ADAPTIVE BREGMAN ALTERNATING PROJECTIONS FOR FEASIBLE GROMOV–WASSERSTEIN LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The Gromov-Wasserstein (GW) problem compares structured distributions without requiring a shared feature space or known correspondences, but its nonconvex objective and coupled marginal constraints make computation challenging. Bregman alternating projected gradient (BAPG) uses inexpensive alternating row and column updates, yet its fixed-penalty relaxation leaves a persistent feasibility gap. We propose Adaptive KL-BAPG (A-KL-BAPG), which combines a finite fixed-penalty burn-in with a guarded increasing-penalty phase. At each tail iteration, the method reuses BAPG's alternating updates and backtracks a delayed-power step until a Sinkhorn-inspired projective-diameter safeguard is satisfied. We prove finite termination of the backtracking at each iteration and show that the feasibility gap vanishes asymptotically. We further establish a best-iterate $O(1/\log N)$ bound for the weighted squared corrected residual and, under a support regularity condition, the existence of a stationary accumulation point for the original GW problem. This distinguishes A-KL-BAPG from fixed-penalty BAPG, whose stationarity guarantees are given for the relaxed problem. Experiments show that A-KL-BAPG achieves a favorable balance of accuracy, objective value, feasibility, and stationarity relative to BAPG variants, projection-based methods, and task-specific baselines. For synthetic and real graph alignment problems, it closely matches the accuracy and objective value of fixed-penalty KL-BAPG while reducing the marginal feasibility gap by 62-99% and the projected stationarity residual by 28-98%. Heterogeneous domain adaptation experiments show a similar pattern: A-KL-BAPG maintains comparable target accuracy and objective values while achieving improved feasibility and stationarity than fixed-penalty KL-BAPG.

1 INTRODUCTION

The Gromov-Wasserstein (GW) problem compares structured distributions based on their internal geometry, without requiring a shared feature space or a known correspondence, making it well-suited to comparing heterogeneous structured data (Mémoli, 2011; Peyré et al., 2016). Its development grew from geometric point-cloud (Mémoli & Sapiro, 2004) and shape comparison (Mémoli, 2009; 2011) to practical discrete formulations (Peyré et al., 2016). Its extension from metric space to weighted networks further supports graph comparison beyond a shared node set (Chowdhury & Mémoli, 2019). More recently, GW has been widely used for graph alignment and representation learning (Xu et al., 2019b; Li et al., 2023) and for transfer across heterogeneous domains (Yan et al., 2018; Demetci et al., 2024; Aritake & Hino, 2026). The resulting optimization, however, is difficult since the squared-loss GW is a nonconvex, quadratic program over the transport polytope.

GW solvers range from Frank-Wolfe methods (CG) (Peyré et al., 2016; Vayer et al., 2019), entropic Bregman projected gradients (EPGD) (Solomon et al., 2016), and Bregamn proximal-point methods (PPA) (Xu et al., 2019b), to recent entropic-GW (EGW) (Rioux et al., 2024) and inexact projected gradient (iPG) (Liang & Yang, 2026). Graph-specific alternatives include the recursive partitioning method ScalaGW (Xu et al., 2019a) and the heat-kernel spectral method SpecGW (Chowdhury & Needham, 2021). These methods serve different purposes, but generally they repeatedly perform a linear transport step or iterative Bregman or Euclidean projections to enforce both marginals. The recent iPG makes inexact projection verifiable, but still requires inner solvers at every iteration. Such routines can dominate the cost of large-scale computations (Li et al., 2023; Liang & Yang, 2026).

Bregman alternating projected gradient (BAPG) offers a faster single-loop alternative. Starting from the one-step BPG-S heuristic (Xu et al., 2019b), BAPG splits the row and column constraints and penalizes disagreement between two transport variables (Li et al., 2023). With KL geometry, each half-step is simply a multiplicative GW-gradient update followed by normalization. These matrix-scaling updates are GPU-friendly (Li et al., 2023) and support parallel, differentiable FGW learning (Deng et al., 2026) with a popular BAPG implementation in POT (Flamary et al., 2021).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The speed of fixed-penalty BAPG comes with a cost: For a penalty ρ, the two split variables may disagree, resulting in marginal errors of order $O(1/\rho)$ that persist throughout iterations (Li et al., 2023). Increasing ρ appears natural; however, if it increases too fast, the associated step $\alpha = 1/\rho$ shrinks too quickly and can lead to finite cumulative progress. If the step size remains too aggressive, the positive scaling maps can become ill-conditioned in a projective sense. Additionally, the existing KL-BAPG also assumes bounded accumulated asymmetric Bregman errors (AAE), which is difficult to verify, and only post-facto empirical evidence of it has been shown. Moreover, singular boundary geometry can yield a stationary point only on a limiting support face (Chen et al., 2024). Therefore, we need to jointly control the penalty growth, cumulative step progress, and projective geometry.

This difficulty motivates a connection between BAPG and Sinkhorn scaling. Sinkhorn alternates exact row and column KL projections (Sinkhorn & Knopp, 1967; Cuturi, 2013; Benamou et al., 2015), and its contraction is governed by the projective diameter (Franklin & Lorenz, 1989; Eveson & Nussbaum, 1995; Idel, 2016). Modern variants retain this scaling structure under changing mirror steps (Ballu & Berthet, 2023) or temperatures (Chizat, 2024). General BAPG also shares the alternating projection structure. Under KL geometry, these projections coincide with the row and column normalizations used by Sinkhorn. Unlike standard Sinkhorn, however, each normalization in KL-BAPG is preceded by a state-dependent multiplicative GW-gradient update. As the penalty increases, these updates become smaller, motivating a schedule that controls the projective geometry of the resulting iterates.

In this paper, we introduce **Adaptive KL-BAPG**(A-KL-BAPG). The method runs a finite, fixed-penalty burn-in phase and then switches to a guarded, increasing-penalty phase. At each tail iteration, it proposes a delayed-power step, performs the same alternating row- and column-marginal normalization update as KL-BAPG (Li et al., 2023), and backtracks until the candidate coupling satisfies a logarithmically expanding projective-diameter bound. We prove that the backtracking terminates finitely at every iteration without requiring an auxiliary projection subproblem. Under this guarded schedule, iterates are asymptotically feasible, and the discrepancy between the alternating iterates vanishes.

Our main contributions are:

- **Vanishing marginal feasibility with explicit BAPG updates.** Existing BAPG methods leave a persistent $O(1/\rho)$ feasibility gap at fixed penalty, assume bounded accumulated asymmetric Bregman errors, and can certify stationarity only on a limiting support face. A-KL-BAPG removes the $O(1/\rho)$ feasibility floor while retaining closed-form updates through an increasing-penalty schedule without an AAE assumption.
- **Stationarity and a rate for the original GW constraints.** We prove finite termination of the guard, control the accumulated KL asymmetry, and derive a best-iterate $O(1/\log N)$ bound for the weighted squared corrected residual. Under an inactive-support nonnegative-residual assumption, we show that the generated iterates share a stationary accumulation point for the original GW problem, i.e., some subsequence converges to a stationary point.
- **Empirical validation on GW and FGW problems.** We evaluate the proposed method on synthetic and real graph alignment and heterogeneous domain adaptation problems. A-KL-BAPG retains matching accuracy and GW objective values comparable to other BAPG methods while significantly reducing the marginal gap and projected stationarity residual. Across the included methods, A-KL-BAPG preserves strong alignment and classification accuracy while improving feasibility and stationarity.

2 Problem Formulation, Algorithm Design and Main Results

Let $D_X \in \mathbb{R}^{n\times n}$ and $D_Y \in \mathbb{R}^{m\times m}$ encode within-domain geometry, and let $\mu \in \mathbb{R}^n_+$ and $\nu \in \mathbb{R}^m_+$ be the desired marginal probability vectors with strictly positive entries. Define $\mathcal{C}_1 = \{\pi \geq \mathbf{0} : \pi \mathbf{1}_m = \mu\}$, $\mathcal{C}_2 = \{\pi \geq \mathbf{0} : \pi^\top \mathbf{1}_n = \nu\}$, and $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$. Let $g_r = \mathbb{I}_{\mathcal{C}_r}$ be the indicator of the corresponding one-marginal set. Up to terms constant on $\mathcal{C}$, the squared-loss GW problem is

$$\min_{\pi\in\mathbb{R}^{n\times m}} F(\pi) + g_1(\pi) + g_2(\pi) = \min_{\pi\in\mathcal{C}} F(\pi), \qquad F(\pi) = -\operatorname{Tr}(D_X \pi D_Y \pi^\top). \tag{1}$$

For clarity, we take D_X, D_Y symmetric with $G(\pi) = D_X \pi D_Y$, $\nabla F(\pi) = -2G(\pi)$. Then a fused linear term can be included without changing the alternating construction. The stationary set of the original constrained problem is $\mathcal{X} := \{x \in \mathcal{C} : \mathbf{0} \in \nabla F(x) + \mathcal{N}_{\mathcal{C}}(x)\}$. To decouple the two marginal constraints, we introduce w to reformulate:

$$\min_{\pi=w} f(\pi, w) + g_1(\pi) + g_2(w) + \rho D_h(\pi, w),$$

with $f(\pi, w) = -\operatorname{Tr}(D_X \pi D_Y w^\top)$ and for the Legendre function h, $D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y\rangle$ is the Bregman divergence. Here we use the Boltzmann-Shannon entropy kernel $h(x) := \sum_{i,j} x_{ij}(\log x_{ij} - 1)$. Then $\nabla h(x) = \log x$ and $D_h(u, v) = \sum_{ij} u_{ij} \log(u_{ij}/v_{ij}) - u_{ij} + v_{ij}$.

Our starting point is to introduce an increasing distance penalty $\{\rho_k\}$ with step size $\alpha_k = 1/\rho_k$. Then, for k-th iteration, the BAPG update aims to solve:

$$\pi^{k+1} = \arg\min f(\pi, w^k) + \rho_k D_h(\pi, w^k) + g_1(\pi),$$
$$w^{k+1} = \arg\min f(\pi^{k+1}, w) + \rho_k D_h(w, \pi^{k+1}) + g_2(w),$$

where $\nabla_\pi f(\pi, w) = -G(w)$, and $\nabla_w f(\pi, w) = -G(\pi)$. Since elements of $\mathcal{C}_1$ and $\mathcal{C}_2$ are nonnegative and have fixed total mass one, $\mathcal{C}_1$ and $\mathcal{C}_2$ are compact. Since $G(X) = D_X X D_Y$ is continuous for fixed finite matrices D_X and D_Y, it is uniformly bounded on the compact set $\mathcal{C}_1 \cup \mathcal{C}_2$. In particular, we use the explicit bound $M_G := |D_X|_{\max}|D_Y|_{\max} \geq \sup_{X \in \mathcal{C}_1 \cup \mathcal{C}_2} |G(X)|\infty$, which follows from the nonnegativity and unit total mass of X.

2.1 Algorithm Design

Let $\rho_0 > 0$ be a fixed penalty used during a finite prefix, and $k_0 < \infty$ be the length of the prefix. For $k \leq k_0$, the step size is $\alpha_k = 1/\rho_0$ and the two KL subproblems have the closed forms

$$z_1^t = w^t \odot \exp\big(\alpha_t G(w^t)\big), \qquad \pi_{ij}^{t+1} = \mu_i \frac{z_{1,ij}^t}{\sum_\ell z_{1,i\ell}^t}, \tag{2a}$$

$$z_2^t = \pi^{t+1} \odot \exp\big(\alpha_t G(\pi^{t+1})\big), \qquad w_{ij}^{t+1} = \nu_j \frac{z_{2,ij}^t}{\sum_\ell z_{2,\ell j}^t}. \tag{2b}$$

By Li et al. (2023), this often yields useful early progress on the objective, but its two split variables have a non-vanishing infeasibility gap. Starting immediately with diminishing steps removes that floor asymptotically, but can suppress early progress. We therefore run $k_0 < \infty$ iterations with fixed penalty ρ_0 and then apply an adaptive diminishing-step tail. After the finite prefix, we reset the tail index to $t = 0$ and relabel the start of the post-burn state as (π^0, w^0).

To control the geometry of the alternating iterates, we introduce Hilbert's projective metric

$$d_H(u, v) = \log\{\max_i(u_i/v_i)/\min_i(u_i/v_i)\},$$

and the associated projective diameter $\Delta(X) = \max_{i,i',j,j'} \log\big(X_{ij}X_{i'j'}/(X_{ij'}X_{i'j})\big)$ where u, v are positive vectors and X is a positive matrix. Then for the tail iterates, we set $\Delta_0 = \Delta(w^0)$. For tail index $t \geq 0$, choose a guard slope $a \in (0, 2/3)$, delayed-power exponent $\beta \in (\frac{2}{3} + \frac{a}{2}, 1)$, and proposal scale $\alpha_0 > 0$. The permitted diameter and nominal step are:

$$\Gamma_t = \Delta_0 + a\log(t+1),\ \bar{\alpha}_t = \alpha_0 \left(\frac{k_0+1}{t+k_0+1}\right)^\beta. \tag{3}$$

The logarithmic threshold grows slowly enough to retain polynomial projective contraction while $\bar{\alpha}_t$ avoids shrinking the proposed step abruptly at the switch. Let $B_\Delta := 8M_G$ denote the conservative projective diameter growth of one cycle, which we established in Lemma 5 (Appendix A.2). The corresponding safe step size of the diameter increase is

$$h_t := \frac{\Gamma_{t+1} - \Gamma_t}{B_\Delta} = \frac{a}{B_\Delta} \log \frac{t+2}{t+1}. \tag{4}$$

Algorithm 1: Adaptive KL-BAPG

Require: $w^0 > 0$, marginals (μ, ν), burn-in k_0, prefix penalty ρ_0, proposal scale α_0, exponent β, guard slope a, growth bound B_Δ, backtracking factor τ

for $k = 0, \ldots, k_0 - 1$ **do** ▷ fixed-penalty burn-in
  $\alpha \leftarrow 1/\rho_0$
  $z_1 \leftarrow w^k \odot \exp(\alpha G(w^k))$; $\pi_{ij}^{k+1} \leftarrow \mu_i z_{1,ij} / \sum_\ell z_{1,i\ell}$
  $z_2 \leftarrow \pi^{k+1} \odot \exp(\alpha G(\pi^{k+1}))$; $w_{ij}^{k+1} \leftarrow \nu_j z_{2,ij} / \sum_\ell z_{2,\ell j}$
end for
$\Delta_0 \leftarrow \Delta(w^{k_0})$; $\alpha_{-1} \leftarrow \infty$
for $t = 0, 1, 2, \ldots$ **do** ▷ guarded increasing-penalty tail
  $k \leftarrow k_0 + t$; $\bar{\alpha}_t \leftarrow \alpha_0\big((k_0+1)/(t+k_0+1)\big)^\beta$
  $h_t \leftarrow (a/B_\Delta)\log((t+2)/(t+1))$; $\ell_t \leftarrow \tau \min\{\bar{\alpha}_t, h_t\}$
  $\alpha \leftarrow \max\{\ell_t, \min(\bar{\alpha}_t, \alpha_{t-1})\}$
  repeat
    $\widetilde{z}_1 \leftarrow w^k \odot \exp(\alpha G(w^k))$; $\widetilde{\pi}_{ij} \leftarrow \mu_i \widetilde{z}_{1,ij} / \sum_\ell \widetilde{z}_{1,i\ell}$
    $\widetilde{z}_2 \leftarrow \widetilde{\pi} \odot \exp(\alpha G(\widetilde{\pi}))$; $\widetilde{w}_{ij} \leftarrow \nu_j \widetilde{z}_{2,ij} / \sum_\ell \widetilde{z}_{2,\ell j}$
    if $\Delta(\widetilde{w}) \leq \Delta_0 + a\log(t+2)$ **then**
      $\alpha_t \leftarrow \alpha$; $\pi^{k+1} \leftarrow \widetilde{\pi}$; $w^{k+1} \leftarrow \widetilde{w}$
    else
      $\alpha \leftarrow \tau\alpha$
    end if
  until $\Delta(\widetilde{w}) \leq \Gamma_{t+1}$
end for

To avoid the accepted step size from becoming arbitrarily small, we define the floor $\ell_t := \tau \min\{\bar{\alpha}_t, h_t\}$, where $\tau \in (0, 1)$ is the backtracking factor. The line search starts with $s_t := \max\{\ell_t, \min\{\bar{\alpha}_t, \alpha_{t-1}\}\}$, with $\alpha_{-1} = +\infty$. The inner minimum guarantees the monotonicity of the accepted tail steps by preventing repeated tests above the previously accepted value, while the outer maximum preserves the lower safeguard. Starting from s_t, the method repeatedly shrinks the nominal step-size by τ and recomputes Equation 2a–Equation 2b until $\Delta(w^{t+1}) \leq \Gamma_{t+1}$. The accepted value is then denoted by α_t. The detailed algorithm is outlined in Algorithm 1.

Remark 1. *After Lemma 7, we show that whenever the current iterate satisfies $\Delta(w^t) \le \Gamma_t$, every $\alpha \le h_t$ is guaranteed to satisfy the next diameter bound $\Delta(w^{t+1}) \le \Gamma_{t+1}$. Thus, the backtracking line search terminates after finitely many steps at each iteration, and the floor $\ell_t = \tau \min\{\bar{\alpha}_t, h_t\}$ provides a controlled lower scale for the accepted step. Detailed proof is provided in Appendix A.2.*

A-KL-BAPG can be viewed as a variant of Sinkhorn scaling with guarded gradient perturbations. Each iteration alternates between row- and column-marginal normalizations as in Sinkhorn, while multiplicative gradient perturbations optimize the objective, and adaptive step sizes balance the optimization progress. The diameter guard controls the changing projective geometry without introducing additional projection subproblems. The formal connection is developed in Appendix A.1.

2.2 Main Results

We are now ready to introduce our main results on the subsequence convergence and stationarity of Algorithm 1. Our result is based on the idea of measuring performance and convergence by diminishing residual, specifically, we define the corrected residual as $R_t := \rho_t\big(\nabla h(w^t) - \nabla h(w^{t+1})\big) + G(\pi^{t+1}) - G(w^t)$. $G(\pi^{t+1}) - G(w^t)$. The value of R_t measures the imbalance between the GW gradient and the marginal constraint terms in the KL updates, with a correction for the difference between $G(w^t)$ and $G(\pi^{t+1})$. Specifically, the first result of Theorem 1 is to shows that its weighted norm vanishes along a subsequence of the generated iterates of the proposed algorithm.

Theorem 1 (Feasibility and residual guarantees)**.** *Consider the GW Problem (1) with symmetric $D_X \in \mathbb{R}^{n\times n}$, $D_Y \in \mathbb{R}^{m\times m}$, and marginals $\mu \in \mathbb{R}^n_{++}$, $\nu \in \mathbb{R}^m_{++}$ satisfying $\mathbf{1}_n^\top \mu = \mathbf{1}_m^\top \nu = 1$, and let $M_G := \sup_{X\in\mathcal{C}_1\cup\mathcal{C}_2} \|G(X)\|_\infty < \infty$. Let the initialization be entrywise positive, let the burn-in length $k_0 \in \mathbb{N}$ be finite, and let the prefix penalty $\rho_0 > 0$, the proposal scale $\alpha_0 > 0$, the backtracking factor $\tau \in (0,1)$, the growth bound $B_\Delta \ge 8M_G$, and the guard slope and delayed-power exponent $a \in (0, 2/3)$, and $\beta \in (2/3 + a/2,\, 1)$. Then, the tail iterates $\{(\pi^t, w^t)\}_{t\ge 0}$ generated by Algorithm 1, indexed from the switch at $k = k_0$, have the following properties:*

1. $\liminf_{t\to\infty} \|R_t\|_{M_t} = 0$, *with $\rho_t := 1/\alpha_t$ and $\|U\|^2_{M_t} := \sum_{i,j} m^t_{ij} U^2_{ij}$, where $m^t_{ij} := \min\{w^t_{ij}, w^{t+1}_{ij}\}$;*
2. $\lim_{t\to\infty} \|x^{t+1} - \pi^{t+1}\|_F + \|\pi^{t+1} - w^{t+1}\|_F = 0$, *where $x^t := \mathcal{P}_{\mathcal{C}}(\pi^t)$ denotes the Euclidean projection of π^t onto $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$; in particular, $\{x^t\}$, $\{\pi^t\}$ and $\{w^t\}$ have the same set of accumulation points.*

Theorem 1 shows that A-KL-BAPG recovers feasibility for the original GW problem and admits a subsequence along which the weighted corrected residual vanishes unconditionally. Later in Theorem 2, with the addition of Assumption 1, we show that the limit of any such convergent subsequence is stationary. Unlike KL-BAPG (Li et al., 2023), which assumes bounded accumulated asymmetric Bregman error (AAE) throughout, A-KL-BAPG guarantees absolute summability of the asymmetric errors over the entire iterate sequence, as shown in Lemma 1.

