BREGMAN–GROMOV–WASSERSTEIN CLUSTERING: FINITE-LENGTH WHOLE-SEQUENCE CONVERGENCE AND LOCAL INITIALIZATION CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Gromov–Wasserstein (GW) graph clustering methods recover communities from pairwise relations by matching nodes to a prototype graph. We propose Bregman–Gromov–Wasserstein (BGW) clustering, which jointly learns the node–prototype transport π, the cluster masses, and the prototype relation D'. The prototype relation admits a closed-form Bregman barycenter update, which we combine with mirror descent on the transport in an alternating solver. We prove finite-length whole-sequence convergence to a constrained critical point and characterize the vanishing-floor limit of the assignment update. We further establish finite-horizon Lipschitz continuity with respect to initialization, showing that the output perturbation is controlled by the initial perturbation up to prototype relabeling. Experiments reveal that different structured initializations can lock BGW into distinct finite-budget solutions, and that their effectiveness varies across graphs. We therefore design MIXGSP, combining GWL, SpecGWL, and product plans to broaden the structured solutions explored by BGW, yielding improvements over some existing GW clustering methods across the benchmarks.

1 INTRODUCTION

Classical graph clustering methods recover communities from pairwise graph relations without requiring node features (Blondel et al., 2008; Clauset et al., 2004; Ng et al., 2001; Rosvall & Bergstrom, 2008). Gromov–Wasserstein (GW) methods compare two graphs by a coupling that matches node-pair relations to node-pair relations (Mémoli, 2011; Peyré et al., 2016). In GW graph clustering, the source graph is matched to a small target graph whose nodes represent clusters (Xu et al., 2019; Chowdhury & Needham, 2021; Vincent-Cuaz et al., 2022). The target has two components: a marginal (the proportion of nodes assigned to each prototype) and a relation (the similarity between prototypes). Existing GW partitioners prescribe the prototype relation and differ in whether the target marginal is fixed. Gromov–Wasserstein Learning (GWL) (Xu et al., 2019) uses a pre-estimated target marginal q with the self-loop-only relation $Q = \mathrm{diag}(q)$; Spectral GWL (SpecGWL) (Chowdhury & Needham, 2021) retains this target construction and compares heat kernels instead of adjacency matrices; semi-relaxed GW (srGW) (Vincent-Cuaz et al., 2022) frees the target marginal while keeping the prototype relation fixed. This leaves open whether the target relation should also adapt to the source graph rather than remain prescribed. We therefore learn it jointly with the assignment.

We propose Bregman–Gromov–Wasserstein (BGW) clustering. BGW learns a node–prototype transport π together with a prototype relation D', with the column sums of π giving the cluster masses. The algorithm alternates entropic mirror descent on π with a closed-form Bregman barycenter update of D'. Different generators induce different prototype means (Appendix K.1), and prescribed prototype components recover existing GW-type formulations. We prove finite-length whole-sequence convergence at a fixed positive transport floor τ, and characterize the behavior as this floor vanishes. Algorithm 1 summarizes the solver.

Convergence does not determine which critical point is reached, and the nonconvex objective makes initialization consequential. Existing GW partitioners also commit to particular starts: GWL uses a rank-one product plan, srGW uses k-means assignments, and product couplings remain a common

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.

default in GW solvers (Xu et al., 2019; Vincent-Cuaz et al., 2022; Chowdhury & Needham, 2021). We therefore study how initialization affects BGW dynamics. We show that, after any fixed number of outer updates, nearby initial states lead to nearby outputs up to relabeling of prototypes. Based on this observation, we consider initializations derived from existing clustering solutions, which we call *structured starts*. Empirically, different structured starts lead reproducibly to different clustering solutions, motivating the initialization scheme of Section 4.2.

This work makes three contributions.

1. We formulate BGW graph clustering, jointly learning the assignment, cluster masses, and prototype relation. It recovers existing GW clustering models through prescribed prototype components and connects, under suitable specializations, to graph cuts, weighted symmetric NMF, and Bregman clustering of pairwise relations.
2. We establish finite-length whole-sequence convergence to a constrained critical point and characterize the behavior as the positive transport floor vanishes.
3. We characterize how BGW depends on initialization, both theoretically and empirically. We show that nearby initializations remain controlled after any fixed number of iterations, and find that different structured initializations can lead to reproducibly different clustering solutions. These observations motivate an initialization scheme combining GWL, SpecGWL, and product plans.

2 Bregman–GW Prototype Learning

The input is a graph G on N nodes, with no node features. From its normalized Laplacian L we form a positive pairwise relation $D = e^{-t_{\mathrm{heat}} L}$; a directed graph uses Chung's Laplacian (Chung, 2005).

BGW clustering learns two objects: a node–prototype transport π, and a $K \times K$ prototype relation D'. The entry π_{ik} is the mass that node i places on prototype k; the column sums of π give the cluster masses $\nu_k = \sum_i \pi_{ik}$. The entry D'_{kl} is the Bregman barycenter of source pairs (i, j) weighted by $\pi_{ik}\pi_{jl}$. The algorithm alternates a D' update at fixed π with an entropic mirror descent step on π at fixed D'. The returned partition is the row-wise maximum of the plan, $\widehat{y}_i = \arg\max_k \pi_{ik}$. Figure 1 shows this pipeline and the target each existing partitioner starts from.

2.1 Objective and structured prototype relation

Write $D \in \mathbb{R}^{N\times N}_{>0}$ for the pairwise relation and let $\mu \in \mathbb{R}^N_{>0}$ satisfy $\sum_i \mu_i = 1$. The statements below keep a general positive μ. Node–prototype transports range over

$$\mathcal{C} = \left\{ \pi \in \mathbb{R}^{N\times K}_{\geq 0} : \sum_{k=1}^{K} \pi_{ik} = \mu_i \right\}. \tag{1}$$

For a strictly convex twice continuously differentiable generator φ, define

$$B_\varphi(a, b) = \varphi(a) - \varphi(b) - \varphi'(b)(a - b).$$

The BGW loss for prototype relation D' is

$$\Psi(\pi, D') = \sum_{i,j,k,l} B_\varphi(D'_{kl}, D_{ij})\pi_{ik}\pi_{jl}. \tag{2}$$

A coupled pair $(i, j) \mapsto (k, l)$ pays the Bregman cost of replacing the source relation D_{ij} by the prototype relation D'_{kl}. The primary theoretical model allows every entry of D' to be optimized over a compact box $\mathcal{D}_{\mathrm{full}}$:

$$F_\lambda(\pi, D') := \Psi(\pi, D') - \lambda\|\pi\|_F^2. \tag{3}$$

Graph clustering uses the structured specialization

$$D'_{kk} = d_k, \qquad D'_{kl} = \delta_\varphi(D) \quad (k \neq l), \tag{4}$$

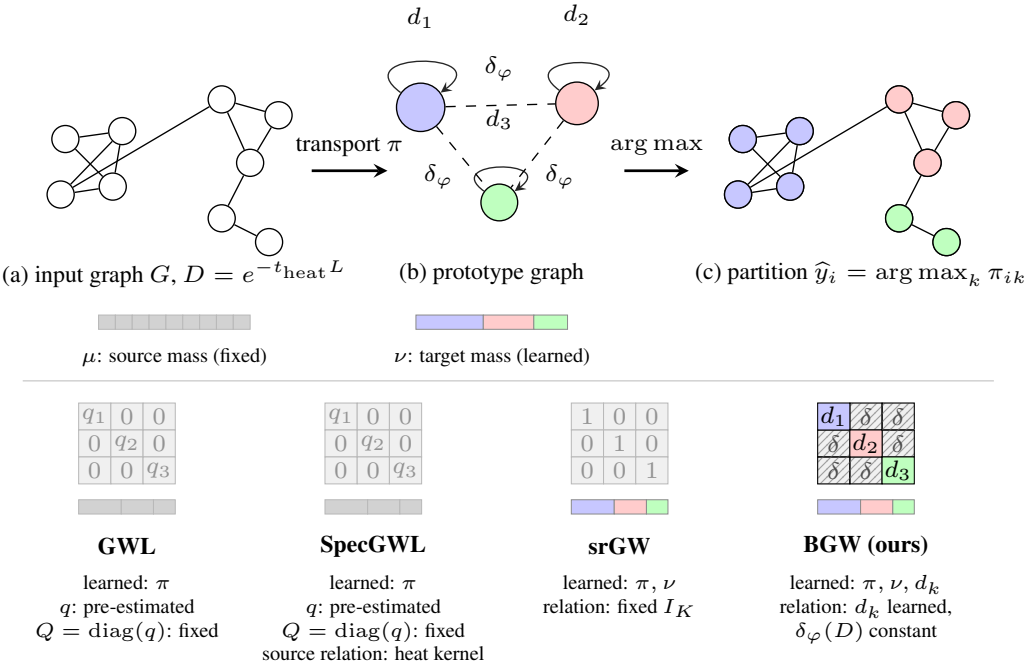

Figure 1: BGW learns the transport π, the cluster masses ν, and the diagonal prototype relation d_k. Top: source graph, learned prototype graph with self-loops d_k and constant off-diagonal $\delta_\varphi(D)$, and the partition $\widehat{y}_i = \arg\max_k \pi_{ik}$ at $K = 3$. Bottom: the target relation (matrices) and target marginal (bars) each method starts from (gray prescribed, hatching constant $\delta_\varphi(D)$, color learned).

where the K prototypes are the clusters, d_k is the learned within-cluster relation, and $\delta_\varphi(D)$ is a constant off-diagonal fill that keeps every first argument of B_φ admissible. Write

$$B_\varphi(0, b) := \lim_{a \downarrow 0} B_\varphi(a, b) \in [0, +\infty]$$

when the limit exists in the extended reals. Then

$$\delta_\varphi(D) := \begin{cases} 0, & B_\varphi(0, D_{ij}) < +\infty \text{ for every } ij, \\ \frac{1}{2} \min\{D_{ij} : D_{ij} > 0\}, & \text{otherwise.} \end{cases}$$

The regularized objective on this family is

$$F_\lambda(\pi, D'(d)) = \Psi(\pi, D'(d)) - \lambda \|\pi\|_F^2. \tag{5}$$

2.2 D' UPDATE

Write $s_k = \nu_k = \sum_i \pi_{ik}$ for the cluster masses. For fixed π, the D' update is the unique minimizer of $\min_{D' \in \mathcal{D}_{\text{full}}} \Psi(\pi, D')$. The objective separates over the K^2 prototype pairs, and each entry has the unique minimizer

$$h_{kl}(\pi) := (\varphi')^{-1} \left(\frac{\sum_{i,j} \pi_{ik} \pi_{jl} \varphi'(D_{ij})}{s_k s_l} \right). \tag{6}$$

Write $h(\pi)$ for the matrix with these entries. The diagonal formula used by the clustering specialization is the restriction

$$h_k(\pi) = (\varphi')^{-1} \left(\frac{\sum_{i,j} \pi_{ik} \pi_{jk} \varphi'(D_{ij})}{s_k^2} \right). \tag{7}$$

For symmetric D, the D' update satisfies $h_{kl}(\pi) = h_{lk}(\pi)$ by exchanging i and j; symmetry is inherited by the update. Minimizing the first argument of $B_\varphi(d_k, D_{ij})$ averages in dual coordinates. Implementation details of the KL-generator smoother are Appendix E. Generator comparisons on the six graphs are Appendix K.1.

Algorithm 1 BGW fixed-λ alternating solver: n_{inner} truncated mirror descent steps on π, then the exact D' update.

Require: Graph G, cluster count K, heat-kernel time t_{heat}, generator φ, fixed λ, floor $\tau > 0$, mirror regularizer ε, inner count $n_{\text{inner}} \geq 1$, stopping tolerance tol, outer budget $q_{\max}$, initial pair (π^0, D'^0) with $D'^0 \in \mathcal{D}_{\text{full}}$.

1: Form $D = e^{-t_{\text{heat}} L}$, the row masses μ, and the truncated polytope $\widetilde{\mathcal{C}}_\tau$ of equation 8.
2: **for** $q = 0, 1, \ldots, q_{\max}$ **do**
3: $\quad \pi^{q,0} \leftarrow \pi^q$.
4: $\quad$ **for** $c = 1, \ldots, n_{\text{inner}}$ **do**
5: $\quad\quad \pi^{q,c} \leftarrow \arg\min_{\pi \in \widetilde{\mathcal{C}}_\tau} \left\{ \langle \nabla_\pi F_\lambda(\pi^{q,c-1}, D'^q), \pi - \pi^{q,c-1} \rangle + \varepsilon D_{\mathrm{KL}}(\pi \| \pi^{q,c-1}) \right\}$ equation 9.
6: $\quad$ **end for**
7: $\quad \pi^{q+1} \leftarrow \pi^{q,n_{\text{inner}}}$.
8: $\quad D'^{q+1} \leftarrow h(\pi^{q+1})$, i.e. $D'^{q+1}_{kl} = (\varphi')^{-1}\big(\sum_{i,j} \pi_{ik}\pi_{jl}\varphi'(D_{ij})\big/(s_k s_l)\big)$ equation 6, with $s_k = \nu_k$ the cluster masses.
9: $\quad$ **if** $\|\pi^{q+1} - \pi^q\|_F + \|D'^{q+1} - D'^q\|_F < \mathrm{tol}$ **then**
10: $\quad\quad$ **break**
11: $\quad$ **end if**
12: **end for**
13: $\widehat{y}_i \leftarrow \arg\max_k \pi^{q+1}_{ik}$.

Ensure: $\widehat{y}$, π^{q+1}, and D'^{q+1}.

2.3 Truncated mirror descent step

The step is an entropic mirror descent update on the truncated polytope

$$\widetilde{\mathcal{C}}_\tau = \{\pi \in \mathcal{C} : \pi_{ik} \geq \tau\}, \qquad 0 < \tau < \min_i \mu_i / K. \tag{8}$$

We refer to the constraint $\tau > 0$ as the *hard floor*. Write $\varepsilon > 0$ for the inverse step size and $\mathrm{KL}(\pi\|\pi^t)$ for the plan-wise Kullback–Leibler divergence.

$$\pi^+ = \arg\min_{\pi \in \widetilde{\mathcal{C}}_\tau} \left\{ \langle \nabla_\pi F_\lambda(\pi^t, D'^t), \pi - \pi^t \rangle + \varepsilon D_{\mathrm{KL}}(\pi \| \pi^t) \right\}. \tag{9}$$

The floor $\tau > 0$ makes $\log \pi$ and ε/τ finite on $\widetilde{\mathcal{C}}_\tau$; finite-length whole-sequence statements use one fixed $\tau > 0$. Setting $\tau = 0$ gives the same step on the closed simplex $\mathcal{C}$, in closed form the row-wise softmax $\pi^+_{i\cdot} = \mu_i \,\mathrm{softmax}(\log \pi_{i\cdot} - \varepsilon^{-1} \nabla_{i\cdot} F_\lambda)$.

For the clustering specialization, line 8 uses the diagonal update (7), while the off-diagonal entries remain fixed as in (4).

3 Convergence and Consistency

This section records finite-length whole-sequence convergence of Algorithm 1 at one fixed λ.

Assume $D_{ij} \in [m, M] \subset (0, \infty)$ and φ is analytic near $[m, M]$ with uniformly positive curvature. Let $\mathcal{D}_{\text{full}} = [m, M]^{K \times K}$ and

$$\Phi_\tau = F_\lambda + \delta_{\widetilde{\mathcal{C}}_\tau} + \delta_{\mathcal{D}_{\text{full}}}.$$

The implemented KL generator uses a 10^{-10} shift (Appendix E). The floor is the fixed $\tau > 0$ of Section 2. After a D' update, $D' \in \mathcal{D}_{\text{full}}$. For $z = (\pi, D')$, define

$$\|z\|_\times := \|\pi\|_F + \|D'\|_F,$$

so $\|z^{t+1} - z^t\|_\times = \|\pi^{t+1} - \pi^t\|_F + \|D'^{t+1} - D'^t\|_F$. For fixed D', let $L_{\text{rel}}(D', \lambda)$ denote the relative curvature bound defined in Appendix A.3, and set

$$\overline{L}_{\text{rel}}(\lambda) := \max_{D' \in \mathcal{D}_{\text{full}}} L_{\text{rel}}(D', \lambda).$$

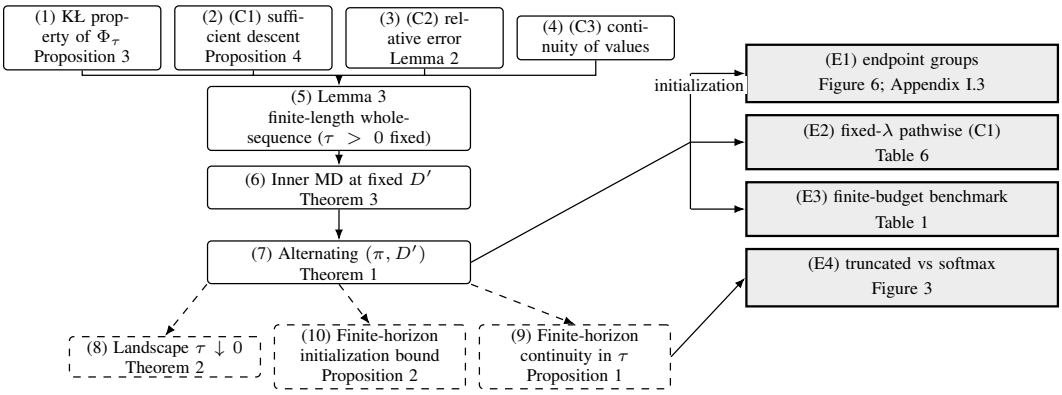

Figure 2: Finite-length whole-sequence chain at one fixed $\tau > 0$ ((1)–(7); Lemma 3 after KŁ and (C1)–(C3)), dashed vanishing-floor and finite-horizon statements ((8)–(10)), and (E1)–(E4).

3.1 Whole-sequence convergence

Theorem 1 (Alternating convergence)**.** *Let $n_{\text{inner}} \geq 1$, and let (π^{t+1}, D'^{t+1}) be given by n_{inner} entropic mirror descent steps for π on $\widetilde{\mathcal{C}}_\tau$ at fixed D'^t, followed by the D' update equation 6. If λ is held constant along this sequence and there exists ℓ such that $\varepsilon > \ell \geq \sup_{t\geq 0} L_{\text{rel}}(D'^t, \lambda)$, then*

$$\sum_{t=0}^{\infty} \|z^{t+1} - z^t\|_\times < \infty,$$

where $z^t = (\pi^t, D'^t)$, and (π^t, D'^t) converges to a constrained critical point of Φ_τ.

The box bound $\varepsilon > \overline{L}_{\text{rel}}(\lambda)$ is sufficient for that pathwise hypothesis.

The proof applies the KŁ finite-length argument to the alternating pair by verifying sufficient descent, relative error, and continuity of objective values; Appendices A–B give the proof and Figure 2 summarizes the dependency chain.

Corollary 1. *Under the assumptions of Theorem 1, the first T fixed-λ outer iterates of Algorithm 1 satisfy, up to constants that depend only on the data, λ, τ, and ε,*

$$\min_{0\leq t<T} \operatorname{dist}\big(0, \partial\Phi_\tau(\pi^{t+1}, D'^{t+1})\big) = O\left(\sqrt{\frac{n_{\text{inner}}}{(\varepsilon - \ell)T}}\right),$$

so for fixed n_{inner} the best outer-iterate stationarity residual is $O(T^{-1/2})$.

Proposition 6 supplies an optional open ray of inverse steps $\varepsilon > \frac{1}{2}\overline{L}_{\text{rel}}(\lambda)$; Table 7 and Appendix J record that prior.

Corollary 2 (Clustering specialization)**.** *If the prototype relation is restricted to $D'_{kk} = d_k$ and $D'_{kl} = \delta_\varphi(D)$ for $k \neq l$, and the diagonal entries are updated by equation 7, the same finite-length whole-sequence argument applies on the corresponding compact domain $[m, M]^K$.*

3.2 Vanishing-floor consistency

The floor $\tau > 0$ affects both the optimization landscape and the solver. We study $\tau \downarrow 0$ at two levels: Theorem 2 treats minimizers and critical points, while Proposition 1 compares truncated and softmax iterates over a fixed finite horizon J. Write

$$Z_\tau = \mathcal{Z}_\tau = \widetilde{\mathcal{C}}_\tau \times \mathcal{D}, \qquad Z_0 = \mathcal{Z}_0 = \mathcal{C} \times \mathcal{D},$$

where $\mathcal{D}$ denotes either $\mathcal{D}_{\text{full}}$ or the structured domain $[m, M]^K$, with constrained objectives Φ_τ and Φ_0, and let $r_\tau = \tau/(\min_i \mu_i/K)$.

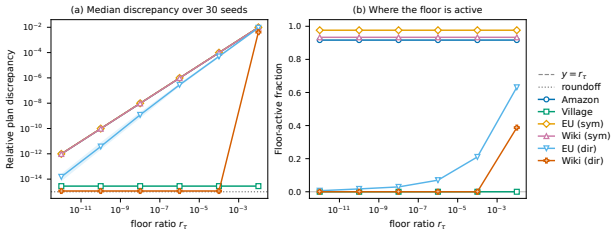

Figure 3: Plan discrepancy $\|\pi_\tau^{20} - \pi_{\text{soft}}^{20}\|_F / \|\pi_{\text{soft}}^{20}\|_F$ tracks r_τ on Amazon, EU(sym), Wiki(sym), and EU(dir), the graphs where the floor is active. The AMI between the two partitions equals 1 at $r_\tau \le 10^{-8}$ on all six graphs (Table 22).

Theorem 2 (Vanishing-floor landscape stability). *Let $0 < \tau_j < \min_i \mu_i / K$ with $\tau_j \downarrow 0$, and define $v(\tau) := \min_{z \in \mathcal{Z}_\tau} F_\lambda(z)$. Then*

(1) $v(\tau_j) \downarrow v(0)$, and every cluster point of global minimizers $z_j \in \arg\min_{\mathcal{Z}_{\tau_j}} F_\lambda$ lies in $\arg\min_{\mathcal{Z}_0} F_\lambda$;

(2) if $z_j \in \mathcal{Z}_{\tau_j}$ satisfies $0 \in \partial\Phi_{\tau_j}(z_j)$, then every cluster point $\bar{z}$ of (z_j) satisfies $0 \in \partial\Phi_0(\bar{z})$.

The proof is in Appendix C.

Proposition 1 (Finite-horizon continuity in the floor). *Within a finite horizon J,*

$$\max_{0 \le j \le J} \|\pi_\tau^J - \pi_{\text{soft}}^J\|_F \le C_J \tau,$$

where π_τ^J and π_{soft}^J are the truncated and softmax plans after J outer blocks.

We verify Proposition 1 numerically: on the six benchmark graphs, at fixed λ and ε, twenty outer updates and no continuation, the plan discrepancy scales with r_τ where the floor is active, and the partition $\widehat{y}_i = \arg\max_k \pi_{ik}$ agrees with the $\tau = 0$ softmax run at every recorded $r_\tau \le 10^{-8}$ (Figure 3; Table 22, Appendix K.2).

3.3 Finite-horizon local initialization consistency

Theorem 1 concerns one trajectory as the outer index tends to infinity. At a finite horizon J, the same J outer blocks of Algorithm 1 are applied to every initial state. The next statement gives a quantitative perturbation bound for that shared iterate with respect to the initial joint state $z = (\pi, D')$.

Proposition 2 (Finite horizon initialization bound). *For every finite horizon J, after alignment of prototype labels,*

$$\min_\sigma \|z^{J,\sigma} - \widehat{z}^J\|_\times \le L_{\text{out}}^J \min_\sigma \|z^\sigma - \widehat{z}\|_\times,$$

where $L_{\text{out}} < \infty$ is a Lipschitz constant for one outer block and

$$\|z\|_\times := \|\pi\|_F + \|D'\|_F \qquad \text{for } z = (\pi, D').$$

The precise permutation action, proof, hard-assignment stability, and extension to a prescribed finite continuation are given in Appendix D. Figure 5 in Appendix I.2 empirically examines this fixed-budget dependence by perturbing structured initializations and tracking the resulting endpoint displacement.

4 Experiments

We use the four prepared graphs of Appendix F (Amazon 1,501 nodes and 12 clusters, Village 1,991 and 12, EU 1,005 and 42, Wikicats 1,998 and 15), score the row-wise maximum $\widehat{y}_i = \arg\max_k \pi_{ik}$ of Section 2 by adjusted mutual information (AMI) (Vinh et al., 2010), and compare three baseline families: community detection (Louvain, Infomap, FastGreedy) (Blondel et al., 2008; Clauset

et al., 2004; Rosvall & Bergstrom, 2008), spectral clustering (Ng et al., 2001), and GW partitioning (GWL, SpecGWL, Bregman Alternating Projected Gradient (BAPG), entropic GW (EGW), srGW). BGW and SpecGWL receive heat kernels, srGW the adjacency, and community methods the raw graph; provenance, protocol labels, the per-method inputs, and the fixed and searched settings are Appendices F and E.

4.1 START-LOCK PHENOMENON AND LOCAL INITIALIZATION CONSISTENCY

We first ask how initialization affects the solution reached under a fixed computational budget. Since the BGW state contains both the transport π and the prototype relation D', we further separate their roles. Under one fixed F_λ and hard-floor domain, with the same solver settings and outer budget, we vary the structured transport start while holding the prototype update fixed, and then hold the transport start fixed while varying the prototype update. We compare a fixed diagonal, a shared learned diagonal value, and the BGW update of equation (7), which learns one d_k per prototype. Full protocols and results are reported in Appendix I.

Across Amazon, Village, and Wiki(dir), changing the structured transport start produces substantially larger changes in the final transport and partition than changing the diagonal prototype update from the same start. On Amazon, for example, the BGW diagonal update reaches AMI 0.409, 0.791, and 0.918 from GWL, SpecGWL, and spectral-clustering starts, respectively. In contrast, learning the diagonal does not consistently improve AMI over the fixed-diagonal variant on any of the three graphs. This is not simply because the learned prototype remains unchanged: for the Amazon GWL and SpecGWL starts, the learned d_k differ across clusters while the final transport and hard partition remain unchanged from the fixed-diagonal run to recorded precision. Thus prototype adaptation need not produce a corresponding assignment change.

