\rangle\end{aligned}$$

Using $\langle \mathrm{e}^u, A\mathrm{e}^{-v}\rangle = \langle \mathrm{e}^{u^\star}\circ \mathrm{e}^{u-u^\star}, A\mathrm{e}^{v^\star - v}\circ \mathrm{e}^{-v^\star}\rangle = \langle \mathrm{e}^{u-u^\star}, \mathrm{e}^{U^\star}A\mathrm{e}^{-V^\star}\mathrm{e}^{v^\star - v}\rangle$, we deduce that

$$\begin{aligned}\langle \mathrm{e}^u, A\mathrm{e}^{-v}\rangle - \langle \mathrm{e}^{u^\star}, A\mathrm{e}^{-v^\star}\rangle &= \langle \mathrm{e}^{u-u^\star}, \mathrm{e}^{U^\star}A\mathrm{e}^{-V^\star}\mathrm{e}^{v^\star - v}\rangle - \langle \mathrm{e}^{u^\star}, A\mathrm{e}^{-v^\star}\rangle\\ &= \textstyle\sum_{i,j} a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}\mathrm{e}^{\Delta u_i - \Delta v_j} - \sum_{i,j} a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}\\ &= \textstyle\sum_{i,j} a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}[\mathrm{e}^{\Delta u_i - \Delta v_j} - 1].\end{aligned}$$

Since $p_i = \sum_j a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}$ and $q_j = \sum_i a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}$, we have

$$\begin{aligned}-\langle p, u - u^\star\rangle + \langle q, v - v^\star\rangle &= -\textstyle\sum_{i,j} a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}\Delta u_i + \sum_{i,j} a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}\Delta v_i\\ &= -\textstyle\sum_{i,j} a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}(\Delta u_i - \Delta v_j).\end{aligned}$$

Plugging in $a_{ij}\mathrm{e}^{u_i^\star - v_j^\star} = a_{ij}^\star$ gives

$$\begin{aligned}\varphi(u, v) - \varphi(u^\star, v^\star) &= \textstyle\sum_{i,j} a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}[\mathrm{e}^{\Delta u_i - \Delta v_j} - 1] - \sum_{i,j} a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}(\Delta u_i - \Delta v_j)\\ &= \textstyle\sum_{i,j} a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}[\mathrm{e}^{\Delta u_i - \Delta v_j} - 1 - (\Delta u_i - \Delta v_j)]\\ &= \textstyle\sum_{i,j} a_{ij}\mathrm{e}^{u_i^\star - v_j^\star}(\mathrm{e}^{\Delta_{ij}} - 1 - \Delta_{ij})\\ &= \textstyle\sum_{i,j} a_{ij}^\star\phi(\Delta_{ij}),\end{aligned}$$

and this completes the proof.

A.5 Proof of Lemma 3.2

Given $\phi(\delta) = \mathrm{e}^\delta - \delta - 1$. We have $\phi'(\delta) = \mathrm{e}^\delta - 1$ and $\phi''(\delta) = \mathrm{e}^\delta$. Hence $\phi'(\delta) - \delta = \phi(\delta)$ and

$$|\phi'(\delta)| = |\phi''(\delta) - 1| = |\mathrm{e}^\delta - 1|$$

If $\delta > 0$, by $\mathrm{e}^\delta - \delta - 1 \ge \frac{\delta^2}{2}$, we have $|\delta| \le \sqrt{2\phi(\delta)}$ and that

$$|\mathrm{e}^\delta - 1| \le |\mathrm{e}^\delta - \delta - 1| + |\delta| \le \phi(\delta) + \sqrt{2\phi(\delta)}.$$

If $\delta < 0$, we deduce that

$$|\mathrm{e}^{\delta} - 1| = 1 - \mathrm{e}^{\delta} \leq \sqrt{2(\mathrm{e}^{\delta} - \delta - 1)} = \sqrt{2\phi(\delta)} \leq \phi(\delta) + \sqrt{2\phi(\delta)},$$

where the first inequality follows from defining

$$h(\delta) \coloneqq (1 - \mathrm{e}^{\delta})^2 - 2(\mathrm{e}^{\delta} - \delta - 1) = \mathrm{e}^{2\delta} - 4\mathrm{e}^{\delta} + 2\delta + 3$$

and noticing that $h(0) = 0$ and that $h'(0) = 2\mathrm{e}^{2\delta} - 4\mathrm{e}^{\delta} + 2 = 2(\mathrm{e}^{\delta} - 1)^2 \geq 0$, giving $h(\delta) \leq 0$ for $\delta \leq 0$. Combining these two cases completes the proof.

A.6 Proof of Lemma 3.3

For brevity, in the proof we will denote $R \coloneqq A^{\star}$ and hence

$$\nabla^2 \varphi(u^{\star}, v^{\star}) = \left(\begin{smallmatrix} P & -R \\ -R^{\top} & Q \end{smallmatrix}\right).$$

Since (u, v) is generated by `AM`, assuming $\nabla_v \lambda(u, v) = 0$, we have

$$v - v^{\star} = Q^{-1} R^{\top} (u - u^{\star}) \tag{11}$$

and consider the preconditioned update $u^{+} = u - P^{-1} \nabla \lambda_u(u, v)$. It gives

$$\begin{aligned} u^{+} - u^{\star} &= u - P^{-1} \nabla \lambda_u(u, v) - u^{\star} \\ &= u - u^{\star} - P^{-1}(P(u - u^{\star}) - R(v - v^{\star})) \\ &= P^{-1} R(v - v^{\star}) \end{aligned} \tag{12}$$

and it is easy to see that $\nabla_u \lambda(u^{+}, v) = P(u^{+} - u^{\star}) - R(v - v^{\star}) = 0$. Hence `AM` applied to λ yields the same u^{+}. Next, we consider the contraction, and we successively deduce that

$$\begin{aligned} \lambda(u, v) - \varphi(u^{\star}, v^{\star}) &= \tfrac{1}{2}\|u - u^{\star}\|_P^2 - \langle u - u^{\star}, R(v - v^{\star}) \rangle + \tfrac{1}{2}\|v - v^{\star}\|_Q^2 \\ &= \tfrac{1}{2}\|u - u^{\star}\|_{P - RQ^{-1}R^{\top}}^2, \end{aligned} \tag{13}$$

where (13) applies (11). Furthermore, plugging (11) into (12) gives $u^{+} - u^{\star} = P^{-1} R Q^{-1} R^{\top} (u - u^{\star})$ and

$$\begin{aligned} \lambda(u^{+}, v) - \varphi(u^{\star}, v^{\star}) &= \tfrac{1}{2}\|u^{+} - u^{\star}\|_P^2 - \langle u^{+} - u^{\star}, R(v - v^{\star}) \rangle + \tfrac{1}{2}\|v - v^{\star}\|_Q^2 \\ &= \tfrac{1}{2}\|u - u^{\star}\|_{RQ^{-1}R^{\top} - RQ^{-1}R^{\top}P^{-1}RQ^{-1}R^{\top}}^2. \end{aligned}$$

Now it suffices to show, for any z, that

$$\|z\|_{RQ^{-1}R^{\top} - RQ^{-1}R^{\top}P^{-1}RQ^{-1}R^{\top}}^2 \leq (1 - \sigma_2)\|z\|_{P - RQ^{-1}R^{\top}}^2.$$

Denote the eigen-decomposition $M \coloneqq P^{-1/2} R Q^{-1} R^{\top} P^{-1/2} = U \Sigma U^{\top}$ and suppose the eigenvalues of Σ are non-increasing (left top is the largest). We may check that $P^{1/2}\mathbf{1}_m$ is an eigenvector with eigenvalue 1:

$$M P^{1/2} \mathbf{1}_m = P^{-1/2} R Q^{-1} R^{\top} P^{-1/2} (P^{1/2}\mathbf{1}) = P^{1/2}\mathbf{1}_m,$$

and therefore $P^{1/2}\mathbf{1}_m \in \mathrm{Nul}(I - M)$. Besides, note that $0 \preceq \Sigma \preceq I$ since $\left(\begin{smallmatrix} P & -R \\ -R^{\top} & Q \end{smallmatrix}\right) \succeq 0$ implies that its Schur complement $P - RQ^{-1}R^{\top} \succeq 0$ is positive semidefinite, which further implies

$$I \succeq P^{-1/2} R Q^{-1} R^{\top} P^{-1/2} \succeq 0$$

Now we deduce that

$$\begin{aligned}\|z\|^2_{RQ^{-1}R^\top - RQ^{-1}R^\top P^{-1}RQ^{-1}R^\top} &= \langle z, (RQ^{-1}R^\top - RQ^{-1}R^\top P^{-1}RQ^{-1}R^\top)z\rangle \\ &= \langle z, (P^{1/2}MP^{1/2} - P^{1/2}M^\top MP^{1/2})z\rangle \\ &= \|P^{1/2}z\|^2_{M-M^\top M}\end{aligned}$$

and $\|z\|^2_{P-RQ^{-1}R^\top} = \langle z, (P - P^{1/2}MP^{1/2})z\rangle = \|P^{1/2}z\|^2_{I-M}$. Since $\mathrm{Nul}(I-M) \subseteq \mathrm{Nul}(M - M^\top M)$, we assume $P^{1/2}z \perp \mathrm{Nul}(I-M)$ and deduce that

$$\begin{aligned}\max_{P^{1/2}z \perp \mathrm{Nul}(I-M)} \frac{\|P^{1/2}z\|^2_{M-M^\top M}}{\|P^{1/2}z\|^2_{I-M}} &= \max_{x \perp \mathrm{Nul}(I-M)} \frac{\langle x, (M - M^\top M)x\rangle}{\langle x, (I-M)x\rangle} \\ &= \max_{x \perp \mathrm{Nul}(I-\Sigma)} \frac{\langle x, (\Sigma - \Sigma^2)x\rangle}{\langle x, (I-\Sigma)x\rangle} \\ &= \max_{x_1 = 0, x \perp \mathrm{Nul}(I-\Sigma)} \frac{\langle x, (\Sigma - \Sigma^2)x\rangle}{\langle x, (I-\Sigma)x\rangle} \\ &\le \lambda_2(\Sigma) \\ &= 1 - \lambda_2(I - \Sigma) = 1 - \sigma_2.\end{aligned}$$

Using $\|z\|^2_{RQ^{-1}R^\top - RQ^{-1}R^\top P^{-1}RQ^{-1}R^\top} \le (1-\sigma_2)\|z\|^2_{P-RQ^{-1}R^\top}$, we have

$$\lambda(u^+, v) - \varphi(u^\star, v^\star) \le (1-\sigma_2)[\lambda(u,v) - \varphi(u^\star, v^\star)],$$

and this completes the proof. Repeating the argument by switching the role of u, v and noticing that

$$\begin{aligned}\lambda_2(I_m - P^{-1/2}RQ^{-1}R^\top P^{-1/2}) &= \lambda_2(I_m - P^{-1/2}RQ^{-1/2}(P^{-1/2}RQ^{-1/2})^\top) \\ &= \lambda_2(I_n - (P^{-1/2}RQ^{-1/2})^\top P^{-1/2}RQ^{-1/2}) \\ &= \lambda_2(I_n - Q^{-1/2}R^\top P^{-1}RQ^{-1/2}),\end{aligned}$$

we have $\lambda(u^+, v^+) - \varphi(u^\star, v^\star) \le (1-\sigma_2)^2[\lambda(u,v) - \varphi(u^\star, v^\star)]$. This completes the proof.