Lemma 1. *Let $(\pi^t, w^t)_{t\ge 0}$ be the iterates generated by Algorithm 1. Under the setting of Theorem 1, $\sum_{t=0}^{\infty} \big|D_h(\pi^{t+1}, w^t) - D_h(w^t, \pi^{t+1})\big| + \big|D_h(w^{t+1}, \pi^{t+1}) - D_h(\pi^{t+1}, w^{t+1})\big| = O(\eta_t^3) < \infty$, i.e., AAE holds.*

While Theorem 1 shows vanishing residuals and asymptotic convergence, to show the existence of a stationary accumulation point, we require the following assumption.

Assumption 1. *For every $\{t_\ell\}$ such that $\|R^{t_\ell}\|_{M_{t_\ell}} \to 0$ and $x^{t_\ell+1} = \mathcal{P}_{\mathcal{C}}(\pi^{t_\ell+1}) \to \bar{x}$, define the active support of $\bar{x}$ as $S(\bar{x}) := \{(i,j) : \bar{x}_{ij} > 0\}$. Then for every $(i,j) \notin S(\bar{x})$, $\liminf_{\ell\to\infty} R^{t_\ell}_{ij} \ge 0$.*

Assumption 1 is needed only to pass from the weighted corrected residual to stationarity at a boundary point. As illustrated in Figure 1, on active entries, the weights in M_t stay positive and force $R^t_{ij} \to 0$. On an inactive entry, the weight may vanish: $R^t_{ij} > 0$ corresponds to an outward descent direction blocked by nonnegativity, while $R^t_{ij} < 0$ reveals an inward feasible descent direction (similar situations discussed in Chen et al. (2024)). Related boundary regularity conditions are standard in constrained optimization, including nondegeneracy and strict complementarity in active-set analyses (Burke & Moré, 1988; Liang et al., 2017), and transversality conditions used to control limiting multipliers (Lewis & Wright, 2011). Our assumption is milder, requiring only that the limiting residuals be nonnegative at inactive entries. In particular, Assumption 1 holds if entries vanishing at the limit do not increase to first order in relative terms along the subsequence, as shown in Proposition 1, which is the main building block for our second main result on the stationarity of the accumulation points.

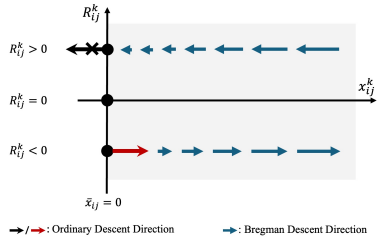


Figure 1: Boundary effect of the Descent. At an inactive entry, an outward descent direction is blocked (black), while an inward direction might remain feasible (red)

Proposition 1 (A posteriori sufficient condition)**.** *Let $\{t_\ell\}$ be subsequence with $x^{t_\ell+1} \to \bar{x}$. Under the setting of Theorem 1, Assumption 1 holds along this subsequence if for every $(i,j) \notin S(\bar{x})$, $\liminf_{\ell\to\infty} \rho_{t_\ell} \log\left(w_{ij}^{t_\ell}/w_{ij}^{t_\ell+1}\right) \geq 0$. In particular, it suffices that $w_{ij}^{t_\ell+1} \leq w_{ij}^{t\ell}(1+\varepsilon_\ell^{ij})$ with $\varepsilon_\ell^{ij} = o(\alpha_{t_\ell})$. Thus, eventual nonincrease of every inactive entry along the subsequence is sufficient.*

Theorem 2 (Stationary accumulation point)**.** *Consider the GW Problem 1 and the A-KL-BAPG iterates $(\pi^t, w^t)_{t\geq 0}$ generated by Algorithm 1 under the conditions of Theorem 1. Let Assumption 1 hold. Then every residual-vanishing subsequence has a limit $\bar{x} \in \mathcal{C}$ satisfying $0 \in \nabla F(\bar{x}) + N_{\mathcal{C}}(\bar{x})$. In particular, the iterates admit at least one stationary accumulation point.*

Theorem 2 shows that projected subsequences and the two alternating sequences have the same accumulation points. Compared with KL-BAPG (Li et al., 2023), our result strengthens the stationarity conclusion. Under bounded accumulated asymmetric Bregman error, KL-BAPG shows that every accumulation point $(\pi^\star, w^\star)$ belongs to the relaxed BAPG fixed-point set, with $\|\pi^\star - w^\star\| = O(\rho^{-1})$ and $\mathrm{dist}((\pi^\star + w^\star)/2, \mathcal{X}) = O(\rho^{-1})$. Thus, KL-BAPG characterizes every accumulation point of the relaxed problem, whereas our result establishes exact subsequence convergence to a stationary point of the original GW problem.

The proof in Section 3 uses the accepted-step lower bound generated by the projective-diameter guard. One might hope to replace it by a lower bound with the same power as the delayed proposal, i.e., $\alpha_t \geq c(t+1)^{-\beta}$ for some $c > 0$ and $\beta < 1$. However, next proposition shows that this stronger lower bound cannot be guaranteed from the logarithmic guard alone.

Proposition 2 (Impossibility of a Universal Power Lower Bound)**.** *Under the conditions of Theorem 1, there exists a valid 2×2 instance, constructed in Appendix A.8, such that any guarded tail of A-KL-BAPG on this instance satisfies $\sum_{s=0}^{T-1} \alpha_s = O(\log T)$. Consequently, no theorem using only the projective-diameter guard can guarantee an eventual lower bound $\alpha_t \geq c(t+1)^{-\beta}$ for any $c > 0$ and any $\beta < 1$.*

3 Convergence Analysis

We now establish Theorem 2. We first introduce some notation and then present auxiliary propositions that will help us complete the proof. First, define $\eta_t := (t+1)^{-(\beta-p)}$. We note that the initial point in the following proof is (π^{k_0}, w^{k_0}), not the original initialization. For $W \in \mathcal{C}_2$, let $r(W) = (W\mathbf{1}_m)/\mu$ so $r(W) = \mathbf{1}_n$ when $W \in \mathcal{C}$.

Proposition 3. *Let $\{(\pi^t, w^t)\}_{t\geq 0}$ be the tail iterates generated by Algorithm 1. Under the conditions of Theorem 1, $\|w^t\mathbf{1}_m - \mu\|_1 = O(\eta_t)$ and $\|(\pi^{t+1})^\top \mathbf{1}_n - \nu\|_1 = O(\eta_t)$. Moreover, for all sufficiently large t, $D_h(\pi^{t+1}, w^t) = O(\eta_t^2)$ and $D_h(\pi^{t+1}, w^{t+1}) = O(\eta_t^2)$.*

Since D_h is not symmetric, the descent analysis generates the asymmetry terms $A_t^{\mathrm{row}} = D_h(\pi^{t+1}, w^t) - D_h(w^t, \pi^{t+1})$ and $A_t^{\mathrm{col}} = D_h(w^{t+1}, \pi^{t+1}) - D_h(\pi^{t+1}, w^{t+1})$ with the total KL asymmetry $A_t = A_t^{\mathrm{row}} + A_t^{\mathrm{col}}$.

Proposition 4. *Let $\{(\pi^t, w^t)\}_{t\geq 0}$ be the tail iterates generated by Algorithm 1. Under the conditions of Theorem 1, for all sufficiently large t, $|A_t^{\mathrm{row}}| + |A_t^{\mathrm{col}}| = O(\eta_t^3)$ and $|A_t| = O(\eta_t^3)$.*

Proposition 5. *For $\{(\pi^t, w^t)\}_{t\geq 0}$ generated by Algorithm 1, under the conditions of Theorem 1, $\sum_{t=1}^{\infty} \rho_t \left[D_h(\pi^t, \pi^{t+1}) + D_h(w^t, w^{t+1})\right] < \infty$.*

Proposition 6. *Let $\{R_t\}_{t\geq 0}$ be the corrected residuals associated with the tail iterates generated by Algorithm 1 and suppose the conditions of Theorem 1 hold. Then there exists an integer K_R such that $\sum_{t=K_R}^{\infty} \alpha_t \|R_t\|_{M_t}^2 < \infty$. Consequently, for every $N \geq K_R + 1$, $\min_{0\leq i\leq N-1} \|R_i\|_{M_i}^2 = O\left(1/\log N\right)$.*

Proof Sketch for Theorems 1 and 2 . Proposition 3 establishes vanishing marginal infeasibility through projective contraction and converts this estimate into quadratic KL cross-divergence, while Proposition 4 shows that the resulting KL asymmetry is cubic. Proposition 5 then incorporates these bounds into the Lyapunov analysis, which in turn yields the residual summability in Proposition 6. Together with the guarded step-size bounds, this summability yields the vanishing residual, while Proposition 3, Pinsker's inequality, and linear regularity show that the two iterates converge to their feasible projections, so all three sequences share the same accumulation points, as established in Theorem 1. Moreover, Assumption 1 guarantees that every residual-vanishing subsequence limit satisfies the KKT conditions and is therefore stationary, as stated in Theorem 2. Complete proofs are provided in Appendix A.

3.1 Discussion on Step-Size Schedules

$1/k$ **schedule:** A natural alternative to the proposed algorithm is the deterministic harmonic schedule $1/k$, i.e., $\alpha_k' = (k+1)^{-1}$ and $\rho_k' = (\alpha_k')^{-1}$. Here, we show that our method has a less restrictive convergence guarantee.

Corollary 1. *For the $1/k$ schedule initialized with $w^0 > 0$ and strictly positive marginals, if $M_G < 1/12$, the feasibility and residual conclusions of Section 3 also hold.*

The proof is deferred in Appendix A.6. $M_G < 1/12$ is a conservative sufficient condition and there is no independent guard-slope or power parameter that can be adjusted when this sufficient condition is violated. For the scaled harmonic schedule $\alpha'_k = c/(k+1)$, the same argument gives the sufficient condition $cM_G < 1/12$ where a smaller c reduces the initial step size. On the other hand, in our proposed method, the projective-diameter slope a and the delayed-power exponent β are design parameters linked by the admissibility condition $\beta > 2/3 + a/2$, which jointly balances the geometric control and step-size decay while providing a convergence guarantee.

Constant Penalty Schedule: The proposed method can be viewed as an extension of the fixed-penalty BAPG method introduced by Li et al. (2023) where the penalty ρ and the corresponding step size α are held constant across iterations. Under their convergence assumption, every accumulation point $(\pi^\star, w^\star)$ belongs to the fixed-point set of the relaxed BAPG problem. By (Li et al., 2023, Proposition 3.3), $\operatorname{dist}\left((\pi^\star + w^\star)/2, \mathcal{X}\right) = O(\rho^{-1})$. For every fixed ρ, it converges to the stationary set of the relaxed split problem rather than the original problem. In particular, the infeasibility gap is controlled by $O(\rho^{-1})$, which only vanishes when $\rho \to \infty$. Our method strengthens the asymptotic guarantee by introducing an increasing penalty in the tail phase, thereby removing the permanent approximation floor associated with the fixed-penalty ρ schedule. The adaptive safeguard does introduce a backtracking procedure within each tail iteration. Nevertheless, as shown in Remark 1, this backtracking loop is well defined and terminates after finitely many reductions at each tail iteration.

Pure Adaptive Power Schedule: A pure adaptive power schedule with a nominal step $\alpha''_k = \alpha_0(k+1)^{-\beta}$, $\rho''_k = \alpha''^{-1}_k$ and the same guard as A-KL-BAPG, still converges since the same convergence analysis in Section 3 applies when $k_0 = 0$. However, it can enter a small-step regime before the coarse optimization phase is complete. For every (i, j), $\exp(\alpha''_k G_{ij}(w^k)) = 1 + \alpha''_k G_{ij}(w^k) + O((\alpha''_k)^2)$, i.e., the objective-driven perturbation is in $O(\alpha''_k)$. When the iterates are still far from this region, the primary task is to obtain substantial objective reduction and approach the stationary set of the relaxed split problem. If the guard is imposed at initialization, these transient changes in the projective diameter may cause repeated backtracking, thereby forcing the accepted step toward the conservative safe scale h_t, despite the need for larger objective-driven steps. Therefore, the finite fixed-penalty prefix helps handle the initial transient and avoid premature step-size decay.

4 Numerical Analysis

We evaluate A-KL-BAPG on two important applications of GW: graph alignment (Xu et al., 2019b; Li et al., 2023) and heterogeneous domain adaptation (Yan et al., 2018; Aritake & Hino, 2026). We compare against major GW/FGW solver families, including conditional-gradient methods (Peyré et al., 2016; Vayer et al., 2019), entropic projected-gradient methods (Solomon et al., 2016), proximal and Bregman methods (Xu et al., 2019b; Li et al., 2023), and recent entropic-GW and inexact projected-gradient algorithms (Rioux et al., 2024; Liang & Yang, 2026). We also include graph-specific and domain adaptation baselines (Vogelstein et al., 2015; Xu et al., 2019a; Chowdhury & Needham, 2021; Courty et al., 2016; 2017; Aritake & Hino, 2022).

4.1 GW Objective, Feasibility, and Stationarity Residual

Given the source and target structure matrices D_X and D_Y, the GW objective evaluated at a transport plan X is $F(X) = -\operatorname{Tr}\left(D_X X D_Y X^\top\right)$. For the fused Gromov-Wasserstein problem involving an additional cross-domain feature cost M, we report the fused objective $F_{\mathrm{FGW}}(X) = -2\lambda\langle D_X X D_Y, X\rangle + (1-\lambda)\langle M, X\rangle$ with $\lambda \in (0, 1)$, which is the nonconstant part of squared-loss FGW. Lower values indicate better minimization of the corresponding GW or FGW criterion. We report prediction accuracy separately to measure downstream performance.

Denote W^t as the second-block iterate returned after iteration k. We report its marginal feasibility gap $r_{\mathrm{feas}}(W^{t+1}) := \left\|W^{t+1}\mathbf{1} - \mu\right\|_1 + \left\|(W^{t+1})^\top \mathbf{1} - \nu\right\|_1$. We evaluate whether the output matrix is close to a stationary point of the constrained GW problem. Since the split iterates produced by BAPG may not exactly lie in $\mathcal{C}$, we first construct a feasible shadow coupling $\widehat{W}_t \in \mathcal{C}$ by rescaling the rows and columns of the reported matrix. We then measure stationarity at $\widehat{W}_t$ by defining the projected-gradient stationarity residual

$$r_{\mathrm{stat}}(\widehat{W}_t) = (1/\eta)\left\|\widehat{W}_t - \mathcal{P}_{\mathcal{C}}\left(\widehat{W}_t - \eta\nabla F(\widehat{W}_t)\right)\right\|_F.$$

Note that for $F(W) = -\mathrm{Tr}(D_X W D_Y W^\top)$, $\nabla F(W) = -\left(D_X^\top W D_Y^\top + D_X W D_Y\right)$. Then $\|\nabla F(W_1) - \nabla F(W_2)\|_F \le \|D_X^\top (W_1 - W_2) D_Y^\top\|_F + \|D_X (W_1 - W_2) D_Y\|_F \le \|D_X^\top\|_2 \|D_Y^\top\|_2 \|W_1 - W_2\|_F +$

Method	Accuracy	GW objective	Feasibility	Stationarity Residual	Runtime (s)
KL-BAPG	0.966	−0.034	0.098	0.470	278.981
KL-BAPG + Sinkhorn	0.966	−0.034	0.005	0.140	279.987
A-KL-BAPG	**0.966**	**−0.034**	$\mathbf{5.524 \times 10^{-6}}$	**0.011**	**225.242**
$1/k$ schedule	0.022	−0.003	1.430×10^{-7}	0.204	54.171
Pure power	0.022	−0.003	1.110×10^{-6}	0.205	51.926
POT CG	5.390	−0.017	2.390×10^{-17}	7.730×10^{-4}	11.1
POT PPA	0.966	−0.034	1.450×10^{-5}	0.002	2112.932
POT EPGD	0.065	−0.003	4.780×10^{-12}	0.233	0.871
POT BAPG	0.966	−0.034	0.098	0.470	298.960
Degree Hungarian	0.003	−0.003	1.310×10^{-15}	0.750	0.009
SciPy FAQ	0.059	−0.015	1.230×10^{-15}	0.333	11.250
iPG	1.000	−0.034	3.100×10^{-10}	3.090×10^{-4}	1927.503

Table 1: Synthetic Graph Alignment Results.

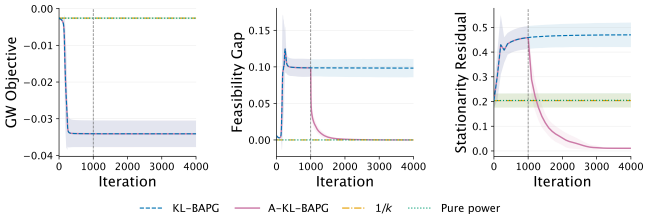


Figure 2: Iteration convergence comparison.

$\|D_X\|_2\|D_Y\|_2\|W_1 - W_2\|_F = 2\|D_X\|_2\|D_Y\|_2\|W_1 - W_2\|_F$. Thus $2\|D_X\|_2\|D_Y\|_2$ is a valid Lipschitz upper bound for the gradient, and stepsize η is defined as $\eta := (2\|D_X\|_2\|D_Y\|_2)^{-1}$.

Remark 2. *Note that $r_{\mathrm{stat}}(\widehat{W}_t) = 0$ is equivalent to $\widehat{W}_t = \mathcal{P}_{\mathcal{C}}\big(\widehat{W}_t - \eta\nabla F(\widehat{W}_t)\big)$, i.e., $0 \in \nabla F(\widehat{W}_t) + \mathcal{N}_{\mathcal{C}}(\widehat{W}_t)$, which is the KKT first-order stationarity condition for the GW problem over $\mathcal{C}$. A similar quantity $\big\|\pi - \mathrm{Proj}_{\mathcal{C}}\left(\pi + D_X\pi D_Y\right)\big\|_F$ is used by Li et al. (2023, Proposition 3.1) to establish a Luo–Tseng error bound for GW. Their quantity is the projected-gradient displacement with $\eta = 1/2$, whereas ours uses a Lipschitz-based step size. When the projected-gradient residual is sufficiently small, the Luo–Tseng error bound (Li et al., 2023) bounds the distance from $\widehat{W}_t$ to the GW stationary set by a constant multiple of $r_{\mathrm{stat}}(\widehat{W}_t)$. Thus, a smaller residual indicates less first-order improvement remained and a closer proximity to the stationary set.*

4.2 Graph Alignment

Synthetic Graph Alignment Following Li et al. (2023), we draw a 250-node Gaussian random partition graph and corrupt it by adding random missing edges and 75 noisy target nodes. We use adjacency matrices as D_X and D_Y, uniform marginals, and average over 30 seeds. We compare A-KL-BAPG with fixed KL-BAPG (Li et al., 2023), KL-BAPG with post-hoc Sinkhorn rounding, $1/k$ schedule, pure-power schedule, and several GW solvers, including conditional-gradient GW (POT CG) (Vayer et al., 2019), proximal-point GW (POT PPA) (Xu et al., 2019b), entropic projected-gradient GW (POT EPGD) (Peyré et al., 2016), and POT's implementation of BAPG. We additionally include the recent inexact projected-gradient (iPG) method of Liang & Yang (2026). For A-KL-BAPG, we use $K = 4000$ total iterations, $k_0 = 1000$, $\rho_0 = 0.1$, and tail parameters $\alpha_0 = 1$, $\beta = 0.75$, $a = 0.1$, and $\tau = 0.5$. For the pure power schedule, $\beta = 0.75$. The hyperparameters for each baseline are set according to the settings in their respective papers, except that POT BAPG uses $\varepsilon = 0.4$ to match the penalty $\rho = 0.1$ of KL-BAPG. Following Li et al. (2023), we use top-1 matching accuracy to evaluate the quality of graph alignment. We also report the mean GW objective, feasibility gap, stationarity residual, and runtime, where higher top-1 accuracy and lower objective, feasibility gap, stationarity residual, and runtime are better. Specific graph construction details, baseline information, and accuracy definition are provided in Appendix B.1.