The broader initialization study gives the same picture across nine families. Structured starts reach distinct solutions (AMI 0.32–0.92 on Amazon and Village), whereas unstructured starts remain near chance (Figure 6; Table 14). After prototype alignment, noisy Mixed starts (Mixed+noise) and spectral retain $30/30$ seeds in their respective Amazon endpoint groups, while GWL-as-π^0 and SpecGWL-as-π^0 retain $23/30$ and $30/30$ (Appendix I). We call this empirical concentration around family-dependent solutions the *start-lock phenomenon*. This empirical start-lock is consistent with the local initialization consistency characterized in Section 3.3.

4.2 A MIXGSP INITIALIZATION CRITERION

The preceding results make the choice of transport initialization consequential, but no single structured start is consistently effective across graphs. The stopping residual is not a reliable selector either: on Amazon, unstructured starts can reach chance-level partitions with smaller residuals than high-AMI structured solutions (Figure 6; Table 14). For example, spectral initialization performs strongly on Amazon but poorly on Wiki(dir), where Mixed+noise reaches a better partition.

We therefore design a ***MixGSP*** of GWL, SpecGWL, and product plans. Rather than committing to one structured family, MixGSP combines complementary structured starts within a single initialization: $\widetilde{\pi}^0 = \rho(w\pi_{\mathrm{GWL}} + (1-w)\pi_{\mathrm{SpecGWL}}) + (1-\rho)\mu\mathbf{1}_K^\top/K$. The prototype initialization is constructed from the corresponding barycenter and prior, followed by the prescribed Stage I continuation before the fixed-λ Stage II solver; full details are given in Algorithm 2 and Appendix E.

The component ablation shows that MixGSP is the main driver of the improvement, while prototype initialization and continuation have graph-dependent effects (Appendix H). We use this initialization in the benchmark evaluation of Table 1, where BGW improves over some existing GW clustering methods across the graph settings.

The distinct solutions reached from structured starts further raise the question of whether they are separated in the objective landscape.

4.3 LANDSCAPE SEPARATION OF STRUCTURED SOLUTIONS

We test this by connecting label-aligned solutions with three paths: an affine path in π with $D' = h(\pi)$, a joint path in (π, D'), and a row-wise KL geodesic (Appendix I.4). We measure the maximum objective rise along each path relative to its endpoints.

Table 1: AMI from a MixGSP. Bold marks the overall column maximum and the maximum within the GW-type rows. BGW-hard-floor is the truncated map of Theorem 1. Selectors and templates: Appendix L.

Method	Amazon	Village	EU(sym)	Wiki(sym)	EU(dir)	Wiki(dir)
Louvain	0.761	**0.937**	0.545	0.548	—	—
Infomap	**0.970**	**0.937**	0.595	0.464	0.574	0.505
Infomap (dir. flow)	—	—	—	—	**0.586**	0.456
FastGreedy	0.772	**0.937**	0.421	0.553	—	—
Spectral	0.696	0.777	**0.629**	0.518	—	—
GWL	0.406	0.657	0.478	0.363	0.440	0.228
SpecGWL	0.749	0.758	0.519	0.439	0.466	0.378
EGW	0.408	0.613	0.476	0.486	0.473	0.476
BAPG	0.127	0.731	0.471	0.556	0.427	0.381
srGW	0.467	0.804	0.502	**0.624**	0.497	**0.626**
BGW ($\tau = 0$)	**0.970**	**0.926**	0.443	0.611	**0.533**	0.323
BGW-hard-floor	**0.970**	**0.926**	**0.565**	0.530	0.485	0.323

Several pairs exhibit positive barriers on the same F_λ. For example, on Amazon the Mixed+noise–spectral pair has relative height $R(\eta^\pi) = 0.127$ along the affine-π path (Figure 4). The separation is not universal: Mixed+noise and GWL are connected without an objective rise along all three tested paths on five of the six graphs. Thus start-lock does not require a barrier along the tested paths, although some endpoint groups show this stronger form of landscape separation. This further supports combining complementary structured starts rather than relying on a single initialization family.

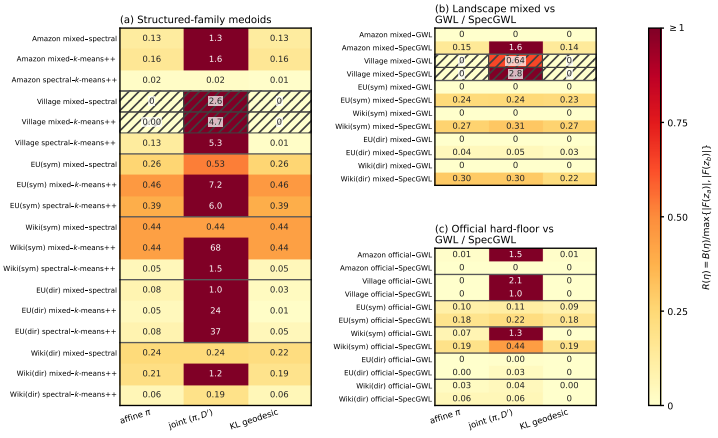

Figure 4: **Landscape separation of structured finite-budget solutions on the same F_λ.** Three paths connect label-aligned solutions: an affine path in π with $D' = h(\pi)$, a joint path in (π, D'), and a row-wise KL geodesic. $R(\eta)$ measures the objective barrier relative to the endpoint scale; zero indicates no objective rise along the tested path.

5 Related Work and Discussion

Convergence guarantees for GW solvers are method-specific. The entropy-regularized scheme of Solomon et al. (2016) is provably convergent, while Rioux et al. (2024) derive non-asymptotic guarantees for entropic GW. BAPG converges to its fixed-point set under a technical assumption and bounds that set relative to the GW critical set (Li et al., 2023). Our analysis establishes finite-length whole-sequence convergence for the alternating BGW transport–prototype updates..

Target-learning variants include GW barycenters of kernel or distance matrices (Peyré et al., 2016; Gong et al., 2022), learned prototype graphs (Deng et al., 2025), and GW quantization of reduced metric-measure spaces (Beier & Eckstein, 2026). DistR jointly learns a low-dimensional target support and its masses, with the target relation parameterized by the embedding geometry, thereby unifying clustering and dimensionality reduction (Van Assel et al., 2025). BGW instead updates prototype relation entries directly as weighted Bregman centroids of source pairwise relations (Banerjee et al., 2005); Appendix M compares the two constructions. Its mirror-descent analysis uses relative-smoothness geometry (Bauschke et al., 2017; Lu et al., 2018).

Model specializations and clustering interpretations. Take $\varphi(x) = \frac{1}{2}x^2$. Then $\delta_\varphi(D) = 0$, so the clustering restriction makes D' diagonal. Writing

$$Q_k(\pi) = \pi_{\cdot,k}^\top D \pi_{\cdot,k}, \qquad C(D,\mu) = \frac{1}{2}\sum_{i,j} D_{ij}^2 \mu_i \mu_j,$$

(7) gives $h_k(\pi) = Q_k(\pi)/s_k^2$. Substitution into (2) yields

$$\Psi(\pi, h(\pi)) = C(D,\mu) - \frac{1}{2}\sum_{k=1}^{K} \frac{Q_k(\pi)^2}{s_k^2}. \tag{10}$$

Since $C(D,\mu)$ is independent of π, minimizing (10) is equivalent to maximizing $\sum_k Q_k(\pi)^2/s_k^2$, the enhanced-balanced within-cluster association objective of Chen et al. (2020). Fixing $D' = I_K$ instead recovers the hard-assignment graph-cut plus exclusive-balancing structure underlying srGW.

On hard assignments with uniform node masses, the same distortion has the matrix-factorization form

$$\min_{T,D'} \|D - TD'T^\top\|_F^2, \tag{11}$$

where $T \in \{0,1\}^{N\times K}$ has one unit per row. Thus this specialization admits a weighted symmetric-NMF interpretation, with T encoding cluster membership and D' the prototype-level relation.

For a general Bregman generator, the clustering interpretation is seen directly by separating intra- and inter-prototype pairs:

$$\Psi(\pi, D') = \underbrace{\sum_{k=1}^{K}\sum_{i,j} B_\varphi(D'_{kk}, D_{ij})\,\pi_{ik}\pi_{jk}}_{\text{intra-prototype}} + \underbrace{\sum_{k\neq l}\sum_{i,j} B_\varphi(D'_{kl}, D_{ij})\,\pi_{ik}\pi_{jl}}_{\text{inter-prototype}}. \tag{12}$$

With $w_{ij}^{(k)} = \pi_{ik}\pi_{jk}$, the intra-prototype term is a weighted Bregman clustering objective on the pairwise relations $\{D_{ij}\}$, whose centroid update is exactly (7) (Banerjee et al., 2005). The inter-prototype term instead measures relations between different clusters against the prescribed off-diagonal prototype relation. Thus BGW couples Bregman clustering of within-cluster pairwise relations with an explicit between-cluster relational term.

6 Conclusion

BGW learns the prototype relation that existing GW partitioners prescribe, admits finite-length whole-sequence convergence, and reveals that the resulting nonconvex clustering dynamics depend strongly but locally stably on structured initialization. Under suitable specializations, BGW recovers existing GW clustering models and connects to graph cuts, weighted symmetric NMF, and Bregman clustering of pairwise relations. Experiments reveal start-lock among structured initializations and motivate the MixGSP used in the benchmark.

Limitations include initialization-dependent nonconvex solutions, prescribed K, and dense heat kernels. Extending the zero-floor theory beyond finite horizons and developing an intrinsic selector from F_λ and D' remain open.

AI USE STATEMENT

We used large language models and other generative AI tools for the purposes disclosed here. We used them to aid and polish writing (wording and copy-editing of text we had already drafted), not to draft sections of the paper. We used them for retrieval and discovery, including searching and screening related work. We used them for research ideation and execution, including formulating experimental protocols and drafting or refactoring analysis, plotting, and solver code. We used them when generating the constructed kernels of Figure 7.

ETHICS STATEMENT

The study reuses public benchmark graphs and does not collect new human-subject data. EU-email is communication metadata; public availability does not remove re-identification concerns, and the artifact does not redistribute the graphs. AMI on these benchmarks is not evidence of fairness for deployment.

REPRODUCIBILITY STATEMENT

Appendix H records the component ablation. Appendix K.3 isolates the optional clipped protocol from the unclipped solver. Appendix J records the fixed-λ solver under precomputed sufficient-descent conditions. Appendix A records the shared lemmas (KŁ, descent, update-map Lipschitz regularity, relative error, finite-length whole-sequence) and inner mirror descent at fixed D'; Appendix B applies the same abstract finite-length whole-sequence lemma to the alternating pair. Appendix M records DistR's clustering–embedding unification and the unrestricted prototype-relation readout. The artifact contains immutable JSON records (Appendix N); graph files are not redistributed.

REFERENCES

Shun-ichi Amari. Information geometry of the EM and em algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995. doi: 10.1016/0893-6080(95)00003-B.

Abhijit Banerjee, Arun G. Chandrasekhar, Esther Duflo, and Matthew O. Jackson. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013. doi: 10.1126/science.1236498.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005.

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. doi: 10.1287/moor.2016.0817.

Florian Beier and Stephan Eckstein. Gromov–Wasserstein quantization and clustering: Structure, rates, and algorithms. arXiv:2608.11016, 2026. Concurrent.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. doi: 10.1088/1742-5468/2008/10/P10008.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007. doi: 10.1137/050644641.

Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018. doi: 10.1137/17M1138558.

Samir Chowdhury and Tom Needham. Generalized spectral clustering via Gromov–Wasserstein learning. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 712–720, 2021.

Fan Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9 (1):1–19, 2005. doi: 10.1007/s00026-005-0237-z.

Andrzej Cichocki and Shun-ichi Amari. Families of alpha-, beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010. doi: 10.3390/e12061532.

Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004. doi: 10.1103/PhysRevE.70.066111.

Bowen Deng, Tong Wang, Lele Fu, Sheng Huang, Chuan Chen, and Tao Zhang. THESAURUS: Contrastive graph clustering by swapping fused Gromov–Wasserstein couplings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):16199–16207, 2025. doi: 10.1609/aaai.v39i15.33779.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1309–1318, 2018.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3259–3269, 2020.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Fengjiao Gong, Yuzhou Nie, and Hongteng Xu. Gromov–Wasserstein multi-modal alignment and clustering. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 603–613, 2022. doi: 10.1145/3511808.3557339.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952. doi: 10.1080/01621459.1952.10483441.

E. A. Leicht and M. E. J. Newman. Community structure in directed networks. *Physical Review Letters*, 100:118703, 2008. doi: 10.1103/PhysRevLett.100.118703.

Jure Leskovec and Andrej Krevl. SNAP datasets: Stanford large network dataset collection. `https://snap.stanford.edu/data/`, 2014.

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 *International Conference on Learning Representations*, 2023.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. doi: 10.1137/16M1099546.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011. doi: 10.1007/s10208-011-9093-5.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, volume 14, pp. 849–856, 2001.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov–Wasserstein averaging of kernel and distance matrices. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 2664–2672, 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.

Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008. doi: 10.1073/pnas.0706851105.

Justin Solomon, Gabriel Peyré, Vladimir G. Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics*, 35(4):1–13, 2016. doi: 10.1145/2897824.2925903.

Hugues Van Assel, Cédric Vincent-Cuaz, Nicolas Courty, Rémi Flamary, Pascal Frossard, and Titouan Vayer. Distributional reduction: Unifying dimensionality reduction and clustering with Gromov–Wasserstein. *Transactions on Machine Learning Research*, 2025. URL `https://mlanthology.org/tmlr/2025/assel2025tmlr-distributional/`.

Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed Gromov–Wasserstein divergence and applications on graphs. In *International Conference on Learning Representations*, 2022.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854, 2010.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable Gromov–Wasserstein learning for graph partitioning and matching. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 555–564, 2017. doi: 10.1145/3097983.3098069.

APPENDIX

Proofs are ordered by dependency (Figure 2 in Section 3). Appendix A records, in order: KŁ of Φ_τ, one-step descent (C1), update-map Lipschitz regularity, relative error (C2), the abstract finite-length whole-sequence lemma, then inner mirror descent at fixed D'. Appendix B applies the same lemma to the alternating pair (π, D'), then the clustering restriction of D', then the optional Stage II bounds. Appendix C records landscape stability as $\tau \downarrow 0$ and finite-horizon continuity of matched solver maps. Appendix D records the finite-horizon Lipschitz bound in the initial joint state. Solver labels used by Table 1, and the dense-gradient timing measurements, are Appendix E. Dataset provenance is Appendix F. Hard-floor replays on the benchmark, pathwise Stage II (C1), and the optional prior ε are Appendix G. The component ablation is Appendix H. Initialization dependence, including the transport/prototype comparison, controlled start perturbations, and the start-lock landscape, is Appendix I (Figure 6, Table 13, Table 14, GWL/SpecGWL-as-π^0 on that F_λ, and interpolating paths in Figure 4, Table 15). Sufficient descent validation under precomputed conditions is Appendix J. Generator scale, floor continuity, and the optional clip are Appendix K, in order: generator ranking by heat-kernel scale (Figure 7, Table 21), the floor sweep, and the continuation-path clip. Selection rules, the recorded AMI ceiling among complete trials, Infomap directed flow, and GW-solver protocols are Appendix L. DistR's clustering–embedding unification, and the corresponding unrestricted prototype-relation readout, are Appendix M. The artifact inventory is Appendix N.

A SHARED LEMMAS: KŁ, DESCENT, UPDATE-MAP LIPSCHITZ REGULARITY, AND INNER MIRROR DESCENT

A.1 PROOF SETTING AND NOTATION

There are N source nodes and K prototypes. The positive row masses $\mu_1, \dots, \mu_N$ satisfy $\sum_i \mu_i = 1$. Every experiment sets $\mu_i \propto \left(\sum_j D_{ij}\right)^{\alpha_{\mathrm{Spec}}}$ with $\alpha_{\mathrm{Spec}} = 0.001$ (Appendix E). For a fixed floor

$$0 < \tau < \min_i \frac{\mu_i}{K},$$

the transport variable $\pi \in \mathbb{R}^{N\times K}$ ranges over

$$\widetilde{\mathcal{C}}_\tau = \left\{ \pi : \sum_{k=1}^{K} \pi_{ik} = \mu_i \text{ and } \pi_{ik} \geq \tau \text{ for every } i, k \right\}.$$

The input relation D has entries $D_{ij} \in [m, M]$, where $[m, M] \subset (0, \infty)$. In the full model the prototype relation satisfies

$$D' \in \mathcal{D}_{\mathrm{full}} = [m, M]^{K\times K}.$$

In the clustering restriction only the diagonal entries are free, so the prototype domain is

$$\mathcal{D}_{\mathrm{clus}} = \{D' : D'_{kk} \in [m, M],\ D'_{kl} = \delta_\varphi(D) \text{ for } k \neq l\}.$$

This set is an affine copy of $[m, M]^K$. We write $\mathcal{D}$ for either $\mathcal{D}_{\mathrm{full}}$ or $\mathcal{D}_{\mathrm{clus}}$, and $\mathcal{Z}_\tau := \widetilde{\mathcal{C}}_\tau \times \mathcal{D}$. Write $Z_\tau := \mathcal{Z}_\tau$. For a joint variable $z = (\pi, D')$, define

$$\|z\|_\times := \|\pi\|_F + \|D'\|_F.$$

Thus $\|z^{t+1} - z^t\|_\times = \|\pi^{t+1} - \pi^t\|_F + \|D'^{t+1} - D'^t\|_F$.

For a set S, its indicator is

$$\delta_S(x) = \begin{cases} 0, & x \in S, \\ +\infty, & x \notin S. \end{cases}$$

The constrained objective is therefore

$$\Phi_\tau(\pi, D') = F_\lambda(\pi, D') + \delta_{\widetilde{\mathcal{C}}_\tau}(\pi) + \delta_{\mathcal{D}}(D').$$

When D' is fixed, $\Phi_\tau(\,\cdot\,, D')$ denotes the objective restricted to the transport variable.

The Bregman generator φ is analytic on an open neighborhood of $[m, M]$ and satisfies

$$\kappa_\varphi := \min_{x\in[m,M]} \varphi''(x) > 0.$$

At outer iteration t, the algorithm holds D'^t fixed, performs n_{inner} mirror descent steps in π, and then applies the D' update. Appendix A.3 analyzes one mirror descent step; there t and $t+1$ label two consecutive mirror descent iterates. The truncated mirror map

$$\mathcal{M}_\tau : Z_\tau \to \widetilde{\mathcal{C}}_\tau$$

sends $z = (\pi, D')$ to one truncated mirror descent step equation 9 (Algorithm 1, line 5). Write $\mathcal{M} = \mathcal{M}_\tau$ for this truncated map. The untruncated softmax map in Appendix C is $\mathcal{M}_0$.

A.2 WHY THE OBJECTIVE HAS THE KŁ PROPERTY

Definition 1 (Kurdyka–Łojasiewicz property)**.** *Let $G : \mathbb{R}^q \to \mathbb{R} \cup \{+\infty\}$ be proper and lower semicontinuous, and let $\bar{z} \in \operatorname{dom} G$. We say that G satisfies the KŁ property at $\bar{z}$ if there exist $\eta > 0$, a neighborhood U of $\bar{z}$, and a function $\psi : [0, \eta) \to \mathbb{R}_{\geq 0}$ such that $\psi(0) = 0$, ψ is continuous on $[0, \eta)$, C^1 on $(0, \eta)$, concave and strictly increasing, and $\psi'(s) > 0$ for $s \in (0, \eta)$, with*

$$\psi'\big(G(z) - G(\bar{z})\big)\, \operatorname{dist}\big(0, \partial G(z)\big) \geq 1$$

whenever $z \in U$ and $G(\bar{z}) < G(z) < G(\bar{z}) + \eta$. Here ∂G is the limiting subdifferential.

Proposition 3 (KŁ property of Φ_τ)**.** *Assume $0 < \tau < \min_i \mu_i/K$. Then Φ_τ is proper, lower semicontinuous, and subanalytic on $\widetilde{\mathcal{C}}_\tau \times \mathcal{D}_{\text{full}}$. It therefore has the KŁ property at every feasible point. The same holds for every fixed-D' slice of Φ_τ and for the clustering restriction.*

Proof of Proposition 3. Bolte et al. (2007, Theorem 3.1) applies to a subanalytic function whose domain is closed and whose restriction to that domain is continuous. At each critical point it gives the nonsmooth Łojasiewicz inequality; Bolte et al. (2007, Remark 3.2) extends the inequality to noncritical points. We verify these assumptions for Φ_τ.

First, $\widetilde{\mathcal{C}}_\tau$ is nonempty because $\pi_{ik} = \mu_i/K > \tau$ is feasible. Its defining row-sum equalities and floor inequalities are polynomial constraints, and

$$\tau \le \pi_{ik} \le \mu_i - (K-1)\tau.$$

Thus $\widetilde{\mathcal{C}}_\tau$ is compact and semialgebraic. The box $\mathcal{D}_{\text{full}}$ is also compact and semialgebraic, so $\mathcal{Z}_\tau = \widetilde{\mathcal{C}}_\tau \times \mathcal{D}_{\text{full}}$ has the same properties.

Second, φ and φ' are analytic near $[m, M]$. For fixed D_{ij}, each cost

$$B_\varphi(D'_{kl}, D_{ij}) = \varphi(D'_{kl}) - \varphi(D_{ij}) - \varphi'(D_{ij})(D'_{kl} - D_{ij})$$

is analytic in D'_{kl}. equation 2 is a finite sum of these costs multiplied by the polynomials $\pi_{ik}\pi_{jl}$, and $-\lambda\|\pi\|_F^2$ is polynomial. Consequently F_λ is analytic on a neighborhood of $\mathcal{Z}_\tau$.

Now $\Phi_\tau = F_\lambda + \delta_{\mathcal{Z}_\tau}$ has the nonempty closed domain $\mathcal{Z}_\tau$. It is proper and lower semicontinuous, and its restriction to the domain is continuous. Its finite-valued graph

$$\big\{\big(z, F_\lambda(z)\big) : z \in \mathcal{Z}_\tau\big\}$$

is subanalytic because $\mathcal{Z}_\tau$ is compact semialgebraic and $z \mapsto (z, F_\lambda(z))$ is analytic. All assumptions of Bolte et al. (2007, Theorem 3.1) are therefore satisfied.

For completeness, define

$$\mathfrak{m}_{\Phi_\tau}(z) := \operatorname{dist}\big(0, \partial\Phi_\tau(z)\big).$$

At a critical point $\bar{z}$, their Theorem 3.1 gives $\theta \in [0, 1)$, $C > 0$, and a neighborhood of $\bar{z}$ on which

$$\big|\Phi_\tau(z) - \Phi_\tau(\bar{z})\big|^\theta \le C\,\mathfrak{m}_{\Phi_\tau}(z).$$

For $s = \Phi_\tau(z) - \Phi_\tau(\bar{z}) > 0$, set

$$\psi(s) = \frac{C}{1-\theta} s^{1-\theta}.$$

Then $\psi'(s) = Cs^{-\theta}$, and the preceding inequality gives

$$\psi'(s)\mathfrak{m}_{\Phi_\tau}(z) \ge 1,$$

which is Definition 1. At a noncritical point, Remark 3.2 states that $\mathfrak{m}_{\Phi_\tau}$ is locally bounded away from zero; a linear ψ gives the same KŁ inequality.

Fixing D' leaves the compact semialgebraic domain $\widetilde{\mathcal{C}}_\tau$ and an analytic objective in π. The clustering domain $\mathcal{D}_{\text{clus}}$ is compact and semialgebraic because it is an affine copy of $[m, M]^K$. The same verification applies in both cases. □

Uniformization used later. Let ω be the cluster set of a sequence satisfying (C1)–(C3). Compactness of $\mathcal{Z}_\tau$ makes ω compact, and monotonicity together with (C3) makes Φ_τ constant on ω. Proposition 3 gives the KŁ property at each point of ω. Therefore Bolte et al. (2018, Appendix 6, Lemma 6.2) supplies one desingularizing function valid near the whole cluster set, as required by Lemma 3.

A.3 ONE-STEP SUFFICIENT DESCENT

For fixed D', the BGW loss is quadratic in π. Write $H_\lambda(D')$ for half the Hessian in π, $\mathcal{T}$ for the subspace of increments with vanishing row sums, and u_i for the per-row upper bound on $\widetilde{\mathcal{C}}_\tau$:

$$H_\lambda(D') = \tfrac{1}{2}\nabla^2_\pi F_\lambda(\pi, D'), \qquad \mathcal{T} = \{\Delta : \textstyle\sum_k \Delta_{ik} = 0\}, \qquad u_i = \mu_i - (K-1)\tau.$$

The relative curvature bound is the positive-part Rayleigh quotient

$$L_{\mathrm{rel}}(D', \lambda) = 2 \sup_{\Delta \in \mathcal{T}\setminus\{0\}} \frac{\big[\langle \Delta, H_\lambda(D')\Delta\rangle\big]_+}{\sum_{ik} \Delta_{ik}^2/u_i},$$

and

$$\overline{L}_{\mathrm{rel}}(\lambda) := \max_{D' \in \mathcal{D}_{\mathrm{full}}} L_{\mathrm{rel}}(D', \lambda).$$

Proposition 4 (Sufficient descent)**.** *Let D'^t be fixed and $L_t := L_{\mathrm{rel}}(D'^t, \lambda)$. If there exists ℓ with $\varepsilon > \ell \geq L_t$, then every mirror descent step satisfies*

$$F_\lambda(\pi^{t+1}, D'^t) \leq F_\lambda(\pi^t, D'^t) - (\varepsilon - \ell) D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t).$$

Define

$$u_i := \mu_i - (K-1)\tau, \qquad u_{\max} := \max_i u_i, \qquad a := \frac{\varepsilon - \ell}{2u_{\max}} > 0.$$

The same mirror descent step also satisfies condition (C1):

$$\Phi_\tau(\pi^{t+1}, D'^t) \leq \Phi_\tau(\pi^t, D'^t) - a\|\pi^{t+1} - \pi^t\|_F^2.$$

Here the two indicator values are zero because both iterates belong to $\widetilde{\mathcal{C}}_\tau$.