A.7 Proof of Lemma 3.4

Recall the notation $\Delta_{ij} \coloneqq \Delta u_i - \Delta v_j$ and define $g(t) \coloneqq \varphi(u + td^u, v + td^v)$ and $h(t) \coloneqq \lambda(u + td^u, v + td^v)$ with $d = (d^u, d^v) \in \mathbb{R}^{m+n}$. We have $g(0) = \varphi(u,v)$, $h(0) = \lambda(u,v)$, and that

$$\begin{aligned}g'(t) &= \tfrac{\mathrm{d}}{\mathrm{d}t}[\textstyle\sum_{i,j} a^\star_{ij}\phi(\Delta_{ij} + t(d^u_i - d^v_j))] \\ &= \textstyle\sum_{i,j} a^\star_{ij}\phi'(\Delta_{ij} + t(d^u_i - d^v_j))(d^u_i - d^v_j) \\ g''(t) &= \textstyle\sum_{i,j} a^\star_{ij}\phi''(\Delta_{ij} + t(d^u_i - d^v_j))(d^u_i - d^v_j)^2. \\ h'(t) &= \textstyle\sum_{i,j} a^\star_{ij}(\Delta_{ij} + t(d^u_i - d^v_j))(d^u_i - d^v_j) \\ h''(t) &= \textstyle\sum_{i,j} a^\star_{ij}(d^u_i - d^v_j)^2\end{aligned}$$

Taking $t = 0$, we have

$$\begin{aligned}g'(0) &= \langle \nabla\varphi(u,v), d\rangle = \textstyle\sum_{i,j} a^\star_{ij}\phi'(\Delta_{ij})(d^u_i - d^v_j) \\ g''(0) &= \langle d, \nabla^2\varphi(u,v)d\rangle = \textstyle\sum_{i,j} a^\star_{ij}\phi''(\Delta_{ij})(d^u_i - d^v_j)^2 \\ h'(0) &= \langle \nabla\lambda(u,v), d\rangle = \textstyle\sum_{i,j} a^\star_{ij}\Delta_{ij}(d^u_i - d^v_j) \\ h''(0) &= \langle d, \nabla^2\lambda(u^\star, v^\star)d\rangle = \textstyle\sum_{i,j} a^\star_{ij}(d^u_i - d^v_j)^2.\end{aligned}$$

Proof of relation 1. We invoke **Lemma 3.2** and deduce that

$$\begin{aligned}
|\langle \nabla\varphi(u,v) - \nabla\lambda(u,v), d\rangle| &= |\sum_{i,j} a^\star_{ij}\phi'(\Delta_{ij})(d_i^u - d_j^v) - \sum_{i,j} a^\star_{ij}\Delta_{ij}(d_i^u - d_j^v)| \\
&= |\sum_{i,j} a^\star_{ij}[\phi'(\Delta_{ij}) - \Delta_{ij}](d_i^u - d_j^v)| \\
&= |\sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})(d_i^u - d_j^v)| \qquad (14)\\
&\le 2[\textstyle\sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})]\|d\|_\infty \\
&= 2\varepsilon\|d\|_\infty, \qquad (15)
\end{aligned}$$

where (14) uses $\phi(\delta) = \phi'(\delta) - \delta$ from **Lemma 3.2**; (15) uses **Proposition 3.1**. Taking $d = \frac{\nabla\varphi(u,v) - \nabla\lambda(u,v)}{\|\nabla\varphi(u,v) - \nabla\lambda(u,v)\|}$ gives $\|\nabla\varphi(u,v) - \nabla\lambda(u,v)\| \le 2\varepsilon\frac{\|d\|_\infty}{\|d\|} \le 2\varepsilon$ and that

$$\|\nabla\varphi(u,v) - \nabla\lambda(u,v)\|_{S^{-1}} = \|S^{-1/2}[\nabla\varphi(u,v) - \nabla\lambda(u,v)]\| \le \tfrac{2\varepsilon}{\sqrt{\|(p,q)\|_{-\infty}}}.$$

Next, consider

$$\begin{aligned}
|\langle \nabla\varphi(u,v), S^{-1/2}d\rangle| &= |\textstyle\sum_{i,j} a^\star_{ij}\phi'(\Delta_{ij})(\frac{d_i^u}{\sqrt{p_i}} - \frac{d_j^v}{\sqrt{q_j}})| \\
&\le \textstyle\sum_{i,j} a^\star_{ij}|\phi'(\Delta_{ij})||\frac{d_i^u}{\sqrt{p_i}} - \frac{d_j^v}{\sqrt{q_j}}| \\
&\le \textstyle\sum_{i,j} a^\star_{ij}(\sqrt{2\phi(\Delta_{ij})} + \phi(\Delta_{ij}))|\frac{d_i^u}{\sqrt{p_i}} - \frac{d_j^v}{\sqrt{q_j}}| \qquad (16)\\
&= \textstyle\sum_{i,j} a^\star_{ij}\sqrt{2\phi(\Delta_{ij})}|\frac{d_i^u}{\sqrt{p_i}} - \frac{d_j^v}{\sqrt{q_j}}| + \sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})|\frac{d_i^u}{\sqrt{p_i}} - \frac{d_j^v}{\sqrt{q_j}}|,
\end{aligned}$$

where (16) invokes $\phi'(\delta) \le \sqrt{2\phi(\delta)} + \phi(\delta)$ from **Lemma 3.2**. Now we bound these two terms respectively by

$$\begin{aligned}
\textstyle\sum_{i,j} a^\star_{ij}\sqrt{2\phi(\Delta_{ij})}|\frac{d_i^u}{\sqrt{p_i}} - \frac{d_j^v}{\sqrt{q_j}}| &= \textstyle\sum_{i,j}\sqrt{2a^\star_{ij}\phi(\Delta_{ij})}\Big|\frac{\sqrt{a^\star_{ij}}d_i^u}{\sqrt{\sum_k a^\star_{ik}}} - \frac{\sqrt{a^\star_{ij}}d_j^v}{\sqrt{\sum_k a^\star_{kj}}}\Big| \\
&\le \sqrt{\textstyle\sum_{i,j} 2a^\star_{ij}\phi(\Delta_{ij})}\sqrt{\textstyle\sum_{i,j}\Big(\frac{\sqrt{a^\star_{ij}}d_i^u}{\sqrt{\sum_k a^\star_{ik}}} - \frac{\sqrt{a^\star_{ij}}d_j^v}{\sqrt{\sum_k a^\star_{kj}}}\Big)^2} \qquad (17)\\
&= \sqrt{2\varepsilon}\sqrt{\textstyle\sum_{i,j}\Big(\frac{\sqrt{a^\star_{ij}}d_i^u}{\sqrt{\sum_k a^\star_{ik}}} - \frac{\sqrt{a^\star_{ij}}d_j^v}{\sqrt{\sum_k a^\star_{kj}}}\Big)^2} \\
&\le \sqrt{2\varepsilon}\sqrt{2\Big[\textstyle\sum_i\sum_j \frac{a^\star_{ij}(d_i^u)^2}{\sum_k a^\star_{ik}} + \sum_j\sum_i \frac{a^\star_{ij}(d_j^v)^2}{\sum_k a^\star_{kj}}\Big]} \qquad (18)\\
&= 2\sqrt{\varepsilon}\sqrt{\textstyle\sum_i (d_i^u)^2 + \sum_j (d_j^v)^2} = 2\sqrt{\varepsilon}\|d\|,
\end{aligned}$$

where (17) uses Cauchy-Schwarz $\sum_{i,j} a_{ij}b_{ij} \le \sqrt{\sum_{i,j} a_{ij}}\sqrt{\sum_{i,j} b_{ij}}$; (18) uses $(a-b)^2 \le 2a^2 + 2b^2$. Moreover, we have $\sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})|\frac{d_i^u}{\sqrt{p_i}} - \frac{d_j^v}{\sqrt{q_j}}| \le \varepsilon\max_{ij}|\frac{d_i^u}{\sqrt{p_i}} - \frac{d_j^v}{\sqrt{q_j}}| \le \frac{2\varepsilon\|d\|_\infty}{\sqrt{\|(p,q)\|_{-\infty}}}$. Taking $d = \frac{S^{-1/2}\nabla\varphi(u,v)}{\|S^{-1/2}\nabla\varphi(u,v)\|}$ gives

$$\|\nabla\varphi(u,v)\|_{S^{-1}} = |\langle \nabla\varphi(u,v), S^{-1/2}d\rangle| \le 2\sqrt{\varepsilon} + \tfrac{2\varepsilon}{\sqrt{\|(p,q)\|_{-\infty}}}.$$

Before concluding, we also prove a bound on $\|\nabla\varphi(u,v)\|$. Consider

$$
\begin{aligned}
|\langle \nabla\varphi(u,v), d\rangle| &= |\textstyle\sum_{i,j} a^\star_{ij}\phi'(\Delta_{ij})(d^u_i - d^v_j)| \\
&\le \textstyle\sum_{i,j} a^\star_{ij}|\phi'(\Delta_{ij})||d^u_i - d^v_j| \\
&\le 2\textstyle\sum_{i,j} a^\star_{ij}[\sqrt{2\phi(\Delta_{ij})} + \phi(\Delta_{ij})]\|d\|_\infty \qquad (19)\\
&= 2[\textstyle\sum_{i,j} a^\star_{ij}\sqrt{2\phi(\Delta_{ij})} + \sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})]\|d\|_\infty \\
&= 2[\sqrt{2}\textstyle\sum_{i,j}\sqrt{a^\star_{ij}}\sqrt{a^\star_{ij}\phi(\Delta_{ij})} + \sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})]\|d\|_\infty \\
&\le 2[\sqrt{2}\sqrt{\textstyle\sum_{i,j} a^\star_{ij}}\sqrt{\textstyle\sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})} + \textstyle\sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})]\|d\|_\infty \qquad (20)\\
&\le (4\sqrt{\|p\|_1\varepsilon} + 2\varepsilon)\|d\|_\infty., \qquad (21)
\end{aligned}
$$

where (19) uses **Lemma 3.2**; (20) uses Cauchy-Schwarz; (21) uses $\|A^\star\|_1 = \|p\|_1$. Taking $d = \frac{\nabla\varphi(u,v)}{\|\nabla\varphi(u,v)\|}$ gives $\|\nabla\varphi(u,v)\| \le 4\sqrt{\|p\|_1\varepsilon} + 2\varepsilon$.