Table 1 and Figure 2 confirm the strength of BAPG for graph alignment: KL-BAPG and POT BAPG obtain high accuracy and low GW objectives, but the returned couplings are far from feasible and stationary. Post-hoc Sinkhorn rounding preserves KL-BAPG's objective and accuracy with reduced feasibility gap, yet still yields a higher stationarity residual, showing that marginal correction alone cannot restore stationarity. A-KL-BAPG retains the same accuracy and objective while reducing the feasibility gap by more than 99% and the stationarity residual by 98% relative to KL-BAPG. $1/k$ and pure-power schedules are nearly feasible but fall short on accuracy and stationarity. POT CG attains a smaller feasibility gap and stationarity residual, yet with substantially worse accuracy and GW objective. POT PPA and iPG achieve high performance across all metrics but require significantly longer runtimes due to inner iterations.

Real Graph Alignment. We evaluate A-KL-BAPG on the PROTEINS and ENZYMES datasets from TUDataset (Borgwardt et al., 2005; Dobson & Doig, 2003; Morris et al., 2020). Following Li et al. (2023), we construct each target graph by adding 10% missing edges and 10% noisy nodes to its source graph. We use unweighted adjacency matrices as the GW structure matrices and uniform node measures. Results are averaged over 30 independent corruptions per graph. Dataset and corruption details are provided in Appendix B.2. We use the same baselines as in Section 4.2 and additionally include Degree Hungarian (Kuhn, 1955), FAQ with ten random restarts (Vogelstein et al., 2015), ScalaGW (Xu et al., 2019a), specGW (Chowdhury & Needham, 2021), and BPG and BPG-S (Xu et al., 2019b). Baseline hyperparameters follow their respective papers, and BAPG schedules use the same hyperparameters as in Section 4.2. In Section 4.2, we report top-1 accuracy for direct comparison. Here, however, several baselines, such as FAQ, directly return discrete assignments, whereas top-1 decoding can map multiple nodes to the same target node. Therefore, we apply the Hungarian decoding for a global one-to-one alignment and report the resulting matching accuracy (Kuhn, 1955). GW objective, feasibility gap, and stationarity residual are also reported if applicable. The decoding and metric definitions are provided in Appendix B.2.

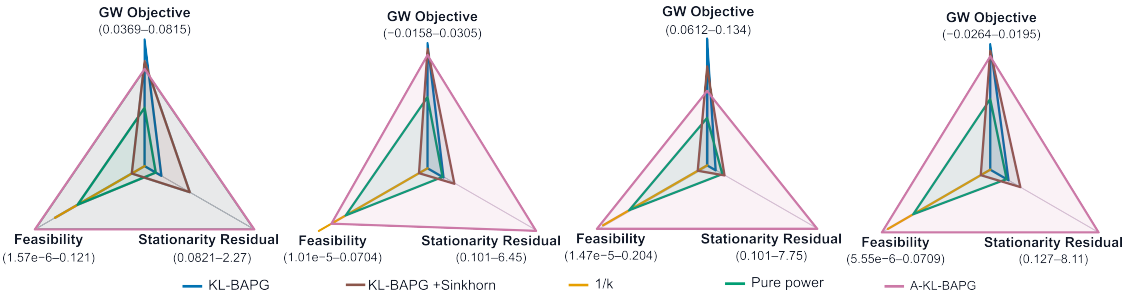
Table 2: ENZYMES and PROTEINS graph alignment performance.

	ENZYMES					PROTEINS				
Method	Accuracy	GW Objective	Feasibility	Stationarity	Time (s)	Accuracy	GW Objective	Feasibility	Stationarity	Time (s)
Degree Hungarian	0.282	−0.041	9.91×10^{-17}	0.406	0.000	0.303	−0.077	1.01×10^{-16}	0.390	0.000
SciPy FAQ	0.503	−0.093	1.12×10^{-16}	0.117	0.040	0.469	−0.131	1.15×10^{-16}	0.100	0.117
ScalaGW	0.137	−0.032	1.38×10^{-16}	–	0.283	0.185	−0.058	1.51×10^{-16}	–	0.249
SpecGW	0.199	−0.067	2.82×10^{-16}	–	0.052	0.204	−0.104	2.86×10^{-16}	–	0.029
BPG	0.706	−0.091	6.89×10^{-6}	0.020	0.175	0.645	−0.119	5.59×10^{-6}	0.034	0.151
BPG-S	0.709	−0.091	5.37×10^{-4}	0.020	0.032	0.649	−0.119	6.59×10^{-4}	0.035	0.038
iPG	0.694	−0.091	2.02×10^{-5}	0.002	2.681	0.673	−0.122	1.95×10^{-5}	0.002	4.060
POT CG	0.204	−0.085	3.40×10^{-16}	0.004	0.074	0.211	−0.122	3.26×10^{-16}	0.008	0.087
POT PPA	0.731	−0.092	2.21×10^{-4}	0.006	52.194	0.693	−0.121	2.20×10^{-4}	0.007	69.666
POT EPGD	0.684	−0.087	7.17×10^{-6}	0.103	77.281	0.632	−0.115	1.03×10^{-5}	0.103	69.272
POT BAPG	0.200	−0.058	0.782	0.440	0.131	0.263	−0.081	0.754	0.436	0.473
$1/k$ schedule	0.326	−0.018	1.07×10^{-6}	0.128	0.380	0.336	−0.040	1.63×10^{-6}	0.147	0.915
Pure power	0.358	−0.021	7.08×10^{-6}	0.148	0.389	0.369	−0.052	9.86×10^{-6}	0.160	0.917
KL-BAPG	0.748	−0.094	0.113	0.178	1.483	0.716	−0.125	0.108	0.184	3.551
KL-BAPG + Sinkhorn	0.752	−0.093	0.092	0.417	1.689	0.719	−0.123	0.079	0.414	3.845
A-KL-BAPG	**0.747**	**−0.091**	**0.043**	**0.130**	**2.445**	**0.718**	**−0.120**	**0.041**	**0.132**	**4.056**

Table 2 confirms the strength of A-KL-BAPG on real graph alignment. It achieves comparable accuracy with fixed KL-BAPG on both ENZYMES and PROTEINS. At the same time, A-KL-BAPG reduces the feasibility gap by approximately 62% on both datasets and the stationarity residual by 27% on ENZYMES and 28% on PROTEINS. Consistent with the synthetic results, Sinkhorn rounding improves KL–BAPG's feasibility but worsens its stationarity, and is worse than A-KL-BAPG on both metrics. $1/k$ and pure-power schedules approach feasibility but yield substantially lower matching accuracy. Projection-based methods attain smaller feasibility and stationarity residuals, but generally at the cost of lower matching accuracy. Thus, A-KL-BAPG achieves the most favorable balance among the BAPG schedules in terms of alignment accuracy, feasibility, and stationarity.

4.3 Domain Adaptation

Synthetic Domain Adaptation We test FGW in synthetic domain adaptation with four binary tasks (Blobs, Two Circles, Two Moons, and Two Spirals), where the source has one common feature, and the target additionally has a target-only feature. The source geometry is label-aware and the target geometry is a 4-nearest-neighbor (4-nn) graph, following Aritake & Hino (2026). We use 400 source and 200 target observations, and repeat each experiment with the same 30 random seeds. Following Aritake & Hino (2026), we set $\lambda = 0.01$. Specific experiment settings are included in Appendix B.3. We compare KL–BAPG ($\rho_k = 0.5$), KL BAPG + Sinkhorn rounding, $1/k$, pure power ($\beta = 0.75$), and our A-KL-BAPG, with $k_0 = 500$, $\alpha_0 = 1$, $\beta = 0.75$, $a = 0.1$, and $\tau = 0.5$. Other baselines include OT-based domain adaptation (Courty et al., 2016), naive OT cluster (Ng et al., 2001), GW (Peyré et al., 2016), FGW (Vayer et al., 2019), and the LA-FGW, GB-FGW, and LAGB-FGW (Aritake & Hino, 2026). OT, GW, and FGW baselines are implemented using POT (Flamary et al., 2021). Hyperparameters for baselines are set according to their respective papers, with details provided in Appendix B.3. Specific accuracy calculation is provided in Appendix B.3.



(a) Blobs (b) Two Circles (c) Two Moons (d) Two Spirals

Figure 3: Optimization diagnostics of synthetic domain adaptation experiment. Outward is better, i.e., lower in values.

Figure 3 shows that fixed-penalty KL–BAPG makes the greatest raw objective progress at the cost of infeasibility split, while Sinkhorn rounding improves its feasibility but remains substantially less stationary than A-KL-BAPG. A-KL-BAPG retains most of the objective gain and achieves the smallest stationarity residuals, whereas the $1/k$ and pure-power schedules improve feasibility at the cost of objective progress. Table 3(a) further confirms the strength of A-KL-BAPG, where it exhibits the highest mean accuracy across the four data sets among all methods. It is therefore the best overall method, while POT LAGB–FGW and naive OT clustering remain competitive on individual data sets. Specific reconstruction visualizations are provided in Figure 5 in Appendix B.3.

Table 3: Synthetic target accuracy and HAR70+ results. $(^*)$ Included for completeness and not directly comparable to the other methods because they use additional task-specific knowledge and assumptions (Aritake & Hino, 2022).

(a) Synthetic target accuracy

Method	Blobs	Circles	Moons	Spirals
KL–BAPG	0.955	0.822	0.917	0.905
KL–BAPG + Sinkhorn	0.960	0.821	0.934	0.907
$1/k$ schedule	0.981	0.667	0.801	0.529
Pure power	0.981	0.696	0.848	0.570
A-KL-BAPG	**0.984**	**0.833**	**0.963**	**0.906**
POT OT	0.913	0.685	0.786	0.597
POT FGW	0.912	0.689	0.786	0.600
POT LA–FGW	0.932	0.708	0.636	0.729
POT GB–FGW	0.913	0.684	0.782	0.596
POT LAGB–FGW	0.940	0.826	0.975	0.859
POT GW	0.502	0.696	0.530	0.489
Naive OT	0.965	0.605	1.000	0.947

(b) HAR70+ performance

Method	Accuracy	GW Objective	Feasibility	Stationarity Residual
KL–BAPG	0.880	-0.012	0.034	1.559
KL–BAPG + Sinkhorn	0.880	-0.011	0.010	1.067
$1/k$ schedule	0.662	0.017	5.04×10^{-6}	4.029
Pure power	0.741	2.97×10^{-4}	4.25×10^{-5}	2.311
A-KL-BAPG	**0.880**	$\mathbf{-0.010}$	$\mathbf{5.07 \times 10^{-4}}$	**0.207**
POT OT	0.827	–	–	0.113
POT FGW	0.827	–	–	0.112
POT LA–FGW	0.829	–	–	0.112
POT GB–FGW	0.854	–	–	0.088
POT LAGB–FGW*	0.879	–	–	0.086
JDOT	0.805	–	–	1.857
DAEVS*	0.895	–	–	–

Real-data Domain Adaptation We evaluate on HAR70+, a human-activity-recognition dataset containing motion-sensor recordings from 18 older adults performing four activities: walking, standing, sitting, and lying with accelerometers placed on the back and thigh to record motion along three spatial axes at 50 Hz (Ustad et al., 2023; Aritake & Hino, 2026). The target graph is constructed by treating windows as nodes and connecting similar windows through a 4-nn graph. We evaluate all 18 subjects over 30 random splits, resulting in 540 trials. Preprocessing, feature construction, and split details are provided in Appendix B.4. Hyperparameters are set according to Section 4.3. We additionally include DAEVS (Aritake & Hino, 2022) and JDOT (Courty et al., 2017), with hyperparameters set according to the corresponding papers. We report target classification accuracy, i.e., the proportion of target samples whose predicted labels match ground-truth labels, for all methods and stationarity for all methods solving an FGW problem. We also include raw GW objective and marginal feasibility gap for KL–BAPG (+ Sinkhorn rounding), $1/k$, pure power, and A-KL-BAPG. Other baselines are not optimizing over the same objective.

Table 3 shows that A-KL-BAPG achieves the strongest overall performance among the BAPG variants. It matches KL–BAPG in target accuracy with a comparable GW objective and reduces the feasibility gap by 99% and the stationarity residual by 85%. Sinkhorn rounding improves KL–BAPG's feasibility at the cost of stationarity, while A-KL-BAPG performs better on both metrics. Among all baselines, A-KL-BAPG achieves one of the highest accuracies, second only to DAEVS, which is tailored to domain adaptation. DAEVS alternates between OT and hard pseudo-labeling and trains an RBF-SVM on the full set of target features, thereby denoising transported labels and exploiting additional sensor information (Aritake & Hino, 2022). This advantage relies on informative initial pseudo-labels, a suitable target classifier, and conditional invariance of the additional features. In contrast, A-KL-BAPG requires neither classifier training nor pseudo-label feedback while achieving comparable accuracy. POT FGW baselines achieve competitive accuracy at much lower runtime. With $\lambda = 0.01$, the objective gives much greater weight to the linear feature cost, which may favor methods that linearize the structural term. However, this speed–accuracy tradeoff is task-dependent: POT CG, SpecGW, and ScalaGW also use linearized structural updates, yet achieve substantially lower accuracy in graph alignment, as shown in Table 1-2. A-KL-BAPG remains competitive across both applications, though sometimes at a higher computational cost. The full comparison against all baselines is reported in Table 4 in Appendix B.4.

5 Conclusion

We proposed Adaptive KL-BAPG to remove the persistent feasibility gap of fixed-penalty BAPG while retaining its inexpensive alternating updates. Combining a finite fixed-penalty burn-in with an increasing-penalty phase and a projective-diameter safeguard to control the evolving coupling geometry, we achieve vanishing marginal infeasibility and absolute summability of the accumulated asymmetric Bregman errors. Under a support regularity condition, the iterates admit a stationary accumulation point of the original constrained GW problem. Synthetic and real experiments in graph alignment and domain adaptation show improved feasibility and stationarity compared to KL-BAPG while maintaining competitive objective values and predictive performance. Future work will focus on strengthening the current subsequence guarantee to full-sequence convergence, weakening the support regularity condition, and reducing the computational cost of the safeguards.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets, implement methods, execute code, polish writing, generative AI tools for retrieval and discovery (e.g., finding related work), and verify proofs. We have not used AI for drafting sections of the paper, and proving mathematical claims. Helping develop theoretical models or conceptual frameworks, formulating mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, interpreting results also didn't use generative AI tools. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Toshimitsu Aritake and Hideitsu Hino. Unsupervised domain adaptation for extra features in the target domain using optimal transport. *Neural Computation*, 34(12):2432–2466, 2022.

Toshimitsu Aritake and Hideitsu Hino. Domain adaptation with additional features via label-aware and graph-based fused Gromov-Wasserstein optimal transport. *Neural Computation*, 38(7):1203–1232, 2026. doi: 10.1162/NECO.a.1525.

Marin Ballu and Quentin Berthet. Mirror sinkhorn: fast online optimization on transport polytopes. In *International Conference on Machine Learning*, pp. 1595–1613. PMLR, 2023.

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

Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schönauer, S. V. N. Vishwanathan, Alex J. Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(Suppl. 1):i47–i56, 2005. doi: 10.1093/bioinformatics/bti1007.

James V Burke and Jorge J Moré. On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988.

He Chen, Jiajin Li, and Anthony Man-Cho So. Spurious stationarity and hardness results for bregman proximal-type algorithms. *arXiv preprint arXiv:2404.08073*, 2024.

Lénaïc Chizat. Annealed sinkhorn for optimal transport: convergence, regularization path and debiasing. *arXiv preprint arXiv:2408.11620*, 2024.

Samir Chowdhury and Facundo Mémoli. The Gromov-Wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

Samir Chowdhury and Tom Needham. Generalized spectral clustering via Gromov-Wasserstein learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 712–720. PMLR, 2021.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Pinar Demetci, Quang Huy Tran, Ievgen Redko, and Ritambhara Singh. Breaking isometric ties and introducing priors in Gromov-Wasserstein distances. In *International Conference on Artificial Intelligence and Statistics*, pp. 298–306. PMLR, 2024.

Bowen Deng, Lele Fu, Sheng Huang, Tianchi Liao, Jialong Chen, Zhang Tao, and Chuan Chen. GLNCD: Graph-level novel category discovery. *Advances in Neural Information Processing Systems*, 38:171058–171087, 2026.

Paul D Dobson and Andrew J Doig. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of molecular biology*, 330(4):771–783, 2003.

Simon P Eveson and Roger D Nussbaum. An elementary proof of the birkhoff-hopf theorem. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 117, pp. 31–55. Cambridge University Press, 1995.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114: 717–735, 1989.

Martin Idel. A review of matrix scaling and sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Adrian S Lewis and Stephen J Wright. Identifying activity. *SIAM Journal on Optimization*, 21(2):597–614, 2011.

Jiajin Li, Jianheng Tang, Lemin Kong, Huikang Liu, Jia Li, Anthony Man-Cho So, and Jose Blanchet. A convergent single-loop algorithm for relaxation of Gromov-Wasserstein in graph data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0jxPyVWmiiF.

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017.

Ling Liang and Lei Yang. A provably convergent and practical algorithm for Gromov–Wasserstein optimal transport. *arXiv preprint arXiv:2605.04175*, 2026.

Facundo Mémoli. Spectral Gromov-Wasserstein distances for shape matching. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 256–263. IEEE, 2009.

Facundo Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

Facundo Mémoli and Guillermo Sapiro. Comparing point clouds. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pp. 32–40, 2004.

Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. *arXiv preprint arXiv:2007.08663*, 2020.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pp. 2664–2672. PMLR, 2016.

Gabriel Rioux, Ziv Goldfeld, and Kengo Kato. Entropic Gromov-Wasserstein distances: Stability and algorithms. *Journal of Machine Learning Research*, 25(363):1–52, 2024.

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

Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

Astrid Ustad, Aleksej Logacjov, Stine Øverengen Trollebø, Pernille Thingstad, Beatrix Vereijken, Kerstin Bach, and Nina Skjæret Maroni. Validation of an activity type recognition model classifying daily physical behavior in older adults: the har70+ model. *Sensors*, 23(5):2368, 2023.

Titouan Vayer, Nicolas Courty, Romain Tavenard, and Rémi Flamary. Optimal transport for structured data with application on graphs. In *International Conference on Machine Learning*, pp. 6275–6284. PMLR, 2019.

Joshua T Vogelstein, John M Conroy, Vince Lyzinski, Louis J Podrazik, Steven G Kratzer, Eric T Harley, Donniell E Fishkind, R Jacob Vogelstein, and Carey E Priebe. Fast approximate quadratic programming for graph matching. *PLOS one*, 10(4):e0121002, 2015.

Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019a.

Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-Wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pp. 6932–6941. PMLR, 2019b.

Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, volume 7, pp. 2969–2975, 2018.

A PROOFS

This appendix records the main technical steps omitted from the main text to establish Theorem 2. The argument first relates the guarded BAPG cycle to exact Sinkhorn contraction, then derives feasibility, Lyapunov, residual, and stationarity statement. All asymptotic statements focus on the tail iteration and a finite prefix changes only constants.