Proof of Proposition 4. The claim is one mirror descent step equation 9 at a fixed $D' = D'^t$. Write $F(\pi) := F_\lambda(\pi, D')$, $\Delta := \pi^{t+1} - \pi^t$, and $H := H_\lambda(D') = \frac{1}{2}\nabla^2_\pi F(\pi)$, the last Hessian being independent of the expansion point. Both π^t and π^{t+1} lie in $\widetilde{\mathcal{C}}_\tau$. If $\Delta = 0$, then $D_{\mathrm{KL}}(\pi^{t+1}\|\pi^t) = 0$ and both displayed inequalities hold as equalities. Assume henceforth $\Delta \neq 0$. If $K = 1$, then $\widetilde{\mathcal{C}}_\tau$ is a singleton and again $\Delta = 0$, so $K \geq 2$ below.

Step 1 (quadratic identity). The coupling equation 2 is bilinear in the two copies of π, and $-\lambda\|\pi\|_F^2$ is quadratic, so F is a polynomial of degree two in π. Taylor expansion of a quadratic polynomial terminates:

$$F(\pi^{t+1}) = F(\pi^t) + \langle \nabla F(\pi^t), \Delta\rangle + \Delta^\top H \Delta.$$

Each row of a point in $\widetilde{\mathcal{C}}_\tau$ sums to μ_i, hence $\sum_k \Delta_{ik} = 0$ for every i, that is $\Delta \in \mathcal{T}$.

Step 2 (KL integral remainder on $\widetilde{\mathcal{C}}_\tau$). Let $\omega(\pi) = \sum_{ik} \pi_{ik} \log \pi_{ik}$. Entrywise,

$$(\nabla\omega(\pi))_{ik} = \log \pi_{ik} + 1, \qquad (\nabla^2\omega(\pi)\Delta)_{ik} = \frac{\Delta_{ik}}{\pi_{ik}}.$$

The Bregman divergence of ω coincides with the KL displayed after equation 9:

$$\begin{aligned}\omega(\pi^{t+1}) - \omega(\pi^t) - \langle \nabla\omega(\pi^t), \Delta\rangle &= \sum_{ik}\Big(\pi_{ik}^{t+1}\log\pi_{ik}^{t+1} - \pi_{ik}^t \log \pi_{ik}^t - (\log\pi_{ik}^t + 1)\Delta_{ik}\Big) \\ &= \sum_{ik}\Big(\pi_{ik}^{t+1}\log\frac{\pi_{ik}^{t+1}}{\pi_{ik}^t} - \Delta_{ik}\Big) = D_{\mathrm{KL}}(\pi^{t+1}\|\pi^t).\end{aligned}$$

To identify the remainder, parameterize the segment by $\pi(s) := \pi^t + s\Delta$ for $s \in [0,1]$, so $\pi(0) = \pi^t$ and $\pi(1) = \pi^{t+1}$. The set $\widetilde{\mathcal{C}}_\tau$ is convex, hence $\pi(s) \in \widetilde{\mathcal{C}}_\tau$ and $\pi_{ik}(s) \geq \tau > 0$. Define the scalar restriction $\gamma(s) := \omega(\pi(s))$. Differentiating under the finite sum gives

$$\gamma'(s) = \sum_{ik}(\log \pi_{ik}(s) + 1)\Delta_{ik} = \langle \nabla\omega(\pi(s)), \Delta\rangle,$$

and a second differentiation gives

$$\gamma''(s) = \sum_{ik} \frac{\Delta_{ik}}{\pi_{ik}(s)} \Delta_{ik} = \sum_{ik} \frac{\Delta_{ik}^2}{\pi_{ik}(s)} = \langle \Delta, \nabla^2 \omega(\pi(s)) \Delta \rangle.$$

The fundamental theorem of calculus yields $\gamma(1) - \gamma(0) = \int_0^1 \gamma'(s)\, ds$. Writing $\gamma'(s) = \gamma'(0) + \int_0^s \gamma''(u)\, du$ and substituting,

$$\gamma(1) - \gamma(0) = \gamma'(0) + \int_0^1 \int_0^s \gamma''(u)\, du\, ds.$$

The region $0 \le u \le s \le 1$ is equivalently $0 \le u \le 1$ and $u \le s \le 1$, so

$$\int_0^1 \int_0^s \gamma''(u)\, du\, ds = \int_0^1 \gamma''(u) \Big(\int_u^1 ds \Big) du = \int_0^1 (1-u)\, \gamma''(u)\, du.$$

Renaming the dummy variable u to s and using $\gamma'(0) = \langle \nabla \omega(\pi^t), \Delta \rangle$ therefore gives

$$\omega(\pi^{t+1}) = \omega(\pi^t) + \langle \nabla \omega(\pi^t), \Delta \rangle + \int_0^1 (1-s)\, \gamma''(s)\, ds.$$

The left-hand side minus the first two terms on the right-hand side is exactly the Bregman divergence computed above, hence

$$D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t) = \int_0^1 (1-s) \sum_{ik} \frac{\Delta_{ik}^2}{\pi_{ik}(s)}\, ds.$$

For each row i and each column k,

$$\pi_{ik}(s) = \mu_i - \sum_{\ell \neq k} \pi_{i\ell}(s) \le \mu_i - (K-1)\tau =: u_i,$$

because the remaining $K-1$ entries are at least τ. Thus $1/\pi_{ik}(s) \ge 1/u_i$, and

$$\sum_{ik} \frac{\Delta_{ik}^2}{\pi_{ik}(s)} \ge \sum_{ik} \frac{\Delta_{ik}^2}{u_i}.$$

The weight $\int_0^1 (1-s)\, ds = \frac{1}{2}$ is independent of (i,k), so

$$D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t) \ge \Big(\sum_{ik} \frac{\Delta_{ik}^2}{u_i} \Big) \int_0^1 (1-s)\, ds = \frac{1}{2} \sum_{ik} \frac{\Delta_{ik}^2}{u_i}.$$

Step 3 (positive-part Rayleigh). By definition of $L_{\mathrm{rel}}(D', \lambda)$ as a supremum over $\mathcal{T} \setminus \{0\}$,

$$L_{\mathrm{rel}}(D', \lambda) \ge 2\, \frac{\big[\Delta^\top H \Delta\big]_+}{\sum_{ik} \Delta_{ik}^2 / u_i},$$

hence

$$\big[\Delta^\top H \Delta\big]_+ \le \frac{L_{\mathrm{rel}}(D', \lambda)}{2} \sum_{ik} \frac{\Delta_{ik}^2}{u_i}.$$

Step 2 rearranges to $\sum_{ik} \Delta_{ik}^2 / u_i \le 2\, D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t)$, and therefore

$$\big[\Delta^\top H \Delta\big]_+ \le L_{\mathrm{rel}}(D', \lambda)\, D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t).$$

If $\Delta^\top H \Delta \le 0$, then $\Delta^\top H \Delta \le 0 \le L_{\mathrm{rel}}(D', \lambda)\, D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t)$, since $L_{\mathrm{rel}} \ge 0$. If $\Delta^\top H \Delta > 0$, then $\Delta^\top H \Delta = \big[\Delta^\top H \Delta\big]_+$ and the same bound holds. In either case

$$\Delta^\top H \Delta \le L_{\mathrm{rel}}(D', \lambda)\, D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t).$$

After a D' update, $D' \in \mathcal{D}_{\mathrm{full}}$, so $L_{\mathrm{rel}}(D', \lambda) \le \overline{L}_{\mathrm{rel}}(\lambda)$. For the clustering specialization the same maximum is taken over $[m, M]^K$.

Step 4 (mirror optimality). By equation 9,

$$\pi^{t+1} \in \arg\min_{\pi \in \widetilde{\mathcal{C}}_\tau} \Big\{ \langle \nabla F(\pi^t), \pi - \pi^t \rangle + \varepsilon D_{\mathrm{KL}}(\pi \| \pi^t) \Big\}.$$

The comparison point π^t is feasible, and $D_{\mathrm{KL}}(\pi^t \| \pi^t) = 0$, so

$$\langle \nabla F(\pi^t), \Delta \rangle + \varepsilon D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t) \le 0.$$

Step 5 (sufficient decrease). Substitute Steps 3–4 into Step 1:

$$\begin{aligned} F(\pi^{t+1}) &= F(\pi^t) + \langle \nabla F(\pi^t), \Delta \rangle + \Delta^\top H \Delta \\ &\le F(\pi^t) + \langle \nabla F(\pi^t), \Delta \rangle + L_{\mathrm{rel}}(D', \lambda)\, D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t) \\ &\le F(\pi^t) - (\varepsilon - L_{\mathrm{rel}}(D', \lambda)) D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t). \end{aligned}$$

The assumption $\ell \ge L_t = L_{\mathrm{rel}}(D', \lambda)$ upgrades the coefficient to $\varepsilon - \ell$.

Step 6 ((C1) coefficient). Write $u_{\max} = \max_i u_i$. Then

$$\sum_{ik} \frac{\Delta_{ik}^2}{u_i} \ge \frac{1}{u_{\max}} \sum_{ik} \Delta_{ik}^2 = \frac{\|\Delta\|_F^2}{u_{\max}}.$$

Step 2 therefore gives $D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t) \ge \|\Delta\|_F^2 / (2u_{\max})$. Combined with Step 5,

$$F(\pi^t) - F(\pi^{t+1}) \ge a \|\Delta\|_F^2, \qquad a = \frac{\varepsilon - \ell}{2u_{\max}}.$$

On $\widetilde{\mathcal{C}}_\tau$ the indicator $\delta_{\widetilde{\mathcal{C}}_\tau}$ is finite at both iterates, so the same decrease holds for the fixed-D' slice $\Phi_\tau(\,\cdot\,, D')$. □

A.4 Lipschitz regularity of the update maps

Lemma 1 (Lipschitz update maps). *The exact prototype map h is Lipschitz from $\widetilde{\mathcal{C}}_\tau$ to $\mathcal{D}_{\mathrm{full}}$, and its clustering restriction is Lipschitz from $\widetilde{\mathcal{C}}_\tau$ to $\mathcal{D}_{\mathrm{clus}}$. For either choice of $\mathcal{D}$, the truncated mirror map*

$$\mathcal{M}_\tau : Z_\tau \to \widetilde{\mathcal{C}}_\tau$$

is Lipschitz.

Proof of Lemma 1. On $\widetilde{\mathcal{C}}_\tau$ one has $s_k = \sum_i \pi_{ik} \ge N\tau > 0$. The argument of $(\varphi')^{-1}$ in equation 6 is the ratio

$$\frac{\sum_{i,j} \pi_{ik} \pi_{jl} \varphi'(D_{ij})}{s_k s_l}.$$

The weights $\pi_{ik}\pi_{jl}/(s_k s_l)$ are nonnegative and sum to one, so the ratio lies in $[\varphi'(m), \varphi'(M)]$. Its denominator is at least $N^2\tau^2$, hence the ratio is C^1 on the compact convex set $\widetilde{\mathcal{C}}_\tau$. Standing strong convexity gives $\varphi'' \ge \kappa_\varphi > 0$ on $[m, M]$, so $(\varphi')^{-1}$ is C^1 on $[\varphi'(m), \varphi'(M)]$. A C^1 map on a compact convex set is Lipschitz. Formula equation 7 is the same composition with $k = l$. Write L_h for a Lipschitz constant of h on $\widetilde{\mathcal{C}}_\tau$, for either choice of $\mathcal{D}$.

It remains to prove that $\mathcal{M} = \mathcal{M}_\tau$ is Lipschitz on Z_τ. Let $z = (\pi, D')$ and $\widehat{z} = (\widehat{\pi}, \widehat{D}')$ lie in Z_τ, and write $x = \mathcal{M}(z)$, $\widehat{x} = \mathcal{M}(\widehat{z})$. The truncated mirror step is the unique minimizer of

$$Q(x; z) = \big\langle \nabla_\pi F_\lambda(\pi, D'), x - \pi \big\rangle + \varepsilon D_{\mathrm{KL}}(x \| \pi)$$

over the hard-floor simplex $\widetilde{\mathcal{C}}_\tau$. As recorded above, every coordinate of a feasible plan on $\widetilde{\mathcal{C}}_\tau$ satisfies $\tau \le x_{ik} \le \mu_i - (K-1)\tau$. Write $u_i := \mu_i - (K-1)\tau$ as in Section 3 and $u_{\max} := \max_i u_i$ as in the proof of Proposition 4. Then every coordinate of a feasible x satisfies $x_{ik} \le u_{\max}$.

The first term of $Q(\,\cdot\,; z)$ is affine in x. Write $\omega(\pi) = \sum_{ik} \pi_{ik} \log \pi_{ik}$, so $\nabla\omega(\pi) = \log \pi + \mathbf{1}$ and $\nabla\omega(x) - \nabla\omega(\pi) = \log(x/\pi)$. This is the Bregman geometry of the mirror step equation 9. Differentiating Q in x therefore gives the entrywise-log gradient

$$\nabla_x Q(x; z) = \nabla_\pi F_\lambda(\pi, D') + \varepsilon \log(x/\pi)$$

and the Hessian

$$\nabla_x^2 Q(x;z) = \varepsilon \operatorname{diag}\left(\frac{1}{x_{ik}}\right).$$

For every direction $V \in \mathbb{R}^{N\times K}$,

$$\left\langle V, \nabla_x^2 Q(x;z)\, V \right\rangle = \varepsilon \sum_{ik} \frac{V_{ik}^2}{x_{ik}} \geq \frac{\varepsilon}{u_{\max}} \|V\|_F^2,$$

because $x_{ik} \leq u_{\max}$. The map $Q(\,\cdot\,;z)$ is C^2 on a neighborhood of $\widetilde{\mathcal{C}}_\tau$. For $x, \widehat{x} \in \widetilde{\mathcal{C}}_\tau$ the fundamental theorem of calculus gives

$$\nabla_x Q(\widehat{x};z) - \nabla_x Q(x;z) = \int_0^1 \nabla_x^2 Q\big(x + t(\widehat{x}-x);z\big)(\widehat{x}-x)\, dt.$$

The segment $x + t(\widehat{x}-x)$ remains in the convex set $\widetilde{\mathcal{C}}_\tau$, so every entry is at most $u_{\max}$. Taking the Frobenius inner product with $\widehat{x}-x$ therefore yields

$$\left\langle \nabla_x Q(\widehat{x};z) - \nabla_x Q(x;z), \widehat{x}-x \right\rangle \geq \frac{\varepsilon}{u_{\max}} \|\widehat{x}-x\|_F^2.$$

Hence $Q(\,\cdot\,;z)$ is uniformly $\varepsilon/u_{\max}$-strongly convex in the Frobenius norm, with a modulus that does not depend on z. The set $\widetilde{\mathcal{C}}_\tau$ is compact and convex, so the minimizer $x = \mathcal{M}(z)$ exists and is unique. Because $Q(\,\cdot\,;z)$ is continuously differentiable and $\widetilde{\mathcal{C}}_\tau$ is closed and convex, this minimizer is characterized by the first-order variational inequality

$$\left\langle \nabla_x Q(x;z), y - x \right\rangle \geq 0 \qquad \text{for every } y \in \widetilde{\mathcal{C}}_\tau.$$

The same characterization at $\widehat{z}$ gives

$$\left\langle \nabla_x Q(\widehat{x};\widehat{z}), y - \widehat{x} \right\rangle \geq 0 \qquad \text{for every } y \in \widetilde{\mathcal{C}}_\tau.$$

Substitute $y = \widehat{x}$ in the first inequality and $y = x$ in the second:

$$\left\langle \nabla_x Q(x;z), \widehat{x}-x \right\rangle \geq 0, \qquad \left\langle \nabla_x Q(\widehat{x};\widehat{z}), x-\widehat{x} \right\rangle \geq 0.$$

Adding these two displays yields

$$\left\langle \nabla_x Q(\widehat{x};\widehat{z}) - \nabla_x Q(x;z), x-\widehat{x} \right\rangle \geq 0.$$

Split the first gradient by inserting $\nabla_x Q(\widehat{x};z)$:

$$\begin{aligned}\big\langle \nabla_x Q(\widehat{x};\widehat{z}) - \nabla_x Q(\widehat{x};z), x-\widehat{x} \big\rangle \\ + \big\langle \nabla_x Q(\widehat{x};z) - \nabla_x Q(x;z), x-\widehat{x} \big\rangle \geq 0.\end{aligned}$$

The strong-convexity inner-product bound already established therefore gives

$$\left\langle \nabla_x Q(\widehat{x};z) - \nabla_x Q(x;z), x-\widehat{x} \right\rangle \leq -\frac{\varepsilon}{u_{\max}} \|x-\widehat{x}\|_F^2.$$

Substitute this bound into the split variational inequality:

$$\left\langle \nabla_x Q(\widehat{x};\widehat{z}) - \nabla_x Q(\widehat{x};z), x-\widehat{x} \right\rangle \geq \frac{\varepsilon}{u_{\max}} \|x-\widehat{x}\|_F^2.$$

The Cauchy–Schwarz inequality on the Frobenius inner product gives

$$\big\|\nabla_x Q(\widehat{x};\widehat{z}) - \nabla_x Q(\widehat{x};z)\big\|_F \|x-\widehat{x}\|_F \geq \frac{\varepsilon}{u_{\max}} \|x-\widehat{x}\|_F^2.$$

If $x = \widehat{x}$ the next display is immediate. If $x \neq \widehat{x}$, divide by $\|x-\widehat{x}\|_F$. In either case

$$\frac{\varepsilon}{u_{\max}} \|x-\widehat{x}\|_F \leq \|\nabla_x Q(\widehat{x};\widehat{z}) - \nabla_x Q(\widehat{x};z)\|_F.$$

The gradient formula already recorded therefore gives

$$\begin{aligned}\nabla_x Q(\widehat{x};\widehat{z}) - \nabla_x Q(\widehat{x};z) &= \nabla_\pi F_\lambda(\widehat{\pi}, \widehat{D}') + \varepsilon \log(\widehat{x}/\widehat{\pi}) \\ &\quad - \nabla_\pi F_\lambda(\pi, D') - \varepsilon \log(\widehat{x}/\pi) \\ &= \nabla_\pi F_\lambda(\widehat{z}) - \nabla_\pi F_\lambda(z) + \varepsilon\big(\log\pi - \log\widehat{\pi}\big),\end{aligned}$$

because $\log(\widehat{x}/\widehat{\pi}) - \log(\widehat{x}/\pi) = \log\widehat{x} - \log\widehat{\pi} - \log\widehat{x} + \log\pi = \log\pi - \log\widehat{\pi}$. The triangle inequality yields

$$\begin{aligned}&\big\|\nabla_x Q(\widehat{x};\widehat{z}) - \nabla_x Q(\widehat{x};z)\big\|_F \\ &\quad\le \big\|\nabla_\pi F_\lambda(z) - \nabla_\pi F_\lambda(\widehat{z})\big\|_F + \varepsilon\|\log\pi - \log\widehat{\pi}\|_F.\end{aligned}$$

On $\widetilde{\mathcal{C}}_\tau$ one has $\pi_{ik} \ge \tau$ and $\widehat{\pi}_{ik} \ge \tau$. For each coordinate (i,k), the mean value theorem supplies a point ξ_{ik} between π_{ik} and $\widehat{\pi}_{ik}$ with $\xi_{ik} \ge \tau$ such that

$$\big|\log\pi_{ik} - \log\widehat{\pi}_{ik}\big| = \frac{|\pi_{ik} - \widehat{\pi}_{ik}|}{\xi_{ik}} \le \tau^{-1}|\pi_{ik} - \widehat{\pi}_{ik}|.$$

Squaring and summing over (i,k) gives

$$\|\log\pi - \log\widehat{\pi}\|_F \le \tau^{-1}\|\pi - \widehat{\pi}\|_F.$$

Appendix A.2 records that F_λ is analytic on a neighborhood of the compact set Z_τ, so $\nabla_\pi F_\lambda$ is C^1 on a neighborhood of Z_τ and therefore Lipschitz on Z_τ. Let L_F be a Lipschitz constant of $\nabla_\pi F_\lambda$ on Z_τ with respect to the product norm $\|z\|_\times$ on the domain and the Frobenius norm on the codomain:

$$\big\|\nabla_\pi F_\lambda(z) - \nabla_\pi F_\lambda(\widehat{z})\big\|_F \le L_F\|z - \widehat{z}\|_\times.$$

The definition of $\|\cdot\|_\times$ also gives $\|\pi - \widehat{\pi}\|_F \le \|z - \widehat{z}\|_\times$. Combining the preceding displays,

$$\begin{aligned}&\big\|\nabla_x Q(\widehat{x};\widehat{z}) - \nabla_x Q(\widehat{x};z)\big\|_F \\ &\quad\le L_F\|z - \widehat{z}\|_\times + \frac{\varepsilon}{\tau}\|\pi - \widehat{\pi}\|_F \\ &\quad\le \Big(L_F + \frac{\varepsilon}{\tau}\Big)\|z - \widehat{z}\|_\times.\end{aligned}$$

Therefore

$$\|x - \widehat{x}\|_F \le \frac{u_{\max}}{\varepsilon}\Big(L_F + \frac{\varepsilon}{\tau}\Big)\|z - \widehat{z}\|_\times.$$

Set

$$L_{\mathcal{M}} = \frac{u_{\max}}{\varepsilon}\Big(L_F + \frac{\varepsilon}{\tau}\Big).$$

Then $\|\mathcal{M}(z) - \mathcal{M}(\widehat{z})\|_F \le L_{\mathcal{M}}\|z - \widehat{z}\|_\times$, so $\mathcal{M}_\tau : Z_\tau \to \widetilde{\mathcal{C}}_\tau$ is Lipschitz. □

A.5 Relative error

Lemma 2 (Relative error)**.** *After an mirror descent step with fixed D', the mirror stationarity identity produces $w_\pi^{t+1} \in \nabla_\pi F_\lambda(\pi^{t+1}, D') + N_{\widetilde{\mathcal{C}}_\tau}(\pi^{t+1})$ satisfying*

$$\|w_\pi^{t+1}\|_F \le \big(L_F + \varepsilon/\tau\big)\|\pi^{t+1} - \pi^t\|_F,$$

where L_F is a Lipschitz constant of $\nabla_\pi F_\lambda(\,\cdot\,, D')$. After an alternating block, the same construction at the last inner step, together with Lemma 1, yields $w_z^{t+1} \in \partial\Phi_\tau(z^{t+1})$ with $\|w_z^{t+1}\|_F \le b\,\delta_t$ and $\delta_t = \sum_s \|\pi^{t,s+1} - \pi^{t,s}\|_F$.

Proof of Lemma 2. **Step 1 (fixed D').** The mirror descent step equation 9 is the unconstrained stationarity condition on the relative interior of the active face, equivalently a normal-cone inclusion on $\widetilde{\mathcal{C}}_\tau$: there exists $\xi^{t+1} \in N_{\widetilde{\mathcal{C}}_\tau}(\pi^{t+1})$ such that

$$0 = \nabla_\pi F_\lambda(\pi^t, D') + \varepsilon\big(\nabla\omega(\pi^{t+1}) - \nabla\omega(\pi^t)\big) + \xi^{t+1}, \qquad \omega(\pi) = \sum_{ik}\pi_{ik}\log\pi_{ik}.$$

Set $w_\pi^{t+1} := \nabla_\pi F_\lambda(\pi^{t+1}, D') + \xi^{t+1}$. Then $w_\pi^{t+1} \in \nabla_\pi F_\lambda(\pi^{t+1}, D') + N_{\widetilde{\mathcal{C}}_\tau}(\pi^{t+1})$, and substituting the stationarity identity gives

$$w_\pi^{t+1} = \nabla_\pi F_\lambda(\pi^{t+1}, D') - \nabla_\pi F_\lambda(\pi^t, D') - \varepsilon\big(\nabla\omega(\pi^{t+1}) - \nabla\omega(\pi^t)\big).$$

On the compact convex set $\widetilde{\mathcal{C}}_\tau$, the gradient $\nabla_\pi F_\lambda(\,\cdot\,, D')$ is Lipschitz with some constant $L_F < \infty$ because F_λ is a quadratic polynomial in π. The entropy gradient has entries $(\nabla\omega(\pi))_{ik} = \log \pi_{ik} + 1$. For $\pi, \pi' \in \widetilde{\mathcal{C}}_\tau$,

$$\big|\log \pi'_{ik} - \log \pi_{ik}\big| = \Big|\int_{\pi_{ik}}^{\pi'_{ik}} \frac{1}{s}\, ds\Big| \le \tau^{-1}\big|\pi'_{ik} - \pi_{ik}\big|,$$

hence $\|\nabla\omega(\pi') - \nabla\omega(\pi)\|_F \le \tau^{-1}\|\pi' - \pi\|_F$. The triangle inequality therefore yields

$$\|w_\pi^{t+1}\|_F \le \big(L_F + \varepsilon/\tau\big)\|\pi^{t+1} - \pi^t\|_F.$$

In particular $\operatorname{dist}\big(0, \nabla_\pi F_\lambda(\pi^{t+1}, D') + N_{\widetilde{\mathcal{C}}_\tau}(\pi^{t+1})\big) \le \|w_\pi^{t+1}\|_F$.