Proof of relation 2. We deduce that

$$
\begin{aligned}
|\langle d, S^{-1/2}[\nabla^2\varphi(u,v) - \nabla^2\lambda(u,v)]S^{-1/2}, d\rangle| &= |\langle S^{-1/2}d, [\nabla^2\varphi(u,v) - \nabla^2\lambda(u,v)], S^{-1/2}d\rangle| \\
&= \big|\textstyle\sum_{i,j} a^\star_{ij}[\phi''(\Delta_{ij}) - 1]\big(\frac{d^u_i}{\sqrt{p_i}} - \frac{d^v_j}{\sqrt{q_j}}\big)^2\big| \\
&\le \textstyle\sum_{i,j} a^\star_{ij}\big|\phi''(\Delta_{ij}) - 1\big|\big(\frac{d^u_i}{\sqrt{p_i}} - \frac{d^v_j}{\sqrt{q_j}}\big)^2 \\
&\le \textstyle\sum_{i,j} a^\star_{ij}\sqrt{2\phi(\Delta_{ij})}\big|\frac{d^u_i}{\sqrt{p_i}} - \frac{d^v_j}{\sqrt{q_j}}\big|^2 + \sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})\big|\frac{d^u_i}{\sqrt{p_i}} - \frac{d^v_j}{\sqrt{q_j}}\big|^2,
\end{aligned}
$$

where the last inequality uses $\phi''(\delta) - 1 = \phi'(\delta)$ and **Lemma 3.2**. Hence, we can similarly bound

$$
\begin{aligned}
&\textstyle\sum_{i,j} a^\star_{ij}\sqrt{2\phi(\Delta_{ij})}|\frac{d^u_i}{\sqrt{p_i}} - \frac{d^v_j}{\sqrt{q_i}}|^2 \qquad (22)\\
=\ &\textstyle\sum_{i,j}\sqrt{2a^\star_{ij}\phi(\Delta_{ij})}\sqrt{a^\star_{ij}}|\frac{d^u_i}{\sqrt{p_i}} - \frac{d^v_j}{\sqrt{q_j}}|^2 \\
\le\ & 2\textstyle\sum_{i,j}\sqrt{2a^\star_{ij}\phi(\Delta_{ij})}\Big[\frac{\sqrt{a^\star_{ij}}(d^u_i)^2}{\sqrt{\sum_k a^\star_{ik}}}\frac{1}{\sqrt{p_i}} + \frac{\sqrt{a^\star_{ij}}(d^v_j)^2}{\sqrt{\sum_k a^\star_{kj}}}\frac{1}{\sqrt{q_j}}\Big] \qquad (23)\\
\le\ & \tfrac{2}{\sqrt{\|(p,q)\|_{-\infty}}}\textstyle\sum_{i,j}\sqrt{2a^\star_{ij}\phi(\Delta_{ij})}\Big[\frac{\sqrt{a^\star_{ij}}(d^u_i)^2}{\sqrt{\sum_k a^\star_{ik}}} + \frac{\sqrt{a^\star_{ij}}(d^v_j)^2}{\sqrt{\sum_k a^\star_{kj}}}\Big] \qquad (24)\\
\le\ & \tfrac{2\sqrt{2}}{\sqrt{\|(p,q)\|_{-\infty}}}\sqrt{\textstyle\sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})}\sqrt{\textstyle\sum_i\sum_j \frac{a^\star_{ij}(d^u_i)^4}{\sum_k a^\star_{ik}} + \sum_j\sum_i \frac{a^\star_{ij}(d^v_j)^4}{\sum_k a^\star_{kj}}} \qquad (25)\\
=\ & \tfrac{2\sqrt{2}}{\sqrt{\|(p,q)\|_{-\infty}}}\sqrt{\textstyle\sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})}\sqrt{\textstyle\sum_i (d^u_i)^4 + \sum_j (d^v_j)^4} = \tfrac{2\sqrt{2\varepsilon}}{\sqrt{\|(p,q)\|_{-\infty}}}\|d\|_4^2 \le \tfrac{2\sqrt{2\varepsilon}}{\sqrt{\|(p,q)\|_{-\infty}}}\|d\|^2 \qquad (26)
\end{aligned}
$$

where (23) uses $(a+b)^2 \le 2a^2 + 2b^2$; (24) uses $p_i, q_j \ge \|(p,q)\|_{-\infty}$; (25) applies Cauchy-Schwarz and (26) uses $\|x\|_4^2 \le \|x\|^2$; We also have $\sum_{i,j} a^\star_{ij}\phi(\Delta_{ij})|\frac{d^u_i}{\sqrt{p_i}} - \frac{d^v_j}{\sqrt{q_j}}|^2 \le \frac{2\varepsilon}{\|(p,q)\|_{-\infty}}\|d\|_\infty^2$. Hence for any $d \ne 0$, we have

$$
|\langle d, S^{-1/2}[\nabla^2\varphi(u,v) - \nabla^2\lambda(u,v)]S^{-1/2}, d\rangle| \le 2\sqrt{\tfrac{2\varepsilon}{\|(p,q)\|_{-\infty}}}\|d\|^2 + \tfrac{2\varepsilon}{\|(p,q)\|_{-\infty}}\|d\|_\infty^2
$$

and $\|S^{-1/2}[\nabla^2\varphi(u,v) - \nabla^2\lambda(u,v)]S^{-1/2}\| \le 2\sqrt{\frac{2\varepsilon}{\|(p,q)\|_{-\infty}}} + \frac{2\varepsilon}{\|(p,q)\|_{-\infty}}$.

Proof of relation 3. Finally, we consider the zeroth-order relation. The proof uses the fact that $\nabla\varphi \approx 0$ to restrict (u,v) and deduce a bound that only depends on σ_2. Since $\varepsilon \le \frac{\sigma_2^2 s}{1296} \le s$, we have $\frac{1}{\sqrt{s}}\varepsilon \le \sqrt{\varepsilon}$ and that $\sqrt{\frac{\varepsilon}{s}} \ge \frac{\varepsilon}{s}$. By Taylor theorem, for any optimal $z^\star = (u^\star, v^\star)$, we define $z_t \coloneqq (u^\star + t(u-u^\star), v^\star + t(v-v^\star))$ and

deduce

$$\begin{aligned}\varphi(u,v)-\lambda(u,v) &= \int_0^1(1-t)\langle z-z^\star,[\nabla^2\varphi(z_t)-\nabla^2\varphi(z^\star)](z-z^\star)\rangle \mathrm{d}t\\ &= \int_0^1(1-t)\langle S^{1/2}(z-z^\star),S^{-1/2}[\nabla^2\varphi(z_t)-\nabla^2\varphi(z^\star)]S^{-1/2}S^{1/2}(z-z^\star)\rangle \mathrm{d}t\\ &\geq -(2\sqrt{\tfrac{\varepsilon}{s}}+\tfrac{\varepsilon}{s})\|z-z^\star\|_S^2 \geq -3\sqrt{\tfrac{\varepsilon}{s}}\|z-z^\star\|_S^2. \end{aligned} \tag{27}$$

Next, using $\|\nabla\varphi(z)-\nabla\lambda(z)\|_{S^{-1}}\leq\frac{2}{\sqrt{s}}\varepsilon$ and relation (1) from **Lemma 3.4**, we have

$$\|\nabla\lambda(z)\|_{S^{-1}}\leq\|\nabla\varphi(z)\|_{S^{-1}}+\|\nabla\lambda(z)-\nabla\varphi(z)\|_{S^{-1}}\leq 2\sqrt{\varepsilon}+\tfrac{2}{\sqrt{s}}\varepsilon+\tfrac{2}{\sqrt{s}}\varepsilon\leq 6\sqrt{\varepsilon}.$$

Given that

$$\nabla\lambda(z)=\begin{pmatrix}P(u-u^\star)-R(v-v^\star)\\ Q(v-v^\star)-R^\top(u-u^\star)\end{pmatrix},$$

without loss of generality, we let $(\delta_u,\delta_v)=\nabla\lambda(z)$ and

$$\begin{aligned}P(u-u^\star)&=R(v-v^\star)+\delta_u\\ Q(v-v^\star)&=R^\top(u-u^\star)+\delta_v,\end{aligned}$$

and that $\|\delta_u\|_{P^{-1}}\leq\|(\delta_u,\delta_v)\|_{S^{-1}}=\|\nabla\lambda(z)\|_{S^{-1}}\leq 6\sqrt{\varepsilon}$. Next, we lower bound $\lambda(z)-\lambda(z^\star)$ and deduce that

$$\begin{aligned}\lambda(u,v)-\lambda(u^\star,v^\star)&=\tfrac{1}{2}\|u-u^\star\|_P^2-\langle u-u^\star,R(v-v^\star)\rangle+\tfrac{1}{2}\|v-v^\star\|_Q^2\\ &=\tfrac{1}{2}\|\delta_u\|_{P^{-1}}^2+\tfrac{1}{2}\|v-v^\star\|_Q^2-\tfrac{1}{2}\|v-v^\star\|_{R^\top P^{-1}R}^2 \qquad (28)\\ &=\tfrac{1}{2}\|\delta_u\|_{P^{-1}}^2+\tfrac{1}{2}\|v-v^\star\|_{Q-R^\top P^{-1}R}^2, \qquad (29)\end{aligned}$$

where (28) plugs in $u-u^\star=P^{-1}(R(v-v^\star)+\delta_u)$. Since for any optimal solution $(u^\star,v^\star)$, $(u^\star+\alpha\mathbf{1}_m,v^\star+\alpha\mathbf{1}_n)$ is also optimal, we can pick $(u^\star,v^\star)$ such that $v-v^\star\perp q\Leftrightarrow Q^{1/2}(v-v^\star)\perp Q^{1/2}\mathbf{1}_n$. With this choice, we have

$$\begin{aligned}\tfrac{1}{2}\|v-v^\star\|_{Q-R^\top P^{-1}R}^2&=\tfrac{1}{2}\|Q^{1/2}(v-v^\star)\|_{I-Q^{-1/2}R^\top P^{-1}RQ^{-1/2}}^2\\ &\geq\min_{w\perp Q^{1/2}\mathbf{1}_n,w\neq 0}\tfrac{1}{2}\tfrac{\langle w,(I-Q^{-1/2}R^\top P^{-1}RQ^{-1/2})w\rangle}{\|w\|^2}\|v-v^\star\|_Q^2\\ &=\tfrac{\sigma_2}{2}\|v-v^\star\|_Q^2. \qquad (30)\end{aligned}$$