Notation To simplify notation, we relabel $(\bar{\pi}^t, \bar{w}^t, \bar{\alpha}_t, \bar{\rho}_t)$ as $(\pi^k, w^k, \alpha_k, \rho_k)$, where α_k is the accepted step from the guarded line search and $\rho_k := 1/\alpha_k$. Thus the line search condition in Algorithm 1 becomes $\Delta(w^{t+1}) \le \Delta_0 + a\log(t+2)$ for $t \ge 0$, which is equivalent to

$$\Delta(w^t) \le \Delta_0 + a\log(t+1), \qquad t \ge 1$$

with $\Delta(w^0) = \Delta_0 = \Delta(w^{k_0})$ for $t = 0$.

A.1 PROJECTIVE CONTRACTION AND EXACT SINKHORN

Definition 1 (Hilbert's projective metric)**.** For $u, v \in \mathbb{R}^r_{++}$, define $d_H(u,v) := \log \frac{\max_i (u_i/v_i)}{\min_i (u_i/v_i)}$.

Definition 2 (Projective diameter)**.** For $K \in \mathbb{R}^{r\times s}_{++}$, define

$$\Delta(K) := \max_{i,i',j,j'} \log \frac{K_{ij}K_{i'j'}}{K_{ij'}K_{i'j}}.$$

Theorem 3 (Birkhoff contraction theorem Eveson & Nussbaum (1995))**.** *For $K \in \mathbb{R}^{r\times s}_{++}$ and for all $u, v \in \mathbb{R}^s_{++}$,*

$$d_H(Ku, Kv) \le \tanh\left(\frac{\Delta(K)}{4}\right) d_H(u,v).$$

Lemma 2. *For positive vectors $u, v \in \mathbb{R}^r_{++}$, $d_H(u^{-1}, v^{-1}) = d_H(u,v)$, where inversion is entrywise. Moreover, if D_1 and D_2 are positive diagonal matrices and X is strictly positive, then $\Delta(D_1 X D_2) = \Delta(X)$.*

Proof. $\max_i u_i^{-1}/v_i^{-1} = \max_i v_i/u_i = \frac{1}{\min_i u_i/v_i}$, and similarly $\min_i u_i^{-1}/v_i^{-1} = \frac{1}{\max_i u_i/v_i}$. Thus the quotient inside the logarithm defining d_H remains unchanged.

Denote $(D_1 X D_2)_{ij} = a_i X_{ij} b_j$ with $a_i, b_j > 0$. For every (i, i', j, j'),

$$\frac{(D_1XD_2)_{ij}(D_1XD_2)_{i'j'}}{(D_1XD_2)_{ij'}(D_1XD_2)_{i'j}} = \frac{a_iX_{ij}b_ja_{i'}X_{i'j'}b_{j'}}{a_iX_{ij'}b_{j'}a_{i'}X_{i'j}b_j} = \frac{X_{ij}X_{i'j'}}{X_{ij'}X_{i'j}}.$$

Thus taking the maxima of logarithms gives $\Delta(D_1 X D_2) = \Delta(X)$. □

For a strictly positive matrix X, we define the exact KL projections onto the row and column marginal constraints by $R(X) := \mathcal{P}^{\mathrm{KL}}_{\mathcal{C}_1}(X)$, $C(X) := \mathcal{P}^{\mathrm{KL}}_{\mathcal{C}_2}(X)$, and $S(X) := C(R(X))$. Thus S denotes one exact Sinkhorn row-column scaling cycle. When $X \in \mathcal{C}_2$, define its row-ratio vector $r(X) := \frac{1}{\mu} X \mathbf{1}_m \in \mathbb{R}^n_{++}$ and $r_i(X) = \frac{1}{\mu_i}\sum_j X_{ij}$. Then $r(X) = \mathbf{1}_n$ if and only if $X \in \mathcal{C}_1$. We also define two row-stochastic matrices associated with X:

$$(A_X)_{ij} := \frac{X_{ij}}{\sum_\ell X_{i\ell}}, \qquad (B_X)_{ji} := \frac{X_{ij}}{\nu_j}.$$

Since $X \in \mathcal{C}_2$, each column of X sums to ν_j and B_X is row-stochastic after the index convention above.

Lemma 3. *Let $X \in \mathcal{C}_2$ be strictly positive. Then $r(S(X)) = A_X\left(B_X(r(X)^{-1})\right)^{-1}$, where inversion is entrywise.*

Proof. $R(X)_{ij} = \frac{\mu_i}{\sum_l X_{il}} X_{ij} = X_{ij}/r_i$. Denote $t_j = \sum_i X_{ij}/r_i$. $t_j/\nu_j = \sum_i \frac{X_{ij}}{\nu_j} r_i^{-1} = \left(B_X(r^{-1})\right)_j$. Then $S(X)_{ij} = \frac{\nu_j}{t_j}\frac{X_{ij}}{r_i}$. Since $X_{ij} = \mu_i r_i (A_X)_{ij}$,

$$\frac{(S(X)\mathbf{1}_m)_i}{\mu_i} = \frac{1}{\mu_i r_i}\sum_j X_{ij}\frac{\nu_j}{t_j} = \sum_j (A_X)_{ij}\frac{\nu_j}{t_j} = \sum_j (A_X)_{ij}\left(B_X(r^{-1})\right)_j^{-1}.$$

□

From Lemma 3, we can view one exact Sinkhorn cycle as two positive linear maps applied to the row-ratio vector. Hence Theorem 3 gives the following estimate.

Lemma 4 (Projective contraction of exact Sinkhorn)**.** *Let $X \in \mathcal{C}_2$ be strictly positive. Then*

$$d_H(r(S(X)), \mathbf{1}_n) \le \tanh^2\left(\frac{\Delta(X)}{4}\right) d_H(r(X), \mathbf{1}_n).$$

Proof. Since $A_X\mathbf{1}_m = \mathbf{1}_n$, $d_H(r(S(X)), \mathbf{1}_n) = d_H\left(A_X\left(B_X(r^{-1})^{-1}\right), A_X(\mathbf{1}_m)\right)$. By Lemma 2, $d_H(B_X(r^{-1})^{-1}, \mathbf{1}_m) = d_H(B_X(r^{-1}), \mathbf{1}_m)$. By Theorem 3 and Lemma 2,

$$d_H(r(S(X)), \mathbf{1}_n) \le \tanh\left(\frac{\Delta(A_X)}{4}\right) d_H(B_X(r^{-1}), \mathbf{1}_m) = \tanh\left(\frac{\Delta(X)}{4}\right) d_H(B_X(r^{-1}), \mathbf{1}_m).$$

Apply Theorem 3 and Lemma 2 again,

$$d_H(B_X(r^{-1}), \mathbf{1}_m) \le \tanh\left(\frac{\Delta(B_X)}{4}\right) d_H(r^{-1}, \mathbf{1}_n) = \tanh\left(\frac{\Delta(X)}{4}\right) d_H(r, \mathbf{1}_n).$$

Therefore, combining all together, we obtain

$$d_H(r(S(X)), \mathbf{1}_n) \le \tanh^2\left(\frac{\Delta(X)}{4}\right) d_H(r(X), \mathbf{1}_n).$$

□

A.2 Controlled Projective Growth and Finite Backtracking

BAPG differs from the exact Sinkhorn only by the exponential perturbation applied before each normalization. In the following Lemma 5, we controls the projective diameter growth of one cycle.

Lemma 5. *For any accepted KL-BAPG step with step size $\alpha > 0$, $\Delta(w^+) \le \Delta(w) + B_\Delta \alpha$, where $B_\Delta := 8M_G$ is a conservative projective-growth constant.*

Proof. For one exponential perturbation $Z = X \odot \exp(\alpha H)$ with $\|H\|_\infty \le M_G$, every log cross-ratio changes by $\alpha\big(H_{ij} + H_{i'j'} - H_{ij'} - H_{i'j}\big)$, whose absolute value is at most $4M_G\alpha$. Thus $\Delta(Z) \le \Delta(X) + 4M_G\alpha$. The row and column KL projections are positive diagonal scalings and therefore do not change cross-ratios. A full KL-BAPG cycle contains two exponential perturbations, so $\Delta(w^+) \le \Delta(w) + 8M_G\alpha \le \Delta(w) + B_\Delta\alpha$. □

We record two elementary properties of the guarded tail. The first guarantees positivity throughout all cycles and the second summarizes the resulting behavior of the backtracking rule.

Lemma 6. *If $\pi^0, w^0 > 0$ entriwise, then for every iterate, π^k, w^k is entrywise positive.*

Proof. Assume $w^k > 0$ entrywise. Then $z_1^k = w^k \odot \exp(\alpha_k G(w^k))$ is positive entrywise. Since $\mu_i > 0$, $\pi_{ij}^{k+1} = (\mu_i z_{1,ij}^k)/\sum_l z_{1,il}^k$ is entrywise positive, i.e., $\pi^{k+1} > \mathbf{0}$. Similarly, $w_{ij}^{k+1} > 0$ and hence $w^{k+1} > 0$ entrywise. By induction, entrywise positivity holds for all iterates. □

Lemma 7. *Assume that the current tail iterate satisfies $\Delta(w^t) \le \Gamma_t$. Then the backtracking loop in Algorithm 1 terminates finitely for each iteration. Moreover, the accepted step satisfies $\ell_t \le \alpha_t \le \bar{\alpha}_t$ and $\alpha_t \le \alpha_{t-1}$. Since $\beta < 1$, there exist constants $c_-, c_+ > 0$ and T_0 such that, for all $t \ge T_0$, $\frac{c_-}{t+2} \le \alpha_t \le c_+(t+1)^{-\beta}$.*

Proof. By the construction of h_t and $\bar{\alpha}_t$, the sequences h_t and $\bar{\alpha}_t$ are nonincreasing. Then the sequence of ℓ_t is also nonincreasing and so is the sequence of s_t. Since $\alpha_{-1} = +\infty$, $\alpha_{-1} \geq \ell_0$ for $t = 0$.

For arbitrary $t \geq 1$, assume $\alpha_{t-1} \geq \ell_{t-1} \geq \ell_t$. Then $\ell_t \leq \alpha_{t-1}$ and $\min\{\bar{\alpha}_t, \alpha_{t-1}\} \leq \alpha_{t-1}$, i.e., $s_t \leq \alpha_{t-1}$. We also know that $\ell_t \leq s_t \leq \bar{\alpha}_t$. If $s_t \leq h_t$, we have showed that at this time the safe guard is satisfied and we can direct accept $a_t = s_t$. Otherwise when $s_t > h_t$, we backtrack until we find the first candidate α_t such that $\alpha_t \leq h_t$. Then by construction, $\alpha_t > \tau h_t$. Since $\ell_t \leq \tau h_t$, $\alpha_t \geq \ell_t$. Since the sequence of s_t decreases,

$$\ell_t \leq \alpha_t \leq s_t \leq \bar{\alpha}_t, \qquad \alpha_t \leq s_t \leq \alpha_{t-1}.$$

For the final bound, we know that $\bar{\alpha}_t \asymp (t+1)^{-\beta}$ and

$$h_t = \frac{a}{B_\Delta} \log\left(1 + \frac{1}{t+1}\right).$$

Since $\beta < 1$, eventually $\bar{\alpha}_t \geq h_t$, so $\ell_t = \tau h_t$ for all sufficiently large t. Since $\log(1 + 1/(t+1)) \geq 1/(t+2)$ and $\alpha_t \leq \bar{\alpha}_t$, there exists $c_- = \tau a / B_\Delta > 0$, $c_+ > 0$ and T_0 such that for all $t \geq T_0$,

$$\frac{c_-}{t+2} \leq \alpha_t \leq c_+ (t+1)^{-\beta}.$$

□

A.3 Marginal Feasibility Estimates and Cross-Divergence Bounds

Fix t and define the exact Sinkhorn cycle launched from w^t by $\widehat{\pi}^{t+1} := R(w^t)$ and $\widehat{w}^{t+1} := S(w^t) = C(\widehat{\pi}^{t+1})$. We compare the actual BAPG iterates (π^{t+1}, w^{t+1}) with $(\widehat{\pi}^{t+1}, \widehat{w}^{t+1})$.

Lemma 8. *For every t and every i, j,*

$$\exp(-2M_G\alpha_t) \leq \frac{\pi_{ij}^{t+1}}{\widehat{\pi}_{ij}^{t+1}} \leq \exp(2M_G\alpha_t), \qquad \exp(-6M_G\alpha_t) \leq \frac{w_{ij}^{t+1}}{\widehat{w}_{ij}^{t+1}} \leq \exp(6M_G\alpha_t).$$

Consequently, $d_H\big(r(w^{t+1}), r(\widehat{w}^{t+1})\big) \leq 12M_G\alpha_t$.

Proof.

$$\frac{\pi_{ij}^{t+1}}{\widehat{\pi}_{ij}^{t+1}} = \exp(\alpha_t G(w^t)_{ij}) \frac{\sum_l w_{il}^t}{\sum_\ell w_{il}^t \exp(\alpha_t G(w^t)_{il})},$$

with $\exp(-M_G\alpha_t)\sum_\ell w_{il}^t \leq \sum_\ell w_{il}^t \exp(\alpha_t G(w^t)_{il}) \leq \exp(M_G\alpha_t)\sum_\ell w_{il}^t$. Thus $\exp(-2M_G\alpha_t) \leq \pi_{ij}^{t+1}/\widehat{\pi}_{ij}^{t+1} \leq \exp(2M_G\alpha_t)$. Recall that $z_2^t = \pi^{t+1} \odot \exp(\alpha_t G(\pi^{t+1}))$ and $\widehat{w}^{t+1} = \mathcal{P}_{\mathcal{C}_2}^{\mathrm{KL}}(\widehat{\pi}^{t+1})$, $w^{t+1} = \mathcal{P}_{\mathcal{C}_2}^{\mathrm{KL}}(z_2^t)$. Hence

$$\frac{w_{ij}^{t+1}}{\widehat{w}_{ij}^{t+1}} = \frac{z_{2,ij}^t / \sum_\ell z_{2,\ell j}^t}{\widehat{\pi}_{ij}^{t+1} / \sum_\ell \widehat{\pi}_{\ell j}^{t+1}} = \frac{z_{2,ij}^t}{\widehat{\pi}_{ij}^{t+1}} \cdot \frac{\sum_\ell \widehat{\pi}_{\ell j}^{t+1}}{\sum_\ell z_{2,\ell j}^t}, \quad \text{and} \quad \frac{z_{2,ij}^t}{\widehat{\pi}_{ij}^{t+1}} = \frac{\pi_{ij}^{t+1} \exp(\alpha_t G(\pi^{t+1})_{ij})}{\widehat{\pi}_{ij}^{t+1}} \in [e^{-3M_G\alpha_t}, e^{3M_G\alpha_t}].$$

Summing over each column, we obtain

$$\frac{\sum_\ell \pi_{\ell j}^{t+1} \exp(\alpha_t G(\pi^{t+1})_{\ell j})}{\sum_\ell \widehat{\pi}_{\ell j}^{t+1}} \in [e^{-3M_G\alpha_t}, e^{3M_G\alpha_t}] \quad \text{and} \quad \frac{\sum_\ell \widehat{\pi}_{\ell j}^{t+1}}{\sum_\ell \pi_{\ell j}^{t+1} \exp(\alpha_t G(\pi^{t+1})_{\ell j})} \in [e^{-3M_G\alpha_t}, e^{3M_G\alpha_t}].$$

Putting these together, we obtain $\frac{w_{ij}^{t+1}}{\widehat{w}_{ij}^{t+1}} \in [e^{-3M_G\alpha_t} \cdot e^{-3M_G\alpha_t}, e^{3M_G\alpha_t} \cdot e^{3M_G\alpha_t}] = [e^{-6M_G\alpha_t}, e^{6M_G\alpha_t}]$. □

With Lemma 8, we can derive a polynomial bound for the projective row error.

Lemma 9. *Assume $\beta > p$. Then there exist constants $\widetilde{C}_\delta > 0$ and $K_\delta \geq 0$ such that, for every $t \geq K_\delta$, $d_H(r(w^t), \mathbf{1}_n) \leq \widetilde{C}_\delta \eta_t$.*

Proof. By Lemma 8, $d_H(r(w^{t+1}), r(\widehat{w}^{t+1})) \leq 12M_G\alpha_t \leq C_\alpha(t+1)^{-\beta}$ for some constant $C_\alpha > 0$ and all sufficiently large t. By the accepted diameter guard, $\Delta(w^t) \leq \Delta_0 + a\log(t+1)$ and hence

$$1 - \tanh^2\left(\frac{\Delta(w^t)}{4}\right) = \operatorname{sech}^2\left(\frac{\Delta(w^t)}{4}\right) \geq \exp(-\frac{\Delta(w^t)}{2}) \geq c_\Delta (t+1)^{-p}$$

for some $c_\Delta > 0$. Then $\tanh^2(\Delta(w^t)/4) \le 1 - c_\Delta(t+1)^{-p}$. By Lemma 4,

$$d_H(r(\widehat{w}^{t+1}), \mathbf{1}_n) = d_H(r(S(w^t)), \mathbf{1}_n) \le \tanh\left(\Delta(w^t)/4\right)^2 d_H(r(w^t), \mathbf{1}_n) \le \left(1 - c_\Delta(t+1)^{-p}\right) d_H(r(w^t), \mathbf{1}_n).$$

By triangle inequality, $d_H(r(w^{t+1}), \mathbf{1}_n) \le d_H(r(w^{t+1}), r(\widehat{w}^{t+1})) + d_H(r(\widehat{w}^{t+1}), \mathbf{1}_n)$. Then

$$d_H(r(w^{t+1}), \mathbf{1}_n) \le \left(1 - c_\Delta(t+1)^{-p}\right) d_H(r(w^t), \mathbf{1}_n) + C_\alpha(t+1)^{-\beta}.$$

Let $s := \beta - p > 0$. If $d_H(r(w^t), \mathbf{1}_n) \le K(t+1)^{-s}$, then $(1 - c_\Delta(t+1)^{-p})\, d_H(r(w^t), \mathbf{1}_n) \le K(t+1)^{-s} - c_\Delta K(t+1)^{-(p+s)}$. Since $p + s = \beta$, the recurrence gives

$$d_H(r(w^{t+1}), \mathbf{1}_n) \le K(t+1)^{-s} - (c_\Delta K - C_\alpha)(t+1)^{-\beta}.$$

Choose K sufficiently large such that $c_\Delta K > C_\alpha$. Since $p = a/2 < 1$, $\beta = p + s < s + 1$ and the correction term $(t+1)^{-\beta}$ decays more slowly than $(t+2)^{-s-1}$, which is the size of $(t+1)^{-s} - (t+2)^{-s}$. Thus the implication $d_H(r(w^t), \mathbf{1}_n) \le K(t+1)^{-s} \Rightarrow d_H(r(w^{t+1}), \mathbf{1}_n) \le K(t+2)^{-s}$ holds for all sufficiently large t. Enlarging K to absorb the finitely many earlier indices gives the claim with $\widetilde{C}_\delta = K$. □

Define the row sums of z_1^t by $r_i^t := (z_1^t \mathbf{1}_m)_i = \sum_{j=1}^m z_{1,ij}^t$. Since π^{t+1} is the KL projection of z_1^t onto the row-marginal constraint set $\mathcal{C}_1$, we have the explicit row-normalization $\pi_{ij}^{t+1} = (\mu_i / r_i^t) z_{1,ij}^t$.

The second half-step is analogous except that now the projection is onto the column marginal constraint set $\mathcal{C}_2$. Define the column sums of z_2^t by $c_j^t := ((z_2^t)^\top \mathbf{1}_n)_j = \sum_{i=1}^n z_{2,ij}^t$. Equivalently, $c_j^t = \sum_i z_{2,ij}^t$. Since w^{t+1} is the KL projection of z_2^t onto $\mathcal{C}_2$, $w_{ij}^{t+1} = (\nu_j / c_j^t) z_{2,ij}^t$.