Step 2 (alternating block). Write the same identity at the last inner step $\pi^{t,n_{\text{inner}}-1} \to \pi^{t+1}$ with fixed D'^t, obtaining $n_\pi^{t+1} \in N_{\widetilde{\mathcal{C}}_\tau}(\pi^{t+1})$ and

$$\begin{aligned} w_\pi^{t+1} &:= \nabla_\pi F_\lambda(\pi^{t+1}, D'^{t+1}) + n_\pi^{t+1} \\ &= \nabla_\pi F_\lambda(\pi^{t+1}, D'^{t+1}) - \nabla_\pi F_\lambda(\pi^{t,n_{\text{inner}}-1}, D'^t) \\ &\quad - \varepsilon\big(\nabla\omega(\pi^{t+1}) - \nabla\omega(\pi^{t,n_{\text{inner}}-1})\big). \end{aligned}$$

Let L_F now be a joint Lipschitz constant of ∇F_λ on the compact product. Lemma 1 gives $\|D'^{t+1} - D'^t\|_F = \|h(\pi^{t+1}) - h(\pi^t)\|_F \le L_h\|\pi^{t+1} - \pi^t\|_F \le L_h\delta_t$, and the last inner displacement satisfies $\|\pi^{t+1} - \pi^{t,n_{\text{inner}}-1}\|_F \le \delta_t$. Hence

$$\big\|\nabla_\pi F_\lambda(\pi^{t+1}, D'^{t+1}) - \nabla_\pi F_\lambda(\pi^{t,n_{\text{inner}}-1}, D'^t)\big\|_F \le L_F(1 + L_h)\delta_t,$$

while the entropy term is at most $(\varepsilon/\tau)\delta_t$. The D' block of the subgradient is zero: the D' update satisfies $0 \in \nabla_{D'} F_\lambda(\pi^{t+1}, D'^{t+1}) + N_{\mathcal{D}_{\text{full}}}(D'^{t+1})$. The joint subgradient is therefore $w_z^{t+1} := (w_\pi^{t+1}, 0) \in \partial\Phi_\tau(z^{t+1})$, and $\|w_z^{t+1}\|_F \le \big(L_F(1 + L_h) + \varepsilon/\tau\big)\delta_t$. This is the joint relative-error bound in Lemma 2. □

A.6 Abstract finite-length whole-sequence lemma

Lemma 3 (Finite-length whole-sequence convergence from KŁ descent)**.** *Let G be proper and lower semicontinuous, and let $\{x^t\}$ be a bounded sequence in $\operatorname{dom} G$. If $\{x^t\}$ is a gradient-like descent sequence—assumptions (C1)–(C3) below—and G has the KŁ property, then*

$$\sum_{t=0}^{\infty} \|x^{t+1} - x^t\| < \infty,$$

the sequence converges to a single point $x^\star$, and $0 \in \partial G(x^\star)$.

The three assumptions are those of Bolte et al. (2018, Appendix 6, Definition 6.1). (C1) sufficient decrease: there exists $a > 0$ such that $G(x^t) - G(x^{t+1}) \ge a\|x^{t+1} - x^t\|^2$. (C2) relative error: there exist $b > 0$ and $w^{t+1} \in \partial G(x^{t+1})$ with $\|w^{t+1}\| \le b\|x^{t+1} - x^t\|$, hence $\operatorname{dist}(0, \partial G(x^{t+1})) \le b\|x^{t+1} - x^t\|$. (C3) if a subsequence $x^{t_j} \to \bar{x}$, then $\limsup_j G(x^{t_j}) \le G(\bar{x})$.

Proof of Lemma 3. The statement is Bolte et al. (2018, Appendix 6, Theorem 6.2). The algebra below is the telescoping used by the two BGW applications.

Let $r_t = G(x^t) - G^\star$ and $s_t = \|x^{t+1} - x^t\|$, where $G^\star$ is the common limiting value on the cluster set. Boundedness, (C1), lower semicontinuity of G, and the original (C3) $\limsup_j G(x^{t_j}) \le G(\bar{x})$ of Bolte et al. (2018, Appendix 6, Definition 6.1) give value convergence along every cluster subsequence, hence a common limit on the cluster set. Condition (C1) gives $r_t - r_{t+1} \ge as_t^2$ for a fixed $a > 0$, so (r_t) is nonincreasing and bounded below, hence convergent, and $\sum_t s_t^2 < \infty$. In particular $s_t \to 0$. If $s_{t_0} = 0$ for some t_0, then $x^{t_0+1} = x^{t_0}$. Condition (C2) at this step yields $w^{t_0+1} \in \partial G(x^{t_0+1})$ with $\|w^{t_0+1}\| \le bs_{t_0} = 0$, hence $0 \in \partial G(x^{t_0})$. From this index onward we set $x^t = x^{t_0}$ (zero further displacement), which is compatible with (C1)–(C3) and makes the sequence eventually constant. The length sum is then a finite sum and the claim holds. Assume henceforth $s_t > 0$ for every t.

After the iterates enter a uniformized KŁ neighborhood of the cluster set (Bolte et al., 2018, Appendix 6, Lemma 6.2), Definition 1 at x^t reads

$$\psi'(r_t)\,\operatorname{dist}\big(0,\partial G(x^t)\big)\geq 1.$$

Condition (C2) at the incoming step $x^{t-1}\to x^t$ gives $\operatorname{dist}(0,\partial G(x^t))\leq bs_{t-1}$, therefore

$$\psi'(r_t)\geq\frac{1}{bs_{t-1}}.$$

Concavity of ψ on $[0,\eta)$ is the first-order inequality $\psi(r_{t+1})\leq\psi(r_t)+\psi'(r_t)(r_{t+1}-r_t)$. Rearrangement and (C1) yield

$$\psi(r_t)-\psi(r_{t+1})\geq\psi'(r_t)(r_t-r_{t+1})\geq a\psi'(r_t)s_t^2\geq\frac{a}{bs_{t-1}}s_t^2.$$

With $C=b/a$, this is $s_t^2\leq Cs_{t-1}(\psi(r_t)-\psi(r_{t+1}))$. The elementary inequality $2\sqrt{uv}\leq u+v$ for $u,v\geq 0$, applied to $u=s_{t-1}$ and $v=C(\psi(r_t)-\psi(r_{t+1}))$, together with $s_t\leq\sqrt{uv}$, produces

$$2s_t\leq s_{t-1}+C\big(\psi(r_t)-\psi(r_{t+1})\big).$$

Summing from $t=\ell+1$ to T and using $\psi\geq 0$ gives

$$\begin{aligned}2\sum_{t=\ell+1}^{T}s_t&\leq\sum_{t=\ell+1}^{T}s_{t-1}+C\big(\psi(r_{\ell+1})-\psi(r_{T+1})\big)\\&\leq s_\ell+\sum_{t=\ell+1}^{T-1}s_t+C\psi(r_{\ell+1}).\end{aligned}$$

The overlapping sums cancel to

$$\sum_{t=\ell+1}^{T}s_t\leq s_\ell+C\psi(r_{\ell+1}).$$

The right-hand side is independent of T, so $\sum_t s_t<\infty$. Partial sums of a convergent series of nonnegative terms form a Cauchy sequence, and $s_t=\|x^{t+1}-x^t\|$, hence $\{x^t\}$ is Cauchy and $x^t\to x^\star$ for a single $x^\star\in\operatorname{dom}G$. Condition (C2) forces $w^{t+1}\to 0$ with $w^{t+1}\in\partial G(x^{t+1})$, and (C3) gives $G(x^t)\to G(x^\star)$. Closedness of the limiting subdifferential along a sequence with convergent values yields $0\in\partial G(x^\star)$. □

Block-control form. The same telescoping argument applies when a nonnegative block control s_t replaces the direct iterate displacement in the descent and relative-error bounds, provided

$$\|x^{t+1}-x^t\|\leq c\,s_t$$

for a fixed $c<\infty$. The proof then gives $\sum_t s_t<\infty$, and hence finite length of the actual iterates.

A.7 Inner mirror descent at fixed D'

Theorem 3 (Inner mirror descent at fixed D')**.** *Let D' be feasible and fixed, and let π^{t+1} be the mirror descent step equation 9 on $\widetilde{\mathcal{C}}_\tau$. If there exists ℓ such that $\varepsilon>\ell\geq\sup_{t\geq 0}L_t$, then $\sum_t\|\pi^{t+1}-\pi^t\|_F<\infty$ and the sequence converges to a point $\pi^\star\in\widetilde{\mathcal{C}}_\tau$ satisfying $0\in\nabla_\pi F_\lambda(\pi^\star,D')+N_{\widetilde{\mathcal{C}}_\tau}(\pi^\star)$.*

This is Lemma 3 applied to the fixed-D' slice of Φ_τ. The Step numbering follows (C1), (C2), then (C3).

Proof of Theorem 3. Fix a feasible D'. The slice

$$\pi\mapsto\Phi_\tau(\pi,D')=F_\lambda(\pi,D')+\delta_{\widetilde{\mathcal{C}}_\tau}(\pi)+\delta_{\mathcal{D}}(D')$$

differs from $F_\lambda(\,\cdot\,,D')$ by the constant $\delta_{\mathcal{D}}(D')$. Assume there exists ℓ with $\varepsilon>\ell\geq\sup_{t\geq 0}L_t$. Since D' is fixed, $L_t=L_{\mathrm{rel}}(D',\lambda)$ for every t.

Step 1 (C1). Proposition 4 gives, writing $\Delta^t = \pi^{t+1} - \pi^t$,

$$\Phi_\tau(\pi^t, D') - \Phi_\tau(\pi^{t+1}, D') \geq (\varepsilon - \ell) D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t) \geq a \|\Delta^t\|_F^2, \tag{13}$$

where $a = (\varepsilon - \ell)/(2u_{\max}) > 0$.

Step 2 (C2). The fixed-D' case of Lemma 2 produces $w_\pi^{t+1} \in \nabla_\pi F_\lambda(\pi^{t+1}, D') + N_{\widetilde{\mathcal{C}}_\tau}(\pi^{t+1})$ with

$$\operatorname{dist}\big(0, \nabla_\pi F_\lambda(\pi^{t+1}, D') + N_{\widetilde{\mathcal{C}}_\tau}(\pi^{t+1})\big) \leq \|w_\pi^{t+1}\|_F \leq b \|\Delta^t\|_F, \qquad b := L_F + \varepsilon/\tau. \tag{14}$$

Step 3 (C3 and conclusion). Every iterate remains in the compact set $\widetilde{\mathcal{C}}_\tau$, on which $F_\lambda(\,\cdot\,, D')$ is continuous. If $\pi^{t_j} \to \bar{\pi} \in \widetilde{\mathcal{C}}_\tau$, then $\Phi_\tau(\pi^{t_j}, D') \to \Phi_\tau(\bar{\pi}, D')$, which is (C3). Proposition 3 gives the KŁ property of every fixed-D' slice. Lemma 3 therefore yields finite-length whole-sequence convergence to a unique limit $\pi^\star \in \widetilde{\mathcal{C}}_\tau$, and $0 \in \nabla_\pi F_\lambda(\pi^\star, D') + N_{\widetilde{\mathcal{C}}_\tau}(\pi^\star)$. □

B Alternating updates

Theorem 1 is Lemma 3 applied to the outer blocks $z^t = (\pi^t, D'^t)$ of Φ_τ. The inner-block (C1) identity is Proposition 4; the last-inner (C2) identity is Lemma 2. This section records the instance of those identities at the outer block. The Step numbering follows (C1), (C2), then (C3).

Proof of Theorem 1. Assume λ is held constant and there exists ℓ with $\varepsilon > \ell \geq \sup_{t \geq 0} L_{\mathrm{rel}}(D'^t, \lambda)$. Write h for the D' map of Lemma 1, and fix an integer $n_{\mathrm{inner}} \geq 1$. An outer block holds D'^t fixed for n_{inner} mirror descent steps, then applies the D' update $D'^{t+1} = h(\pi^{t+1})$. Write $\pi^{t,0} = \pi^t$, $\pi^{t,n_{\mathrm{inner}}} = \pi^{t+1}$, $z^t = (\pi^t, D'^t)$, and

$$\delta_t := \sum_{s=0}^{n_{\mathrm{inner}}-1} \|\pi^{t,s+1} - \pi^{t,s}\|_F.$$

Apply the block-control form of Lemma 3 with $G = \Phi_\tau$, $x^t = z^t$, and scalars $s_t = \delta_t$. Lemma 1 gives $\|z^{t+1} - z^t\|_\times \leq (1 + L_h)\delta_t$. The telescoping in Appendix A.6 uses only these scalars.

Step 1 (C1). Sum the fixed-D' decrease of Proposition 4 over the inner indices $s = 0, \ldots, n_{\mathrm{inner}} - 1$. Cauchy–Schwarz gives $\delta_t^2 \leq n_{\mathrm{inner}} \sum_s \|\pi^{t,s+1} - \pi^{t,s}\|_F^2$. The D' update is the closed-form minimizer of the D'-block, so it does not increase F_λ. On $\mathcal{Z}_\tau$ the indicators are finite, hence

$$\Phi_\tau(z^t) - \Phi_\tau(z^{t+1}) \geq a \delta_t^2, \qquad a := \frac{\varepsilon - \ell}{2u_{\max} n_{\mathrm{inner}}} > 0. \tag{15}$$

Lemma 1 also gives $\|D'^{t+1} - D'^t\|_F \leq L_h \delta_t$. Combined with $\|z^{t+1} - z^t\|_\times = \|\pi^{t+1} - \pi^t\|_F + \|D'^{t+1} - D'^t\|_F$ and $\|\pi^{t+1} - \pi^t\|_F \leq \delta_t$,

$$\|z^{t+1} - z^t\|_\times \leq \delta_t + L_h \delta_t = (1 + L_h)\, \delta_t.$$

Step 2 (C2). The last-inner case of Lemma 2 produces $w_z^{t+1} \in \partial \Phi_\tau(z^{t+1})$ with

$$\operatorname{dist}\big(0, \partial \Phi_\tau(z^{t+1})\big) \leq b \delta_t, \qquad b := L_F(1 + L_h) + \frac{\varepsilon}{\tau}. \tag{16}$$

Step 3 (C3 and conclusion). Every outer iterate remains in the compact product $\widetilde{\mathcal{C}}_\tau \times \mathcal{D}_{\mathrm{full}}$, on which F_λ is continuous. Thus if $z^{t_j} \to \bar{z}$ in that product, then $\Phi_\tau(z^{t_j}) \to \Phi_\tau(\bar{z})$, which is (C3). Proposition 3 gives the KŁ property of Φ_τ. Lemma 3 with the scalars $s_t = \delta_t$ and the constants (a, b) from equation 15–equation 16 therefore yields $\sum_t \delta_t < \infty$, a unique limit $z^\star \in \mathcal{Z}_\tau$, and $0 \in \partial \Phi_\tau(z^\star)$. The Lipschitz bound on h upgrades $\sum_t \delta_t < \infty$ to

$$\sum_t \|z^{t+1} - z^t\|_\times < \infty.$$

This proves Theorem 1 for every fixed $n_{\mathrm{inner}} \geq 1$. □

Proof of Corollary 1. **Step 1 (telescope sufficient decrease).** Summing equation 15 from $t = 0$ to $T - 1$ and using the definition of Δ_Φ gives

$$a \sum_{t=0}^{T-1} \delta_t^2 \le \Phi_\tau(z^0) - \Phi_\tau(z^T) \le \Delta_\Phi.$$

Step 2 (select the smallest residual). Let $r_{t+1} := \operatorname{dist}(0, \partial\Phi_\tau(z^{t+1}))$. equation 16 gives $r_{t+1}^2 \le b^2\delta_t^2$. Therefore

$$\min_{0\le t<T} r_{t+1}^2 \le \frac{1}{T}\sum_{t=0}^{T-1} r_{t+1}^2 \le \frac{b^2}{T}\sum_{t=0}^{T-1}\delta_t^2 \le \frac{b^2\Delta_\Phi}{aT}.$$

Step 3 (substitute the block constant). Using $a = (\varepsilon - \ell)/(2u_{\max}n_{\text{inner}})$ from equation 15 and taking square roots yields the stated bound. □

B.1 Clustering restriction of D'

If D' is restricted to $D'_{kk} = d_k$ and $D'_{kl} = \delta_\varphi(D)$ for $k \ne l$, the D' map is the diagonal map of Lemma 1, with values in $[m, M]^K$. Substituting this Lipschitz map for the map h in Steps 1–3 leaves the same pair of constants (a, b) on $\widetilde{\mathcal{C}}_\tau \times [m, M]^K$. Lemma 3 therefore applies, which is the main-text corollary. For the clustering restriction we identify the prototype domain with its free diagonal coordinates $d \in [m, M]^K$. The off-diagonal values $\delta_\varphi(D)$ are fixed finite constants; analyticity is required only with respect to the free prototype coordinates.

B.2 Stage II directional and precomputable bounds

Proposition 4 turns a uniform bound $\varepsilon > \ell \ge L_{\text{rel}}(D'^t, \lambda)$ into (C1) for one mirror descent step, through the one-sided Kullback–Leibler remainder. The next statement reads the same (C1) from the realized increment $\Delta^t = \pi^{t+1} - \pi^t$, without taking a supremum over the tangent space $\mathcal{T}$. The statement after that bounds the realized ratio by the Rayleigh maximum on the box, which produces one sufficient choice of ε that can be computed before the path is run. After an clustering D' update, D' lies in $[m, M]^K$.

Recall condition (C1) of Lemma 3 on a slice $\pi \mapsto \Phi_\tau(\pi, D')$: there exists $a > 0$ such that

$$\Phi_\tau(\pi^t, D') - \Phi_\tau(\pi^{t+1}, D') \ge a\|\pi^{t+1} - \pi^t\|_F^2.$$

On $\widetilde{\mathcal{C}}_\tau$ both indicators in Φ_τ vanish, so the left-hand side equals $F_\lambda(\pi^t, D') - F_\lambda(\pi^{t+1}, D')$. Write $u_i = \mu_i - (K-1)\tau$ and $u_{\max} = \max_i u_i$, as in Section 3.

Proposition 5 (Pathwise (C1) at one mirror descent step)**.** *Hold $\lambda = \lambda_{\text{II}}$ fixed, and let D'^t be given by a D' update. Let π^{t+1} be one mirror descent step equation 9 on $\widetilde{\mathcal{C}}_\tau$ at this frozen D'^t, with inverse step size $\varepsilon > 0$. Write*

$$\Delta^t = \pi^{t+1} - \pi^t, \qquad H = H_{\lambda_{\text{II}}}(D'^t) = \tfrac{1}{2}\nabla_\pi^2 F_{\lambda_{\text{II}}}(\pi, D'^t),$$

the Hessian being independent of the expansion point because $F_{\lambda_{\text{II}}}(\,\cdot\,, D'^t)$ is quadratic in π. Write the Jeffreys divergence

$$J_t = D_{\text{KL}}(\pi^{t+1}\|\pi^t) + D_{\text{KL}}(\pi^t\|\pi^{t+1}).$$

Set $\ell_t = 0$ if $\Delta^t = 0$, and

$$\ell_t = \frac{\big[\langle \Delta^t, H\Delta^t\rangle\big]_+}{J_t}$$

if $\Delta^t \ne 0$. If $\ell_t < \varepsilon$, this mirror descent step satisfies (C1) on the slice $\pi \mapsto \Phi_\tau(\pi, D'^t)$ with

$$a_t = \frac{\varepsilon - \ell_t}{u_{\max}} > 0.$$

If a Stage II path has $\bar{\ell} := \sup_t \ell_t < \varepsilon$, then every such step satisfies (C1) with the same coefficient $a = (\varepsilon - \bar{\ell})/u_{\max}$.

On a recorded Stage II budget of T outer blocks, write $\overline{\ell}_{\mathrm{II},T}$ for the maximum of ℓ_t over the generated inner mirror descent steps, and $R_{\mathrm{II},T} = \overline{\ell}_{\mathrm{II},T}/\varepsilon$. Then $R_{\mathrm{II},T} < 1$ is $\overline{\ell}_{\mathrm{II},T} < \varepsilon$: it is a posteriori / pathwise verification of sufficient decrease on that budget.

Proof. Fix one mirror descent step at frozen D'^t. Write $F(\pi) := F_{\lambda_{\mathrm{II}}}(\pi, D'^t)$ and $g^t = \nabla F(\pi^t)$. Both π^t and π^{t+1} lie in $\widetilde{\mathcal{C}}_\tau$. If $\Delta^t = 0$, then $\ell_t = 0$, $J_t = 0$, and both sides of (C1) vanish, so the claim holds. Assume $\Delta^t \neq 0$ for the rest of the argument, hence $J_t > 0$.

Step 1 (quadratic identity). As in Step 1 of Proposition 4, F is a polynomial of degree two in π, so the Taylor expansion terminates:

$$F(\pi^{t+1}) - F(\pi^t) = \langle g^t, \Delta^t \rangle + \langle \Delta^t, H\Delta^t \rangle.$$

Step 2 (Jeffreys identity). Let $\omega(\pi) = \sum_{ik} \pi_{ik} \log \pi_{ik}$, so $(\nabla\omega(\pi))_{ik} = \log \pi_{ik} + 1$. The Bregman divergence of ω on $\widetilde{\mathcal{C}}_\tau$ is the Kullback–Leibler divergence of Proposition 4:

$$D_{\mathrm{KL}}(\pi^{t+1} \| \pi^t) = \omega(\pi^{t+1}) - \omega(\pi^t) - \langle \nabla\omega(\pi^t), \Delta^t \rangle,$$
$$D_{\mathrm{KL}}(\pi^t \| \pi^{t+1}) = \omega(\pi^t) - \omega(\pi^{t+1}) - \langle \nabla\omega(\pi^{t+1}), \pi^t - \pi^{t+1} \rangle.$$

Adding these two identities cancels $\omega(\pi^{t+1}) - \omega(\pi^t)$ and uses $\pi^t - \pi^{t+1} = -\Delta^t$, hence

$$J_t = \langle \nabla\omega(\pi^{t+1}) - \nabla\omega(\pi^t), \Delta^t \rangle.$$

Entrywise this is

$$J_t = \sum_{ik} \big(\pi_{ik}^{t+1} - \pi_{ik}^t\big)\big(\log \pi_{ik}^{t+1} - \log \pi_{ik}^t\big).$$

Step 3 (mirror KKT and the normal cone). The step equation 9 is the Bregman proximal map of ϕ on $\widetilde{\mathcal{C}}_\tau$. First-order optimality therefore yields

$$g^t + \varepsilon\big(\nabla\omega(\pi^{t+1}) - \nabla\omega(\pi^t)\big) + \eta^t = 0$$

for some $\eta^t \in N_{\widetilde{\mathcal{C}}_\tau}(\pi^{t+1})$. The normal cone consists of those η with $\langle \eta, \pi' - \pi^{t+1} \rangle \leq 0$ for every $\pi' \in \widetilde{\mathcal{C}}_\tau$. The point π^t is feasible, so $\langle \eta^t, \pi^t - \pi^{t+1} \rangle \leq 0$, that is $\langle \eta^t, \Delta^t \rangle \geq 0$. Take the inner product of the stationarity identity against Δ^t and substitute Step 2:

$$\langle g^t, \Delta^t \rangle + \varepsilon J_t + \langle \eta^t, \Delta^t \rangle = 0.$$

The normal-cone sign therefore gives $\langle g^t, \Delta^t \rangle \leq -\varepsilon J_t$.

Step 4 (directional decrease). Substitute Step 3 into Step 1:

$$F(\pi^{t+1}) - F(\pi^t) \leq -\varepsilon J_t + \langle \Delta^t, H\Delta^t \rangle.$$

The quadratic remainder is at most its positive part, $\langle \Delta^t, H\Delta^t \rangle \leq \big[\langle \Delta^t, H\Delta^t \rangle\big]_+$. By definition of ℓ_t, $\big[\langle \Delta^t, H\Delta^t \rangle\big]_+ = \ell_t J_t$, hence

$$F(\pi^{t+1}) - F(\pi^t) \leq -(\varepsilon - \ell_t) J_t.$$

If $\ell_t < \varepsilon$, the coefficient of J_t is positive.

Step 5 (scalar log inequality). For $x, y > 0$,

$$(x - y)(\log x - \log y) = \int_{\min\{x,y\}}^{\max\{x,y\}} \frac{|x - y|}{s}\, ds \geq \frac{(x - y)^2}{\max\{x, y\}},$$

because $s \leq \max\{x, y\}$ throughout the interval. Apply this coordinatewise to $(x, y) = (\pi_{ik}^{t+1}, \pi_{ik}^t)$ in the expression of J_t from Step 2:

$$J_t \geq \sum_{ik} \frac{(\Delta_{ik}^t)^2}{\max\{\pi_{ik}^{t+1}, \pi_{ik}^t\}}.$$

Step 6 ((C1) coefficient). On $\widetilde{\mathcal{C}}_\tau$ every coordinate satisfies $\pi_{ik} \le \mu_i - (K-1)\tau = u_i$, because the remaining $K-1$ entries of the same row are at least τ. Thus $\max\{\pi_{ik}^{t+1}, \pi_{ik}^t\} \le u_i$, and Step 5 gives

$$J_t \ge \sum_{ik} \frac{(\Delta_{ik}^t)^2}{u_i} \ge \frac{1}{u_{\max}} \sum_{ik} (\Delta_{ik}^t)^2 = \frac{\|\Delta^t\|_F^2}{u_{\max}}.$$

Combined with Step 4 and $\ell_t < \varepsilon$,

$$F(\pi^t) - F(\pi^{t+1}) \ge (\varepsilon - \ell_t) J_t \ge \frac{\varepsilon - \ell_t}{u_{\max}} \|\Delta^t\|_F^2.$$

The indicators in Φ_τ vanish at both iterates, which is (C1) on the slice with $a_t = (\varepsilon - \ell_t)/u_{\max}$. If a uniform ℓ satisfies $\varepsilon > \ell \ge \sup_s \ell_{t,s}$ on the n_{inner} inner steps of an outer block, write $\Delta_{t,s} = \pi^{t,s+1} - \pi^{t,s}$ and

$$\delta_t = \sum_{s=0}^{n_{\text{inner}}-1} \|\Delta_{t,s}\|_F.$$

Cauchy–Schwarz on these n_{inner} nonnegative summands gives

$$\delta_t^2 \le n_{\text{inner}} \sum_{s=0}^{n_{\text{inner}}-1} \|\Delta_{t,s}\|_F^2.$$

Each inner step then satisfies

$$F_\lambda(\pi^{t,s}, D'^t) - F_\lambda(\pi^{t,s+1}, D'^t) \ge \frac{\varepsilon - \ell}{u_{\max}} \|\Delta_{t,s}\|_F^2.$$

The D' update does not increase F_λ. Summing the inner decreases and substituting the Cauchy–Schwarz bound yields the outer (C1)

$$\Phi_\tau(z^t) - \Phi_\tau(z^{t+1}) \ge \frac{\varepsilon - \ell}{u_{\max}} \sum_{s=0}^{n_{\text{inner}}-1} \|\Delta_{t,s}\|_F^2 \ge \frac{\varepsilon - \ell}{u_{\max} n_{\text{inner}}} \delta_t^2.$$

□

Proposition 6 (Precomputable uniform bound on the pathwise ratio)**.** *Hold $\lambda = \lambda_{\text{II}}$ fixed, and let each D'^t be given by a D' update, so $D'^t \in [m, M]^K$. Let ℓ_t and J_t be as in Proposition 5. Then every mirror increment satisfies*

$$\ell_t \le \tfrac{1}{2} L_{\text{rel}}(D'^t, \lambda_{\text{II}}) \le \tfrac{1}{2} \overline{L}_{\text{rel}}(\lambda_{\text{II}}).$$

The first inequality uses the realized Δ^t; the second uses only that D'^t lies in the box. Consequently the set

$$\{\varepsilon > 0 : \varepsilon > \tfrac{1}{2} \overline{L}_{\text{rel}}(\lambda_{\text{II}})\}$$

is a nonempty open ray of inverse step sizes: each such ε forces $\ell_t < \varepsilon$ at every Stage II mirror descent step, and Proposition 5 applies with a uniform a.