Next, we deduce that

$$\begin{aligned}\varepsilon&\geq\varphi(u,v)-\varphi(u^\star,v^\star)\\ &=\varphi(u,v)-\lambda(u,v)+\lambda(u,v)-\varphi(u^\star,v^\star)\\ &\geq -3\sqrt{\tfrac{\varepsilon}{s}}\|z-z^\star\|_S^2+\tfrac{1}{2}\|\delta_u\|_{P^{-1}}^2+\tfrac{1}{2}\|v-v^\star\|_{Q-R^\top P^{-1}R}^2 \qquad (31)\\ &\geq -3\sqrt{\tfrac{\varepsilon}{s}}\|z-z^\star\|_S^2+\tfrac{1}{2}\|\delta_u\|_{P^{-1}}^2+\tfrac{\sigma_2}{2}\|v-v^\star\|_Q^2, \qquad (32)\end{aligned}$$

where (31) plugs in (27) and (29); (32) applies (30). Using $P(u-u^\star)=R(v-v^\star)+\delta_u$, we have

$$\begin{aligned}\|u-u^\star\|_P^2&=\|P^{-1}R(v-v^\star)+P^{-1}\delta_u\|_P^2\\ &=\langle P^{-1}R(v-v^\star)+P^{-1}\delta_u,R(v-v^\star)+\delta_u\rangle\\ &=\|v-v^\star\|_{R^\top P^{-1}R}^2+2\langle P^{-1/2}R(v-v^\star),P^{-1/2}\delta_u\rangle+\|\delta_u\|_{P^{-1}}^2\\ &\leq(1+\theta)\|v-v^\star\|_{R^\top P^{-1}R}^2+(1+\tfrac{1}{\theta})\|\delta_u\|_{P^{-1}}^2\end{aligned}$$

for any $\theta > 0$. Hence

$$\begin{aligned}\|z - z^\star\|_S^2 &= \|u - u^\star\|_P^2 + \|v - v^\star\|_Q^2 \\ &\le (1+\theta)\|v - v^\star\|_{R^\top P^{-1} R}^2 + \|v - v^\star\|_Q^2 + (1+\tfrac{1}{\theta})\|\delta_u\|_{P^{-1}}^2 \\ &= (1+\theta)\|Q^{1/2}(v - v^\star)\|_{Q^{-1/2}R^\top P^{-1} R Q^{-1/2}}^2 + (1+\tfrac{1}{\theta})\|\delta_u\|_{P^{-1}}^2 \\ &\le (2+\theta)\|v - v^\star\|_Q^2 + (1+\tfrac{1}{\theta})\|\delta_u\|_{P^{-1}}^2.\end{aligned}$$

since $I \succeq Q^{-1/2}R^\top P^{-1} R Q^{-1/2}$. Plugging the relation back into (32), we have

$$\begin{aligned}\varepsilon \ge\ & -3\sqrt{\tfrac{\varepsilon}{s}}\|z - z^\star\|_S^2 + \tfrac{1}{2}\|\delta_u\|_{P^{-1}}^2 + \tfrac{\sigma_2}{2}\|v - v^\star\|_Q^2 \\ \ge\ & -3\sqrt{\tfrac{\varepsilon}{s}}\big[(2+\theta)\|v - v^\star\|_Q^2 + (1+\tfrac{1}{\theta})\|\delta_u\|_{P^{-1}}^2\big] + \tfrac{1}{2}\|\delta_u\|_{P^{-1}}^2 + \tfrac{\sigma_2}{2}\|v - v^\star\|_Q^2 \\ =\ & \big[\tfrac{\sigma_2}{2} - (6+3\theta)\sqrt{\tfrac{\varepsilon}{s}}\big]\|v - v^\star\|_Q^2 + \big[\tfrac{1}{2} - 3\sqrt{\tfrac{\varepsilon}{s}}(1+\tfrac{1}{\theta})\big]\|\delta_u\|_{P^{-1}}^2\end{aligned}$$

Taking $\theta = 1$ and using $\varepsilon \le \frac{s\sigma_2^2}{1296}$, we have

$$\tfrac{1}{2} - 6\sqrt{\tfrac{\varepsilon}{s}} \ge 0, \quad \text{and} \quad 9\sqrt{\tfrac{\varepsilon}{s}} \le \tfrac{\sigma_2}{4},$$

giving $\varepsilon \ge \frac{\sigma_2}{4}\|v - v^\star\|_Q^2$ and that $\|v - v^\star\|_Q^2 \le \frac{4\varepsilon}{\sigma_2}$. Finally,

$$\|z - z^\star\|_S^2 \le 3\|v - v^\star\|_Q^2 + 2\|\delta_u\|_{P^{-1}}^2 \le \tfrac{12}{\sigma_2}\varepsilon + 72\varepsilon = 12(\tfrac{1}{\sigma_2} + 6)\varepsilon \le \tfrac{84}{\sigma_2}\varepsilon.$$

and that

$$|\varphi(u,v) - \lambda(u,v)| \le (2\sqrt{\tfrac{\varepsilon}{s}} + \tfrac{\varepsilon}{s})\|z - z^\star\|_S^2 \le \tfrac{168}{\sigma_2}(\tfrac{1}{\sqrt{s}}\varepsilon^{3/2} + \tfrac{1}{s}\varepsilon^2).$$

This completes the proof.

Remark 7. Symmetrically, we can also show that $\|u - u^\star\|_P^2 \le \frac{4\varepsilon}{\sigma_2}$ by picking $u^\star$ such that $u - u^\star \perp p$. But note that we may not simultaneously enforce both $u - u^\star \perp p$ and $v - v^\star \perp q$ with the same $z^\star$.

A.8 Proof of Lemma 3.5

Let $\varepsilon = \varphi(u,v) - \varphi(u^\star, v^\star)$. Fix v and define $\alpha(\cdot) \coloneqq \varphi(\cdot, v)$ and $\ell(\cdot) \coloneqq \lambda(\cdot, v)$. We have $\nabla^2\alpha(u) = \nabla_{uu}^2\varphi(u,v)$. By **Lemma 3.3**, we have

$$\ell(u - P^{-1}\nabla\ell(u)) - \varphi(u^\star, v^\star) \le (1 - \sigma_2)[\ell(u) - \varphi(u^\star, v^\star)].$$

By **Lemma 3.4**, we have

$$\begin{aligned}\|\nabla\alpha(u) - \nabla\ell(u)\|_{S^{-1}} &= \|\nabla_u\varphi(u,v) - \nabla_u\lambda(u,v)\|_{S^{-1}} \\ &\le \|\nabla\varphi(u,v) - \nabla\lambda(u,v)\|_{S^{-1}} \le \tfrac{2}{\sqrt{s}}\varepsilon.\end{aligned}$$

Then we deduce that

$$\begin{aligned}
\varphi(u^+, v) - \varphi(u^\star, v^\star) &= \alpha(u^+) - \varphi(u^\star, v^\star) \\
&= \min_u \alpha(u) - \varphi(u^\star, v^\star) \\
&\le \alpha(u - P^{-1}\nabla\alpha(u)) - \varphi(u^\star, v^\star) \\
&= \alpha(u - P^{-1}\nabla\alpha(u)) - \ell(u - P^{-1}\nabla\alpha(u)) + \ell(u - P^{-1}\nabla\alpha(u)) - \varphi(u^\star, v^\star) \\
&= \underbrace{\alpha(u - P^{-1}\nabla\alpha(u)) - \ell(u - P^{-1}\nabla\alpha(u))}_{\Delta_1} + \underbrace{\ell(u - P^{-1}\nabla\alpha(u)) - \ell(u - P^{-1}\nabla\ell(u))}_{\Delta_2} \\
&\quad + \underbrace{\ell(u - P^{-1}\nabla\ell(u)) - \varphi(u^\star, v^\star)}_{\Delta_3}.
\end{aligned}$$

Now we bound Δ_1, Δ_2, and Δ_3 respectively. For Δ_1, if $\alpha(u - P^{-1}\nabla\alpha(u)) \le \alpha(u)$, then $\varphi(u - P^{-1}\nabla\alpha(u), v) \le \varphi(u, v)$, and we have

$$\Delta_1 \le \tfrac{168}{\sigma_2}(\tfrac{1}{\sqrt{s}}\varepsilon^{3/2} + \tfrac{1}{s}\varepsilon^2).$$