Lemma 10 (Exact Row/Column Projection Identities)**.** *For every t, for row projection, the following identities hold:* $\|\pi^{t+1} - z_1^t\|_1 = \|z_1^t \mathbf{1}_m - \mu\|_1$,

$$D_h(\pi^{t+1}, z_1^t) = \sum_{i=1}^n \left[\mu_i \log \frac{\mu_i}{r_i^t} - \mu_i + r_i^t\right], \quad \textit{and} \quad D_h(z_1^t, \pi^{t+1}) = \sum_{i=1}^n \left[r_i^t \log \frac{r_i^t}{\mu_i} - r_i^t + \mu_i\right].$$

Similarly, for column projection, $\|w^{t+1} - z_2^t\|_1 = \|(z_2^t)^\top \mathbf{1}_n - \nu\|_1$,

$$D_h(w^{t+1}, z_2^t) = \sum_{j=1}^m \left[\nu_j \log \frac{\nu_j}{c_j^t} - \nu_j + c_j^t\right], \quad \textit{and} \quad D_h(z_2^t, w^{t+1}) = \sum_{j=1}^m \left[c_j^t \log \frac{c_j^t}{\nu_j} - c_j^t + \nu_j\right].$$

Proof. We know that $\sum_{j=1}^m |\pi_{ij}^{t+1} - z_{1,ij}^t| = \left|\frac{\mu_i}{r_i^t} - 1\right| \sum_{j=1}^m z_{1,ij}^t = |\mu_i - r_i^t|$. Summing over i, we obtain $\|\pi^{t+1} - z_1^t\|_1 = \|z_1^t \mathbf{1}_m - \mu\|_1$.

$$D_h(\pi^{t+1}, z_1^t) = \sum_{i,j} \left(\pi_{ij}^{t+1} \log \frac{\pi_{ij}^{t+1}}{z_{1,ij}^t} - \pi_{ij}^{t+1} + z_{1,ij}^t\right) = \sum_{ij} z_{1,ij}^t \left(\frac{\mu_i}{r_i^t} \log \frac{\mu_i}{r_i^t} - \frac{\mu_i}{r_i^t} + 1\right)$$
$$= \sum_{i=1}^n \left[\mu_i \log \frac{\mu_i}{r_i^t} - \mu_i + r_i^t\right].$$

The derivation of formula for $D_h(z_1^t, \pi^{t+1})$ is identical. The proof for column projection identities is analogous to the above proof for row projection identities. □

With the exact row projection identities, we can now convert the projective row-ratio estimate into ordinary marginal infeasibility estimates.

Proof for Proposition 3

Proof. For each t, define the row ratio vector $\theta_i^t := (w^t \mathbf{1}_m)_i / \mu_i$. Since $(w^t)^\top \mathbf{1}_n = \nu$, $\sum_i (w^t \mathbf{1}_m)_i = \sum_{i,j} w_{ij}^t = \sum_j \nu_j = 1$ and $\sum_i \theta_i^t \mu_i = \sum_i (w^t \mathbf{1}_m)_i = 1$. Since $\mu_i > 0$ and $\sum_i \mu_i = 1$, $\min_i(\theta_i^t) \le 1 \le \max_i(\theta_i^t)$. Then for every $i \in [n]$, $\exp(-d_H(\theta^t, \mathbf{1}_n)) \le \theta_i^t \le \exp(d_H(\theta^t, \mathbf{1}_n))$ and

$$\|w^t \mathbf{1}_m - \mu\|_1 = \sum_i \mu_i |\theta_i^t - 1| \le \exp(d_H(\theta^t, \mathbf{1}_n)) - 1.$$

Define $K_{\mathrm{row}} := \min\left\{t \geq K_\delta : \widetilde{C}_\delta \eta_t \leq 1\right\}$. Then for every $t \geq K_{\mathrm{row}}$, $0 \leq d_H(r(w^t), \mathbf{1}_n) \leq \widetilde{C}_\delta \eta_t \leq 1$. Define $C_{\mathrm{row}} := e\widetilde{C}_\delta$. By the mean value theorem, $\exp(d_H(r(w^t), \mathbf{1}_n)) - 1 \leq e d_H(r(w^t), \mathbf{1}_n)$ and thus $\|w^t \mathbf{1}_m - \mu\|_1 \leq e\widetilde{C}_\delta \eta_t = C_{\mathrm{row}} \eta_t$. Define $C_z := \exp(M_G c_+) M_G c_+$. Since $|G_{ij}(w^t)| \leq M_G$ and $\alpha_t \leq c_+(t+1)^{-\beta} \leq c_+$, the mean value theorem gives $|\exp(\alpha_t G_{ij}(w^t)) - 1| \leq e^{M_G c_+} M_G \alpha_t \leq C_z (t+1)^{-\beta}$.

Since $z_1^t - w^t = w^t \odot (\exp(\alpha_t G(w^t)) - 1)$, $\sum_{i,j} w_{ij}^t = 1$ and $p \geq 0$, $\|z_1^t - w^t\|_1 \leq C_z (t+1)^{-\beta} \leq C_z \eta_t$. Define $C_{\mathrm{fea}} := \exp(M_G c_+) M_G c_+ + e\widetilde{C}_\delta$. Then $\|z_1^t \mathbf{1}_m - \mu\|_1 \leq \|z_1^t - w^t\|_1 + \|w^t \mathbf{1}_m - \mu\|_1 \leq (C_{\mathrm{row}} + C_z)\eta_t = C_{\mathrm{fea}} \eta_t$. Since $w^t \in \mathcal{C}_2$, by Lemma 10, $\|\pi^{t+1} - z_1^t\|_1 = \|z_1^t \mathbf{1}_m - \mu\|_1$ and

$$\begin{aligned}\|(\pi^{t+1})^\top \mathbf{1}_n - \nu\|_1 \leq \|\pi^{t+1} - w^t\|_1 &\leq \|\pi^{t+1} - z_1^t\|_1 + \|z_1^t - w^t\|_1 \\ &\leq \|z_1^t \mathbf{1}_m - \mu\|_1 + \|z_1^t - w^t\|_1 \leq 2C_{\mathrm{fea}} \eta_t\end{aligned}$$

We hence show that $d_H(r(w^t), \mathbf{1}_n) = O(\eta_t)$, $\|z_1^t \mathbf{1}_m - \mu\|_1 = O(\eta_t)$, and $\|(\pi^{t+1})^\top \mathbf{1}_n - \nu\|_1 = O(\eta_t)$.

Now define $K_{\mathrm{cross},1} := \min\left\{t \geq K_{\mathrm{row}} : C_{\mathrm{fea}} \eta_t \leq \frac{\mu_{\min}}{2}\right\}$. Then for every $t \geq K_{\mathrm{cross},1}$, $|r_i^t - \mu_i| \leq C_{\mathrm{fea}} \eta_t \leq \mu_{\min}/2$ where $\mu_{\min} := \min_i \mu_i$. Hence for every i, $r_i^t \in [\mu_i/2, 3\mu_i/2]$. Since $z_1^t = w^t \odot \exp(\alpha_t G(w^t))$, $\log z_1^t - \log w^t = \alpha_t G(w^t)$. Since h is strictly convex, by the three-point identity of Bregman divergence,

$$\begin{aligned}D_h(\pi^{t+1}, w^t) &= D_h(\pi^{t+1}, z_1^t) + D_h(z_1^t, w^t) + \langle \nabla h(z_1^t) - \nabla h(w^t), \pi^{t+1} - z_1^t \rangle \\ &= D_h(\pi^{t+1}, z_1^t) + D_h(z_1^t, w^t) + \alpha_t \langle G(w^t), \pi^{t+1} - z_1^t \rangle.\end{aligned}$$

By Lemma 10, for $t \geq K_{\mathrm{cross},1}$, $|\mu_i - r_i^t| \leq \|\pi^{t+1} - z_1^t\|_1 = \|z_1^t \mathbf{1}_m - \mu\|_1 \leq C_{\mathrm{fea}} \eta_t \leq \mu_{\min}/2$. To satisfy this, we need $r_i^t \in \left[\frac{\mu_i}{2}, \frac{3\mu_i}{2}\right]$ for every $i \in [n]$. Define $\phi_i(s) := \mu_i \log \frac{\mu_i}{s} - \mu_i + s$. By Lemma 10, $D_h(\pi^{t+1}, z_1^t) = \sum_i \phi_i(r_i^t)$. Since $\phi_i(\mu_i) = 0$, $\phi_i'(\mu_i) = 0$, and $\phi_i''(s) = \frac{\mu_i}{s^2}$, for sufficiently large t,

$$\phi_i''(s) = \frac{\mu_i}{s^2} \leq \frac{4}{\mu_i} \leq \frac{4}{\mu_{\min}},$$

for all $s \in \left[\frac{\mu_i}{2}, \frac{3\mu_i}{2}\right]$. By Taylor's theorem, around $s = \mu_i$, $\phi_i(r_i^t) \leq \frac{2}{\mu_{\min}}(r_i^t - \mu_i)^2$. Summing over $i \in [n]$, we obtain

$$D_h(\pi^{t+1}, z_1^t) \leq \frac{2}{\mu_{\min}} \|z_1^t \mathbf{1}_m - \mu\|_2^2 \leq \frac{2}{\mu_{\min}} \|z_1^t \mathbf{1}_m - \mu\|_1^2 \leq \frac{2C_{\mathrm{fea}}^2}{\mu_{\min}} \eta_t^2.$$

Define $\xi_{ij}^t := \alpha_t G_{ij}(w^t)$. Then $z_{1,ij}^t = w_{ij}^t \exp(\xi_{ij}^t)$ and

$$D_h(z_1^t, w^t) = \sum_{i,j} w_{ij}^t \left[\exp(\xi_{ij}^t)\xi_{ij}^t - \exp(\xi_{ij}^t) + 1\right] = \sum_{i,j} w_{ij}^t \psi(\xi_{ij}^t),$$

where $\psi(s) := \exp(s)(s-1) + 1$. Since $\psi(0) = 0$, $\psi'(0) = 0$, and $\psi''(s) = \exp(s)(1+s)$. By Taylor's theorem, for $|s| \leq M_G c_+$, $\psi(s) \leq \frac{1}{2}\exp(M_G c_+)(1 + M_G c_+)s^2$. Then

$$D_h(z_1^t, w^t) \leq \frac{1}{2}\exp(M_G c_+)(1 + M_G c_+) \sum_{ij} w_{ij}^t \left(\alpha_t G_{ij}(w^t)\right)^2 \leq \frac{1}{2}\exp(M_G c_+)(1 + M_G c_+)\alpha_t^2 M_G^2 \sum_{ij} w_{ij}^t.$$

Since $w^t \in \mathcal{C}_2$, $\sum_{ij} w_{ij}^t = \sum_j \nu_j = 1$. Then $D_h(z_1^t, w^t) = O(\alpha_t^2) = O(\eta_t^2)$.

Finally, by Lemma 10, $\|\pi^{t+1} - z_1^t\|_1 = \|z_1^t \mathbf{1}_m - \mu\|_1$. Then $|\alpha_t \langle G(w^t), \pi^{t+1} - z_1^t \rangle| \leq \alpha_t M_G \|z_1^t \mathbf{1}_m - \mu\|_1 \leq M_G C_{\mathrm{fea}} \alpha_t \eta_t = O(\eta_t^2)$. Combining all three terms together, we have

$$D_h(\pi^{t+1}, w^t) \leq \frac{2C_{\mathrm{fea}}^2}{\mu_{\min}} \eta_t^2 + O(\alpha_t^2) + O(\alpha_t \eta_t).$$

Since $\alpha_t = O((t+1)^{-\beta}) = O(\eta_t)$, there exists a constant $C_{1,\mathrm{cross}}$ such that the whole right-hand side is bounded by $C_{1,\mathrm{cross}} \eta_t^2$, i.e., $D_h(\pi^{t+1}, w^t) \leq C_{1,\mathrm{cross}} \eta_t^2$.

The estimate for $D_h(\pi^{t+1}, w^{t+1})$ is obtained analogously from the second half-step. Define $u_j^t := ((z_2^t)^\top \mathbf{1}_n)_j$. Since $w^{t+1} = \mathcal{P}_{\mathcal{C}_2}^{\mathrm{KL}}(z_2^t)$, $w_{ij}^{t+1} = \frac{\nu_j}{u_j^t} z_{2,ij}^t$ and $\log z_{2,ij}^t - \log w_{ij}^{t+1} = \log \frac{u_j^t}{\nu_j}$. Again by the three-point identity for Bregman divergences, $D_h(\pi^{t+1}, w^{t+1}) = D_h(\pi^{t+1}, z_2^t) + D_h(z_2^t, w^{t+1}) + \langle \log z_2^t - \log w^{t+1}, \pi^{t+1} - z_2^t \rangle$.

Since $z_2^t = \pi^{t+1} \odot \exp(\alpha_t G(\pi^{t+1}))$ and $|G(\pi^{t+1})|_\infty \le M_G$, the same Taylor argument used for $D_h(z_1^t, w^t)$ gives $D_h(\pi^{t+1}, z_2^t) = O(\alpha_t^2) = O(\eta_t^2)$. We know that $|u^t - \nu|_1 = |(z_2^t)^\top \mathbf{1}_n - \nu|_1 \le |(\pi^{t+1})^\top \mathbf{1}_n - \nu|_1 + |z_2^t - \pi^{t+1}|_1$. Since $|(\pi^{t+1})^\top \mathbf{1}_n - \nu|_1 \le 2C_{\text{fea}}\eta_t$, by the mean-value theorem, $|z_2^t - \pi^{t+1}|_1 = O(\alpha_t) = O(\eta_t)$. Therefore, there exists a constant $C_{\text{col}} > 0$ such that $|u^t - \nu|_1 \le C_{\text{col}}\eta_t$. For sufficiently large t, this implies $u_j^t \in [\nu_j/2, 3\nu_j/2]$ for every j. By Lemma 10 and Taylor's theorem, we obtain

$$D_h(z_2^t, w^{t+1}) \le \frac{1}{\nu_{\min}} |u^t - \nu|_1^2 \le \frac{C_{\text{col}}^2}{\nu_{\min}} \eta_t^2.$$

Since $u_j^t \in [\nu_j/2, 3\nu_j/2]$ for sufficiently large t, the mean-value theorem gives

$$\left| \log \frac{u_j^t}{\nu_j} \right| \le \frac{2|u_j^t - \nu_j|}{\nu_j} \le \frac{2|u^t - \nu|_1}{\nu_{\min}}.$$

Hence $|\log z_2^t - \log w^{t+1}|_\infty = O(\eta_t)$. Also, $|\pi^{t+1} - z_2^t|_1 = O(\alpha_t) = O(\eta_t)$. Thus $|\langle \log z_2^t - \log w^{t+1}, \pi^{t+1} - z_2^t \rangle| = O(\eta_t^2)$. Combining the three estimates, we show that there exist an integer $K_{\text{cross},2} \ge K_{\text{cross},1}$ and a constant $C_{2,\text{cross}} > 0$ such that, for every $t \ge K_{\text{cross},2}$, $D_h(\pi^{t+1}, w^{t+1}) \le C_{2,\text{cross}}\eta_t^2$. Thus $D_h(\pi^{t+1}, w^t) = O(\eta_t^2)$ and $D_h(\pi^{t+1}, w^{t+1}) = O(\eta_t^2)$. □

Proof for Proposition 4

Proof. For notation convenience, we define the row error and first-order row moments: $s_i^t := \sum_j w_{ij}^t$, $e_i^t := s_i^t - \mu_i$, $a_i^t := \sum_j w_{ij}^t G_{ij}(w^t)$, and $b_i^t := \sum_j w_{ij}^t \exp(\alpha_t G_{ij}(w^t)) G_{ij}(w^t)$. Recall that $r_i^t = \sum_j z_{1,ij}^t = \sum_j w_{ij}^t \exp(\alpha_t G_{ij}(w^t))$ and $\pi_{ij}^{t+1} = \frac{\mu_i}{r_i^t} z_{1,ij}^t$. Then for every t, by the definition of KL divergence,

$$A_t^{\text{row}} = \sum_{i,j} \left[(\pi_{ij}^{t+1} + w_{ij}^t) \log \frac{\pi_{ij}^{t+1}}{w_{ij}^t} + 2(w_{ij}^t - \pi_{ij}^{t+1}) \right] = \sum_{i,j} \left[(\pi_{ij}^{t+1} + w_{ij}^t) \left(\log \frac{\mu_i}{r_i^t} + \alpha_t G_{ij}(w^t) \right) + 2(w_{ij}^t - \pi_{ij}^{t+1}) \right].$$

Since $\sum_j \pi_{ij}^{t+1} = \mu_i$ and $\sum_j w_{ij}^t = s_i^t$, $\sum_j (\pi_{ij}^{t+1} + w_{ij}^t) \log \frac{\mu_i}{r_i^t} = (\mu_i + s_i^t) \log \frac{\mu_i}{r_i^t}$ where $\mu_i + s_i^t = 2\mu_i + e_i^t$. Also by the definition above, $\alpha_t \sum_j (\pi_{ij}^{t+1} + w_{ij}^t) G_{ij}(w^t) = \alpha_t \left(\frac{\mu_i}{r_i^t} b_i^t + a_i^t \right)$. Since $\sum_j w_{ij}^t - \pi_{ij}^{t+1} = s_i^t - \mu_i = e_i^t$, summing over i,

$$A_t^{\text{row}} = \sum_i \left[(2\mu_i + e_i^t) \log \frac{\mu_i}{r_i^t} + \alpha_t \left(\frac{\mu_i b_i^t}{r_i^t} + a_i^t \right) + 2e_i^t \right]. \tag{5}$$

Define $\zeta_i^t := (2\mu_i + e_i^t) \log \frac{\mu_i}{r_i^t} + \alpha_t \left(\frac{\mu_i b_i^t}{r_i^t} + a_i^t \right) + 2e_i^t$. From equation 5, $A_t^{\text{row}} = \sum_i \zeta_i^t$. Define $u_i^t := \sum_j w_{ij}^t G_{ij}(w^t)^2$. Since $\|G(w^t)\|_\infty \le M_G$ and $\sum_j w_{ij}^t \le 1$, the quantities a_i^t, b_i^t, and u_i^t are uniformly bounded in absolute value by constants depending only on M_G.