The open ray $\varepsilon > \frac{1}{2}\overline{L}_{\text{rel}}(\lambda_{\text{II}})$ forces $\ell_t < \varepsilon$ at every step, so Proposition 5 gives a uniform (C1). Theorem 1 uses the stronger pathwise hypothesis $\varepsilon > \sup_t L_{\text{rel}}(D'^t, \lambda_{\text{II}})$; the box bound $\varepsilon > \overline{L}_{\text{rel}}(\lambda_{\text{II}})$ is sufficient for that hypothesis. Proposition 6 and Corollary 3 use the sharper realized-step bound and require only $\varepsilon > \frac{1}{2}\overline{L}_{\text{rel}}(\lambda_{\text{II}})$. The recorded Table 1 check is the pathwise ratio of Proposition 5. The sufficient-descent validation under precomputed conditions uses the optional point $\varepsilon = 1.05\overline{L}_{\text{rel},0}$ of that ray.

Proof. If $\Delta^t = 0$, then $\ell_t = 0$ and both claimed inequalities hold because $L_{\text{rel}} \ge 0$. Assume $\Delta^t \ne 0$.

Step 1 (Jeffreys versus the weighted ℓ_2 quotient). Step 6 of Proposition 5 gives $J_t \ge \sum_{ik} (\Delta_{ik}^t)^2 / u_i$. The definition of ℓ_t therefore yields

$$\ell_t = \frac{\left[\langle \Delta^t, H\Delta^t \rangle\right]_+}{J_t} \le \frac{\left[\langle \Delta^t, H\Delta^t \rangle\right]_+}{\sum_{ik} (\Delta_{ik}^t)^2 / u_i}.$$

Step 2 (definition of L_{rel}). The pointwise Rayleigh quotient in Section 3 is

$$L_{\mathrm{rel}}(D', \lambda) = 2 \sup_{\Delta \in \mathcal{T} \setminus \{0\}} \frac{\big[\langle \Delta, H_\lambda(D')\Delta \rangle\big]_+}{\sum_{ik} \Delta_{ik}^2 / u_i}.$$

The factor 2 is the same weight $\int_0^1 (1-s)\, ds = \frac{1}{2}$ that appears in the one-sided Kullback–Leibler expansion of Proposition 4. The realized increment Δ^t lies in $\mathcal{T}$, so

$$\frac{\big[\langle \Delta^t, H\Delta^t \rangle\big]_+}{\sum_{ik} (\Delta_{ik}^t)^2 / u_i} \le \frac{L_{\mathrm{rel}}(D'^t, \lambda_{\mathrm{II}})}{2}.$$

Combined with Step 1, $\ell_t \le \frac{1}{2} L_{\mathrm{rel}}(D'^t, \lambda_{\mathrm{II}})$.

Step 3 (update in the box). Lemma 1 places an clustering D' update in $[m, M]^K$. In this subsection $\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})$ is the maximum of $L_{\mathrm{rel}}(\,\cdot\,, \lambda_{\mathrm{II}})$ on that box, which is the theorem domain of Corollary 2. Hence $L_{\mathrm{rel}}(D'^t, \lambda_{\mathrm{II}}) \le \overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})$, the second claimed inequality.

Step 4 (the maximum is finite, so the ray is nonempty). The set $\{\Delta \in \mathcal{T} : \sum_{ik} \Delta_{ik}^2 / u_i = 1\}$ is closed and bounded in a finite-dimensional space, hence compact, and is nonempty for $K \ge 2$. The map $D' \mapsto H_{\lambda_{\mathrm{II}}}(D')$ is continuous on $[m, M]^K$ because $F_{\lambda_{\mathrm{II}}}$ depends continuously on D' there. The Rayleigh quotient is therefore a maximum of a continuous function of (D', Δ) over a compact product, and $D' \mapsto L_{\mathrm{rel}}(D', \lambda_{\mathrm{II}})$ is continuous on the compact box $[m, M]^K$. It attains a finite maximum $\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})$. The set of ε strictly larger than half that maximum is then a nonempty open ray in $(0, \infty)$. Any ε in the ray satisfies $\ell_t \le \frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}}) < \varepsilon$ at every t, which is the hypothesis of Proposition 5 with uniform $\overline{\ell} \le \frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})$. □

Corollary 3 (Stage II under a precomputable ε)**.** *Hold $\lambda = \lambda_{\mathrm{II}}$ fixed, take every D' update to be exact, and choose*

$$\varepsilon > \tfrac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}}).$$

Then the same finite-length whole-sequence argument as in the proof of Theorem 1 applies from the first Stage II iterate.

Proof. Proposition 6 gives $\ell_t \le \frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}}) < \varepsilon$ at every Stage II mirror descent step. Proposition 5 therefore yields (C1) on each inner slice with

$$a_t = \frac{\varepsilon - \ell_t}{u_{\max}} \ge \frac{\varepsilon - \frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})}{u_{\max}}.$$

Write $\Delta_{t,s} = \pi^{t,s+1} - \pi^{t,s}$ and $\delta_t = \sum_{s=0}^{n_{\mathrm{inner}}-1} \|\Delta_{t,s}\|_F$. Cauchy–Schwarz on these n_{inner} nonnegative summands gives $\delta_t^2 \le n_{\mathrm{inner}} \sum_s \|\Delta_{t,s}\|_F^2$. Each inner decrease is at least $\big(\varepsilon - \frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})\big)/u_{\max}$ times $\|\Delta_{t,s}\|_F^2$, and the D' update does not increase F_λ. The outer (C1) coefficient is therefore

$$a_{\mathrm{out}} = \frac{\varepsilon - \frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})}{u_{\max} n_{\mathrm{inner}}} > 0.$$

Lemma 2 and compactness of the feasible product are unchanged, so (C2) and (C3) hold. Lemma 3 with the KŁ property from Proposition 3 is then the same argument as Steps 1–3 of Theorem 1, started at the first Stage II iterate. □

C TECHNICAL DETAILS FOR THE VANISHING-FLOOR LIMIT

This appendix supports Theorem 2, Proposition 1, and the main-text Figure 3. It separates stability of the optimization landscape from continuity at a finite horizon J.

The three limits used in the paper are distinct. Theorem 1 holds for one fixed $\tau > 0$ and concerns the iteration index $t \to \infty$. Theorem 2 concerns global minimizers and constrained critical points of a family of problems with $\tau_j \downarrow 0$. Proposition 1 concerns a finite horizon J of outer blocks of Algorithm 1. None of these results exchanges $t \to \infty$ with $\tau \downarrow 0$.

Floor-selection rule. The hard-floor experiments use the scale-normalized rule

$$\tau = r_\tau \frac{\mu_{\min}}{K}, \qquad \mu_{\min} := \min_i \mu_i, \qquad r_\tau = 10^{-12}.$$

This keeps the same dimensionless floor ratio across datasets with different row-marginal scales. The ratio is chosen small to reduce the truncation perturbation, but it remains strictly positive so that the entropy gradient and the truncated mirror descent step are defined on the theorem domain. The recorded ratio is $r_\tau = 10^{-12}$. The practical check is feasibility and finite projection residuals: start at 10^{-12} and verify $\min_i \mu_i/K > \tau$, $\min_{i,k} \pi_{ik} \geq \tau$, and finite projection residuals. Sufficient descent may be read from Proposition 5 via $\overline{\ell}_{\mathrm{II},T} < \varepsilon$, or from Corollary 3 via $\varepsilon > \frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda)$. If numerical underflow or unstable active-set behavior occurs, increase r_τ to the smallest stable value. Decreasing r_τ is not free: the relative-error constant contains ε/τ, so the fixed-floor proof becomes less uniform even though the feasible domains approach the boundary.

Proof of Theorem 2. The two statements follow the numbered claims of Theorem 2.

Step 1 (values and global minimizers). If $0 < \tau' < \tau < \min_i \mu_i/K$, then $\widetilde{\mathcal{C}}_\tau \subset \widetilde{\mathcal{C}}_{\tau'} \subset \mathcal{C}$, so $\mathcal{Z}_{\tau_j} \subset \mathcal{Z}_{\tau_{j+1}} \subset \mathcal{Z}_0$. Hence $v(\tau_j)$ is nonincreasing and $v(\tau_j) \geq v(0)$.

Fix $(\pi, D') \in \mathcal{Z}_0$ and $0 < \tau < \min_i \mu_i/K$. The convex combination

$$\pi^\tau = (1 - a_\tau)\pi + a_\tau q, \qquad a_\tau = \frac{K\tau}{\min_i \mu_i}, \qquad q_{ik} = \frac{\mu_i}{K},$$

has the same row sums as π and satisfies $\pi^\tau_{ik} \geq \tau$, so $(\pi^\tau, D') \in \mathcal{Z}_\tau$. As $\tau \downarrow 0$, one has $a_\tau \to 0$ and therefore $\pi^\tau \to \pi$. equation 2 is a polynomial in π whose coefficients are continuous in D' on the compact box $[m, M]$ with $m > 0$, so F_λ is continuous on $\mathcal{Z}_0$. Hence $F_\lambda(\pi^\tau, D') \to F_\lambda(\pi, D')$. Feasibility yields $v(\tau) \leq F_\lambda(\pi^\tau, D')$, so $\limsup_j v(\tau_j) \leq F_\lambda(\pi, D')$. The pair (π, D') was arbitrary, hence $\limsup_j v(\tau_j) \leq v(0)$. Combined with $v(\tau_j) \geq v(0)$ this is $v(\tau_j) \to v(0)$.

Now let $z_j \in \arg\min_{\mathcal{Z}_{\tau_j}} F_\lambda$. Compactness of $\mathcal{Z}_0$ supplies a subsequence $z_{j_r} \to \bar{z} \in \mathcal{Z}_0$. Continuity of F_λ and $F_\lambda(z_{j_r}) = v(\tau_{j_r}) \to v(0)$ give $F_\lambda(\bar{z}) = v(0)$, so $\bar{z}$ is a global minimizer on $\mathcal{Z}_0$. Every cluster point of the truncated global minimizers therefore lies in $\arg\min_{\mathcal{Z}_0} F_\lambda$.

Step 2 (constrained critical points). Let $z_j = (\pi_j, D'_j) \in \mathcal{Z}_{\tau_j}$ satisfy $0 \in \partial\Phi_{\tau_j}(z_j)$, and pass to a convergent subsequence $z_j \to \bar{z} = (\bar{\pi}, \bar{D}') \in \mathcal{Z}_0$. Since F_λ is C^1 and $\mathcal{Z}_{\tau_j}$ is closed and convex,

$$-\nabla F_\lambda(z_j) \in N_{\widetilde{\mathcal{C}}_{\tau_j}}(\pi_j) \times N_{\mathcal{D}}(D'_j).$$

Write $(v_j, w_j) := -\nabla F_\lambda(z_j)$. Continuity of the gradient gives $(v_j, w_j) \to (-\nabla_\pi F_\lambda(\bar{z}), -\nabla_{D'} F_\lambda(\bar{z}))$.

Take any $\pi \in \operatorname{relint} \mathcal{C}$. All entries of π are positive, so $\pi \in \widetilde{\mathcal{C}}_{\tau_j}$ for all large j, and $\langle v_j, \pi - \pi_j \rangle \leq 0$. Passing to the limit,

$$\langle -\nabla_\pi F_\lambda(\bar{z}), \pi - \bar{\pi} \rangle \leq 0 \qquad \text{for every } \pi \in \operatorname{relint} \mathcal{C}.$$

The relative interior is dense in $\mathcal{C}$, so the inequality extends to every $\pi \in \mathcal{C}$. Thus $-\nabla_\pi F_\lambda(\bar{z}) \in N_{\mathcal{C}}(\bar{\pi})$. For every $D' \in \mathcal{D}$ one has $\langle w_j, D' - D'_j \rangle \leq 0$; the limit is $-\nabla_{D'} F_\lambda(\bar{z}) \in N_{\mathcal{D}}(\bar{D}')$. Therefore

$$0 \in \nabla F_\lambda(\bar{z}) + N_{\mathcal{C}\times\mathcal{D}}(\bar{z}) = \partial\Phi_0(\bar{z}).$$

This is the outer-limit inclusion $\limsup_{\tau\downarrow 0} \operatorname{crit} \Phi_\tau \subseteq \operatorname{crit} \Phi_0$. □

Proof of Proposition 1. **Step 1 (uniform error of one mirror descent step).** On a row of mass $\mu_i > 0$, write q for the softmax step and x for the truncated mirror descent step. On $\{x \geq \tau,\ \mathbf{1}^\top x = \mu_i\}$ the linear-plus-KL objective differs from $D_{\mathrm{KL}}(x\|q)$ by a constant in x, so x uniquely minimizes $D_{\mathrm{KL}}(\cdot\,\|q)$ there. The KKT form is $x_k = \max\{\tau, cq_k\}$ with $c \in (0, 1]$ chosen so that $\mathbf{1}^\top x = \mu_i$. Let $A = \{k : cq_k \leq \tau\}$. Equating row sums gives $(1 - c)\sum_{k\notin A} q_k = |A|\tau - \sum_{k\in A} q_k \geq 0$, hence $\sum_{k\in A} q_k \leq |A|\tau$. Therefore

$$\begin{aligned}\|x - q\|_1 &= \sum_{k\in A} |\tau - q_k| + (1 - c)\sum_{k\notin A} q_k \\ &\leq \sum_{k\in A} (\tau + q_k) + \big(|A|\tau - \sum_{k\in A} q_k\big) = 2|A|\tau \leq 2K\tau.\end{aligned}$$

Summing over the N rows and using $\|X\|_F \le \|X\|_1$ gives, for the same input pair $z = (\pi, D')$,

$$\|\mathcal{M}_\tau(z) - \mathcal{M}_0(z)\|_F \le 2NK\tau,$$

where $\mathcal{M}_\tau$ denotes one mirror descent step with floor τ, including evaluation of $\nabla_\pi F_\lambda(z)$, and $\mathcal{M}_0$ denotes the untruncated softmax step.

Step 2 (a Lipschitz neighborhood of the finite softmax path). Consider the finitely many iterates of the $\tau = 0$ run. The initial plan is entrywise positive, and every softmax update is entrywise positive. Hence

$$\alpha_{\min} := \min\big\{(\pi^j_{\text{soft}})_{ik} : 0 \le j \le J,\ 1 \le i \le N,\ 1 \le k \le K\big\} > 0.$$

Every column mass on this finite path is therefore at least $N\alpha_{\min}$. Choose a closed neighborhood $\mathcal{U}$ of the finite path in which all entries are at least $\alpha_{\min}/2$; all column masses in $\mathcal{U}$ are at least $N\alpha_{\min}/2$.

On the compact set $\mathcal{U}$, the gradient $\nabla_\pi F_\lambda$ is Lipschitz. The row softmax formula is smooth in its positive current row and gradient, so the composed map $\mathcal{M}_0$ is Lipschitz there. The D' update h in equation 6, or its clustering restriction equation 7, is Lipschitz there because its denominators are bounded below by $(N\alpha_{\min}/2)^2$ and $(\varphi')^{-1}$ is continuously differentiable on the compact image under consideration. Thus finite constants $L_{\mathcal{M},0}, L_h$ exist such that

$$\|\mathcal{M}_0(z) - \mathcal{M}_0(\widetilde{z})\|_F \le L_{\mathcal{M},0}\|z - \widetilde{z}\|_\times, \qquad \|h(\pi) - h(\widetilde{\pi})\|_F \le L_h\|\pi - \widetilde{\pi}\|_F,$$

in the product norm $\|z\|_\times$ of Appendix A. Set $L_{1,0} := L_{\mathcal{M},0} + 1$, $L_2 := 1 + L_h$, and $B_1 := 2NK$.

Step 3 (error propagation through alternating updates). For $z, \widetilde{z} \in \mathcal{U}$, Step 1 and the first Lipschitz bound in Step 2 give

$$\begin{aligned}\|\mathcal{M}_\tau(z) - \mathcal{M}_0(\widetilde{z})\|_F &\le \|\mathcal{M}_\tau(z) - \mathcal{M}_0(z)\|_F + \|\mathcal{M}_0(z) - \mathcal{M}_0(\widetilde{z})\|_F \\ &\le 2NK\tau + L_{\mathcal{M},0}\|z - \widetilde{z}\|_\times.\end{aligned}$$

At a D' update, the second Lipschitz bound gives

$$\|(\pi, h(\pi)) - (\widetilde{\pi}, h(\widetilde{\pi}))\|_\times \le (1 + L_h)\|\pi - \widetilde{\pi}\|_F.$$

Let J be the number of outer iterations in this sequence, each consisting of n_{inner} mirror descent steps followed by one D' update, and let e_j be the product-norm discrepancy after outer iteration j. On $\mathcal{U}$, one mirror descent step adds $B_1\tau$ by Step 1 and multiplies the running discrepancy by at most $L_{1,0}$, while one D' update multiplies it by at most L_2. Applying these $n_{\text{inner}} + 1$ bounds gives

$$e_{j+1} \le Le_j + B\tau, \qquad L := L_{1,0}^{n_{\text{inner}}} L_2, \qquad B := B_1 n_{\text{inner}} L_{1,0}^{n_{\text{inner}}-1} L_2,$$

both finite and independent of τ. Because both runs start from the same pair, $e_0 = 0$ and $z^0_\tau = z^0_{\text{soft}} \in \mathcal{U}$. We argue by induction on j that $e_j \le B\tau\sum_{r=0}^{j-1} L^r$ and $z^j_\tau \in \mathcal{U}$ hold together. Fix $\tau_* > 0$ small enough that $C_J\tau_*$ is smaller than the radius of $\mathcal{U}$. If $z^j_\tau \in \mathcal{U}$, the recursion gives $e_{j+1} \le C_J\tau_*$, and z^{j+1}_τ lies within that distance of $z^{j+1}_{\text{soft}} \in \mathcal{U}$ in the norm $\|\cdot\|_\times$, hence in $\mathcal{U}$ and the recursion applies again. With $C_J := B\sum_{r=0}^{J-1} L^r < +\infty$,

$$e_j \le B\tau\sum_{r=0}^{j-1} L^r \le C_J\tau$$

for every $j \le J$. Taking the maximum over the iterates of the sequence proves the stated finite-horizon estimate. □

Proposition 1 justifies the finite-horizon comparison with the $\tau = 0$ softmax path.

D Finite-horizon Lipschitz dependence on the initialization

This appendix proves Proposition 2 and records the two corollaries. Write $\|z\|_\times = \|\pi\|_F + \|D'\|_F$ for $z = (\pi, D')$, as in Appendix A. Let $z = (\pi, D')$ and $\widehat{z} = (\widehat{\pi}, \widehat{D}')$ lie in Z_τ. A permutation σ of the K prototype labels acts by $(\pi^\sigma)_{i,\sigma(k)} = \pi_{ik}$ and $(D'^\sigma)_{\sigma(k)\sigma(l)} = D'_{kl}$. Let

$$T : Z_\tau \to Z_\tau$$

send z to one outer block of Algorithm 1: n_{inner} truncated mirror steps at fixed D', followed by the exact update $D' \leftarrow h(\pi)$. Write

$$z^J = T^{\circ J}(z), \qquad z^{J,\sigma} = (z^J)^\sigma.$$

The proof combines the Lipschitz regularity of the two update maps from Lemma 1 with prototype-label equivariance.

Corollary 4 (Finite-horizon stability of the hard assignment)**.** *Fix an initial state z and a finite horizon J, and write*

$$T^{\circ J}(z) = (\pi^J, D'^J).$$

Suppose every row of π^J has a unique maximizer, and set

$$\gamma_J = \min_i \left[\pi^J_{i,k_i^\star} - \max_{l \neq k_i^\star} \pi^J_{il}\right] > 0, \qquad k_i^\star = \arg\max_k \pi^J_{ik}.$$

If

$$\min_\sigma \|z^\sigma - \widehat{z}\|_\times < \frac{\gamma_J}{2L^J_{\text{out}}},$$

then the hard assignments after J outer blocks agree up to the corresponding prototype permutation.

Corollary 5 (Prescribed finite continuation)**.** *Let $\lambda_1, \dots, \lambda_{n_{\text{cont}}}$ be a prescribed finite sequence, and let T_{λ_r} denote one outer block at λ_r. Then*

$$T_{\lambda_{n_{\text{cont}}}} \circ \cdots \circ T_{\lambda_1}$$

is Lipschitz and equivariant under prototype-label permutations. Consequently, the finite-horizon Lipschitz dependence of Proposition 2 extends to any prescribed finite Stage I continuation followed by a finite Stage II horizon.

Proof of Proposition 2. Write L_h and $L_{\mathcal{M}}$ for the Lipschitz constants of h and of the truncated map $\mathcal{M} = \mathcal{M}_\tau$ furnished by Lemma 1. Let $z = (\pi, D')$ and $\widehat{z} = (\widehat{\pi}, \widehat{D}')$ lie in Z_τ. Lemma 1 gives

$$\|\mathcal{M}(z) - \mathcal{M}(\widehat{z})\|_F \le L_{\mathcal{M}} \|z - \widehat{z}\|_\times.$$

Define the one-inner-step joint map $I : Z_\tau \to Z_\tau$ by $I(\pi, D') := \big(\mathcal{M}(\pi, D'), D'\big)$. Since $\mathcal{M}(z) \in \widetilde{\mathcal{C}}_\tau$ and $D' \in \mathcal{D}$, the value $I(z)$ lies in Z_τ. Then

$$\begin{aligned}
\|I(z) - I(\widehat{z})\|_\times &= \|\mathcal{M}(z) - \mathcal{M}(\widehat{z})\|_F + \|D' - \widehat{D}'\|_F \\
&\le L_{\mathcal{M}} \|z - \widehat{z}\|_\times + \|D' - \widehat{D}'\|_F \\
&\le L_{\mathcal{M}} \|z - \widehat{z}\|_\times + \|z - \widehat{z}\|_\times = L_1 \|z - \widehat{z}\|_\times,
\end{aligned}$$

where $L_1 := L_{\mathcal{M}} + 1$ and the last step uses $\|D' - \widehat{D}'\|_F \le \|z - \widehat{z}\|_\times$. Iterating the L_1-Lipschitz map I exactly n_{inner} times yields

$$\big\|I^{\circ n_{\text{inner}}}(z) - I^{\circ n_{\text{inner}}}(\widehat{z})\big\|_\times \le L_1^{n_{\text{inner}}} \|z - \widehat{z}\|_\times.$$

Write $I^{\circ n_{\text{inner}}}(z) = (\pi_{\text{out}}, D')$ and $I^{\circ n_{\text{inner}}}(\widehat{z}) = (\widehat{\pi}_{\text{out}}, \widehat{D}')$. The outer block is $T(z) = (\pi_{\text{out}}, h(\pi_{\text{out}}))$. Set $L_2 := 1 + L_h$. The exact D' update and the Lipschitz bound on h give

$$\begin{aligned}
\|T(z) - T(\widehat{z})\|_\times &= \|\pi_{\text{out}} - \widehat{\pi}_{\text{out}}\|_F + \|h(\pi_{\text{out}}) - h(\widehat{\pi}_{\text{out}})\|_F \\
&\le \|\pi_{\text{out}} - \widehat{\pi}_{\text{out}}\|_F + L_h \|\pi_{\text{out}} - \widehat{\pi}_{\text{out}}\|_F \\
&= L_2 \|\pi_{\text{out}} - \widehat{\pi}_{\text{out}}\|_F \\
&\le L_2 \big\|I^{\circ n_{\text{inner}}}(z) - I^{\circ n_{\text{inner}}}(\widehat{z})\big\|_\times \\
&\le L_1^{n_{\text{inner}}} L_2 \|z - \widehat{z}\|_\times.
\end{aligned}$$

Set $L_{\text{out}} := L_1^{n_{\text{inner}}} L_2$. Then T is L_{out}-Lipschitz, and J outer blocks satisfy

$$\|T^{\circ J}(z) - T^{\circ J}(\widehat{z})\|_\times \le L^J_{\text{out}} \|z - \widehat{z}\|_\times.$$

The map T is prototype-equivariant. A permutation σ of the K prototype labels acts by $(\pi^\sigma)_{i,\sigma(k)} = \pi_{ik}$ and $(D'^\sigma)_{\sigma(k)\sigma(l)} = D'_{kl}$, equivalently $\pi^\sigma_{ik} = \pi_{i,\sigma^{-1}(k)}$ and $D'^\sigma_{kl} = D'_{\sigma^{-1}(k)\sigma^{-1}(l)}$. Substitute into the Bregman pairing:

$$\Psi(\pi^\sigma, D'^\sigma) = \sum_{i,j,k,l} B_\varphi\big(D'_{\sigma^{-1}(k)\sigma^{-1}(l)}, D_{ij}\big)\pi_{i,\sigma^{-1}(k)}\pi_{j,\sigma^{-1}(l)}.$$

The change of dummy indices $a = \sigma^{-1}(k)$, $b = \sigma^{-1}(l)$ is a bijection of $\{1,\dots,K\}^2$, so $\Psi(\pi^\sigma, D'^\sigma) = \Psi(\pi, D')$. The regularizer $\|\pi\|_F^2$ is likewise column-permutation invariant, hence $F_\lambda(z^\sigma) = F_\lambda(z)$. Differentiating in π and transporting columns by σ gives $(\nabla_\pi F_\lambda(z^\sigma))_{i,\sigma(k)} = (\nabla_\pi F_\lambda(z))_{ik}$. The set $\widetilde{\mathcal{C}}_\tau$ is invariant under column permutations, and $D_{\mathrm{KL}}(x^\sigma\|\pi^\sigma) = D_{\mathrm{KL}}(x\|\pi)$ because the sum over columns is invariant under the same bijection. Hence $Q(x^\sigma; z^\sigma) = Q(x; z)$ with the Q of Lemma 1. Uniqueness of the truncated minimizer gives $\mathcal{M}(z^\sigma) = \mathcal{M}(z)^\sigma$, and therefore $I(z^\sigma) = I(z)^\sigma$. Repeating through the fixed inner count yields $I^{\circ n_{\mathrm{inner}}}(z^\sigma) = I^{\circ n_{\mathrm{inner}}}(z)^\sigma$. The barycenter formula satisfies $s_{\sigma(k)}(\pi^\sigma) = s_k(\pi)$ and $h(\pi^\sigma) = h(\pi)^\sigma$, hence $T(z^\sigma) = T(z)^\sigma$. Induction on the finite horizon J gives $T^{\circ J}(z^\sigma) = T^{\circ J}(z)^\sigma$, that is $z^{J,\sigma} = T^{\circ J}(z^\sigma)$.