To show $\alpha(u - P^{-1}\nabla\alpha(u)) \le \alpha(u)$, we first note that $\nabla^2\alpha(u) = \nabla^2_{uu}\varphi(u, v)$ and with $\varepsilon \le \frac{s\sigma_2^2}{1296}$, we have $-2\sqrt{\frac{\varepsilon}{s}} + \frac{2}{s}\varepsilon \le \frac{1}{2}$, which implies

$$\begin{aligned}
P^{-1/2}\nabla^2\alpha(u)P^{-1/2} &= P^{-1/2}[\nabla^2\ell(u) + \nabla^2\alpha(u) - \nabla^2\ell(u)]P^{-1/2} \\
&\preceq I + \|P^{-1/2}[\nabla^2\alpha(u) - \nabla^2\ell(u)]P^{-1/2}\|I \\
&\preceq I + \|S^{-1/2}[\nabla^2\varphi(u, v) - \nabla^2\lambda(u, v)]S^{-1/2}\|I \\
&\preceq \tfrac{3}{2}I
\end{aligned}$$

and $\nabla^2\alpha(u) \preceq \frac{3}{2}P$. Next, consider $\gamma(\theta) \coloneqq \alpha(u - \theta P^{-1}\nabla\alpha(u))$ and let $\theta_{\max} = \sup_\theta\{\theta > 0 : \gamma(\theta) = \alpha(u)\}$. $\theta_{\max}$ is well-defined since $P^{-1} \succ 0$ and $P^{-1}\nabla\alpha(u)$ is a descent direction of α at u. By convexity, we have $\alpha(u - \theta P^{-1}\nabla\alpha(u)) \le \alpha(u)$ for all $\theta \le \theta_{\max}$. Then we deduce that

$$\begin{aligned}
\alpha(u) &= \gamma(\theta_{\max}) \\
&= \alpha(u) - \theta_{\max}\langle\nabla\alpha(u), P^{-1}\nabla\alpha(u)\rangle \\
&\quad + \theta^2_{\max}\textstyle\int_0^1\langle\nabla\alpha(u), P^{-1}\nabla^2\alpha(u - t\theta_{\max}P^{-1}\nabla\alpha(u))P^{-1}\nabla\alpha(u)\rangle(1 - t)\mathrm{d}t \\
&\le \alpha(u) - \theta_{\max}\langle\nabla\alpha(u), P^{-1}\nabla\alpha(u)\rangle + \theta^2_{\max}\textstyle\int_0^1\langle\nabla\alpha(u), P^{-1}(\frac{3}{2}P)P^{-1}\nabla\alpha(u)\rangle(1 - t)\mathrm{d}t, \\
&= \alpha(u) - \theta_{\max}\|\nabla\alpha(u)\|^2_{P^{-1}} + \tfrac{3}{4}\theta^2_{\max}\|\nabla\alpha(u)\|^2_{P^{-1}},
\end{aligned}$$

where the inequality holds since $\alpha(u - t\theta_{\max}P^{-1}\nabla\alpha(u)) \le \alpha(u)$. Given

$$\tfrac{3}{4}\theta^2_{\max}\|\nabla\alpha(u)\|^2_{P^{-1}} \ge \theta_{\max}\|\nabla\alpha(u)\|^2_{P^{-1}}, \tag{33}$$

we have $\theta_{\max} \ge \frac{4}{3}$. Therefore, $\alpha(u - P^{-1}\nabla\alpha(u)) = \gamma(1) \le \gamma(\theta_{\max}) = \alpha(u)$, and it implies

$$\Delta_1 \le \tfrac{336}{\sigma_2\sqrt{s}}[\varphi(u, v) - \varphi(u^\star, v^\star)]^{3/2}.$$

For Δ_2, since $\ell(u)$ is a quadratic function minimized at $u - P^{-1}\nabla\ell(u)$, we have

$$\begin{aligned}
\Delta_2 &= \ell(u - P^{-1}\nabla\alpha(u)) - \ell(u - P^{-1}\nabla\ell(u)) \\
&= \tfrac{1}{2}\|\nabla\alpha(u) - \nabla\ell(u)\|^2_{P^{-1}} \\
&\le \tfrac{1}{2}\|\nabla\varphi(u, v) - \nabla\lambda(u, v)\|^2_{S^{-1}} \le \tfrac{2}{s}\varepsilon^2.
\end{aligned}$$

For Δ_3, we have

$$\begin{aligned}\ell(u - P^{-1}\nabla\ell(u)) - \varphi(u^\star, v^\star) &\le (1-\sigma_2)[\ell(u) - \varphi(u^\star, v^\star)] \\ &= (1-\sigma_2)[\ell(u) - \alpha(u) + \alpha(u) - \varphi(u^\star, v^\star)] \\ &\le (1-\sigma_2)[\alpha(u) - \varphi(u^\star, v^\star)] + |\ell(u) - \alpha(u)| \\ &\le (1-\sigma_2)[\alpha(u) - \varphi(u^\star, v^\star)] + \tfrac{168}{\sigma_2}(\tfrac{1}{\sqrt{s}}\varepsilon^{3/2} + \tfrac{1}{s}\varepsilon^2)\end{aligned}$$

Putting the relations together, we have

$$\begin{aligned}&\varphi(u^+, v) - \varphi(u^\star, v^\star) \\ &\le (1-\sigma_2)[\varphi(u,v) - \varphi(u^\star, v^\star)] + \tfrac{336}{\sqrt{s}\sigma_2}[\varphi(u,v) - \varphi(u^\star, v^\star)]^{3/2} + \tfrac{338}{s\sigma_2}[\varphi(u,v) - \varphi(u^\star, v^\star)]^2,\end{aligned}$$

This completes the whole proof.

A.9 Proof of Theorem 3.2

For $\varepsilon \le \frac{s\sigma_2^4}{1344^2}$, we have

$$\tfrac{336}{\sigma_2\sqrt{s}}[\varphi(u,v) - \varphi(u^\star, v^\star)]^{3/2} \le \tfrac{\sigma_2}{4}[\varphi(u,v) - \varphi(u^\star, v^\star)].$$

For $\varepsilon \le \frac{s\sigma_2^2}{1352}$, we have

$$\tfrac{338}{s\sigma_2}[\varphi(u,v) - \varphi(u^\star, v^\star)]^2 \le \tfrac{\sigma_2}{4}[\varphi(u,v) - \varphi(u^\star, v^\star)].$$

Hence $\varphi(u^+, v) - \varphi(u^\star, v^\star) \le (1 - \frac{\sigma_2}{2})[\varphi(u,v) - \varphi(u^\star, v^\star)]$, and it remains to bound

$$\begin{aligned}\varphi(u^1, v^1) - \varphi(u^\star, v^\star) &= \varphi(\mathbf{0}_m, \mathbf{0}_n) - \varphi(u^\star, v^\star) \\ &\le \|\nabla\varphi(\mathbf{0}_m, \mathbf{0}_n)\|_1 \|(u^\star, v^\star)\|_\infty \\ &\le [\|A\mathbf{1}_m - p\|_1 + \|A^\top \mathbf{1}_n - q\|_1]D \\ &\le 2[\|A\|_1 + \|p\|_1]D.\end{aligned}$$

It a total complexity of

$$\begin{aligned}K_\varepsilon &\le \tfrac{2\|p\|_1 D^2}{\frac{\sigma_2^4}{1344^2}} + \tfrac{2}{\sigma_2}\log(\tfrac{2[\|A\|_1 + \|p\|_1]D}{\varepsilon}) \\ &= \tfrac{2\cdot 1344^2 \|p\|_1 D^2}{\sigma_2^4} + \tfrac{2}{\sigma_2}\log(2[\|A\|_1 + \|p\|_1]D) + \tfrac{2}{\sigma_2}\log(\tfrac{1}{\varepsilon})\end{aligned}$$

To ensure $\|\nabla\varphi(u,v)\| \le \varepsilon'$, it suffices to have $\|\nabla\varphi(u,v)\| \le 4\sqrt{\|p\|_1\varepsilon} + 2\varepsilon \le \varepsilon'$, and it suffices to take

$$\varepsilon \le \min\{\tfrac{(\varepsilon')^2}{64\|p\|_1}, \tfrac{\varepsilon'}{4}\}.$$

Since each step we analyze is a half-iteration, dividing the complexity by 2 completes the proof.

A.10 Proof of Theorem 3.2

Fix parameters $\theta \in (0,1)$ and $\rho_1, \rho_2, c_1, c_2 > 0$. Consider the matrix scaling problem with

$$A = \begin{pmatrix} \frac{1}{\rho_1 c_1} & \frac{\theta}{\rho_1 c_2} \\ \frac{\theta}{\rho_2 c_1} & \frac{1}{\rho_2 c_2} \end{pmatrix}, \qquad p = q = \begin{pmatrix} 1+\theta \\ 1+\theta \end{pmatrix},$$

so that the solution is given by

$$A^\star = D_1^\star A D_2^\star = \left(\begin{smallmatrix} 1 & \theta \\ \theta & 1 \end{smallmatrix}\right) \quad \text{with} \quad D_1^\star = \left(\begin{smallmatrix} \rho_1 & 0 \\ 0 & \rho_2 \end{smallmatrix}\right), \qquad D_2^\star = \left(\begin{smallmatrix} c_1 & 0 \\ 0 & c_2 \end{smallmatrix}\right).$$

Let $M = P^{-1/2} A^\star Q^{-1} (A^\star)^\top P^{-1/2}$ and $L = I - M$. Since $P = Q = (1+\theta)I$,

$$M = \tfrac{A^\star (A^\star)^\top}{(1+\theta)^2} = \tfrac{1}{(1+\theta)^2} \left(\begin{smallmatrix} 1+\theta^2 & 2\theta \\ 2\theta & 1+\theta^2 \end{smallmatrix}\right).$$

Its top eigenvector is $\mathbf{1}_2$ with eigenvalue 1, and its second largest eigenvalue is

$$\mu_2 = \operatorname{tr}(M) - 1 = \tfrac{2(1+\theta^2)-(1+\theta)^2}{(1+\theta)^2} = (\tfrac{1-\theta}{1+\theta})^2.$$

Therefore,

$$\sigma_2(\theta) = \lambda_2(L) = 1 - \mu_2 = \tfrac{4\theta}{(1+\theta)^2}.$$

By the additive shift invariance of the dual potential, we can parametrize the deviation from the optimum by a single scalar s, taking $u(s) = u^\star + (s, -s)$ and letting $v(s)$ be the exact minimizer $v(s) = \arg\min_v \varphi(u(s), v)$ (a half Sinkhorn step). Next, starting at A and (u, v) is equivalent to starting at $A^\star$ and $(u - u^\star, v - v^\star)$, with identical dual potential gaps and imbalance sequences. We adopt this change of variables, and substituting $v(s)$ into the dual potential gap $G(s) := \varphi(u(s), v(s)) - \varphi(u^\star, v^\star)$ gives the closed form

$$G(s) = (1+\theta) \log \tfrac{(1+\theta)^2}{(e^s + \theta e^{-s})(\theta e^s + e^{-s})} = \tfrac{4\theta}{1+\theta} s^2 + O(s^4)$$

In particular, G is locally a positive-definite quadratic in the imbalance s. Now we proceed to updating u from $v(s)$ by $u^+ = \log p - \log(A^\star \mathrm{e}^{-v(s)})$, and compute $s^+ := \frac{1}{2} \log(\mathrm{e}^{u_1^+} / \mathrm{e}^{u_2^+})$. A direct differentiation of the mapping $s \to s^+$ at $s = 0$ gives, for a single half-step,

$$s^+ = -\big(\tfrac{1-\theta}{1+\theta}\big)^2 s + \mathcal{O}(s^2) = -(1-\sigma_2)s + \mathcal{O}(s^2),$$

and substituting into $G(s)$, the dual potential gap $G_{k+1} = G(s^+)$ after one full Sinkhorn iteration is given by

$$G_{k+1} = (1-\sigma_2)^2 G_k (1 + o(1)).$$

This matches **Lemma 3.5** with equality, confirming that the local factor $(1-\sigma_2)^2$ per iteration is attained on this family and cannot be improved. In conclusion, after a finite, ε-independent burn-in $c(\theta)$ (controlled by **Theorem 3.1** and the threshold of **Lemma 3.5**) the iterates enter the local regime, where $\log(1/G_k)$ grows by at most $-2\log(1-\sigma_2) + o(1) = 2\sigma_2 + o(\sigma_2)$ per cycle. Hence reaching $G_k \le \varepsilon$ requires

$$k \;\ge\; c(\theta) + \tfrac{\log(1/\varepsilon)}{-2\log(1-\sigma_2)} \;=\; c(\theta) + \tfrac{1}{2\sigma_2(\theta)} \log(\tfrac{1}{\varepsilon})(1 + o(1)),$$

matching the upper bound $\mathcal{O}\big(c + \frac{1}{\sigma_2} \log(\frac{1}{\varepsilon})\big)$ of **Theorem 3.2** up to the additive constant.