From Taylor expanding the exponential, $\exp(\alpha_t G_{ij}(w^t)) = 1 + \alpha_t G_{ij}(w^t) + \frac{\alpha_t^2}{2} G_{ij}(w^t)^2 + O(\alpha_t^3)$. Then the perturbed row sum admits the expansion

$$r_i^t = \sum_j w_{ij}^t \exp(\alpha_t G_{ij}(w^t)) = s_i^t + \alpha_t a_i^t + \frac{\alpha_t^2}{2} u_i^t + O(\alpha_t^3) = \mu_i + e_i^t + \alpha_t a_i^t + \frac{\alpha_t^2}{2} u_i^t + O(\alpha_t^3).$$

Similarly, $b_i^t = \sum_j w_{ij}^t \exp(\alpha_t G_{ij}(w^t)) G_{ij}(w^t) = a_i^t + \alpha_t u_i^t + O(\alpha_t^2)$. Set $\lambda_i^t := r_i^t - \mu_i$. By the boundedness of a_i^t, u_i^t and diminishing α_t $\lambda_i^t = e_i^t + \alpha_t a_i^t + \frac{\alpha_t^2}{2} u_i^t + O(\alpha_t^3) = O(|e_i^t| + \alpha_t)$. For sufficiently large t, $|\lambda_i^t|$ is small enough to Taylor expand around μ_i. Thus by Taylor expansion,

$$\log \frac{\mu_i}{r_i^t} = \log \frac{\mu_i}{\mu_i + \lambda_i^t} = -\frac{\lambda_i^t}{\mu_i} + \frac{{\lambda_i^t}^2}{2\mu_i^2} + O\left({\lambda_i^t}^3\right), \qquad \frac{\mu_i}{r_i^t} = \frac{\mu_i}{\mu_i + \lambda_i^t} = 1 - \frac{\lambda_i^t}{\mu_i} + O\left({\lambda_i^t}^2\right).$$

Since $\lambda_i^t = O(|e_i^t| + \alpha_t)$,

$$\begin{aligned} \zeta_i^t &= (2\mu_i + e_i^t) \log \frac{\mu_i}{r_i^t} + \alpha_t \left(\frac{\mu_i b_i^t}{r_i^t} + a_i^t \right) + 2e_i^t \\ &= (2\mu_i + e_i^t) \left(-\frac{\lambda_i^t}{\mu_i} + \frac{(\lambda_i^t)^2}{2\mu_i^2} \right) + \alpha_t \left[(a_i^t + \alpha_t u_i^t) \left(1 - \frac{\lambda_i^t}{\mu_i} \right) + a_i^t \right] + 2e_i^t + O((|e_i^t| + \alpha_t)^3) \end{aligned}$$

$$= -2\lambda_i^t - \frac{e_i^t\lambda_i^t}{\mu_i} + \frac{(\lambda_i^t)^2}{\mu_i} + 2\alpha_t a_i^t + \alpha_t^2 u_i^t - \frac{\alpha_t a_i^t \lambda_i^t}{\mu_i} + 2e_i^t + O((|e_i^t| + \alpha_t)^3)$$
$$= \left(-2\lambda_i^t + 2e_i^t + 2\alpha_t a_i^t + \alpha_t^2 u_i^t\right) + \frac{\lambda_i^t(\lambda_i^t - e_i^t - \alpha_t a_i^t)}{\mu_i} + O((|e_i^t| + \alpha_t)^3),$$

where the first bracket satisfies $-2\lambda_i^t+2e_i^t+2\alpha_t a_i^t+\alpha_t^2 u_i^t = -2\left(e_i^t + \alpha_t a_i^t + \frac{\alpha_t^2}{2}u_i^t + O(\alpha_t^3)\right)+2e_i^t+2\alpha_t a_i^t+\alpha_t^2 u_i^t = O(\alpha_t^3)$. Since $\lambda_i^t - e_i^t - \alpha_t a_i^t = \frac{\alpha_t^2}{2}u_i^t + O(\alpha_t^3) = O(\alpha_t^2)$ and $\lambda_i^t = O(|e_i^t| + \alpha_t)$, we have

$$\frac{\lambda_i^t(\lambda_i^t - e_i^t - \alpha_t a_i^t)}{\mu_i} = O((|e_i^t| + \alpha_t)\alpha_t^2) = O((|e_i^t| + \alpha_t)^3).$$

Therefore, $\zeta_i^t = O((|e_i^t| + \alpha_t)^3)$. By Proposition 3, there exist constants $C_{\text{row}} > 0$ and $K_{\text{row}} \geq 0$ such that, for all $t \geq K_{\text{row}}$, $\|w^t\mathbf{1}_m - \mu\|_1 \leq C_{\text{row}}\eta_t$. Since $e_i^t = s_i^t - \mu_i$, this implies $|e_i^t| \leq C_{\text{row}}\eta_t$. Also, because $\alpha_t \leq c_+(t+1)^{-\beta}$ and $p \geq 0$, $\alpha_t \leq c_+\eta_t$. Hence $|\zeta_i^t| \leq C_{\text{row}}\eta_t^3$. Summing over finitely many rows gives $|A_t^{\text{row}}| \leq C_{\text{row,asym}}\eta_t^3$, for $C_{\text{row,asym}} > 0$. Analogously, we can repeat the preceding argument with columns, with ν_j in place of μ_i, and with the column infeasibility estimate from Proposition 3. Thus for all sufficiently large t, we can show there exists some $C_{\text{col,asym}} > 0$ such that $|A_t^{\text{col}}| \leq C_{\text{col,asym}}\eta_t^3$. Summing the row and column bounds yields the $|A_t| \leq C_{\text{asym}}\eta_t^3$, for constant $C_{\text{asym}} > 0$ and sufficiently large t. We thus show that $|A_t| = O\left((t+1)^{-3(\beta-p)}\right)$. □

A.4 Lyapunov and Corrected Residual Estimates

Proof for Proposition 5 Define row and column scaling potentials

$$p_i^{t+1} := -\rho_t \log\left(\frac{\mu_i}{\sum_\ell w_{i\ell}^t \exp\left(G_{i\ell}(w^t)/\rho_t\right)}\right), \qquad q_j^{t+1} := -\rho_t \log\left(\frac{\nu_j}{\sum_\ell \pi_{\ell j}^{t+1} \exp\left(G_{\ell j}(\pi^{t+1})/\rho_t\right)}\right),$$

with lifted matrices defined as $p_{ij}^{t+1} := p_i^{t+1}$ and $q_{ij}^{t+1} := q_j^{t+1}$ for $i = 1, \ldots, n,\ j = 1, \ldots, m$. Then p^{t+1} is row-constant and q^{t+1} is column-constant. Since row-constant matrices are exactly normals to the affine row-constraint manifold and similarly for column-constant matrices, we have $p^{k+1} \in \mathcal{N}_{\mathcal{C}_1}(\pi^{t+1})$ and $q^{t+1} \in \mathcal{N}_{\mathcal{C}_2}(w^{t+1})$. Also, since

$$\log \pi_{ij}^{t+1} = \frac{G_{ij}(w^t)}{\rho_t} + \log\frac{\mu_i}{\sum_l w_{il}^t \exp(G_{il}(w^t)/\rho_t)} + \log w_{ij}^t = \frac{G_{ij}(w^t)}{\rho_t} - \frac{p_{ij}^{t+1}}{\rho_t} + \log w_{ij}^t,$$

at iteration t, the exact half-step optimality conditions are:

$$-G(w^t) + \rho_t(\log \pi^{t+1} - \log w^t) + p^{t+1} = \mathbf{0}, \tag{6}$$
$$-G(\pi^{t+1}) + \rho_t(\log w^{t+1} - \log \pi^{t+1}) + q^{t+1} = \mathbf{0}. \tag{7}$$

Define $\Lambda_t = f(\pi^t, w^t) + \rho_t D_h(\pi^t, w^t)$. Since $\pi^{t+1}, \pi^t \in \mathcal{C}_1$ and p^{t+1} is row constant, $\sum_j \pi_{ij}^{t+1} = \sum_j \pi_{ij}^t = \mu_i$ and $p_{ij}^{t+1} = p_i^{t+1}$. Thus $\langle p^{t+1}, \pi^{t+1} - \pi^t\rangle = 0$. Then

$$-\langle G(w^t), \pi^{t+1} - \pi^t\rangle + \rho_t\langle \log \pi^{t+1} - \log w^t, \pi^{t+1} - \pi^t\rangle = 0.$$

Since $f(\pi, w) = -\langle \pi, G(w)\rangle$, $f(\pi^t, w^t) - f(\pi^{t+1}, w^t) = \langle G(w^t), \pi^{t+1} - \pi^t\rangle$. By three-point identity for Bregman divergences, $\langle \log \pi^{t+1} - \log w^t, \pi^{t+1} - \pi^t\rangle = D_h(\pi^t, \pi^{t+1}) + D_h(\pi^{t+1}, w^t) - D_h(\pi^t, w^t)$. Thus

$$f(\pi^t, w^t) - f(\pi^{t+1}, w^t) = \rho_t\left(D_h(\pi^t, \pi^{t+1}) + D_h(\pi^{t+1}, w^t) - D_h(\pi^t, w^t)\right). \tag{8}$$

Similarly, since $w^t, w^{t+1} \in \mathcal{C}_2$ and q^{t+1} is column constant, $\langle q^{t+1}, w^t - w^{t+1}\rangle = 0$ and we can obtain

$$f(\pi^{t+1}, w^t) - f(\pi^{t+1}, w^{t+1}) = \rho_t\left(D_h(w^t, w^{t+1}) + D_h(w^{t+1}, \pi^{t+1}) - D_h(w^t, \pi^{t+1})\right). \tag{9}$$

Adding equation 8 and equation 9 together, we obtain

$$f(\pi^t, w^t) - f(\pi^{t+1}, w^{t+1}) = \rho_t\big(D_h(\pi^t, \pi^{t+1}) + D_h(w^t, w^{t+1}) + D_h(\pi^{t+1}, w^t) + D_h(w^{t+1}, \pi^{t+1}) - D_h(\pi^t, w^t) - D_h(w^t, \pi^{t+1})\big).$$

Recall that $A_t = A_t^{\text{row}} + A_t^{\text{col}} = D_h(\pi^{t+1}, w^t) - D_h(w^t, \pi^{t+1}) + D_h(w^{t+1}, \pi^{t+1}) - D_h(\pi^{t+1}, w^{t+1})$. Combining this, we obtain

$$\Lambda_t - f(\pi^{t+1}, w^{t+1}) - \rho_t D_h(\pi^{t+1}, w^{t+1}) = \rho_t(D_h(\pi^t, \pi^{t+1}) + D_h(w^t, w^{t+1}) + A_t). \tag{10}$$

From equation 10, $\Lambda_t - \Lambda_{t+1} = \rho_t \left(D_h(\pi^t, \pi^{t+1}) + D_h(w^t, w^{t+1}) + A_t\right) - (\rho_{t+1} - \rho_t) D_h(\pi^{t+1}, w^{t+1})$. By Lemma 7, α_t is nonincreasing and $\alpha_t \geq \frac{c_-}{t+2}$ for all sufficiently large t. Then $\rho_t = 1/\alpha_t$ is nondecreasing and $\rho_t \leq \frac{t+2}{c_-}$, i.e., $\rho_t = O(t+2)$. By Proposition 4, $\rho_t |A_t| = O\big((t+1)^{1-3(\beta-p)}\big)$, which is summable when $\beta > 2/3+p$. By Proposition 3, $D_h(\pi^{t+1}, w^{t+1}) = O(\eta_t^2) = O\left((t+1)^{-2(\beta-p)}\right)$ with $r := 2(\beta-p) > 4/3 > 1$. Since ρ_t is nondecreasing and $\rho_t = O(t)$. Let T be sufficiently large. For $N > T$,

$$\sum_{t=T}^{N} (\rho_{t+1} - \rho_t)(t+1)^{-r} = \rho_{N+1}(N+1)^{-r} - \rho_T (T+1)^{-r} + \sum_{t=T+1}^{N} \rho_t \left(t^{-r} - (t+1)^{-r}\right).$$

We know that $\rho_{N+1}(N+1)^{-r} = O((N+2)(N+1)^{-r}) = O(N^{1-r})$, and $r > 1$. By the mean-value theorem, $t^{-r} - (t+1)^{-r} \leq C_r t^{-r-1}$ for a constant $C_r > 0$. Then $\rho_t \left(t^{-r} - (t+1)^{-r}\right) = O(t \cdot t^{-r-1}) = O(t^{-r})$, and $\sum_t t^{-r} < \infty$ because $r > 1$. Hence the partial sums are uniformly bounded. Since the terms $(\rho_{t+1} - \rho_t)(t+1)^{-r}$ are nonnegative, the series converges. Thus

$$\sum_t (\rho_{t+1} - \rho_t) D_h(\pi^{t+1}, w^{t+1}) < \infty.$$

Since $\mathcal{C}_1 \times \mathcal{C}_2$ is compact, f is continuous and D_h is nonnegative, Λ_t is bounded below. Summing the Lyapunov identity above, we obtain

$$\sum_{t=1}^{\infty} \rho_t (D_h(\pi^t, \pi^{t+1}) + D_h(w^t, w^{t+1})) < \infty.$$

Lemma 11. *There exist an integer K_R and a constant $C_R > 0$ such that for every $t \geq K_R$,*

$$\alpha_t \|R_t\|_{M_t}^2 \leq 4\rho_t D_h(w^t, w^{t+1}) + C_R \alpha_t \eta_t^2.$$

Proof. Entrywise: $\min\{a,b\}(\log a - \log b)^2 \leq 2(a \log \frac{a}{b} - a + b) \Rightarrow \|\log w^t - \log w^{t+1}\|_{M_t}^2 \leq 2D_h(w^t, w^{t+1})$.

By Pinsker's Inequality, $\|\pi^{t+1} - w^t\|_F^2 \leq \|\pi^{t+1} - w^t\|_1^2 \leq 2D_h(\pi^{t+1}, w^t)$. Then for $L_G := \|D_X\|_2 \|D_Y\|_2$,

$$\begin{aligned}
\|R_t\|_{M_t}^2 &\leq 2\rho_t^2 \|\nabla h(w^t) - \nabla h(w^{t+1})\|_{M_t}^2 + 2\|G(\pi^{t+1}) - G(w^t)\|_{M_t}^2 \\
&\leq 2\rho_t^2 \|\nabla h(w^t) - \nabla h(w^{t+1})\|_{M_t}^2 + 2\|D_X\|_2^2 \|D_Y\|_2^2 \|\pi^{t+1} - w^t\|_F^2 \\
&\leq 4\rho_t^2 D_h(w^t, w^{t+1}) + 4L_G^2 D_h(\pi^{t+1}, w^t).
\end{aligned}$$

By Proposition 3, there exist $K_{\text{cross},1}$ and $C_{1,\text{cross}} > 0$ such that, for all $t \geq K_{\text{cross},1}$, $D_h(\pi^{t+1}, w^t) \leq C_{1,\text{cross}} \eta_t^2$. Hence

$$4L_G^2 \alpha_t D_h(\pi^{t+1}, w^t) \leq 4L_G^2 C_{1,\text{cross}} \alpha_t \eta_t^2.$$

Define $C_R := 4L_G^2 C_{1,\text{cross}}$ and choose $K_R \geq K_{\text{cross},1}$ sufficiently large. Combining all above, for every $t \geq K_R$,

$$\alpha_t \|R_t\|_{M_t}^2 \leq 4\rho_t D_h(w^t, w^{t+1}) + C_R \alpha_t \eta_t^2.$$

□

Proof of Proposition 6

Proof. From equation 6, $p^{t+1} + q^{t+1} = G(\pi^{t+1}) + G(w^t) + \rho_t(\nabla h(w^t) - \nabla h(w^{t+1}))$, i.e., $R_t = -2G(w^t) + p^{t+1} + q^{t+1}$. Since $w_{ij}^t \geq 0$ and $\sum_{ij} w_{ij}^t = 1$, $w_{ij}^t \leq 1$ and thus for any matrix $U \in \mathbb{R}^{n \times m}$, $\|U\|_{M_t}^2 = \sum_{ij} m_{ij}^t U_{ij}^2 \leq \|U\|_F^2$. By Lemma 11, there exist constants $C_R > 0$ and K_R such that, for every $t \geq K_R$, $\alpha_t \|R_t\|_{M_t}^2 \leq 4\rho_t D_h(w^t, w^{t+1}) + C_R \alpha_t \eta_t^2$. Summing from $t = K_R$ to $N-1$,

$$\sum_{t=K_R}^{N-1} \alpha_t \|R_t\|_{M_t}^2 \leq 4 \sum_{t=K_R}^{N-1} \rho_t D_h(w^t, w^{t+1}) + C_R \sum_{t=K_R}^{N-1} \alpha_t \eta_t^2.$$

By the Lyapunov summability estimate: $\sum_{t=K_R}^{\infty} \rho_t D_h(w^t, w^{t+1}) < \infty$ from Proposition 5 . Since $\alpha_t = O((t+1)^{-\beta})$ and $\eta_t^2 = (t+1)^{-2(\beta-p)}$, the second series is bounded by a constant times $\sum_t (t+1)^{-3\beta+2p}$, which is summable under $\beta > 2/3 + p$. Then $\sum_{t=K_R}^{N-1} \alpha_t \|R_t\|_{M_t}^2$ are uniformly bounded in N. Passing to the limit,

$$\sum_{t=K_R}^{\infty} \alpha_t \|R_t\|_{M_t}^2 < \infty.$$

Define $C_\star := \sum_{t=K_R}^{\infty} \alpha_t \|R_t\|_{M_t}^2 < \infty$. Then for every $N > K_R$,

$$\left(\min_{K_R \le i \le N-1} \|R_i\|_{M_i}^2\right) \sum_{t=K_R}^{N-1} \alpha_t \le \sum_{t=K_R}^{N-1} \alpha_t \|R_t\|_{M_t}^2 \le C_\star.$$

Thus $\min_{K_R \le i \le N-1} \|R_i\|_{M_i}^2 \le C_\star / \sum_{t=K_R}^{N-1} \alpha_t$. Since

$$\sum_{t=K_R}^{N-1} \alpha_t \ge c_- \sum_{t=K_R}^{N-1} \frac{1}{t+2} \ge c_- \log\left(\frac{N+1}{K_R+2}\right),$$

we have showed

$$\min_{K_R \le i \le N-1} \|R_i\|_{M_i}^2 \le \frac{C_\star}{c_- \log\big((N+1)/(K_R+2)\big)},$$

which is equivalent to $\min_{0 \le i \le N-1} \|R_i\|_{M_i}^2 = O\left(\frac{1}{\log N}\right)$. □

A.5 Subsequence Convergence

Lemma 12 (Linear regularity of the marginal sets)**.** *There exists $\kappa > 0$, depending only on n and m, such that every $\pi \in \mathbb{R}^{n \times m}$ satisfies* $\operatorname{dist}_F(\pi, \mathcal{C}) \le \kappa \max\{\operatorname{dist}_F(\pi, \mathcal{C}_1), \operatorname{dist}_F(\pi, \mathcal{C}_2)\}$.

Proof. Let $U = \operatorname{proj}_{\mathcal{C}_1}(\pi)$ and $V = \operatorname{proj}_{\mathcal{C}_2}(\pi)$. Denote the column sums of U by $c = U^\top \mathbf{1}_n$. We first construct a coupling $Y \in \mathcal{C}$ close to U.

For each column j, define $a_j := (c_j - \nu_j)_+$ and $b_j := (\nu_j - c_j)_+$ where a_j is the amount by which the j-th column sum of U exceeds its prescribed marginal ν_j, while b_j is the corresponding shortage. Since $\mathbf{1}_m^\top c = \mathbf{1}_n^\top \mu = \mathbf{1}_m^\top \nu = 1$, the total excess equals the total shortfall, i.e., $d := \sum_{j=1}^m a_j = \sum_{j=1}^m b_j$. If $d = 0$, then $U \in \mathcal{C}$. Otherwise, remove the excess mass from each column proportionally:

$$H_{ij} = \begin{cases} U_{ij} a_j / c_j, & c_j > 0, \\ 0, & c_j = 0, \end{cases} \qquad r = H \mathbf{1}_m,$$

and redistribute the removed mass within each row among the deficient columns with $Y = U - H + \frac{r b^\top}{d}$. Then $Y \ge 0$, $Y \mathbf{1}_m = \mu$, and $Y^\top \mathbf{1}_n = \nu$. Thus $Y \in \mathcal{C}$ and $\|Y - U\|_1 \le 2d = \|c - \nu\|_1$. Because $V^\top \mathbf{1}_n = \nu$, $\|c - \nu\|_1 = \|(U - V)^\top \mathbf{1}_n\|_1 \le \|U - V\|_1 \le \sqrt{nm}\, \|U - V\|_F$. Therefore,

$$\operatorname{dist}_F(\pi, \mathcal{C}) \le \|\pi - U\|_F + \|U - Y\|_F \le (1 + \sqrt{nm}) \operatorname{dist}_F(\pi, \mathcal{C}_1) + \sqrt{nm} \operatorname{dist}_F(\pi, \mathcal{C}_2).$$

With $\kappa = 1 + 2\sqrt{nm}$, $\operatorname{dist}_F(\pi, \mathcal{C}) \le \kappa \max\{\operatorname{dist}_F(\pi, \mathcal{C}_1), \operatorname{dist}_F(\pi, \mathcal{C}_2)\}$. □

Proof for Lemma 1

Proof. Proposition 4 gives $\big|D_h(\pi^{t+1}, w^t) - D_h(w^t, \pi^{t+1})\big| + \big|D_h(w^{t+1}, \pi^{t+1}) - D_h(\pi^{t+1}, w^{t+1})\big| = O(\eta_t^3)$. Because $\eta_t = (t+1)^{-(\beta - p)}$ and $\beta - p > 2/3$, the resulting series is summable. □

Proof for Theorem 1

Proof. Since $k_0 < \infty$, the finite prefix before k_0 does not affect any asymptotic statement. After relabeling the post-burn tail, the accepted steps satisfy the guarded bounds in Lemma 7; all estimates above apply with constants depending on the switched point (π^{k_0}, w^{k_0}).