For every prototype permutation σ,

$$\begin{aligned}\min_{\sigma'} \|z^{J,\sigma'} - \widehat{z}^J\|_\times &\le \|(z^J)^\sigma - \widehat{z}^J\|_\times = \|T^{\circ J}(z^\sigma) - T^{\circ J}(\widehat{z})\|_\times \\ &\le L^J_{\mathrm{out}}\|z^\sigma - \widehat{z}\|_\times.\end{aligned}$$

Minimizing the right-hand side over σ yields the aligned bound of Proposition 2. □

Proof of Corollary 4. Let σ be a prototype permutation that attains the minimum in $\min_{\sigma'} \|z^{\sigma'} - \widehat{z}\|_\times$. The hypothesis and Proposition 2, together with the equivariance $z^{J,\sigma} = T^{\circ J}(z^\sigma)$ recorded in that proof, give

$$\|z^{J,\sigma} - \widehat{z}^J\|_\times = \|T^{\circ J}(z^\sigma) - T^{\circ J}(\widehat{z})\|_\times \le L^J_{\mathrm{out}}\|z^\sigma - \widehat{z}\|_\times < \frac{\gamma_J}{2}.$$

Write $z^{J,\sigma} = (\pi^{J,\sigma}, D'^{J,\sigma})$ and $\widehat{z}^J = (\widehat{\pi}^J, \widehat{D}'^J)$. Then

$$\|\widehat{\pi}^J - \pi^{J,\sigma}\|_F \le \|z^{J,\sigma} - \widehat{z}^J\|_\times < \frac{\gamma_J}{2}.$$

The entrywise maximum norm $\|A\|_\infty := \max_{i,k}|A_{ik}|$ satisfies $\|A\|_\infty \le \|A\|_F$, hence

$$\|\widehat{\pi}^J - \pi^{J,\sigma}\|_\infty < \frac{\gamma_J}{2}.$$

For every row i, write $k_i^\star = \arg\max_k \pi^J_{ik}$. The action of σ gives $(\pi^{J,\sigma})_{i,\sigma(k_i^\star)} = \pi^J_{i,k_i^\star}$ and $(\pi^{J,\sigma})_{i,\sigma(l)} = \pi^J_{il}$ for every $l \ne k_i^\star$. For every competitor $l \ne k_i^\star$,

$$\begin{aligned}\widehat{\pi}^J_{i,\sigma(k_i^\star)} - \widehat{\pi}^J_{i,\sigma(l)} &> \big((\pi^{J,\sigma})_{i,\sigma(k_i^\star)} - \tfrac{\gamma_J}{2}\big) - \big((\pi^{J,\sigma})_{i,\sigma(l)} + \tfrac{\gamma_J}{2}\big) \\ &= \pi^J_{i,k_i^\star} - \pi^J_{il} - \gamma_J \ge \gamma_J - \gamma_J = 0,\end{aligned}$$

where the first inequality is strict because $\|\widehat{\pi}^J - \pi^{J,\sigma}\|_\infty < \gamma_J/2$. Thus $\sigma(k_i^\star)$ is the unique maximizer of row i of $\widehat{\pi}^J$. The hard assignments after J outer blocks therefore agree up to this prototype permutation. □

Proof of Corollary 5. For each fixed λ_r, Lemma 1 and the proof of Proposition 2 give a finite Lipschitz constant L_{λ_r} for T_{λ_r}. Hence the prescribed Stage I composition is Lipschitz with constant

$$\prod_{r=1}^{n_{\mathrm{cont}}} L_{\lambda_r}.$$

Each factor is prototype-equivariant, so the composition is also equivariant. Composing with a Stage II horizon of J outer blocks, which is Lipschitz with constant $L^J_{\mathrm{out,II}}$, gives the composite Lipschitz bound $L^J_{\mathrm{out,II}}\prod_{r=1}^{n_{\mathrm{cont}}} L_{\lambda_r}$ and the claim. □

Successive-iterate stopping. Proposition 2 applies the same prescribed finite horizon J to both initial states, and the Lipschitz bound is only for that shared J. Algorithm 1 indexes outer blocks by q and also exits when $\|\pi^{q+1} - \pi^q\|_F + \|D'^{q+1} - D'^q\|_F < \mathrm{tol}$. Let $r : \mathcal{Z}_\tau \to \mathbb{R}$ be continuous and write

$$n(z) = \min\{t \le q_{\max} : r(T^{\circ t}(z)) \le \mathrm{tol}\},$$

with $n(z) = q_{\max}$ if the set is empty, and $S(z) = T^{\circ n(z)}(z)$. If $z_\star$ admits sequences $z_m^-, z_m^+ \to z_\star$ with

$$r(z_m^-) \le \mathrm{tol}, \qquad r(z_m^+) > \mathrm{tol}, \qquad r(T(z_m^+)) \le \mathrm{tol},$$

and $T(z_\star) \neq z_\star$, then $S(z_m^-) \to z_\star$ while $S(z_m^+) \to T(z_\star)$. The map S need not be continuous, and the Lipschitz bound of Proposition 2 does not apply to S. The successive-iterate test of Algorithm 1 is this form after the state is augmented from z^q to (z^{q-1}, z^q).

Relative plan distances. Every $\pi \in \widetilde{\mathcal{C}}_\tau$ satisfies $\tau\sqrt{NK} \le \|\pi\|_F \le \|\pi\|_1 = 1$, hence

$$\|\pi - \widehat{\pi}\|_F \le \frac{\|\pi - \widehat{\pi}\|_F}{\|\pi\|_F} \le \frac{1}{\tau\sqrt{NK}}\|\pi - \widehat{\pi}\|_F.$$

Absolute Frobenius distance and the relative Frobenius distance used in Section 4.1 induce the same topology on $\widetilde{\mathcal{C}}_\tau$.

Initialization weights. If a scalar weight α produces $z^0(\alpha)$ and, locally,

$$\min_\sigma \|z^0(\alpha)^\sigma - z^0(\alpha_0)\|_\times \le L_\alpha|\alpha - \alpha_0|,$$

then Proposition 2 gives

$$\min_\sigma \|z^J(\alpha)^\sigma - z^J(\alpha_0)\|_\times \le L_{\mathrm{out}}^J L_\alpha|\alpha - \alpha_0|.$$

If the reference terminal plan has row margin $\gamma_J > 0$, then

$$|\alpha - \alpha_0| < \frac{\gamma_J}{2L_{\mathrm{out}}^J L_\alpha}$$

is sufficient to preserve the hard assignment up to prototype relabeling.

E Solver Protocols and Timing

This appendix names the maps used in Table 1. The BGW-hard-floor row is the truncated solver of Theorem 1 at Stage II. The $\tau = 0$ softmax and clipped rows are outside that theorem. Spectral EU(sym) uses the same heat-kernel Ng protocol as Wiki(sym) and occupies every prescribed cluster ($K = 42$).

The inner hard-floor step is a truncated mirror descent step on the truncated row simplex, implemented by `project_row_simplex_lower_bound`. After the unconstrained softmax reference q, each row solves $x_k = \max(\tau, cq_k)$ by an active-set loop of at most K passes, which is $O(K^2)$ per row. The loop stops when a multiplier c makes the truncated row sum to μ_i; in exact arithmetic this is the unique truncated mirror descent step. Float64 residuals of the row-sum and floor constraints are numerical diagnostics.

Implementation KL uses the smoothed generator

$$\varphi_{\epsilon_0}(x) = (x + \epsilon_0)\log(x + \epsilon_0) - (x + \epsilon_0),$$

with D' update

$$h_k(\pi) = \exp\left(\frac{\sum_{i,j} \pi_{ik}\pi_{jk}\log(D_{ij} + \epsilon_0)}{s_k^2}\right) - \epsilon_0.$$

This is the minimizer of the stated smoothed generator. The KL smoother is $\epsilon_0 = 10^{-10}$. Raw heat kernels may contain nonpositive entries from floating-point evaluation of $e^{-t_{\mathrm{heat}}L}$ (Amazon fraction ≤ 0 is 0.847; Table 20). The shift $\epsilon_0 = 10^{-10}$ places every entry in $(0, \infty)$, so the KL generator meets $[m, M] \subset (0, \infty)$. The implemented objective and the precomputed Rayleigh bounds are evaluated on the shifted generator domain: $m = \min_{ij}(D_{ij} + \epsilon_0)$ and $M = \max_{ij}(D_{ij} + \epsilon_0)$.

The $\tau = 0$ protocol replaces the truncated projection in Algorithm 1 by the softmax step on $\mathcal{C}$,

$$\pi_{i\cdot}^{q+1} = \mu_i \operatorname{softmax}\big(\log(\max(\pi_{i\cdot}^{q}, 10^{-300})) - \varepsilon^{-1}\nabla_{i\cdot}F_\lambda(\pi^q, D'^q)\big)\,. \tag{17}$$

Gradient clipping $|\varepsilon^{-1}\nabla F| \le 50$ is an optional numerical bound; with it the subtracted term is $\operatorname{clip}(\varepsilon^{-1}\nabla_{i\cdot}F_\lambda(\pi^q, D'^q), -50, 50)$. The six-graph comparison on a matched Stage I continuation is Appendix K.3.

The Table 1 hard-floor protocol uses $\tau = r_\tau \mu_{\min}/K$ with $r_\tau = 10^{-12}$ and no clipping. A smaller r_τ reduces the domain perturbation and worsens the (C2) factor ε/τ (Appendix C). Write $\operatorname{norm}(x) := x/(\mathbf{1}^\top x)$ whenever $\mathbf{1}^\top x > 0$.

The source weights use SpecGWL's fixed exponent $\alpha_{\mathrm{Spec}} = 0.001$; p_t is that method's K-atom target from μ. Throughout, d denotes the diagonal of D', its only learned part in the clustering specialization, while the off-diagonal is $\delta_\varphi(D)$.

Algorithm 2 defines the MixGSP transport initialization and the prototype seed. The MixGSP weights are (w, ρ), while α_D controls prototype seeding. The searched parameters are $w, \rho, \alpha_D, t_{\mathrm{heat}}, \lambda, \varepsilon$; fixed across datasets are α_{Spec}, r_τ, and $n_{\mathrm{inner}} = 4$.

Algorithm 2 Stage I initialization: MixGSP and finite λ-continuation

Require: D, μ, K, $\tau > 0$, λ_{tgt}, n_{cont}, δ_λ, γ_λ, λ_{II}, n_{inner}, MixGSP weights (w, ρ), prototype-seeding weight α_D, column prior $p_t \in \Delta^{K-1}$, and a seed.

1: Construct GWL and SpecGWL soft plans, then set $\widetilde{\pi}^0 \leftarrow \rho(w\pi_{\mathrm{GWL}} + (1-w)\pi_{\mathrm{SpecGWL}}) + (1-\rho)\mu\mathbf{1}_K^\top/K$.
2: Compute $d^0_{\mathrm{data}} \leftarrow \operatorname{norm}(\operatorname{diag}(h(\widetilde{\pi}^0)))$ using equation 7, set $d^0_{\mathrm{prior}} \leftarrow (p_t)_{:,1}$, and form $d^0 \leftarrow \operatorname{norm}(\alpha_D d^0_{\mathrm{data}} + (1-\alpha_D)d^0_{\mathrm{prior}})$.
3: Project each row of $\widetilde{\pi}^0$ onto its μ_i-simplex with lower bound τ, obtaining π^0.
4: **for** $r = 1, \ldots, n_{\mathrm{cont}}$ **do**
5: Set $\lambda_r \leftarrow \delta_\lambda \gamma_\lambda^{r-1}\lambda_{\mathrm{tgt}}$.
6: **for** $s = 1, \ldots, n_{\mathrm{inner}}$ **do**
7: Truncated mirror descent step equation 9 for π on $\widetilde{\mathcal{C}}_\tau$ at fixed d and this λ_r.
8: **end for**
9: Update $D' \leftarrow h(\pi)$, so $d \leftarrow \operatorname{diag}(D')$.
10: **end for**
11: Pass $\lambda = \lambda_{\mathrm{II}}$ to Stage II (λ_{II} need not equal λ_{tgt} or $\lambda_{n_{\mathrm{cont}}}$).

Ensure: (π, D') as the Stage II start, and λ_{II}.

Stage I is a finite λ-continuation of n_{cont} outer blocks at $\lambda_r = \delta_\lambda \gamma_\lambda^{r-1}\lambda_{\mathrm{tgt}}$ for $r = 1, \ldots, n_{\mathrm{cont}}$. In every experiment we set $n_{\mathrm{cont}} = 3$, $\delta_\lambda = 0.05$, $\gamma_\lambda = 1.5$. Stage II then holds $\lambda = \lambda_{\mathrm{II}}$ fixed; λ_{II} need not equal λ_{tgt} or $\lambda_{n_{\mathrm{cont}}}$. Because λ changes across Stage I, it is a finite sequence of warm-started objectives; Theorem 1 applies to the subsequent fixed-λ Stage II dynamics. Table 2 names the protocol rows used later in this appendix.

Table 2: BGW experiment protocols. The column r_τ is the normalized floor ratio $\tau = r_\tau \mu_{\min}/K$; $r_\tau = 0$ is softmax. SH/map/mod = Silhouette / map equation / modularity. The first three rows are Table 1 clustering searches (Stage I/II, search ε, graph-type selectors). The remaining rows are theory checks at fixed λ: prior $\varepsilon = 1.05\overline{L}_{\mathrm{rel},0}$ (Appendix J), recorded ε (landscape), and the paired r_τ check (Appendix K.2). Theorem 1 applies when λ is fixed, $\tau > 0$, and there exists ℓ with $\varepsilon > \ell \ge \sup_t L_t$.

Protocol	r_τ	λ	ε	Selection
Table 1 hard-floor	10^{-12}	Stage I/II	search	SH/map/mod
Table 1 $\tau = 0$	0	Stage I/II	search	SH/map/mod
Table 1 clipped	0, clip	Stage I/II	search	SH/map/mod
Table 16	10^{-12}	fixed	$1.05\overline{L}_{\mathrm{rel},0}$	none
Figure 6 / Table 13	10^{-12}	fixed	recorded	none
Appendix K.2 r_τ	varied	fixed	recorded	none

The three Table 1 clustering rows use 150 outer iterations and four mirror descent steps per outer block. Landscape Amazon uses 20 outer steps at the same inner count (Appendix I); Appendix J uses 20 outer steps and one inner step; the floor sweep uses twenty outer by four inner steps (Appendix K.2). Search ε is the TPE value used in Table 1. Tables 6 and 7 record the pathwise Stage II ratio $R_{\mathrm{II},T}$ and the optional precomputable ray $\varepsilon > \frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})$ (Proposition 6). The locked Table 1 searches report the full trial pool.

The graph-type selector assignment is Table 24. When assigned, Silhouette maximizes heat-kernel Silhouette and holds labels out of the objective. Undirected graphs use kernel PCA of D with Euclidean Silhouette. Directed graphs use ℓ_2-normalized Chung heat-kernel rows with cosine Silhouette. The MixGSP search is the three-plan simplex of GWL, SpecGWL, and the product prior; the spectral weight is zero.

SpecGWL uses the Chowdhury raw t-grid with coverage selection (`graph_partition_gd2`). GWL uses the Chowdhury artifact's `graph_partition_gd` on adjacency. That function does not read its `ot_dict`; it calls `gromov_wasserstein_asym_fixed_initialization` and then `gwa_cg` (product initialization, conditional gradient with Armijo line search, at most 200 iterations).

E.1 Timing and input-coverage details

Each dense clustering gradient costs $\mathcal{O}(N^2K)$ and stores an $N \times N$ relation matrix. Repeated clustering updates with $K = 10$ and five outer by four inner steps take 0.012 seconds at $N = 500$ and 0.137 seconds at $N = 3000$; heat-kernel construction is excluded. The fixed-τ sweep crosses $K \in \{5, 10, 20\}$ and $N \in \{500, 1000, 1500, 2000, 3000\}$ under a fixed five-outer by four-inner budget. All 15 trajectories are objective-monotone and use the prescribed positive floor. These measurements compare solver cost.

F Dataset Provenance and Processing

The four clustering graphs are obtained from public sources and are **not** redistributed with the artifact. Table 3 reports properties of the files actually loaded.

Table 3: Prepared graph files used in the experiments.

Graph	nodes	edges	classes	directed	label interpretation
Amazon	1,501	4,626	12	no	product category
Village	1,991	8,423	12	no	village/community
EU-email	1,005	25,571	42	yes	department
Wikicats	1,998	3,621	15	yes	page category

Table 4 records sources, file pins, and terms. Acquisition follows the public SNAP and SpecGWL sources in Table 4.

Table 4: Verified provenance. SHA-256 prefixes follow the table; full hashes and URLs are in the acquisition manifest.

Graph	Source and file	Terms
EU-email	SNAP email-Eu-core (Leskovec & Krevl, 2014; Yin et al., 2017). File `email-Eu-core.txt.gz`.	SNAP posting; metadata
Amazon	SNAP com-Amazon; SpecGWL 12-cat. subgraph (Chowdhury & Needham, 2021). File `amazon.p`.	SNAP posting; MIT is code-only
Wikicats	SNAP wiki-topcats; SpecGWL 15-cat. subgraph (Chowdhury & Needham, 2021). File `wikicats.p`.	SNAP posting; no page text
Village	Dataverse v9.4 (Banerjee et al., 2013); SpecGWL extract. File `India_database.p`.	original deposit CC0 1.0

SHA-256 prefixes: `email-Eu-core.txt.gz 4b47acdb`; labels `e5abe5b4`; `amazon.p 63999953`; `wikicats.p 105b123a`; `India_database.p 0e8a6ae6`. SpecGWL commit `52b17c43610f` (2021-03-02). Dataverse DOI `10.7910/DVN/U3BIHX`.

EU-email is reconstructed from the SNAP edge and department-label files. Those files match the prepared SpecGWL pickle on the edge set and the $1{,}005$ department labels. Amazon, Wikicats, and Village are fetched from the pinned SpecGWL commit and checked against the SHA-256 values above. The SpecGWL software license is MIT; it does not replace the dataset terms in Table 4. Swiss-roll and S-curve embeddings are generated; digits and breast cancer are the scikit-learn copies of the UCI records. Prepared graph files and the SpecGWL pickles are excluded from the artifact ZIP.

For undirected and "sym" experiments, we construct a symmetric normalized Laplacian after replacing a directed adjacency by $(A + A^\top)/2$ when applicable, then form its heat kernel. For "dir", we use the symmetric Chung directed Laplacian (Chung, 2005), which retains stationary-flow weights while satisfying the symmetric-relation assumption. Labels enter AMI evaluation and the AMI-oracle baseline search. They are held out of the BGW objective and of Silhouette selection. The recorded Infomap EU(dir) and Wiki(dir) scores use one-way edges with the library default undirected flow.

G Hard-Floor Replays and Stage II Descent

This appendix records three Table 1 objects: finite-budget stationarity residuals (Table 5), the recorded Stage II pathwise (C1) ratio $R_{\mathrm{II},T}$ (Table 6), and the optional prior ε of Proposition 6 (Table 7).

Under the Table 1 hard-floor protocol of Appendix E, all six runs fix $\tau = 10^{-12} \min_i \mu_i / K$, use the active-set entropic projection, and disable gradient clipping. Write

$$r_{\mathrm{pg}} = \frac{\|\Pi_{\widetilde{\mathcal{C}}_\tau}(\pi - \nabla_\pi F_\lambda) - \pi\|_F}{\|\nabla_\pi F_\lambda\|_F}, \qquad r_{D'} = \|h(\pi) - D'\|_F.$$

The joint residual is $\sqrt{r_{\mathrm{pg}}^2 + r_{D'}^2}$. A small joint residual is a finite-budget stationarity check of the replayed winner. Table 5 reports the six winner replays.

Table 5: Hard-floor winner replays. The joint residual equals r_{pg} because $r_{D'} = 0$ on every row. The column active is the fraction of entries equal to the floor τ.

Setting	absolute τ	AMI	active	r_{pg}	$r_{D'}$
Amazon	5.55×10^{-17}	0.970	0.845	2.29×10^{-5}	0
Village	4.18×10^{-17}	0.926	0.840	1.27×10^{-6}	0
EU(sym)	2.37×10^{-17}	0.565	0.802	1.55×10^{-2}	0
Wiki(sym)	3.33×10^{-17}	0.530	0.290	8.60×10^{-4}	0
EU(dir)	2.37×10^{-17}	0.485	0.814	1.87×10^{-2}	0
Wiki(dir)	3.33×10^{-17}	0.321	0.000	2.30×10^{-2}	0

Every run attains its prescribed floor and has D' update residual zero, so the joint residual equals r_{pg}. Reported AMI is computed on the partition $\widehat{y}_i = \arg\max_k \pi_{ik}$. Amazon, Village, and Wiki(sym) have $r_{\mathrm{pg}} \le 8.60\times10^{-4}$. EU(sym), EU(dir), and Wiki(dir) retain finite-budget remainders $r_{\mathrm{pg}} = 1.55\times10^{-2}$, 1.87×10^{-2}, and 2.30×10^{-2}. Five rows reproduce the selected AMI exactly. Wiki(dir) reports the replayed endpoint AMI 0.321. All six winner replays have no objective increase above the logging tolerance 10^{-12}. Wiki(dir) leaves the floor inactive.

The Table 1 Stage II check is pathwise (C1) on the recorded budget. Proposition 5 requires $\ell_t < \varepsilon$ at every generated Stage II step. Write $R_{\mathrm{II},T} = \overline{\ell}_{\mathrm{II},T}/\varepsilon$. The inequality $R_{\mathrm{II},T} < 1$ is exactly that finite-budget hypothesis. Table 6 records the six ratios.

Table 6: Hard-floor Stage II pathwise (C1) check. $R_{\mathrm{II},T} = \overline{\ell}_{\mathrm{II},T}/\varepsilon < 1$ on every recorded Stage II budget (Proposition 5).

Setting	λ_{II}	ε	$\overline{\ell}_{\mathrm{II},T}$	$R_{\mathrm{II},T}$
Amazon	0.0203	0.0492	0	0
Village	0.972	0.0972	1.78×10^{-5}	1.83×10^{-4}
EU(sym)	0.00861	2.87×10^{-3}	0	0
Wiki(sym)	0.0112	0.0552	4.74×10^{-5}	8.58×10^{-4}
EU(dir)	0.00696	2.40×10^{-3}	0	0
Wiki(dir)	0.840	0.0433	0	0

Amazon, EU(sym), EU(dir), and Wiki(dir) have a nonpositive quadratic remainder at every moving Stage II iterate, so $R_{\mathrm{II},T} = 0$. Village and Wiki(sym) have $\overline{\ell}_{\mathrm{II},T} = 1.78\times10^{-5}$ and 4.74×10^{-5}, hence $R_{\mathrm{II},T} = 1.83\times10^{-4}$ and 8.58×10^{-4}. All six recorded Stage II budgets therefore satisfy the pathwise hypothesis of Proposition 5.

Proposition 6 supplies an optional prior: any $\varepsilon > \frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})$ places the whole Stage II ray under Corollary 3. Table 7 compares that threshold with the search ε already used in Tables 1 and 6, and with one constructive point $\varepsilon = 1.05\overline{L}_{\mathrm{rel},0}$ on $\mathcal{S}_0 = \{d^0\} \cup [m, M]^K$. D' updates stay in $[\min D, \max D]$ by monotonicity of φ'. On all six graphs the search ε already used in Tables 1 and 6 lies below $\frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})$, so the Table 1 Stage II check is the pathwise ratio.

Table 7: Optional prior ε at the start of Stage II (Proposition 6). $\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})$ is the maximum of the Rayleigh quotient on the theorem domain $[m, M]^K$, after D' updates. Any $\varepsilon > \frac{1}{2}\overline{L}_{\mathrm{rel}}(\lambda_{\mathrm{II}})$ places Stage II under Corollary 3. The prior column is the constructive point $\varepsilon = 1.05\overline{L}_{\mathrm{rel},0}$.