B Proof of results in Section 4

B.1 Auxiliary results

Lemma B.1. *Suppose f is L-smooth and convex. Then there exists an algorithm that outputs x^K such that $\|\nabla f(x^K)\|^2 \le \varepsilon$ in $K_\varepsilon := \lceil (\frac{528 L^2 \|x^1 - x^\star\|^2}{\varepsilon})^{1/4} \rceil$ iterations, where $x^\star$ is an optimal solution.*

Proof. The algorithm is `FISTA` + `FISTA-G`. Invoking Corollary 1 from [26] completes the proof. □

Lemma B.2. *Suppose $f(x) = \frac{1}{n}\sum_{j=1}^n f_j(x)$ with each f_j is $L_{\max}$-smooth and σ-strongly convex. Then there exists an algorithm that outputs x^K such that $\mathbb{E}[f(x^K)] - f(x^\star) \le \varepsilon$ in $K_\varepsilon := \mathcal{O}\big(\big(n + \sqrt{\frac{nL_{\max}}{\sigma}}\big)\log(\frac{f(x^1)-f(x^\star)}{\varepsilon})\big)$ stochastic gradient oracle calls.*

Proof. The algorithm is `Katyusha` [1]. Invoking Theorem 2.1 of [1] completes the proof. □

B.2 Proof of Lemma 4.1

Let $a_{[:,j]}$ denote the j-th column of A. Through direct calculation, we have

$$\begin{aligned}\zeta(u) &= \varphi(u, -\log(q/A^\top \mathrm{e}^u)) \\ &= \textstyle\sum_{j=1}^n q_j \log\langle a_{[:,j]}, \mathrm{e}^u\rangle - \langle p, u\rangle - \sum_{j=1}^n q_j \log q_j + \langle \mathbf{1}_n, q\rangle\end{aligned}$$

and that

$$\begin{aligned}\nabla\zeta(u) &= \textstyle\sum_{j=1}^n \frac{q_j}{\langle a_{[:,j]}, \mathrm{e}^u\rangle}\mathrm{e}^U a_{[:,j]} - p \\ \nabla^2\zeta(u) &= \mathcal{D}(\textstyle\sum_{j=1}^n q_j \frac{\mathrm{e}^U a_{[:,j]}}{\langle a_{[:,j]}, \mathrm{e}^u\rangle}) - \sum_{j=1}^n q_j \frac{\mathrm{e}^U a_{[:,j]} a_{[:,j]}^\top \mathrm{e}^U}{\langle a_{[:,j]}, \mathrm{e}^u\rangle^2} \\ &= \mathcal{D}(\textstyle\sum_{j=1}^n q_j \frac{\mathrm{e}^U a_{[:,j]}}{\langle \mathbf{1}_m, \mathrm{e}^U a_{[:,j]}\rangle}) - \sum_{j=1}^n q_j \frac{\mathrm{e}^U a_{[:,j]} a_{[:,j]}^\top \mathrm{e}^U}{\langle \mathbf{1}_m, \mathrm{e}^U a_{[:,j]}\rangle^2} \\ &=: \textstyle\sum_{j=1}^n q_j[\mathcal{D}(\sigma_j) - \sigma_j\sigma_j^\top],\end{aligned}$$

where we define $\sigma_j := \frac{\mathrm{e}^U a_{[:,j]}}{\langle \mathbf{1}_m, \mathrm{e}^U a_{[:,j]}\rangle}$ and $\mathbf{0}_m \le \sigma_j \le \mathbf{1}_m$ by nonnegativity of A and e^U, and the fact that A must have at least nonzero per row. To establish smoothness, we have

$$\|\nabla^2\zeta(u)\| \le \textstyle\sum_{j=1}^n q_j \|\mathcal{D}(\sigma_j) - \sigma_j\sigma_j^\top\| \le \frac{\|q\|_1}{2} = \frac{\|p\|_1}{2},$$

where we use Lipschitzness of softmax functions [30]: $\|\mathcal{D}(x) - xx^\top\| \le \frac{1}{2}$ for $x \in \{x \ge \mathbf{0}_m : \langle \mathbf{1}_m, x\rangle = 1\}$. To establish Lipschitz Hessian, we similarly deduce, for any x, y, that

$$\begin{aligned}\|\nabla^2\zeta(x) - \nabla^2\zeta(y)\| &= \|\textstyle\sum_{j=1}^n q_j[\mathcal{D}(\sigma_j) - \sigma_j\sigma_j^\top] - \sum_{j=1}^n q_j[\mathcal{D}(\sigma_j') - \sigma_j'(\sigma_j')^\top]\| \\ &\le \textstyle\sum_{j=1}^n q_j\|\sigma_j - \sigma_j'\|_\infty + \sum_{j=1}^n q_j\|\sigma_j\sigma_j^\top - \sigma_j'(\sigma_j')^\top\| \\ &= \textstyle\sum_{j=1}^n q_j\|\sigma_j - \sigma_j'\|_\infty + \sum_{j=1}^n q_j\|\sigma_j\sigma_j^\top - \sigma_j(\sigma_j')^\top + \sigma_j(\sigma_j')^\top - \sigma_j'(\sigma_j')^\top\| \\ &\le \textstyle\sum_{j=1}^n q_j\|\sigma_j - \sigma_j'\|_\infty + \sum_{j=1}^n q_j[\|\sigma_j\| + \|\sigma_j'\|]\|\sigma_j - \sigma_j'\| \\ &\le 3\textstyle\sum_{j=1}^n q_j\|\sigma_j - \sigma_j'\|, \qquad (34)\end{aligned}$$

where $\sigma_j = \frac{\mathrm{e}^X a_{[:,j]}}{\langle \mathbf{1}_m, \mathrm{e}^X a_{[:,j]}\rangle}$ and $\sigma_j' = \frac{\mathrm{e}^Y a_{[:,j]}}{\langle \mathbf{1}_m, \mathrm{e}^Y a_{[:,j]}\rangle}$ are both outcomes of weighted softmax; (34) uses the fact that $\|\sigma_j\| \le 1$, and Lipschitzness of softmax implies

$$\|\sigma_j - \sigma_j'\| \le \tfrac{1}{2}\|x - y\|.$$

Plugging back, we have $\|\nabla^2\zeta(x) - \nabla^2\zeta(y)\| \le \frac{3}{2}\sum_{j=1}^n q_j\|x-y\| = \frac{3}{2}\|p\|_1\|x-y\|$. Finally, the quadratic growth $\frac{\sigma_2}{4}\|\Pi(u-u^\star)\|_P^2 \le \zeta(u) - \zeta(u^\star)$ follows from the symmetric argument in **Remark 7**. To show the PL inequality,

we deduce that, taking $u^\star$ such that $\Pi(u-u^\star)=u-u^\star$, that,

$$\begin{aligned}\zeta(u)-\zeta(u^\star) &\le \langle \nabla\zeta(u), u-u^\star\rangle && (35)\\ &= \langle P^{-1/2}\nabla\zeta(u), P^{1/2}(u-u^\star)\rangle \\ &\le \|\nabla\zeta(u)\|_{P^{-1}}\|\Pi(u-u^\star)\|_P && (36)\\ &\le \|\nabla\zeta(u)\|_{P^{-1}}\sqrt{\tfrac{4}{\sigma_2}[\zeta(u)-\zeta(u^\star)]}, && (37)\end{aligned}$$

where (35) uses convexity of ζ; (36) uses Cauchy-Schwarz and (37) uses quadratic growth. Squaring and dividing both sides by $\zeta(u)-\zeta(u^\star)$ gives the desired relation. Finally, with **Lemma 3.4**, we have

$$\|\nabla\zeta(u)\|^2_{P^{-1}} = \|\nabla\varphi(u,v(u))\|^2_{S^{-1}} \le 4\varepsilon + \tfrac{8}{\sqrt{s}}\varepsilon^{3/2} + \tfrac{4}{s}\varepsilon^2$$

and with $\varepsilon \le s$, we have $\frac{8}{\sqrt{s}}\varepsilon^{3/2} \le 8\varepsilon$ and $\frac{4}{s}\varepsilon^2 \le 4\varepsilon$, giving $\|\nabla\zeta(u)\|^2_{P^{-1}} \le 16\varepsilon = 16[\zeta(u)-\zeta(u^\star)]$. To establish lower and upper bounds on the spectrum, using $\|\Pi(u-u^\star)\|_P \le \sqrt{\frac{4\varepsilon}{\sigma_2}}$ and by Weyl's inequality,

$$\lambda_2(P^{-1/2}\nabla^2\zeta(u)P^{-1/2}) \ge \lambda_2(P^{-1/2}\nabla^2\zeta(u^\star)P^{-1/2}) - \tfrac{H}{s}\|\Pi(u-u^\star)\|_P \ge \sigma_2 - \sqrt{\tfrac{4H^2\varepsilon}{s^2\sigma_2}}.$$

$$\lambda_m(P^{-1/2}\nabla^2\zeta(u)P^{-1/2}) \le \lambda_m(P^{-1/2}\nabla^2\zeta(u^\star)P^{-1/2}) + \tfrac{H}{s}\|\Pi(u-u^\star)\|_P \le 1 + \sqrt{\tfrac{4H^2\varepsilon}{s^2\sigma_2}}$$

Hence for $\varepsilon \le \frac{\sigma_2^3 s^2}{16H^2}$, $\sigma_2 - \sqrt{\frac{4H^2\varepsilon}{s^2\sigma_2}} \ge \frac{\sigma_2}{2}$ and this completes the proof.