(1) By Proposition 6, $\sum_{t=K_R}^{\infty} \alpha_t \|R_t\|_{M_t}^2 < \infty$. Since $\sum_{t=K_R}^{\infty} \alpha_t = \infty$, if $\liminf_{t \to \infty} \|R_t\|_{M_t} \ne 0$, then there exists $\varepsilon > 0$ and $K \ge K_R$ such that for all $t \ge K$, $\|R_t\|_{M_t} \ge \varepsilon$. Then

$$\sum_{t=K}^{\infty} \alpha_t \|R_t\|_{M_t}^2 \ge \varepsilon^2 \sum_{t=K}^{\infty} \alpha_t = \infty,$$

which contradicts Proposition 6. Thus $\liminf_{t \to \infty} \|R_t\|_{M_t} = 0$.

(2) We know that $D_h(\pi^{t+1}, w^{t+1}) \le C_{2,\mathrm{cross}}\eta_t^2$. Since $\eta_t \to 0$, $D_h(\pi^{t+1}, w^{t+1}) \to 0$. Since π^{t+1} and w^{t+1} have total mass one, by Pinsker's inequality $\|\pi^{t+1} - w^{t+1}\|_1^2 \le 2D_h(\pi^{t+1}, w^{t+1})$, and thus $\varepsilon_t := \|\pi^{t+1} - w^{t+1}\|_1 \to 0$. Define shadow sequence $x^{t+1} := \mathcal{P}_{\mathcal{C}}(\pi^{t+1})$. Since $\pi^{t+1} \in \mathcal{C}_1$ and $w^{t+1} \in \mathcal{C}_2$, we have $\mathrm{dist}(\pi^{t+1}, \mathcal{C}_1) = 0$ and $\mathrm{dist}(\pi^{t+1}, \mathcal{C}_2) \le \|\pi^{t+1} - w^{t+1}\|$. Since $\mathcal{C}_1, \mathcal{C}_2$ are polyhedral, by Lemma 12, $\mathrm{dist}(\pi^{t+1}, \mathcal{C}) \le \kappa \max\{\mathrm{dist}(\pi^{t+1}, \mathcal{C}_1), \mathrm{dist}(\pi^{t+1}, \mathcal{C}_2)\}$ and

$$\|x^{t+1} - \pi^{t+1}\| + \|x^{t+1} - w^{t+1}\| \le 2\|x^{t+1} - \pi^{t+1}\| + \|\pi^{t+1} - w^{t+1}\| = 2\mathrm{dist}(\pi^{t+1}, \mathcal{C}) + \varepsilon_t \le (2\kappa + 1)\varepsilon_t.$$

As $\varepsilon_t \to 0$, $\|x^{t+1} - \pi^{t+1}\| \to 0$ and $\|x^{t+1} - w^{t+1}\| \to 0$. Then the sequences $\{x^t\}$, $\{\pi^t\}$, and $\{w^t\}$ have the same set of accumulation points. □

Lemma 13. *Let $x^{t_\ell+1} = \mathcal{P}_{\mathcal{C}}(\pi^{t_\ell+1}) \to \bar{x}$ be a residual-vanishing convergent shadow subsequence, and let $S := \{(i,j) : \bar{x}_{ij} > 0\}$. The following conditions hold along the subsequence:*

- *(a)* ***Active-support nondegeneracy.*** *There exists $c_S > 0$ such that for all sufficiently large ℓ, for all $(i,j) \in S$, $\min\{w_{ij}^{t_\ell}, w_{ij}^{t_\ell+1}\} \ge c_S$.*
- *(b)* ***Bounded normalized row/column potentials.*** *Suppose Assumption 1 holds along this subsequence. Then after fixing one global gauge for a fixed row index i_0, $\{p^{t_\ell+1}\}$ and $\{q^{t_\ell+1}\}$ are bounded.*

Proof. **(a)** By Proposition 3, $D_h(\pi^{t_\ell+1}, w^{t_\ell+1}) \to 0$. Since both $\pi^{t_\ell+1}$ and $w^{t_\ell+1}$ have total mass one, by Pinsker's inequality, $\|\pi^{t_\ell+1} - w^{t_\ell+1}\|_1 \to 0$. Since $\pi^{t_\ell+1} \in \mathcal{C}_1$ and $w^{t_\ell+1} \in \mathcal{C}_2$, $\mathrm{dist}(\pi^{t_\ell+1}, \mathcal{C}_1) = 0$ and $\mathrm{dist}(\pi^{t_\ell+1}, \mathcal{C}_2) \le \|\pi^{t_\ell+1} - w^{t_\ell+1}\|$. Since $\mathcal{C}_1$ and $\mathcal{C}_2$ are polyhedral, by Lemma 12, there exists a constant $\kappa > 0$ such that

$$\mathrm{dist}(\pi^{\mathrm{t}_\ell+1}, \mathcal{C}_1 \cap \mathcal{C}_2) \le \kappa \max\{\mathrm{dist}(\pi^{\mathrm{t}_\ell+1}, \mathcal{C}_1), \mathrm{dist}(\pi^{\mathrm{t}_\ell+1}, \mathcal{C}_2)\}.$$

Since $x^{t_\ell+1} = \mathcal{P}_{\mathcal{C}}(\pi^{t_\ell+1})$ and $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$, $\|x^{t_\ell+1} - \pi^{t_\ell+1}\| = \mathrm{dist}(\pi^{\mathrm{t}_\ell+1}, \mathcal{C}) \to 0$. Since $x^{t_\ell+1} \to \bar{x}$, $\pi^{t_\ell+1} \to \bar{x}$. By Proposition 5, $\sum_{k=1}^{\infty} \rho_k D_h(w^k, w^{k+1}) < \infty$. Since ρ_k is bounded below by a positive constant, $D_h(w^k, w^{k+1}) \to 0$. Again by Pinsker's inequality $\|w^{k+1} - w^k\|_1 \to 0$. Therefore $w^{t_\ell} \to \bar{x}$ as well. If $(i,j) \in S$, then $\bar{x}_{ij} > 0$. Hence, for all sufficiently large ℓ, $w_{ij}^{t_\ell} \ge \frac{\bar{x}_{ij}}{2}$ and $w_{ij}^{t_\ell+1} \ge \frac{\bar{x}_{ij}}{2}$. Since S is finite, for all $(i,j) \in S$ and for sufficiently large ℓ, $\min\{w_{ij}^{t_\ell}, w_{ij}^{t_\ell+1}\} \ge c_S$ where $c_S := \frac{1}{2}\min_{(i,j)\in S} \bar{x}_{ij} > 0$.

(b) By **(a)**, for every $(i,j) \in S$, $m_{ij}^{t_\ell} = \min\{w_{ij}^{t_\ell}, w_{ij}^{t_\ell+1}\} \ge c_S$ for sufficiently large ℓ. Since $\|R_{t_\ell}\|_{M_{t_\ell}} \to 0$, for $(i,j) \in S$,

$$c_S R_{t_\ell,ij}^2 \le m_{ij}^{t_\ell} R_{t_\ell,ij}^2 \le \sum_{a,b} m_{ab}^{t_\ell} R_{t_\ell,ab}^2 = \|R_{t_\ell}\|_{M_{t_\ell}}^2 \to 0,$$

so $R_{t_\ell,ij} \to 0$. Since $R_k = -2G(w^k) + p^{k+1} + q^{k+1}$, for every $(i,j) \in S$, $p_i^{t_\ell+1} + q_j^{t_\ell+1} = 2G(w^{t_\ell})_{ij} + R_{t_\ell,ij}$. Since $w^{t_\ell} \to \bar{x}$ and G is continuous, the right-hand side is bounded. Hence the active-edge sums $p_i^{t_\ell+1} + q_j^{t_\ell+1}$ are bounded for all $(i,j) \in S$.

Define $\mathcal{G}_S$ as a bipartite graph with row vertices $\{1,\dots,n\}$ and column vertices $\{1,\dots,m\}$, and edge set S. Since $\bar{x} \in \mathcal{C}$ and the μ, ν have strictly positive entries, every row and column is incident to at least one edge of S. Let $\mathcal{C}_1, \dots, C_r$ be the connected components of $\mathcal{G}_S$. For each component, write $C_a = (I_a, J_a)$, where I_a and J_a denote its row and column index sets. Choose one reference row index $i_a \in I_a$ and define

$$\widehat{p}_i^{t_\ell+1} := p_i^{t_\ell+1} - p_{i_a}^{t_\ell+1}, \qquad \widehat{q}_j^{t_\ell+1} := q_j^{t_\ell+1} + p_{i_a}^{t_\ell+1},$$

for every $i, j \in C_a$. Then $\widehat{p}_{i_a}^{t_\ell+1} = 0$ since $p_i^{t_\ell+1}$ is row-constant and on every active edge $(i,j) \in S$, $\widehat{p}_i^{t_\ell+1} + \widehat{q}_j^{t_\ell+1} = p_i^{t_\ell+1} + q_j^{t_\ell+1}$. Hence these normalized sums remain bounded on all active entries. Because C_a is connected, as showed in Figure 4, every row or column index in C_a can be reached from i_a through a finite alternating sequence of row and column indices with corresponding entries belonging to S. Starting from $\widehat{p}_{i_a}^{t_\ell+1} = 0$, boundedness of the active-entry sums successively implies boundedness of the adjacent normalized column multipliers and then of the adjacent normalized row multipliers. Proceeding along such a sequence, we show that $\{\widehat{p}_i^{t_\ell+1}\}_\ell$ and $\{\widehat{q}_j^{t_\ell+1}\}_\ell$ bounded for every $i \in I_a$ and $j \in J_a$.

Define the component offset $c_a^\ell := p_{i_a}^{t_\ell+1}$. We fix one global gauge, say $c_1^\ell = 0$, and prove that the relative offsets $c_a^\ell - c_b^\ell$ are bounded. For two distinct components C_a and C_b, choose a row vertex $i_a \in C_a$ and a column vertex $j_b \in C_b$. Then $(i_a, j_b) \notin S$. By Assumption 1, $\liminf_{\ell\to\infty} R_{t_\ell,i_a j_b} \ge 0$. By construction,

$$R_{t_\ell,i_a j_b} = -2G(w^{t_\ell})_{i_a j_b} + \widehat{p}_{i_a}^{t_\ell+1} + \widehat{q}_{j_b}^{t_\ell+1} + c_a^\ell - c_b^\ell.$$

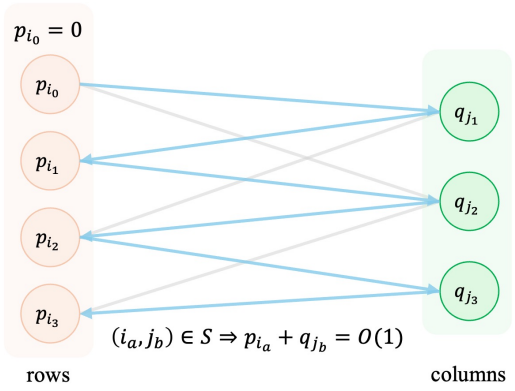


Figure 4: Bipartite Active-Support Graph. Each edge represents $(i_a, j_b) \in S$. Fixing $p_{i_0} = 0$, the bounded edge sums propagate boundedness along the blue alternating path; gray lines denote the remaining active-support edges.

Since the first three terms on the right-hand side are bounded, by Assumption 1, $c_a^\ell - c_b^\ell$ is bounded from below. Similarly, reverse and choose a row vertex in $\mathcal{C}_b$ and a column vertex in $\mathcal{C}_a$. The same argument shows that $c_b^\ell - c_a^\ell$ is bounded from below. Therefore $c_a^\ell - c_b^\ell$ is bounded above and below. Since there are finitely many connected components and $c_1^\ell = 0$, all relative gauges $\{c_a^\ell\}_\ell$ are bounded. Combining this with the boundedness of $\{\widehat{p}^{t_\ell+1}\}$ and $\{\widehat{q}^{t_\ell+1}\}$, we obtain the boundedness of $\{p^{t_\ell+1}\}_\ell$ and $\{q^{t_\ell+1}\}_\ell$. □

Proof for Theorem 2

Proof. Let $x^{t_\ell+1} \to \bar{x}$ be any residual-vanishing convergent shadow subsequence. By Theorem 1, $\pi^{t_\ell+1} \to \bar{x}$ and $w^{t_\ell+1} \to \bar{x}$. Lemma 13 further gives $w^{t_\ell} \to \bar{x}$ and, after fixing a gauge and passing to a further subsequence, $p^{t_\ell+1} \to \bar{p}$ and $q^{t_\ell+1} \to \bar{q}$. Thus the identity $R_t = -2G(w^t) + p^{t+1} + q^{t+1}$ passes to the limit. Writing the limit as $\bar{R}$, Lemma 13 gives $\bar{R}_{ij} = 0$ on $\mathrm{supp}(\bar{x})$, while Assumption 1 gives $\bar{R}_{ij} \geq 0$ on its complement. Therefore, $0 = -2G(\bar{x}) + \bar{p} + \bar{q} - \bar{R}$ and $\bar{p} + \bar{q} - \bar{R} \in \mathcal{N}_{\mathcal{C}}(\bar{x})$. Since $\nabla F(\bar{x}) = -2G(\bar{x})$, $\mathbf{0} \in \nabla F(\bar{x}) + \mathcal{N}_{\mathcal{C}}(\bar{x})$. Since Theorem 1 provides a residual-vanishing subsequence, and compactness of $\mathcal{C}$ provides a convergent further subsequence, the shadow sequence has at least one stationary accumulation point. □

A.6 Proof of Corollary 1

Proof. For any (i, i', j, j'),

$$\log \frac{(z_1^k)_{ij}(z_1^k)_{i'j'}}{(z_1^k)_{i'j}(z_1^k)_{ij'}} = \log \frac{w_{ij}^k w_{i'j'}^k}{w_{i'j}^k w_{ij'}^k} + \alpha_k'(G(w^k)_{ij} + G(w^k)_{i'j'} - G(w^k)_{i'j} - G(w^k)_{ij'}).$$

Since $\|G(w^k)\|_\infty \leq M_G$, $G(w^k)_{ij} + G(w^k)_{i'j'} - G(w^k)_{i'j} - G(w^k)_{ij'} \leq 4M_G$. Taking maxima on both sides, we obtain $\Delta(z_1^k) \leq \Delta(w^k) + 4M_G\alpha_k'$. Since the row projection onto $\mathcal{C}_1$ is a diagonal row scaling, $\Delta(\pi^{k+1}) = \Delta(z_1^k)$. Similarly, $\Delta(z_2^k) \leq \Delta(\pi^{k+1}) + 4M_G\alpha_k'$ with $\Delta(w^{k+1}) = \Delta(z_2^k)$. Thus $\Delta(w^{k+1}) \leq \Delta(\pi^{k+1}) + 4M_G\alpha_k' \leq \Delta(w^k) + 8M_G\alpha_k' = \Delta(w^k) + B_\Delta\alpha_k'$. Summing the changes up until iteration k,

$$\Delta(w^k) \leq \Delta(w^0) + B_\Delta' \sum_{t=0}^{t-1} \frac{1}{t+1} \leq \Delta(w^0) + B_\Delta'(1 + \log k).$$

Similar to the proof in Lemma 9,

$$1 - \tanh^2\left(\frac{\Delta(w^k)}{4}\right) \geq \exp(-\frac{\Delta(w^k)}{2}) \geq \exp\left(-\frac{\Delta(w^0) + B_\Delta'}{2}\right)(k+1)^{-4M_G} =: c_\Delta'(k+1)^{-4M_G},$$

with $c_\Delta' := \exp\left(-\frac{\Delta(w^0)+B_\Delta'}{2}\right)$. Hence $\tanh^2(\Delta(w^k)/4) \leq 1 - c_\Delta'(k+1)^{-\gamma}$ with $\gamma := 4M_G$. Thus the harmonic trajectory produces the same polynomial contraction-loss form used in Lemma 9 with p replaced by γ.

By Lemma 8, $d_H\left(r(w^{k+1}, r(\widehat{w}^k)\right) \le 12M_G\alpha'_k = O((k+1)^{-1})$. Combining this with Lemma 4, we obtain

$$d_H(r(w^{k+1}), \mathbf{1}_n) \le (1 - c'_\Delta(k+1)^{-\gamma})d_H(r(w^k), \mathbf{1}_n) + C'(k+1)^{-1},$$

which is exactly the recursion we see in Lemma 9 with $\beta = 1, p = \gamma$, and $\beta - p = 1 - \gamma$. When $B'_\Delta < \frac{2}{3}, \gamma < \frac{1}{3} < \beta$. Thus the comparison condition in Lemma 9 applies and gives $d_H(r(w^k), \mathbf{1}_n) = O((k+1)^{-(1-\gamma)}) = O(\eta'_k)$ for $\eta'_k := (k+1)^{-(1-\gamma)}$. Similar to Proposition 3, $\|w^k\mathbf{1}_m - \mu\|_1 = O(\eta'_k)$ and $\|z_1^k - w^k\| = O(\alpha'_k)$. Since $\gamma \ge 0$, $\alpha'_k = (k+1)^{-1} \le (k+1)^{-(1-\gamma)} = \eta'_k$. Then

$$\|z_1^k\mathbf{1}_m - \mu\|_1 \le \|z_1^k - w^k\|_1 + \|w^k\mathbf{1}_m - \mu\|_1 = O(\eta'_k) + O(\alpha'_k) = O(\eta'_k).$$

Similarly, we obtain $\|(\pi^k)^\top\mathbf{1}_n - \nu\|_1 = O(\eta'_k)$, thus yielding corresponding row- and column-marginal feasibility estimates of order $O(\eta'_k)$.

Since the proof for Proposition 3 and 4 require only the marginal errors and the step size are both $O(\eta'_k)$ with $\gamma < 1$. The same proof structure gives $D_h(\pi^{k+1}, w^k) = O({\eta'_k}^2)$, $D_h(\pi^{k+1}, w^{k+1}) = O({\eta'_k}^2)$, and $|A_t| = O({\eta'_k}^3)$.