Setting	λ_{II}	$\overline{L}_{\mathrm{rel}}$	$\frac{1}{2}\overline{L}_{\mathrm{rel}}$	search ε	prior ε
Amazon	0.0203	1.37	0.686	0.0492	9.18
Village	0.972	41.1	20.6	0.0972	43.2
EU(sym)	0.00861	15.9	7.97	2.87×10^{-3}	16.7
Wiki(sym)	0.0112	41.8	20.9	0.0552	43.9
EU(dir)	0.00696	10.4	5.19	2.40×10^{-3}	10.9
Wiki(dir)	0.840	0.0894	0.0447	0.0433	0.909

H COMPONENT ABLATION

The MixGSP transport initialization, prototype seeding (Init), and short λ-continuation (Cont) are ablated with three recorded runs per cell. Table 8 reports family means.

Table 8: Component ablation under recorded dataset-specific configurations (hard-partition AMI; three runs per cell).

MixGSP	Init	Cont	Amazon	Village	EU(dir)	Wiki(dir)
✓	✓	✓	0.967	0.833	0.546	0.323
✓	✓		0.784	0.828	0.459	0.323
✓		✓	0.586	0.901	0.548	0.323
✓			0.880	0.894	0.460	0.323
	✓	✓	0.000	0.000	0.012	0.123
	✓		0.000	0.000	0.000	0.123
		✓	0.000	0.000	0.012	0.096
			0.000	0.255	0.000	0.029

MixGSP is required for AMI above chance on Amazon, Village, and EU(dir): every MixGSP-off cell is at most 0.123, except the Village all-off cell at 0.255. On Amazon, with MixGSP on, adding Init alone gives 0.784 and adding Cont alone gives 0.586, both below the MixGSP-only cell at 0.880; using both reaches 0.967. Wiki(dir) stays at 0.323 whenever MixGSP is on. The retained pipeline is MixGSP+Init+Cont. The ablation configurations were selected by AMI. Three repeated executions per cell have identical stored outputs up to recorded precision; the table reports the cell mean.

I INITIALIZATION DEPENDENCE AND LANDSCAPE EXPERIMENTS

I.1 TRANSPORT INITIALIZATION AND PROTOTYPE ADAPTATION

We first separate the effect of the initial transport from subsequent adaptation of the diagonal prototype relation. All comparisons use the same hard-floor objective, prescribed outer budget, and solver settings; there is no Stage I continuation or residual-based early stopping.

For a fixed transport start, we compare three prototype variants. FIXED keeps the initial diagonal fixed throughout the run. SHARED learns one common diagonal value across prototypes. DIAGONAL is the clustering update of equation (7), with one learned d_k per prototype.

Conversely, within each prototype variant we compare GWL, SpecGWL, spectral, and MixGSP starts. The outer budget is $J = 20$ on Amazon and $J = 150$ on Village/Wiki(dir), with $n_{\text{inner}} = 4$. MixGSP is the initialization of Section 4.2; no prototype-column alignment is applied before its addition. Spectral denotes the sklearn spectral-clustering start.

Table 9: Terminal AMI on one hard-floor F_λ after the prescribed outer budget. Ground-truth labels are used only for scoring. Diagonal is the clustering update equation 7.

Graph	π^0	Fixed	Shared	Diagonal
Amazon	GWL	0.4085	0.4061	0.4085
Amazon	SpecGWL	0.7910	0.8748	0.7910
Amazon	spectral	0.9141	0.9149	0.9175
Amazon	MixGSP	0.4056	0.4056	0.4088
Village	GWL	0.6773	0.6718	0.6773
Village	SpecGWL	0.7652	0.7648	0.7652
Village	spectral	0.7484	0.7505	0.7484
Village	MixGSP	0.7648	0.7648	0.7648
Wiki(dir)	GWL	0.2611	0.2291	0.2546
Wiki(dir)	SpecGWL	0.2108	0.2108	0.2108
Wiki(dir)	spectral	0.0980	0.0980	0.0980
Wiki(dir)	MixGSP	0.3244	0.3243	0.3244

Using the product norm $\|\cdot\|_\times$ of Appendix A, define

$$d_A(z, \widehat{z}) = \min_{\sigma} \|z^\sigma - \widehat{z}\|_\times,$$

and

$$\Delta_{\text{init}} = d_A\big(z^J(\pi_A^0), z^J(\pi_B^0)\big)$$

for a common prototype variant, while

$$\Delta_{\text{proto}} = d_A\big(z^J_{\text{fixed}}(\pi^0), z^J_{\text{diag}}(\pi^0)\big)$$

for a common transport start. Write $\mathcal{P}_K$ for the set of $K \times K$ permutation matrices. For transport plans, define

$$d_A^\pi(\pi, \widehat{\pi}) := \min_{P \in \mathcal{P}_K} \|\pi - \widehat{\pi} P\|_F, \qquad \text{relF}_A(\pi, \widehat{\pi}) := \frac{d_A^\pi(\pi, \widehat{\pi})}{\|\pi\|_F}.$$

The Hungarian adjusted Rand index (ARI) of the hard assignments does not use ground-truth labels.

Table 10: Largest aligned endpoint gaps on the Diagonal arm (Δ_{init}) against Fixed–Diagonal (Δ_{proto}) on the same π^0. relF denotes relF_A. Nearby in the main text is this aligned relative Frobenius distance.

Graph	pair	Δ_{init} d_A	relF(π)	hard-label disc.	Δ_{proto} d_A (relF, ARI)
Amazon	GWL–SpecGWL	0.0447	1.113	0.620	0.0160 (0.171, 0.971; spectral)
Village	SpecGWL–spectral	0.0330	1.068	0.571	0.0061 (0.005, 1.000; GWL)
Wiki(dir)	SpecGWL–spectral	0.0327	1.327	0.880	0.0044 (0.181, 0.987; GWL)

On Amazon, GWL and MixGSP remain close on the Diagonal arm ($d_A = 0.0023$, relF 0.089, hard-label disagreement 0.004). The largest Shared d_A is 0.0182 on Amazon SpecGWL (ARI 0.869, $\Delta\text{AMI} = +0.084$).

To distinguish a nearly shared learned prototype from genuine cluster-dependent adaptation, we report

$$\text{CV}(d) = \frac{\text{std}_k(d_k)}{\text{mean}_k(d_k)}.$$

Shared has $\text{CV}(d) = 0$ by construction. Table 11 compares each learned arm to Fixed on the same π^0.

Table 11: Prototype adaptation versus assignment change. relF denotes relF_A. ARI and relF compare Fixed to the learned arm after Hungarian alignment, without ground-truth labels. ΔAMI is only a score against ground truth.

Graph	π^0	learned	CV(d)	relF(π)	ARI	ΔAMI
Amazon	GWL	Diagonal	3.316	0.000	1.000	+0.000
Amazon	GWL	Shared	0	0.052	0.997	−0.002
Amazon	SpecGWL	Diagonal	0.562	0.000	1.000	+0.000
Amazon	SpecGWL	Shared	0	0.379	0.869	+0.084
Amazon	spectral	Diagonal	0.545	0.171	0.971	+0.003
Amazon	MixGSP	Diagonal	3.316	0.073	0.994	+0.003
Village	GWL	Diagonal	1.742	0.005	1.000	+0.000
Village	MixGSP	Diagonal	0	0.000	1.000	+0.000
Wiki(dir)	GWL	Diagonal	1.030	0.181	0.987	−0.006
Wiki(dir)	MixGSP	Diagonal	1.039	0.032	0.999	+0.000

Two mechanisms appear. On Village MixGSP, $\text{CV}(d) = 0$, so the learned diagonal is a shared scale. On Amazon GWL and SpecGWL, the per-cluster d_k adapt ($\text{CV}(d) = 3.316$ and 0.562) while the Fixed–Diagonal transport and hard assignment remain unchanged to recorded precision ($\text{relF}_A = 0$, ARI $= 1$). The second mechanism is the main finding: prototype adaptation need not move the finite-budget assignment.

I.2 CONTROLLED PERTURBATIONS OF STRUCTURED STARTS

To probe a structured start under a prescribed horizon J, form

$$\pi_\alpha = (1-\alpha)\pi^0 + \alpha\pi_{\mathrm{prod}},$$

project onto $\widetilde{\mathcal{C}}_\tau$, and recompute the Diagonal $D'^0(\alpha)$. The grid is $\alpha \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 0.05, 0.1, 0.3, 1\}$. Write $x = d_A(z^0_\alpha, z^0)$ and $y = d_A(z^J_\alpha, z^J)$. The reference starts are GWL, SpecGWL, and spectral, all on the Diagonal arm.

Table 12: Controlled product-plan perturbations of structured starts. Small α leave y small; $\alpha = 1$ is the product plan.

Graph	π^0	y ($\alpha \le 10^{-2}$)	hard-label disc. ($\alpha = 0.3$)	hard-label disc. ($\alpha = 1$)
Amazon	GWL	0.0014	0.001	0.917
Amazon	SpecGWL	0.0068	0.027	0.917
Amazon	spectral	0.0000	0.011	0.905
Village	GWL	0.0028	0.005	0.913
Village	SpecGWL	0.0002	0.001	0.916
Village	spectral	0.0001	0.001	0.755
Wiki(dir)	GWL	0.0001	0.000	0.927
Wiki(dir)	SpecGWL	0.0000	0.000	0.933
Wiki(dir)	spectral	0.0000	0.000	0.366

For $\alpha \le 10^{-2}$, the terminal displacement y remains small and the hard assignment is essentially unchanged. At $\alpha = 1$, the hard-label disagreement is about 0.9 on most Amazon, Village, and Wiki(dir) GWL/SpecGWL pairs (Figure 5). These observations are consistent with Proposition 2 and Corollary 4.

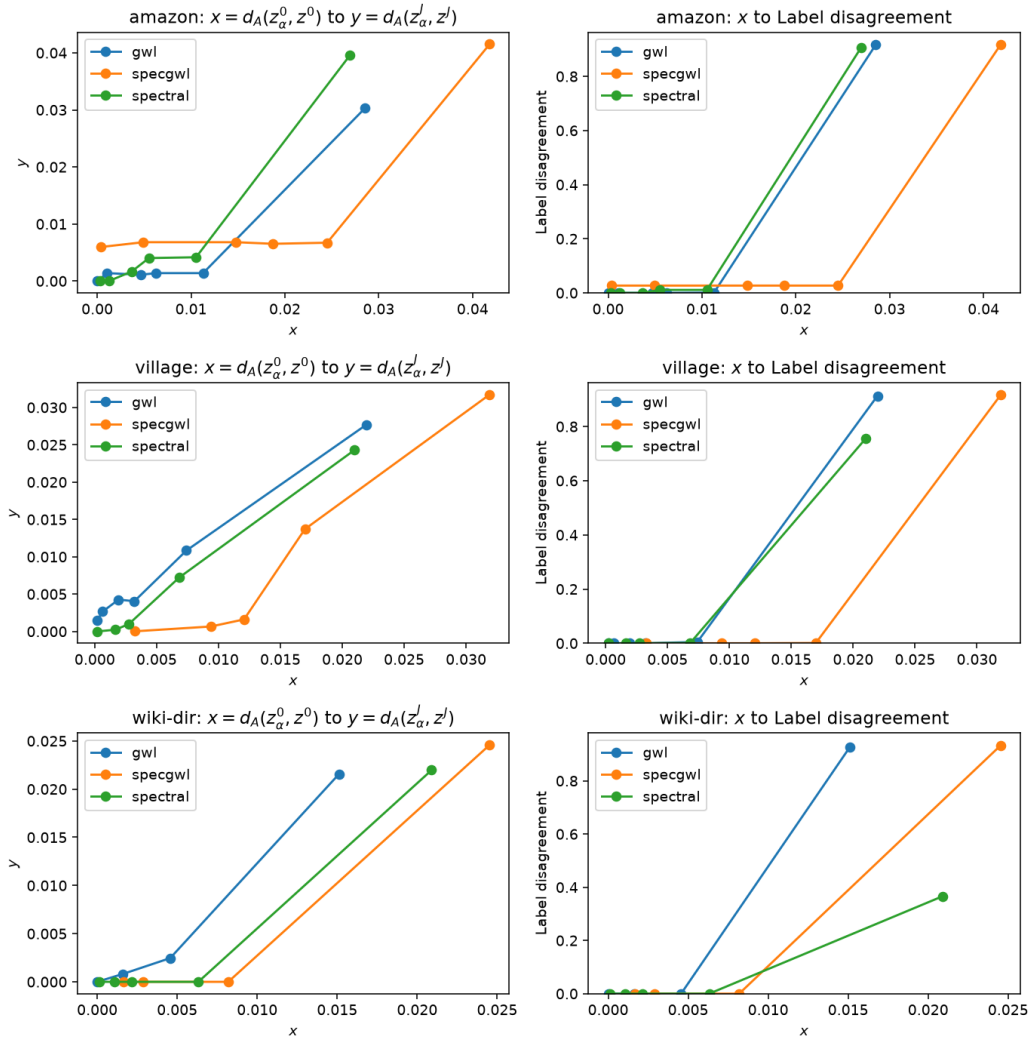

Figure 5: Product-plan perturbations of structured starts on the Diagonal arm. Left: initial aligned gap $x = d_A(z^0_\alpha, z^0)$ against terminal gap $y = d_A(z^J_\alpha, z^J)$. Right: x against hard-label disagreement. Small x leave y and the hard assignment nearby; the product plan ($\alpha = 1$) flips most labels.

I.3 Start-lock families and endpoint groups

The hard-floor initialization study uses the saved thirty-seed record. The solver uses the active-set entropic projection, $\tau = 10^{-12} \min_i \mu_i / K$, no gradient clip, $D' = h(\pi^0)$, and a fixed λ with no hybrid continuation. Stopping is the recorded outer budget: Amazon 20 steps, Village and Wiki(dir) 150. Every initialization family uses the same floor, projection, λ, ε, and stopping rule; only the initial plan changes.

Families construct π^0 as follows, then project onto $\widetilde{\mathcal{C}}_\tau$. Mixed+noise adds Dirichlet row noise of level $\mathrm{Unif}(0.05, 0.2)$ to the GWL/SpecGWL mix. GWL-as-π^0 applies the same noise to the adjacency GWL coupling ($w = 1, \rho = 1$), while SpecGWL-as-π^0 applies it to the adaptive-t SpecGWL coupling ($w = 0, \rho = 1$). Spectral uses spectral clustering on the symmetrized cost and softens the hard assignment by 0.05. k-means++ assigns on the recorded embedding with the same softening. $\mathrm{Dir}(\alpha)$ draws independent Dirichlet($\alpha \mathbf{1}_K$) rows and rescales each row to μ_i. Logit-normal draws independent Gaussian logits with $\sigma \in \{0.5, 1, 2\}$, applies softmax, and rescales. Stick-breaking uses $\mathrm{Beta}(1, 1)$ weights and rescales. Each sampled family has thirty seeds; the product plan is a

single deterministic reference and is excluded from tests. Dirichlet(10) and stick-breaking are additional row-stochastic samplers beyond Table 13. Kruskal–Wallis (Kruskal & Wallis, 1952) and Mann–Whitney (Mann & Whitney, 1947) p-values are not pooled across graphs.

Table 13: Hard-floor initialization study with thirty seeds per sampled family; only π^0 changes. Bootstrap intervals, residuals, and additional families are reported in Table 14.

Family	Amazon	Village	Wiki(dir)
Mixed+noise	0.409	0.765	0.323
GWL-as-π^0	0.442	0.699	0.329
SpecGWL-as-π^0	0.812	0.748	0.226
Spectral	0.917	0.768	0.106
k-means++	0.895	0.322	0.007
Dir(1)	0.000	0.000	0.000
Logit-normal	0.000	0.001	0.000
Dir(0.1)	0.000	0.000	0.000

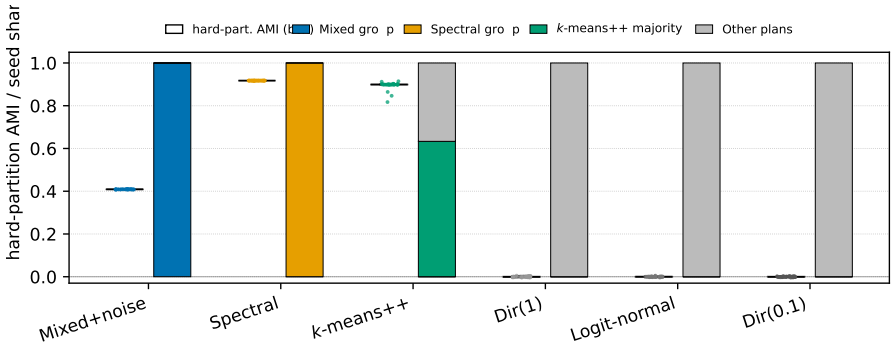

Figure 6: Start-lock on Amazon under one hard-floor F_λ. Structured initialization families concentrate around distinct finite-budget solutions after prototype alignment, whereas unstructured starts remain near chance. Full statistics and additional initialization families are reported in Table 14.

Table 14: Hard-floor initialization families (thirty seeds; product $n = 1$). AMI is computed on the hard partition. Success is $\Pr(\mathrm{AMI} \geq 0.5)$ with Wilson 95% confidence intervals (Wilson, 1927). Residuals are median joint projected-gradient residuals; floor is the fraction of entries equal to τ at the endpoint.

Graph	family	AMI	F_λ	success	residual	floor
Amazon	Mixed+noise	0.409	-3.74×10^{-4}	0 [0.00, 0.11]	3.1×10^{-5}	0.917
Amazon	GWL-as-π^0	0.442	-3.74×10^{-4}	0 [0.00, 0.11]	3.7×10^{-5}	0.917
Amazon	SpecGWL-as-π^0	0.812	-7.15×10^{-4}	1 [0.89, 1]	9.0×10^{-5}	0.915
Amazon	Spectral	0.917	-8.01×10^{-4}	1 [0.89, 1]	3.6×10^{-4}	0.915
Amazon	k-means++	0.895	-7.57×10^{-4}	1 [0.89, 1]	5.9×10^{-4}	0.915
Amazon	$\mathrm{Dir}(1)$	0.000	-3.33×10^{-4}	0 [0.00, 0.11]	1.6×10^{-16}	0.917
Amazon	Logit-normal	0.000	-3.33×10^{-4}	0 [0.00, 0.11]	1.6×10^{-16}	0.917
Amazon	$\mathrm{Dir}(0.1)$	0.000	-3.33×10^{-4}	0 [0.00, 0.11]	1.6×10^{-16}	0.917
Amazon	$\mathrm{Dir}(10)$	0.000	-3.33×10^{-4}	0 [0.00, 0.11]	1.6×10^{-16}	0.917
Amazon	Stick-breaking	0.000	-3.33×10^{-4}	0 [0.00, 0.11]	1.6×10^{-16}	0.917
Amazon	Product	0.000	-2.78×10^{-5}	0	4.1×10^{-16}	0.000
Village	Mixed+noise	0.765	-6.88×10^{-6}	1 [0.89, 1]	0.72	0.000
Village	GWL-as-π^0	0.699	-1.60×10^{-4}	1 [0.89, 1]	2.7×10^{-3}	0.222
Village	SpecGWL-as-π^0	0.748	-2.39×10^{-4}	1 [0.89, 1]	2.1×10^{-3}	0.419
Village	Spectral	0.768	-1.12×10^{-4}	1 [0.89, 1]	9.2×10^{-4}	0.172
Village	k-means++	0.322	-8.84×10^{-5}	0 [0.00, 0.11]	5.3×10^{-4}	0.161
Village	$\mathrm{Dir}(1)$	0.000	-6.81×10^{-6}	0 [0.00, 0.11]	0.67	0.000
Village	Logit-normal	0.001	-9.78×10^{-6}	0 [0.00, 0.11]	0.73	0.000
Village	$\mathrm{Dir}(0.1)$	0.000	-2.29×10^{-5}	0 [0.00, 0.11]	0.60	0.052
Village	$\mathrm{Dir}(10)$	0.000	-3.70×10^{-6}	0 [0.00, 0.11]	0.31	0.000
Village	Stick-breaking	0.000	-2.29×10^{-5}	0 [0.00, 0.11]	0.61	0.000
Village	Product	0.000	-3.35×10^{-6}	0	1.1×10^{-15}	0.000
Wiki(dir)	Mixed+noise	0.323	-3.11×10^{-4}	0 [0.00, 0.11]	5.9×10^{-11}	0.931
Wiki(dir)	GWL-as-π^0	0.329	-6.65×10^{-4}	0 [0.00, 0.11]	6.4×10^{-17}	0.457
Wiki(dir)	SpecGWL-as-π^0	0.226	-7.04×10^{-4}	0 [0.00, 0.11]	8.1×10^{-17}	0.491
Wiki(dir)	Spectral	0.106	-4.25×10^{-4}	0 [0.00, 0.11]	5.1×10^{-17}	0.933
Wiki(dir)	k-means++	0.007	-4.31×10^{-4}	0 [0.00, 0.11]	3.0×10^{-8}	0.931
Wiki(dir)	$\mathrm{Dir}(1)$	0.000	-3.02×10^{-4}	0 [0.00, 0.11]	2.2×10^{-6}	0.668
Wiki(dir)	Logit-normal	0.000	-3.02×10^{-4}	0 [0.00, 0.11]	2.5×10^{-7}	0.616
Wiki(dir)	$\mathrm{Dir}(0.1)$	0.000	-3.01×10^{-4}	0 [0.00, 0.11]	1.9×10^{-10}	0.925
Wiki(dir)	$\mathrm{Dir}(10)$	0.001	-3.01×10^{-4}	0 [0.00, 0.11]	1.3×10^{-2}	0.004
Wiki(dir)	Stick-breaking	0.010	-4.30×10^{-4}	0 [0.00, 0.11]	5.3×10^{-3}	0.922
Wiki(dir)	Product	0.000	-3.15×10^{-5}	0	3.3×10^{-16}	0.000

On Amazon, every Holm-adjusted Mann–Whitney comparison (Holm, 1979; Mann & Whitney, 1947) between a structured family (mixed, spectral, k-means++) and an unstructured family, and every comparison among the three structured families, has $p < 10^{-9}$. The three unstructured core families are not separated from one another (Holm $p = 1$ for AMI and for F_λ).[1]

The prototype labels of π are unlabeled. A label alignment of two endpoint plans is a permutation matrix $P \in \mathcal{P}_K$. Because

$$\|\pi - \widehat{\pi}P\|_F = \|\pi P^{-1} - \widehat{\pi}\|_F,$$

d_A^π is symmetric up to exchanging the aligned arguments. The denominator of relF_A is the first argument. Two recorded endpoints are grouped when $\mathrm{relF}_A(\pi, \widehat{\pi})$ is at most five times the median within-family pairwise value, clipped to $[0.02, 0.25]$ (0.25 on Amazon). An endpoint group is one such group of recorded plans.

On Amazon, all thirty Mixed+noise seeds occupy one endpoint group and all thirty spectral seeds another; $19/30$ k-means++ seeds share a third group, while each of the 150 unstructured seeds

[1] Core-family Kruskal–Wallis H (Kruskal & Wallis, 1952) on AMI is 154.3 (Amazon), 151.4 (Village), and 152.1 (Wiki(dir)), each with $p < 10^{-30}$. Unstructured AMI is identically 0, so those H values are inflated by zero-variance cells.

forms a singleton group. Amazon has 161 groups in total, 156 of them singletons. Under the same rule, SpecGWL-as-π^0 retains $30/30$ seeds and GWL-as-π^0 retains $23/30$ at the 0.02 clip floor, with pairwise partition ARI 1. Village retains $30/30$ for GWL-as-π^0 and $24/30$ for SpecGWL-as-π^0; Wiki(dir) retains $30/30$ and $26/30$, respectively (ARI 0.996). EU(sym), Wiki(sym), and EU(dir) lock $30/30$ at AMI 0.469–0.488.

I.4 Interpolating paths between structured endpoints

Here *mixed* denotes the Mixed+noise family of Section I.3. The path letter is η (the interpolating path on F_λ). The Stage I multiplier remains γ_λ (Algorithm 2). On the same fixed F_λ as Figure 6, three interpolating paths join each pair of structured-family endpoints (Draxler et al., 2018; Frankle et al., 2020; Garipov et al., 2018). A path $\eta : [0,1] \to \{(\pi, D')\}$ between endpoints z_a, z_b has barrier

$$B(\eta) = \max_{s\in[0,1]} F_\lambda\big(\eta(s)\big) - \max\big\{F_\lambda(z_a), F_\lambda(z_b)\big\}$$

and relative height

$$R(\eta) = \frac{B(\eta)}{\max\big\{|F_\lambda(z_a)|, |F_\lambda(z_b)|\big\}}.$$

The affine-π path is

$$\eta^\pi(s) = \big((1-s)\pi_a + s\tilde{\pi}_b,\, h\big((1-s)\pi_a + s\tilde{\pi}_b\big)\big),$$

with $\tilde{\pi}_b = \pi_b P$ for the label-alignment permutation P that realises $d_A^\pi(\pi_a, \pi_b)$. The joint path is the product-space segment

$$\eta^{\text{joint}}(s) = \big((1-s)\pi_a + s\tilde{\pi}_b,\, (1-s)D'_a + s\widetilde{D}'_b\big),$$

so the prototype is the linear blend of the endpoint prototypes. The row-wise e-geodesic (Amari, 1995) interpolates the endpoint log-probabilities, restores the row masses μ, and then sets $D' = h(\pi)$. Figure 4 reports $R(\eta)$. Table 15 records the medoid $B(\eta)$, with the matched-pair median in parentheses.

Table 15: Medoid path barrier $B(\eta)$; parentheses report the median over five seed pairs matched by index.