B.3 Proof of Theorem 4.1

Given that ζ is L-smooth convex with $L = \frac{1}{2}\|p\|_1$, invoking **Lemma B.1** with $L = \frac{1}{2}\|p\|_1$ completes the proof. The inequality uses $\|u^\star\| \le \sqrt{n}\|u^\star\|_\infty$. Computing $\nabla\zeta(u)$ involves finding $v = p/A^\top u$, whose arithmetic complexity is the same as half of a `SK` iteration.

B.4 Proof of Theorem 4.2

It suffices to adopt Nesterov's regularization technique for making gradient small [32]. Define

$$\zeta_\sigma(u) \coloneqq \zeta(u) + \tfrac{\sigma}{2}\|u\|^2.$$

Since ζ is L-smooth and convex, ζ_σ is $(L+\sigma)$-smooth and σ-strongly convex. Denote $u^\star_\sigma = \arg\min_u \zeta_\sigma(u)$. By the optimality condition, $\nabla\zeta_\sigma(u^\star_\sigma) = \nabla\zeta(u^\star_\sigma) + \sigma u^\star_\sigma = 0$ and with quadratic growth,

$$\zeta_\sigma(u^\star) - \zeta_\sigma(u^\star_\sigma) \ge \tfrac{\sigma}{2}\|u^\star_\sigma - u^\star\|^2.$$

By definition, the above relation implies

$$\tfrac{\sigma}{2}\|u^\star_\sigma - u^\star\|^2 \le \zeta_\sigma(u^\star) - \zeta_\sigma(u^\star_\sigma) = \zeta(u^\star) + \tfrac{\sigma}{2}\|u^\star\|^2 - \zeta(u^\star_\sigma) - \tfrac{\sigma}{2}\|u^\star_\sigma\|^2 \le \tfrac{\sigma}{2}\|u^\star\|^2 - \tfrac{\sigma}{2}\|u^\star_\sigma\|^2$$

since $\zeta(u^\star) \le \zeta(u^\star_\sigma)$. Rearranging, we have $\|u^\star_\sigma\|^2 \le \|u^\star_\sigma - u^\star\|^2 + \|u^\star_\sigma\|^2 \le \|u^\star\|^2$ and $\|\nabla\zeta(u^\star_\sigma)\| = \sigma\|u^\star_\sigma\| \le \sigma\|u^\star\|$. Now suppose we run `Katyusha` on ζ_σ, which, by **Lemma B.2**, outputs $\hat{u}$ such that

$$\mathbb{E}[\zeta_\sigma(\hat{u})] - \zeta_\sigma(u^\star_\sigma) \le \hat{\varepsilon}$$

in $\mathcal{O}\big((n+\sqrt{\frac{nL_{\max}}{\sigma}})\log(\frac{\zeta_\sigma(u^1)-\zeta_\sigma(u_\sigma^\star)}{\hat{\varepsilon}})\big)$ stochastic gradient calls, where $L_{\max}=\frac{n}{2}\|p\|_\infty$. Now we deduce that

$$\begin{aligned}
\mathbb{E}[\|\nabla\zeta(\hat{u})\|] &= \mathbb{E}[\|\nabla\zeta(\hat{u})-\nabla\zeta(u_\sigma^\star)+\nabla\zeta(u_\sigma^\star)\|] \\
&\le \|\nabla\zeta(u_\sigma^\star)\|+\mathbb{E}[\|\nabla\zeta(\hat{u})-\nabla\zeta(u_\sigma^\star)\|] &(38)\\
&\le \sigma\|u^\star\|+L\mathbb{E}[\|\hat{u}-u_\sigma^\star\|] &(39)\\
&\le \sigma\|u^\star\|+L\sqrt{\mathbb{E}[\|\hat{u}-u_\sigma^\star\|^2]} &(40)\\
&\le \sigma\|u^\star\|+L\sqrt{\tfrac{2}{\sigma}\mathbb{E}[\zeta_\sigma(\hat{u})-\zeta_\sigma(u_\sigma^\star)]} &(41)\\
&= \sigma\|u^\star\|+L\sqrt{\tfrac{2}{\sigma}\hat{\varepsilon}},
\end{aligned}$$

where (38) uses triangle inequality; (39) uses the previous relation $\|\nabla\zeta(u_\sigma^\star)\|\le\sigma\|u^\star\|$ and L-smoothness; (40) uses Jensen's inequality $\mathbb{E}[|X|]\le\sqrt{\mathbb{E}[X^2]}$; (41) uses quadratic growth. Taking $\sigma=\frac{\varepsilon}{2\|u^\star\|}$ and $\hat{\varepsilon}=\frac{\sigma}{8L^2}=\frac{\varepsilon}{16L^2\|u^\star\|}$, we have

$$\mathbb{E}[\|\nabla\zeta(\hat{u})\|]\le\tfrac{\varepsilon}{2\|u^\star\|}\|u^\star\|+L\sqrt{\tfrac{2}{\sigma}\tfrac{\sigma}{8L^2}}=\tfrac{\varepsilon}{2}+\tfrac{\varepsilon}{2}=\varepsilon,$$

which gives a total complexity of

$$\begin{aligned}
\mathcal{O}\big((n+\sqrt{\tfrac{nL_{\max}}{\sigma}})\log\big(\tfrac{\zeta_\sigma(u^1)-\zeta_\sigma(u_\sigma^\star)}{\hat{\varepsilon}}\big)\big) &= \mathcal{O}\big((n+\sqrt{\tfrac{2nL_{\max}\|u^\star\|}{\varepsilon}})\log\big(\tfrac{16L^2\|u^\star\|[\zeta_\sigma(u^1)-\zeta_\sigma(u_\sigma^\star)]}{\varepsilon}\big)\big)\\
&= \mathcal{O}\big((n+n^{3/4}\sqrt{\tfrac{2L_{\max}\|u^\star\|_\infty}{\varepsilon}})\log\big(\tfrac{16L^2\|u^\star\|[\zeta_\sigma(u^1)-\zeta_\sigma(u_\sigma^\star)]}{\varepsilon}\big)\big)\\
&= \tilde{\mathcal{O}}\big(n+n^{3/4}\sqrt{\tfrac{2L_{\max}\|u^\star\|_\infty}{\varepsilon}}\big),
\end{aligned}$$

where we hide the terms in log since with $u^1=\mathbf{0}$, we have

$$\zeta_\sigma(u^1)-\zeta_\sigma(u_\sigma^\star)=\zeta(\mathbf{0}_m)-\zeta_\sigma(u_\sigma^\star)\le\zeta(\mathbf{0}_m)-\zeta(u_\sigma^\star)\le\zeta(\mathbf{0}_m)-\zeta(u^\star)\le\|u^\star\|_\infty\|\nabla\zeta(\mathbf{0}_m)\|_1\le\|u^\star\|_\infty(\|p\|_1+\|A\|_1)$$

Plugging in $L_{\max}=\frac{n}{2}\|p\|_\infty$ gives $\tilde{\mathcal{O}}(n+n^{5/4}\sqrt{\frac{\|p\|_\infty\|u^\star\|_\infty}{\varepsilon}})$ complexity of stochastic gradients. Given that the amortized complexity of each iteration is $\mathcal{O}(\frac{\mathrm{nnz}(A)}{n})$, we have the expected total arithmetic complexity given by

$$\tilde{\mathcal{O}}(\mathrm{nnz}(A)+\mathrm{nnz}(A)\cdot n^{1/4}\sqrt{\tfrac{\|p\|_\infty\|u^\star\|_\infty}{\varepsilon}}).$$

Noticing that $\|u^\star\|\le D$, this completes the proof.

B.5 Proof of Lemma 4.2

Convexity of h_u follows from the composition rule of convex functions. To show smoothness, it suffices to show that $P^{-1/2}\nabla^2 h_u(w)P^{-1/2}\preceq L\cdot I$, which holds since

$$\nabla^2 h_u(w)=\mathcal{D}\big(\tfrac{\nabla\zeta(u)}{\|P^{-1/2}\nabla\zeta(u)\|}\big)\nabla^2\zeta(u-\mathcal{D}(w)\nabla\zeta(u))\mathcal{D}\big(\tfrac{\nabla\zeta(u)}{\|P^{-1/2}\nabla\zeta(u)\|}\big)$$

and for some $\hat{w}$, we have

$$P^{-1/2}\nabla^2 h_u(w)P^{-1/2}=\mathcal{D}\big(\tfrac{P^{-1/2}\nabla\zeta(u)}{\|P^{-1/2}\nabla\zeta(u)\|}\big)\nabla^2\zeta(\hat{w})\mathcal{D}\big(\tfrac{P^{-1/2}\nabla\zeta(u)}{\|P^{-1/2}\nabla\zeta(u)\|}\big)\preceq L\cdot I.$$

Given $\zeta(u)-\zeta(u^\star)\le\frac{s\sigma_2^2}{1296}$, by **Lemma 3.4**, $\|\nabla\zeta(u)\|_{P^{-1}}=\|\nabla\varphi(u,v(u))\|_{S^{-1}}\le 2\sqrt{\varepsilon}+\frac{2}{\sqrt{s}}\varepsilon$ and

$$
\begin{aligned}
P^{-1/2}\nabla^2\zeta(u)P^{-1/2} &\preceq P^{-1/2}(\mathcal{D}(\nabla\zeta(u)+p))P^{-1/2} \\
&\preceq \mathcal{D}(P^{-1}\nabla\zeta(u)) + I \\
&\preceq \tfrac{1}{\sqrt{s}}\|\nabla\zeta(u)\|_{P^{-1}} I + I \\
&\preceq (\tfrac{2}{\sqrt{s}}\sqrt{\varepsilon} + \tfrac{2}{s}\varepsilon + 1) I \preceq \tfrac{3}{2} I
\end{aligned}
$$

Denote $u_t \coloneqq u - tP^{-1}\nabla\zeta(u)$. With the same reasoning as (33), we can show that $\zeta(u - P^{-1}\nabla\zeta(u)) \le \zeta(u)$ and $\zeta(u_t) - \zeta(u^\star) \le \varepsilon$ for all t, giving $\nabla^2\zeta(u_t) \preceq \frac{3}{2}P$. Then we deduce that

$$
\begin{aligned}
\|\nabla\zeta(u)\|^2_{P^{-1}} h_u(P^{-1}\mathbf{1}_m) &= \zeta(u - P^{-1}\nabla\zeta(u)) - \zeta(u) \\
&= -\langle\nabla\zeta(u), P^{-1}\nabla\zeta(u)\rangle + \textstyle\int_0^1 (1-t)\langle P^{-1}\nabla\zeta(u), \nabla^2\zeta(u_t)P^{-1}\nabla\zeta(u)\rangle \mathrm{d}t \\
&\le -\|\nabla\zeta(u)\|^2_{P^{-1}} + \textstyle\int_0^1 (1-t)\langle P^{-1}\nabla\zeta(u), (\frac{3}{2}P)P^{-1}\nabla\zeta(u)\rangle \mathrm{d}t \\
&= -\|\nabla\zeta(u)\|^2_{P^{-1}} + \tfrac{3}{4}\|\nabla\zeta(u)\|^2_{P^{-1}} = -\tfrac{1}{4}\|\nabla\zeta(u)\|^2_{P^{-1}}.
\end{aligned}
$$

Dividing both sides by $\|\nabla\zeta(u)\|^2_{P^{-1}}$ completes the proof.