For the harmonic schedule, $\rho'_{k+1} - \rho'_k = 1$ and $\rho'_k|A_k| = O\left((k+1)^{1-3(1-\gamma)}\right) = O\left((k+1)^{-2+3\gamma}\right)$. When $\gamma < \frac{1}{3}$,

$$\sum_{k=1}^{\infty} \rho'_k|A_t| < \infty.$$

When $\gamma < \frac{1}{2}$, the series of penalty-variation error satisfies

$$\sum_{k=1}^{\infty}(\rho'_{k+1} - \rho'_k)D(\pi^{k+1}, w^{k+1}) = \sum_{k=1}^{\infty} O({\eta'_k}^2) = \sum_{k=1}^{\infty} O\left((k+1)^{-2(1-\gamma)}\right) < \infty.$$

By the Lyapunov identity and lower-boundedness similar to Proposition 5, we obtain

$$\sum_{k=1}^{\infty} \rho'_k\left(D_h(\pi^k, \pi^{k+1}) + D_h(w^k, w^{k+1})\right) < \infty.$$

Lemma 11 only uses the half-step optimality conditions and the cross-divergence bound. Thus, Lemma 11 remains valid for $(\alpha'_k, \rho'_k, \eta'_k)$ where $\alpha'_k\|R_k\|^2_{M_k} = O\left(\rho'_k D_h(w^k, w^{k+1})\right) + O(\alpha'_k{\eta'_k}^2)$. When $\gamma < \frac{1}{3}$, $\alpha'_k{\eta'_k}^2 = (k+1)^{-1-2(1-\gamma)} = (k+1)^{-3+2\gamma}$ is summable. Thus

$$\sum_{k=1}^{\infty} \alpha'_k\|R_k\|^2_{M_k} = \sum_{k=1}^{\infty} O\left(\rho'_k D_h(w^k, w^{k+1})\right) + \sum_{k=1}^{\infty} O(\alpha'_k{\eta'_k}^2) < \infty.$$

Since $\sum_{k=1}^{N-1} \alpha'_k = \Theta(\log N)$, similar to the proof in Proposition 6, we obtain

$$\min_{0 \le i \le N-1} \|R_i\|^2_{M_i} = O\left(\frac{1}{\log N}\right).$$

The proof of Theorem 2 uses the divergence of $\sum_k \alpha'_k$, compactness, and Assumption 1, residual summability, and vanishing cross-divergence, all of which established above. We thus show that under the restraint $\gamma < \frac{1}{3}$, i.e., $M_G < \frac{1}{12}$, the same subsequence-convergence conclusions follow. □

A.7 Proof of Proposition 1

Proof. Since $R_t = \rho_t\big(\log w^t - \log w^{t+1}\big) + G(\pi^{t+1}) - G(w^t)$, by Proposition 3 and Pinsker's inequality, $\|\pi^{t+1} - w^t\|_F \le \|\pi^{t+1} - w^t\|_1 \le \sqrt{2D_h(\pi^{t+1}, w^t)} = O(\eta_t)$. Since G is linear, $\|G(\pi^{t+1}) - G(w^t)\|_\infty \le \|D_X(\pi^{t+1} - w^t)D_Y\|_F \le \|D_X\|_2\|D_Y\|_2\|\pi^{t+1} - w^t\|_F = O(\eta_t)$. Since D_X and D_Y are fixed, along $\{t_\ell\}$, $R_{t_\ell,ij} = \rho_{t_\ell}\log\frac{w_{ij}^{t_\ell}}{w_{ij}^{t_\ell+1}} + O(\eta_{t_\ell})$. Fix $(i,j) \notin S(\bar{x})$. By hypothesis, there exists a sequence $\varepsilon_\ell = o(\alpha_{t_\ell})$ such that $\left(w_{ij}^{t_\ell+1}/w_{ij}^{t_\ell}\right) \le 1 + \varepsilon_\ell^{ij}$. Consequently, since $\varepsilon_\ell^{ij} = o(\alpha_{t_\ell})$ and $\log(1+\varepsilon_\ell^{ij}) = \varepsilon_\ell^{ij} + o(\varepsilon_\ell^{ij})$,

$$\rho_{t_\ell}\log\frac{w_{ij}^{t_\ell}}{w_{ij}^{t_\ell+1}} \ge -\rho_{t_\ell}\log(1+\varepsilon_\ell^{ij}) = -\frac{\log(1+\varepsilon_\ell^{ij})}{\alpha_{t_\ell}} = o(1).$$

Since $\eta_{t_\ell} \to 0$, the residual decomposition therefore yields $\liminf_{\ell\to\infty} R_{t_\ell,ij} \geq 0$, which holds for every $(i,j) \notin S(\bar{x})$. Therefore, Assumption 1 holds along the subsequence.

If $w_{ij}^{t_\ell+1} \leq w_{ij}^{t_\ell}$ eventually, then $\rho_{t_\ell} \log\left(w_{ij}^{t_\ell}/w_{ij}^{t_\ell+1}\right) \geq 0$, so the same conclusion follows immediately. The first condition then implies $\liminf_{\ell\to\infty} R_{ij}^{t_\ell} \geq 0$. Moreover, since $\varepsilon_\ell^{ij} = o(\alpha_{t_\ell})$ and $\rho_{t_\ell} = 1/\alpha_{t_\ell}$, $-\rho_{t_\ell}\log(1+\varepsilon_\ell^{ij}) = o(1)$, which proves the second claim. □

A.8 Proof of Proposition 2

Proof. Let

$$\mu = \nu = \left(\frac{1}{2}, \frac{1}{2}\right)^\top, \qquad D_X = D_Y = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} =: J.$$

Consider the one-parameter family

$$W(u) = \begin{pmatrix} u & \frac{1}{2}-u \\ \frac{1}{2}-u & u \end{pmatrix}, \qquad u \in \left(\frac{1}{4}, \frac{1}{2}\right).$$

Then $W(u) \in \mathcal{C}_1 \cap \mathcal{C}_2$ and $G(W(u)) = JW(u)J = W(u)$. Its projective diameter is

$$\Delta(W(u)) = \log \frac{W_{11}(u)W_{22}(u)}{W_{12}(u)W_{21}(u)} = 2\log\frac{u}{\frac{1}{2}-u} > 0.$$

During the first exponential perturbation, the log cross-ratio changes by

$$\alpha\big(G_{11}(W(u)) + G_{22}(W(u)) - G_{12}(W(u)) - G_{21}(W(u))\big) = \alpha(4u-1).$$

The row normalization is a positive diagonal scaling and therefore preserves cross-ratios. Moreover, the normalized matrix remains in the same family, say $W(u_1)$, with $u_1 > u$. The second exponential perturbation changes the log cross-ratio by $\alpha(4u_1 - 1) \geq \alpha(4u-1)$, and the column normalization again preserves cross-ratios. Hence one full alternating update cycle satisfies

$$\Delta(w^{t+1}) - \Delta(w^t) \geq 2\alpha_t(4u_t - 1).$$

Since u_t is nondecreasing and $u_0 > 1/4$, there exists $c_0 > 0$ such that $2(4u_t - 1) \geq c_0$ for all t. Therefore

$$\Delta(w^T) - \Delta(w^0) \geq c_0 \sum_{s=0}^{T-1} \alpha_s.$$

If the projective guard holds, then $\Delta(w^T) - \Delta(w^0) \leq a\log(T+1)$. Thus

$$\sum_{s=0}^{T-1} \alpha_s \leq \frac{a}{c_0}\log(T+1).$$

Now suppose, for contradiction, that there exist $c > 0$, $\beta < 1$, and T_0 such that $\alpha_t \geq c(t+1)^{-\beta}$, for every $t \geq T_0$. Then for every $T > T_0$,

$$\sum_{s=0}^{T-1} \alpha_s \geq c \sum_{s=T_0}^{T-1} (s+1)^{-\beta} \geq c'\big(T^{1-\beta} - T_0^{1-\beta}\big)$$

for some constant $c' > 0$. Since $\beta < 1$, this lower bound grows faster than $\log(T+1)$, contradicting

$$\sum_{s=0}^{T-1} \alpha_s \leq \frac{a}{c_0}\log(T+1).$$

□

This proposition explains why Lemma 7 is stated in the asymmetric form $\frac{c_-}{t+2} \leq \alpha_t \leq c_+(t+1)^{-\beta}$. The logarithmic guard gives only the harmonic accepted-step floor $\alpha_t \geq c_-/(t+2)$, while the delayed-power proposal gives the upper control $\alpha_t \leq c_+(t+1)^{-\beta}$.

B EXPERIMENTS

B.1 SYNTHETIC GRAPH ALIGNMENT EXPERIMENT DETAILS

Graph Construction The source graph contains $n = 250$ nodes and is sampled from a Gaussian random partition model with parameters $s = 50$, $v = 5$, within-community edge probability $p_{\text{in}} = 0.2$, and between-community edge probability $p_{\text{out}} = 0.02$. Here s is the nominal mean community size and v controls its variation. Community sizes are sampled until 250 nodes are assigned, with the final community taking the remaining nodes. Given the source graph, we construct the target graph using the noisy-node procedure of Li et al. (2023). For noise level $q = 0.3$, we first add $\lfloor q|E| \rfloor$ randomly sampled edges that are absent from the source graph and append $\lfloor qn \rfloor = 75$ noisy nodes, resulting in a target graph with 325 nodes. Each added node is connected randomly to the existing graph. We use the unweighted adjacency matrices as the structural matrices D_X and D_Y, together with uniform node distributions $\mu = \mathbf{1}_{250}/250$ and $\nu = \mathbf{1}_{325}/325$. All results are averaged over 30 independent random seeds. Runtime denotes the wall-clock time from initialization to the returned coupling.

FGW stationarity residual. For synthetic and real domain-adaptation experiments, we evaluate the stationarity residual using the FGW objective $F_{\text{FGW}}(X) = 2\lambda F_{\text{GW}}(X) + (1-\lambda)\langle M, X\rangle$. Thus, $\nabla F_{\text{FGW}}(X) = 2\lambda \nabla F_{\text{GW}}(X) + (1-\lambda)M$. The linear term does not affect the gradient Lipschitz bound, so we use $\eta = \max\{1, 4\lambda|D_X|_2|D_Y|_2\}^{-1}$.

KL-BAPG + Sinkhorn Scaling We compare fixed-penalty KL-BAPG with a postprocessed variant that applies alternating row and column scalings to its final coupling, without further GW optimization. Scaling stops when the sum of the marginal L_1 errors is at most 10^{-12} or after 1,000 cycles. We apply no subsequent exact marginal correction. We report accuracy, GW objective, and marginal infeasibility for the returned coupling, and evaluate projected stationarity on its feasible shadow using the same diagnostic procedure as for other infeasible outputs. This comparison measures the effect of post-hoc scaling on feasibility and projected stationarity.

Top-1 Accuracy Let $W \in \mathbb{R}+^{n\times m}$ denote the estimated coupling and let $\tau(i)$ be the ground-truth target node corresponding to source node i. The predicted correspondence is the target node receiving the largest coupling weight, $\widehat{\tau}(i) = \arg\max_{1\le j\le m} W_{ij}$ and top-1 accuracy is $\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\widehat{\tau}(i) = \tau(i)\}$.

B.2 REAL GRAPH-ALIGNMENT DETAILS

In PROTEINS and ENZYMES, nodes represent secondary protein-structure elements such as helices, sheets, and turns, while edges encode sequential or spatial proximity. ENZYMES contains proteins from six enzyme classes and PROTEINS distinguishes enzymes from non-enzymes; these graph-level labels are not used. We retain graphs with at most 250 nodes, yielding 598 ENZYMES graphs and 1106 PROTEINS graphs. For each source graph $G_X = (V_X, E_X)$ and corruption level $q = 0.1$, following Li et al. (2023), we add $\lfloor q|E_X| \rfloor$ randomly sampled absent edges and $\lfloor q|V_X| \rfloor$ noisy nodes and randomly connect them to the augmented graph. This produces the target graph $G_Y = (V_Y, E_Y)$. We set $D_X = A_X$ and $D_Y = A_Y$, where A_X and A_Y are the corresponding unweighted symmetric adjacency matrices, and use uniform node measures.

Given a feasible coupling $\widehat{W}$, Hungarian decoding produces an injective assignment $\widehat{\tau}$. We measure Hungarian one-to-one matching accuracy as

$$\text{Accuracy} = \frac{1}{|V_X|}\sum_{i\in V_X} \mathbf{1}\{\widehat{\tau}(i) = i\}.$$

B.3 SYNTHETIC DOMAIN ADAPTATION DETAILS

Graph Construction For source sample i and target sample j, let $x_i^s \in \mathbb{R}$ denote the common source feature and $x_j^t = (x_{j,c}^t, x_{j,a}^t) \in \mathbb{R}^2$ denote the target observation, where $x_{j,c}^t$ is the common feature and $x_{j,a}^t$ is the target-only feature. Source and target class proportions are both $1/2$, and we use uniform marginals $\mu_i = 1/n_s$ and $\nu_j = 1/n_t$. Target labels are hidden from every method and are accessed only after training to compute accuracy. Following Aritake & Hino (2026), the cross-domain feature cost is

$$M_{ij} = \frac{|x_i^s - x_{j,c}^t|^2}{\max_{i',j'} |x_{i'}^s - x_{j',c}^t|^2}.$$

Let $\widetilde{D}_X$ be the max-normalized squared-Euclidean source distance. We form the label-aware source matrix

$$(D_X)_{ii'} = 0.9(\widetilde{D}_X)_{ii'} + 0.1\mathbf{1}\{y_i^s \neq y_{i'}^s\}.$$

The target matrix D_Y is the all-pairs shortest-path distance on the symmetrized 4-NN graph of the full target features with squared-Euclidean edge weights and clipping distances at 10. This is different from Aritake & Hino (2026) where they assign disconnected pairs ten times their largest finite shortest-path distance. All BAPG variants minimize $F_{\mathrm{FGW}}(X)$ with $\lambda = 10^{-2}$. Here we use the squared-loss FGW for the objective function while Aritake & Hino (2026) use an absolute distance discrepancy.

Given the returned coupling X, we form a feasible coupling $\widehat{X}$ by alternating the row scaling $X \leftarrow \mathrm{Diag}(\mu/(X\mathbf{1}))X$ and column scaling $X \leftarrow X\,\mathrm{Diag}(\nu/(X^\top\mathbf{1}))$. We stop when $\mathrm{feas}(X) \leq 10^{-12}$ or after 500 scaling cycles, and use the resulting $\widehat{X}$ to transport source labels: $\widehat{y}_j^t = \arg\max_c \frac{1}{\nu_j}\sum_i \widehat{X}_{ij}\mathbf{1}\{y_i^s = c\}$. The accuracy is then defined as $\mathrm{Accuracy} = \frac{1}{n_t}\sum_j \mathbf{1}\{\widehat{y}_j^t = y_j^t\}$.

Baseline Information Baselines are POT implementations of exact OT, GW, FGW, LA–FGW, GB–FGW, LAGB–FGW, and naive OT clustering (Aritake & Hino, 2026). Naive OT clustering averages the transported class scores within each target spectral cluster and assigns the clusterwise to its maximum. The BAPG variants start from $X^0 = \mu\nu^\top$ and run for 4,000 iterations. The POT GW and FGW-based baselines use a maximum of 200 outer iterations with tolerance 10^{-7}.

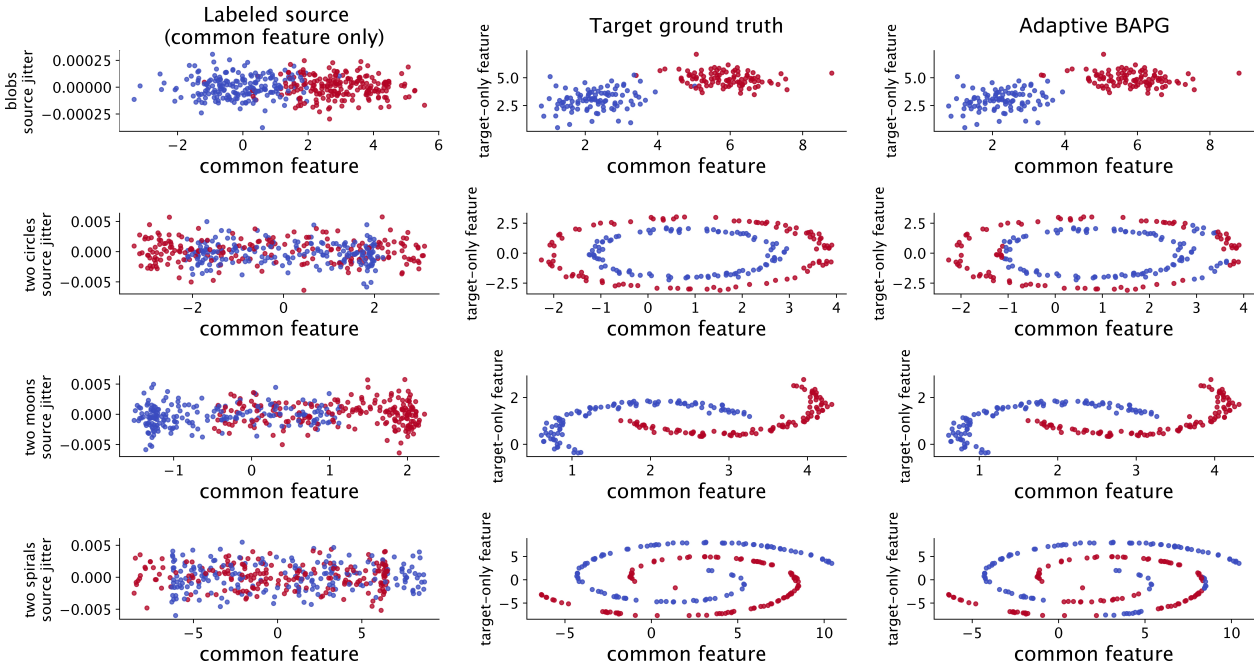


Figure 5: Visualization of heterogeneous domain adaptation for seed 0. Left: labeled source samples plotted by their single observed common feature; the vertical displacement of source points is jitter for visualization only and is not used as method input. Middle: target samples in the common and target-only features, colored by ground-truth labels hidden during training. Right: the same target samples colored by A-KL-BAPG predictions. Quantitative comparisons with KL-BAPG and other baselines are reported in Table 3(a).

Results Visualization Figure 5 provides a qualitative visualization of the four synthetic domain-adaptation settings: blobs, two circles, two moons, and two spirals. The first column shows the labeled source samples, where only the common feature is observed while the second column shows the target samples colored by their ground-truth labels. The target samples contain an additional, target-specific feature revealing their two-dimensional geometry. The third column shows the labels predicted by A-KL-BAPG. Despite the shift between the source and target distributions, A-KL-BAPG separates the two classes almost perfectly for the blobs and two-moons. For the two-circles setting, most points are classified correctly with errors concentrated in portions of the rings that are difficult to distinguish using the common feature alone. The two-spirals setting is particularly challenging because the two classes become intertwined and overlap substantially when projected onto the common feature. Nevertheless, the predicted labels largely follow

the ground truth. These examples demonstrate the ability of A-KL-BAPG to produce coherent target predictions across a range of heterogeneous domain-adaptation settings.

B.4 Real Domain Adaptation Details

Graph Construction Each recording is divided into 5-second windows and represented by the mean and standard deviation of each axis, giving six back-sensor features shared by both domains and six thigh-sensor features available only in the target domain. For each subject, we use a stratified 80/20 split where the labeled training windows that form the source domain use only the six shared features and the test windows that form the unlabeled target domain use all twelve features. Following Aritake & Hino (2026), we introduce a controlled shift in the target data. For each split, we fit PCA to the unshifted target samples and perturb the first three principal directions sequentially. Along each direction, we sort the sample projections and perform 50 sweeps, updating samples in random order and drawing each updated projection uniformly between its current neighbors. Boundary positions are padded by the original projection range. We use the split seed for these draws and leave the source samples unchanged. After shifting the target observations, we construct D_X, D_Y, and M as in Appendix B.3. In particular, D_Y is the all-pairs shortest-path distance on a symmetrized 4-NN graph built from all twelve target features, with squared-Euclidean edge weights and distances capped at 10; D_X uses the six shared source features and source labels, while M uses only the six features shared across domains. λ is also set to 0.01, following Aritake & Hino (2026).

Table 4: Complete HAR70+ domain adaptation results. Accuracy values are retained from the previous experiments. Runtimes are updated using 54 jobs and reported as mean $\pm$ standard deviation. All values are rounded to three decimal places. A dash indicates an unavailable updated runtime.

Method	Accuracy	Runtime (s)
KL–BAPG	0.880 ± 0.074	11.739 ± 7.283
KL–BAPG + Sinkhorn	0.880 ± 0.074	11.967 ± 7.531
$1/k$	0.662 ± 0.068	11.635 ± 7.216
Pure power	0.741 ± 0.068	11.627 ± 7.215
A-KL-BAPG	0.880 ± 0.074	67.665 ± 58.035
POT OT	0.827 ± 0.079	0.008 ± 0.006
POT GW	0.803 ± 0.211	0.102 ± 0.127
POT FGW	0.827 ± 0.079	0.025 ± 0.017
POT LA-FGW	0.829 ± 0.078	0.028 ± 0.020
POT GB-FGW	0.854 ± 0.079	0.047 ± 0.036
LAGB-FGW	0.879 ± 0.077	0.056 ± 0.048
Naive OT cluster	0.767 ± 0.085	0.011 ± 0.008
JDOT	0.805 ± 0.095	0.112 ± 0.079
DAEVS	0.895 ± 0.091	0.255 ± 0.181