B.6 Proof of Lemma 4.3

A good scaling vector w makes $h_u(w)$ small. Since h_u is convex, its negative gradient aligns with any direction that points some scaling matrix better than w being used, say $\hat{w}$:

$$
h_u(\hat{w}) \ge h_u(w) + \langle\nabla h_u(w), \hat{w} - w\rangle \quad \Rightarrow \quad \langle -P^{-1/2}\nabla h_u(w), P^{1/2}(\hat{w} - w)\rangle \ge h_u(w) - h_u(\hat{w}) \ge 0
$$

Hence, preconditioned gradient descent on w reduces $\frac{1}{2}\|w - \hat{w}\|^2_P$: define $w^+ = w - \frac{1}{L}P^{-1}\nabla h_u(w)$. We have

$$
\begin{aligned}
\tfrac{1}{2}\|w^+ - \hat{w}\|^2_P &\le \tfrac{1}{2}\|w - \hat{w}\|^2_P - \tfrac{1}{L}[h_u(w) - h_u(\hat{w})] + \tfrac{1}{2L^2}\|\nabla h_u(w)\|^2_{P^{-1}} \\
&\le \tfrac{1}{2}\|w - \hat{w}\|^2_P - \tfrac{1}{L}\underbrace{[h_u(w^+) - h_u(\hat{w})]}_{\Delta h},
\end{aligned} \tag{42}
$$

where $h_u(w^+) \le h_u(w) - \frac{1}{2L^2}\|\nabla h_u(w)\|^2_{P^{-1}}$ by L-smoothness in **Lemma 4.2**. Next, we have, by definition, that

$$
\begin{aligned}
\log(\zeta(u^+) - \zeta(u^\star)) - \log(\zeta(u) - \zeta(u^\star)) &= \log(\tfrac{\zeta(u^+) - \zeta(u^\star)}{\zeta(u) - \zeta(u^\star)}) \\
&= \log(\tfrac{\zeta(u^+) - \zeta(u) + \zeta(u) - \zeta(u^\star)}{\zeta(u) - \zeta(u^\star)}) \\
&= \log(1 + \tfrac{\zeta(u^+) - \zeta(u)}{\zeta(u) - \zeta(u^\star)}) \\
&= \log(1 + \tfrac{\zeta(u^+) - \zeta(u)}{\|\nabla\zeta(u)\|^2_{P^{-1}}} \tfrac{\|\nabla\zeta(u)\|^2_{P^{-1}}}{\zeta(u) - \zeta(u^\star)}) \\
&\le \log(1 + \tfrac{\sigma_2}{4}\min\{h_u(w^+), 0\}) \qquad (43) \\
&\le \log(1 + \tfrac{\sigma_2}{4} h_u(w^+)) \qquad (44) \\
&\le \tfrac{\sigma_2}{4} h_u(w^+), \qquad (45)
\end{aligned}
$$

where (43) uses the PL inequality from **Lemma 4.2** and the fact that $\zeta(u^+) \le \min\{\zeta(u), \zeta(u - \mathcal{D}(w^+)\nabla\zeta(u))\}$; (44) uses monotonicity of log and (45) uses $\log(1+x) \le x$. Together with $\frac{L\sigma_2}{8}\|w^+ - \hat{w}\|^2_P - \frac{L\sigma_2}{8}\|w - \hat{w}\|^2_P \le$

$-\frac{\sigma_2}{4}[h_u(w^+) - h_u(\hat{w})]$ from (42), we have

$$\begin{aligned}\Omega_{\hat{w}}(u^+, w^+) &= \log(\zeta(u^+) - \zeta(u^\star)) + \tfrac{L\sigma_2}{8}\|w^+ - \hat{w}\|_P^2 \\ &\le \log(\zeta(u) - \zeta(u^\star)) + \tfrac{L\sigma_2}{8}\|w - \hat{w}\|_P^2 - \tfrac{\sigma_2}{4}[h_u(w^+) - h_u(\hat{w})] + \tfrac{\sigma_2}{4}h_u(w^+) \\ &= \log(\zeta(u) - \zeta(u^\star)) + \tfrac{L\sigma_2}{8}\|w - \hat{w}\|_P^2 + \tfrac{\sigma_2}{4}h_u(\hat{w}) \\ &= \Omega_{\hat{w}}(u, w) + \tfrac{\sigma_2}{4}h_u(\hat{w}),\end{aligned}$$

and this completes the proof.

B.7 Proof of Lemma 4.4

For brevity, we drop the iteration index and use $u, u^{1/2}, w, y, u^+, w^+$ to denote $u^k, u^{k+1/2}, w^k, y^k, u^{k+1}, w^{k+1}$. Given that $\max\{\zeta(u), \zeta(w)\} - \zeta(u^\star) \le \min\{\frac{s\sigma_2^2}{1296}, \frac{\sigma_2^3 s^2}{24\|p\|_1}\} =: \tau$, the point $y = u + \frac{\sqrt{\sigma_2}}{2+\sqrt{\sigma_2}}(w-u) = \frac{\sqrt{\sigma_2}}{2+\sqrt{\sigma_2}}w + (1 - \frac{\sqrt{\sigma_2}}{2+\sqrt{\sigma_2}})u$ is the convex combination between u and w, and by convexity, $\zeta(y) \le \frac{\sqrt{\sigma_2}}{2+\sqrt{\sigma_2}}\zeta(w) + (1 - \frac{\sqrt{\sigma_2}}{2+\sqrt{\sigma_2}})\zeta(u) \le \zeta(u^\star) + \tau$. By **Lemma 4.1**, for all $t \in [0,1]$, $\lambda_2(P^{-1/2}\nabla^2\zeta(y + t(u^\star - y))P^{-1/2}) \ge \frac{\sigma_2}{2}$, and

$$\begin{aligned}&\zeta(u^\star) - \zeta(y) - \langle\nabla\zeta(y), u^\star - y\rangle \\ =\ &\textstyle\int_0^1 (1-t)\langle u^\star - y, \nabla^2\zeta(y + t(u^\star - y))(u^\star - y)\rangle \mathrm{d}t \\ =\ &\textstyle\int_0^1 (1-t)\langle P^{1/2}\Pi(u^\star - y), P^{-1/2}\nabla^2\zeta(y + t(u^\star - y))P^{-1/2}P^{1/2}\Pi(u^\star - y)\rangle \mathrm{d}t \\ \ge\ &\textstyle\int_0^1 (1-t)\frac{\sigma_2}{2}\|\Pi(u^\star - y)\|_P^2 \mathrm{d}t = \frac{\sigma_2}{4}\|\Pi(u^\star - y)\|_P^2.\end{aligned}$$

Next, by $\zeta(y) \le \zeta(u^\star) + \tau$, we have $\zeta(u^{1/2}) \le \zeta(y) \le \zeta(u^\star) + \tau$. Hence $P^{-1/2}\nabla^2\zeta(y + t(u^+ - y))P^{-1/2} \preceq 2$ for all $t \in [0,1]$, giving

$$\begin{aligned}&\zeta(u^{1/2}) - \zeta(y) - \langle\nabla\zeta(y), u^{1/2} - y\rangle \\ =\ &\textstyle\int_0^1 (1-t)\langle u^{1/2} - y, \nabla^2\zeta(y + t(u^{1/2} - y))(u^{1/2} - y)\rangle \mathrm{d}t \le \|\Pi(u^{1/2} - y)\|_P^2.\end{aligned}$$

Finally, by convexity, we have $\zeta(u) \ge \zeta(y) + \langle\nabla\zeta(y), u - y\rangle$. By [9, Lemma 4.14], adding the three inequalities

$$\begin{aligned}\zeta(u^\star) - \zeta(y) - \langle\nabla\zeta(y), u^\star - y\rangle &\ge \tfrac{\sigma_2}{4}\|\Pi(u^\star - y)\|_P^2 \\ \zeta(u^{1/2}) - \zeta(y) - \langle\nabla\zeta(y), u^{1/2} - y\rangle &\le \|\Pi(u^{1/2} - y)\|_P^2 \\ \zeta(u) &\ge \zeta(y) + \langle\nabla\zeta(y), u - y\rangle\end{aligned}$$

with weights $\frac{\sqrt{\sigma_2}}{2-\sqrt{\sigma_2}}, 1, \frac{2}{2-\sqrt{\sigma_2}}$ gives $f(u^{1/2}, w^+) - \zeta(u^\star) \le (1 - \frac{1}{2}\sqrt{\sigma_2})[f(u,w) - \zeta(u^\star)]$. Using $\zeta(u^+) \le \zeta(u^{1/2})$ shows $f(u^+, w^+) - \zeta(u^\star) \le (1 - \frac{1}{2}\sqrt{\sigma_2})[f(u,w) - \zeta(u^\star)]$.

Finally we show that w^+ satisfy $\zeta(w^+) - \zeta(u^\star) \le \tau$. Given that $f(u,w) - \zeta(u^\star) \le \min\{4\|p\|_1^{-1}, 1\}\tau$, we have then $\frac{2}{\sigma_2}\|\Pi(w^+ - u^\star)\|_P^2 \le f(u,w) - \zeta(u^\star) \le \min\{4\|p\|_1^{-1}, 1\}\tau$ and

$$\|\Pi(w^+ - u^\star)\|_P^2 \le \tfrac{\sigma_2 \min\{4\|p\|_1^{-1}, 1\}}{2}\tau.$$

By L-smoothness, we have

$$\zeta(w^+) - \zeta(u^\star) \le \tfrac{L}{2}\|\Pi(w^+ - u^\star)\|_P^2 \le \tfrac{\sigma_2 L}{4} \le \tfrac{\|p\|_1}{4}\min\{4\|p\|_1^{-1}, 1\}\tau \le \tau$$

and this completes the proof.