Graph	Pair	affine π, $D' = h(\pi)$	joint (π, D')	KL geodesic
Amazon	mixed–spectral	1.02×10^{-4} (1.02×10^{-4})	1.06×10^{-3} (1.06×10^{-3})	1.01×10^{-4} (1.01×10^{-4})
	mixed–k-means++	1.21×10^{-4} (1.21×10^{-4})	1.24×10^{-3} (1.24×10^{-3})	1.21×10^{-4} (1.21×10^{-4})
	spectral–k-means++	1.37×10^{-5} (1.37×10^{-5})	1.37×10^{-5} (1.53×10^{-5})	7.56×10^{-6} (7.56×10^{-6})
Village	mixed–spectral	0 (0)	4.45×10^{-4} (7.96×10^{-4})	0 (0)
	mixed–k-means++	1.47×10^{-7} (9.85×10^{-8})	3.74×10^{-4} (3.80×10^{-4})	0 (0)
	spectral–k-means++	2.23×10^{-5} (5.78×10^{-5})	9.02×10^{-4} (4.59×10^{-4})	2.00×10^{-6} (1.94×10^{-6})
EU(sym)	mixed–spectral	5.56×10^{-4} (5.56×10^{-4})	1.13×10^{-3} (1.28×10^{-3})	5.56×10^{-4} (5.56×10^{-4})
	mixed–k-means++	9.97×10^{-4} (9.76×10^{-4})	1.57×10^{-2} (1.56×10^{-2})	9.97×10^{-4} (9.76×10^{-4})
	spectral–k-means++	8.48×10^{-4} (7.74×10^{-4})	1.30×10^{-2} (1.25×10^{-2})	8.48×10^{-4} (7.74×10^{-4})
Wiki(sym)	mixed–spectral	1.27×10^{-3} (1.28×10^{-3})	1.27×10^{-3} (1.28×10^{-3})	1.27×10^{-3} (1.28×10^{-3})
	mixed–k-means++	1.27×10^{-3} (1.27×10^{-3})	1.98×10^{-1} (1.89×10^{-1})	1.27×10^{-3} (1.27×10^{-3})
	spectral–k-means++	1.49×10^{-4} (1.55×10^{-4})	4.25×10^{-3} (2.96×10^{-3})	1.49×10^{-4} (1.55×10^{-4})
EU(dir)	mixed–spectral	1.56×10^{-5} (1.60×10^{-5})	2.08×10^{-4} (1.99×10^{-4})	7.05×10^{-6} (6.65×10^{-6})
	mixed–k-means++	2.39×10^{-5} (2.43×10^{-5})	1.28×10^{-2} (1.28×10^{-2})	6.25×10^{-6} (7.80×10^{-6})
	spectral–k-means++	4.21×10^{-5} (4.23×10^{-5})	1.96×10^{-2} (1.99×10^{-2})	2.75×10^{-5} (2.79×10^{-5})
Wiki(dir)	mixed–spectral	1.02×10^{-4} (1.01×10^{-4})	1.02×10^{-4} (1.02×10^{-4})	9.55×10^{-5} (9.51×10^{-5})
	mixed–k-means++	9.82×10^{-5} (1.10×10^{-4})	5.49×10^{-4} (2.69×10^{-4})	8.77×10^{-5} (1.03×10^{-4})
	spectral–k-means++	2.91×10^{-5} (6.43×10^{-5})	9.04×10^{-5} (1.11×10^{-4})	2.65×10^{-5} (4.91×10^{-5})

EU(sym) structured AMI medians are 0.478 (mixed), 0.644 (spectral), and 0.303 (k-means++), with joint residuals of order 10^{-17}. Wiki(sym) medians are 0.259, 0.119, and 0.056, residuals of order 10^{-17}. EU(dir) medians are 0.473, 0.549, and -0.003, with mixed residual 3.39×10^{-3} and spectral residual 2.07×10^{-3}. Village mixed has residual 0.72, substantially larger than the other reported endpoint residuals. A positive $B(\eta)$ rules out a downhill segment along the corresponding tested

path. Amazon Mixed+noise–spectral has positive barriers along all three tested paths, whereas the Village Mixed+noise–spectral affine-π path and KL geodesic have zero barrier, which licenses the hatching in Figure 4. A parenthetical value equal to the medoid entry means that all five matched seed pairs agree. As a numerical sanity check, a constant interpolant $\eta(s) = z_a$ returns $B = 8.7\times10^{-19}$, below the smallest reported nonzero barrier. On the symmetric graphs, affine-π and the KL geodesic agree to six digits at $s = 0.5$ and differ by up to 45% off that midpoint.

J Sufficient Descent Validation under Precomputed Conditions

This experiment evaluates the precomputed sufficient-descent condition on Amazon and Village with three seeds each. It fixes $r_\tau = 10^{-12}$, holds the target λ fixed, uses one mirror descent step per outer block, and disables clipping and continuation. $\overline{L}_{\mathrm{rel},0}$ is the prior maximum of the positive-part Rayleigh quotient of H_λ on $\mathcal{S}_0 = \{d^0\} \cup [m, M]^K$ at that fixed target λ. The protocol sets $\varepsilon = 1.05\overline{L}_{\mathrm{rel},0}$. D' updates remain in $[m, M]^K$, so the restriction of that maximum to the box covers every D' update. The recorded L_t along each trajectory is the same Rayleigh quotient evaluated at d^t. The shifted smoothed-KL domain in Appendix E handles nonpositive heat-kernel entries; the corresponding unsmoothed KL operator-norm bound is not finite for the implemented objective. Table 16 records the protocol.

Table 16: Optional prior ε sufficient-decrease check on Amazon and Village (three seeds each), 20 mirror descent steps with no clipping, and fixed target λ. The mirror descent step is $1/\varepsilon$. This choice of ε is sufficient on the recorded budget.

Setting	$\overline{L}_{\mathrm{rel},0}$	ε	AMI mean	residual max	$\min_t(\varepsilon - L_t)$
Amazon	27.73	29.12	0.406	0.456	29.12
Village	41.14	43.19	0.765	0.681	43.19

All three Amazon seeds and all three Village seeds are objective-monotone. This $\varepsilon = 1.05\overline{L}_{\mathrm{rel},0}$ choice validates the sufficient-decrease inequality on a 20-step budget.

K Generator Scale, Floor Continuity, and the Optional Clip

K.1 Generator ranking follows heat-kernel scale

Table 1 uses KL. The same hard-floor solver was also run with the generators in Table 17: squared Euclidean (L_2), Itakura–Saito (IS), and the Beta family. On the Beta family, L_2, KL, and IS are the members at $\beta = 2$, $\beta = 1$, and $\beta = 0$ (Cichocki & Amari, 2010), with $\varphi''(x) = x^{\beta-2}$.

Table 17: Generators used in the hard-floor screen.

Family	generator $\varphi(x)$	update mean
L_2	$\frac{1}{2}x^2$	arithmetic
KL	$x \log x - x$	geometric
IS	$-\log x$	harmonic
Beta	$\frac{x^\beta - \beta x + \beta - 1}{\beta(\beta-1)}$	power

Replacing only φ changes the recovered partition when the pairwise geometry changes. Figure 7 uses three constructed kernels ($N = 90$, $K = 3$): the additive block kernel, called ℓ_2-signal in the figure, and two multiplicative kernels, called IS-signal at $10\times$ and $100\times$.

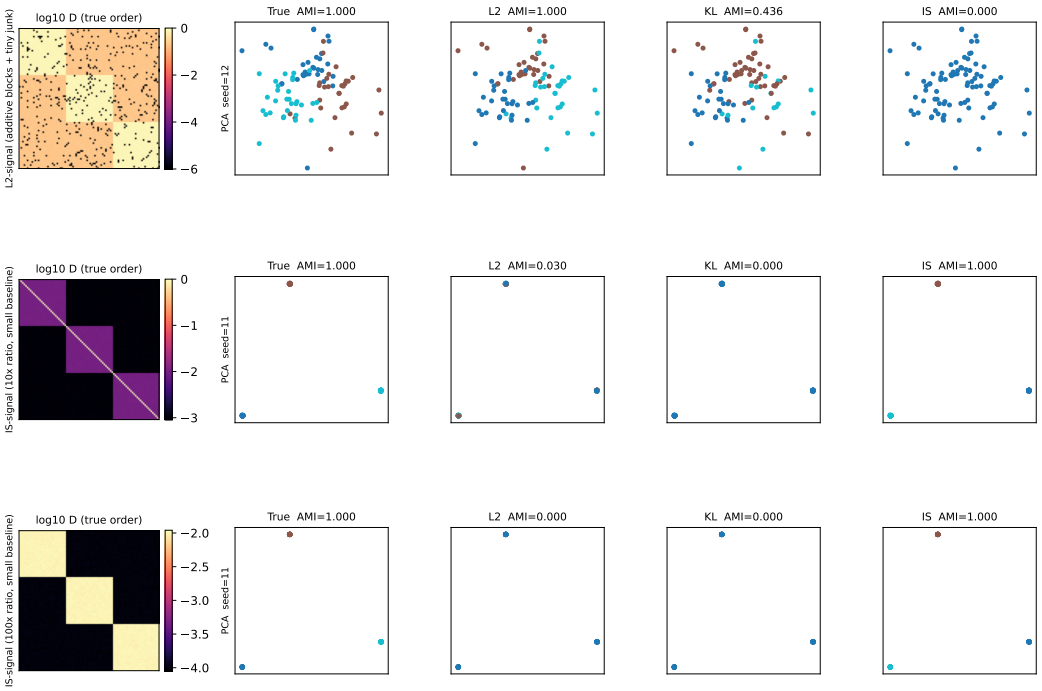

Figure 7: Additive versus relative pairwise geometry changes which generator recovers the labels. On the ℓ_2-signal kernel the displayed L_2 seed has AMI 1 and IS has AMI 0; on both IS-signal kernels the displayed IS seed has AMI 1 while L_2 stays at most 0.030.

On the additive block kernel the displayed L_2 seed has AMI 1 and the IS seed has AMI 0; on either multiplicative kernel the displayed IS seed has AMI 1 and L_2 stays at most 0.030. A single seed can be lucky. Table 18 therefore records twelve-seed means.

Table 18: Mean AMI of the partition $\widehat{y}_i = \arg\max_k \pi_{ik} \pm$ sample standard deviation over twelve Dirichlet seeds on constructed kernels. Additive ℓ_2-signal: L_2 0.615, IS 0. Multiplicative IS-signal: IS 0.167, L_2 at most 0.005.

Kernel	L_2	KL	IS
ℓ_2-signal	0.615 ± 0.388	0.310 ± 0.249	0.000 ± 0.000
IS-signal ($10\times$)	0.004 ± 0.011	0.000 ± 0.000	0.167 ± 0.389
IS-signal ($100\times$)	0.005 ± 0.012	0.083 ± 0.289	0.167 ± 0.389

The twelve-seed means keep the same ranking: L_2 AMI 0.615 against IS AMI 0 on the additive kernel, and IS AMI 0.167 against L_2 AMI at most 0.005 on the multiplicative kernels. The recovered partition therefore tracks additive versus relative pairwise geometry.

Those AMI gaps follow how each cost weights a relative difference at different scales of D. For every generator in Table 17,

$$B_\varphi(a,b) = (a-b)^2 \int_0^1 (1-t)\,\varphi''(b+t(a-b))\,dt,$$

so the local weight is $\varphi''(b)$, and on the Beta family that weight is $b^{\beta-2}$. Table 19 holds the relative gap at 0.1 and changes only the scale of D.

Table 19: The same relative gap 0.1 yields costs in $[0.004690, 0.005000]$ at scale 1, and splits from $5.000 \cdot 10^{-11}$ (L_2) to 0.004690 (IS) at scale 10^{-4} (Cichocki & Amari, 2010).

(D', D)	L_2	KL	$\beta = 0.5$	IS
$(1.1, 1)$	0.005000	0.004841	0.004765	0.004690
$(1.1 \cdot 10^{-4}, 10^{-4})$	$5.000 \cdot 10^{-11}$	$4.841 \cdot 10^{-7}$	$4.765 \cdot 10^{-5}$	0.004690

At scale 1 the four costs lie in $[0.004690, 0.005000]$. At scale 10^{-4}, L_2 is $5.000 \cdot 10^{-11}$, KL is $4.841 \cdot 10^{-7}$, $\beta = 0.5$ is $4.765 \cdot 10^{-5}$, and IS remains 0.004690. The same relative gap therefore weights the generators together at scale 1 and apart at scale 10^{-4}. The six benchmark heat kernels occupy that window (Table 20).

Table 20: Benchmark heat kernels span the scale range that separates the generators. Undirected log-span is 14.79–31.94; directed Chung log-span is 1.78 and 1.50. Log-span uses positive entries only, $\log_{10}(q_{0.99}/q_{0.01})$, and the two fraction columns are over all kernel entries.

Dataset	N	nonpositive	log-span$^+$	below 10^{-4}
Amazon	1501	0.847	17.06	0.855
Village	1991	0.912	15.16	0.916
EU(sym)	1005	0.019	14.79	0.518
Wiki(sym)	1998	0.686	31.94	0.937
EU(dir)	1005	0.000	1.78	0.004
Wiki(dir)	1998	0.000	1.50	0.359

Undirected kernels have positive-entry log-span from 14.79 to 31.94 and a large mass below 10^{-4}. Directed Chung kernels are strictly positive, with log-span 1.78 and 1.50. The six kernels therefore sit in the scale window that separates the generators. Replacing only φ on these kernels is a scale test of the same mechanism (Table 21).

Table 21: With only φ replaced, AMI ranking changes with the graph. Amazon KL/IS 0.970 versus L_2 0.791; Village $\beta = 0.5$ 0.120 versus KL 0.927; Wiki(dir) L_2 0.308 versus KL 0.260.

Dataset	L_2	KL	IS	$\beta = 0.5$	$\beta = 1.5$
Amazon	0.791	0.970	0.970	0.970	0.952
Village	0.755	0.927	0.927	0.120	0.875
EU(sym)	0.478	0.637	-0.004	—	—
Wiki(sym)	0.446	0.612	0.590	—	—
EU(dir)	0.447	0.541	—	0.551	—
Wiki(dir)	0.308	0.260	—	—	—

Amazon: KL and IS 0.970, L_2 0.791. Village: KL and IS 0.927, $\beta = 0.5$ 0.120. Wiki(dir): L_2 0.308, KL 0.260. The ranking therefore changes with the graph under a common Silhouette screen. EU(sym) IS reaches Silhouette 0.964 with AMI -0.004 on 20 nonempty clusters; most mass sits in one cluster, so a high Silhouette can accompany chance-level AMI. Directed IS and Wiki(dir) $\beta = 0.5$ have no recorded trial.

K.2 CONTROLLED FLOOR SWEEP

Proposition 1 concerns a finite horizon J of outer blocks at a *fixed* floor: from the same positive start, the truncated iterates and the softmax iterates satisfy $\max_{0\le j\le J}\|\pi_\tau^j - \pi_{\mathrm{soft}}^j\|_F \le C_J\tau$. The numerical sweep varies τ; each paired comparison is one fixed-floor instance of that bound. Where the floor is active, the paired discrepancy at outer step 20 tracks r_τ.

The numerical check holds λ and ε fixed, uses twenty outer blocks and four inner updates, and omits Stage I continuation. Both arms start from the same pair $(\pi^0, h(\pi^0))$ with a 5% product prior. One arm uses the positive floor $\tau = r_\tau \min_i \mu_i/K$; the other uses the softmax step on $\mathcal{C}$. Table 22 reports the relative discrepancy at outer step 20, over thirty paired seeds. The same records are Figure 3 in Section 3.2.

Table 22: Median (q95) of $\|\pi_\tau^{20} - \pi_{\mathrm{soft}}^{20}\|_F/\|\pi_{\mathrm{soft}}^{20}\|_F$ over thirty paired seeds. The same records appear in Figure 3.

Graph	$r_\tau = 10^{-8}$	$r_\tau = 10^{-12}$
Amazon	9.66×10^{-9} (9.85×10^{-9})	9.65×10^{-13} (9.84×10^{-13})
Village	2.80×10^{-15} (2.92×10^{-15})	2.80×10^{-15} (2.92×10^{-15})
EU(sym)	9.87×10^{-9} (9.87×10^{-9})	9.87×10^{-13} (9.87×10^{-13})
Wiki(sym)	9.65×10^{-9} (9.65×10^{-9})	9.65×10^{-13} (9.65×10^{-13})
EU(dir)	1.19×10^{-9} (1.99×10^{-9})	1.52×10^{-14} (2.66×10^{-14})
Wiki(dir)	1.15×10^{-15} (1.27×10^{-15})	1.15×10^{-15} (1.27×10^{-15})

Where the floor is active, the median discrepancy tracks r_τ: Amazon and EU(sym) fall from order 10^{-9} at $r_\tau = 10^{-8}$ to order 10^{-13} at $r_\tau = 10^{-12}$. Wiki(sym) and EU(dir), where the floor is also active, scale the same way.

The row-wise maximum of π_τ^{20} agrees with the softmax run at every recorded $r_\tau \le 10^{-8}$, on all six graphs. At $r_\tau = 10^{-2}$, eleven EU(dir) seeds have AMI 0.998 against the softmax run. The reported window $r_\tau \le 10^{-8}$ is the range in which the partitions coincide.

This finite-budget check is the numerical counterpart of Proposition 1. On these no-continuation paths the optional clip is inactive. The clip on a matched Stage I continuation is Appendix K.3.

K.3 OPTIONAL GRADIENT CLIP AT $\tau = 0$

On a matched Stage I continuation, clipping forks Amazon AMI 0.923 versus 0.970 and leaves four graphs unchanged. This comparison holds the MixGSP start, ε, and the three-outer λ-continuation fixed, and toggles only $|\varepsilon^{-1}\nabla F| \le 50$. There is no second AMI search. Table 23 records the six-graph comparison.

Table 23: Same MixGSP start and Stage I continuation; only $|\varepsilon^{-1}\nabla F| \le 50$ is toggled. Four graphs coincide. Amazon AMI 0.923 and 0.970 is a continuation-path fork ($\lambda_{\mathrm{eff}} = 0.025$ and 0.056); endpoint $\mathrm{clip_fraction} = 0$. Wiki(sym) $|\Delta\mathrm{AMI}| = 0.056$. Mean $|\Delta\mathrm{AMI}| = 0.017$. The column label-mismatch is the fraction of nodes whose hard-assignments differ after matching prototype columns.

Graph	AMI $\tau = 0$	AMI clip	$\lvert\Delta\mathrm{AMI}\rvert$	AMI ($\tau = 0$ vs clip)	label-mismatch
Amazon	0.923	0.970	0.047	0.953	0.086
Village	0.927	0.927	0	1	0
EU(sym)	0.637	0.637	0	1	0
Wiki(sym)	0.608	0.664	0.056	0.888	0.088
EU(dir)	0.536	0.536	0	1	0
Wiki(dir)	0.277	0.277	0	1	0

Four of six graphs coincide (relative Frobenius 0, label disagreement 0). On Amazon the clip acts in Stage I and the trajectories fork: without clip, $\lambda_{\mathrm{eff}} = 0.025$ and AMI 0.923 (7 occupied clusters);

with clip, $\lambda_{\text{eff}} = 0.056$ and AMI 0.970 (8 occupied clusters). Endpoint $\mathrm{clip_fraction}$ is 0 on every graph, including Amazon, so the bound is inactive at the recorded endpoint. Wiki(sym) differs by 0.056. Mean $|\Delta\mathrm{AMI}| = 0.017$. The two partitions agree at AMI at least 0.888; the largest disagreement after matching prototype columns is 0.088 of Wiki(sym) nodes.

L Selection Fairness, Recorded AMI Ceiling, Infomap Flow, and GW-Solver Protocols

L.1 Infomap directed flow

Table 1 retains the SpecGWL-artifact undirected-flow Infomap scores (EU(sym) 0.595, Wiki(sym) 0.464). A matched two-level library call on the same directed graphs gives undirected-flow AMI 0.583 (EU) and 0.493 (Wiki), and directed-flow AMI 0.586 and 0.456.

Section 4.2 reports the selector AMI of the MixGSP start. Table 1 gives each method its published input and a fifty-evaluation budget: heat kernel for BGW and SpecGWL, adjacency for GWL, EGW, BAPG, and srGW. Heat BGW uses truncated mirror descent after a finite λ-continuation and the graph-type selectors of Table 24. SpecGWL uses the Chowdhury coverage t-grid; GWL, EGW, BAPG, and srGW use published adjacency templates. The fifty-trial AMI pool is Table 25; Stage II pathwise ratios $R_{\mathrm{II},T}$ are Table 6; wall-clock is Appendix E.1. A dash is a method–graph mismatch: Louvain and FastGreedy (Blondel et al., 2008; Clauset et al., 2004) omit the directed columns; the Infomap directed-flow row omits the undirected columns (Appendix L.1); spectral dashes on the two directed Chung kernels are the Ng partition occupying a single cluster (Ng et al., 2001). Table 24 gives the BGW criterion assignment. The mapping depends on observable graph and representation characteristics; within each locked search, labels enter only the final AMI evaluation.

Table 24: Internal criterion assigned from observed graph and representation characteristics. Directed modularity is Leicht–Newman (Leicht & Newman, 2008); the directed flow criterion is the map equation on the directed graph.

Input	Graph	Observed characteristic	Internal criterion
Heat	Amazon, Village	native undirected heat-kernel geometry	Silhouette on $D = e^{-t_{\mathrm{heat}} L}$
Heat	EU (sym, dir)	native directed communication flow	directed flow $-L$
Heat	Wiki (sym, dir)	native directed hyperlink topology	Newman / Leicht–Newman Q
Adjacency	Amazon, Village, Wiki	community structure	Newman modularity
Adjacency	EU (sym, dir)	native directed communication flow	directed flow $-L$

Table 25: Highest AMI among complete trials of the Table 1 heat TPE searches (fifty trials). AMI is recorded on every complete trial and is unused as the TPE objective. Selector AMI uses Table 24. Hard-floor Wiki(sym) reaches 0.652; $\tau = 0$ EU(sym) reaches 0.608.

Record	Amazon	Village	EU(sym)	Wiki(sym)	EU(dir)	Wiki(dir)
BGW ($\tau = 0$), selector	0.970	0.926	0.443	0.611	0.533	0.323
BGW ($\tau = 0$), max AMI	0.970	0.928	0.608	0.640	0.541	0.325
BGW-hard-floor, selector	0.970	0.926	0.565	0.530	0.485	0.323
BGW-hard-floor, max AMI	0.970	0.928	0.610	0.652	0.539	0.325

Amazon reaches 0.970 on both the selector trial and the complete-trial maximum; Village is 0.926 against 0.928. On EU(sym) and Wiki(sym) the complete-trial maximum exceeds the selector AMI; those ceilings license a selector derived from F_λ and D' (Section 6).

Appendix E specifies the SpecGWL and GWL maps; Table 1 reports their selected scores. Community methods used library defaults and did not receive a matched resolution search.

L.2 GW solvers: TPE selection and repository templates

The Table 1 EGW, BAPG, and srGW rows use published adjacency templates. EGW is POT proximal-point (PPA) entropic GW (Flamary et al., 2021) on the GWL pair $C_t = \mathrm{diag}(p_t)$

with scaled square loss, at $\varepsilon = 10^{-3}$. BAPG uses $C_t = I_K$ and the published adjacency $\rho \in \{0.1, 0.05, 0.01\}$ occupancy grid (Li et al., 2023). srGW is the official adjacency solver of Vincent-Cuaz et al. (2022) ($C_t = I_K$, degree-power h, k-means on $\mathrm{diag}(h)C$, modularity; seed 20260831).

Heat-kernel TPE for EGW, BAPG, and srGW, and repository-template runs at each paper's default hyperparameters, are artifact rows. The heat BAPG arm uses the heat ρ schedule of Li et al. (2023), so that the baseline keeps the schedule of its source paper rather than ours; heat EGW and BAPG select by Silhouette over fifty TPE trials, at the same trial budget as our own search. AMI uses the same arithmetic normalization as Table 1, so the two tables are comparable. Table 26 records those heat-kernel TPE scores. Adjacency EGW, BAPG, and srGW are the official templates of Table 1.

Table 26: Heat-kernel AMI for EGW, BAPG, and srGW under Silhouette TPE (fifty trials). Adjacency templates are Table 1.

Input	Method	Amazon	Village	EU(sym)	Wiki(sym)	EU(dir)	Wiki(dir)
heat	EGW	0.894	0.793	0.362	0.602	0.455	0.159
heat	BAPG	0.851	0.926	0.387	0.530	0.485	0.160
heat	srGW	0.836	0.000	0.470	0.002	0.446	0.157

Amazon heat EGW reaches 0.894 and heat BAPG 0.851. Village heat BAPG reaches 0.926. Wiki(dir) heat srGW is 0.157.

Under one fixed F_λ, the tuning parameters t_{heat}, the MixGSP weights (w, ρ), the prototype-seeding weight α_D, and ε influence which endpoint group a run reaches. Decreasing λ along the continuation can merge the GWL and SpecGWL endpoint groups; Figure 6 records the corresponding initialization landscape.

Sweeping the seeding weight at the recorded settings isolates its effect: over $\alpha_D \in \{0, 0.25, 0.5, 0.75, 1\}$ the reported endpoint moves on five of the six graphs, by 0.29 AMI on Amazon and 0.12 on Wiki(sym), and coincides on EU(sym) (AMI span 0.001, pairwise ARI ≥ 0.99). The direction is graph-dependent: AMI is highest at the data-driven end on Amazon, falls with α_D on Village, EU(dir), and Wiki(dir), and peaks at $\alpha_D = 0.25$ on Wiki(sym) (Table 27).

Table 27: Seeding-weight sweep under one fixed F_λ: AMI of the hard partition from the MixGSP start at five values of α_D, with all other recorded settings held fixed, and the range of pairwise ARI over the five endpoints. The searched values are $\alpha_D = 0.94$ (Amazon), 0.20 (Village), 0.05 (EU(sym)), 0.69 (EU(dir)), 0.11 (Wiki(sym)), and 0.46 (Wiki(dir)).

α_D	Amazon	Village	EU(sym)	EU(dir)	Wiki(sym)	Wiki(dir)
0.00	0.429	0.924	0.478	0.497	0.508	0.324
0.25	0.581	0.924	0.478	0.488	0.569	0.307
0.50	0.591	0.896	0.479	0.483	0.503	0.289
0.75	0.724	0.880	0.480	0.478	0.498	0.287
1.00	0.674	0.859	0.480	0.473	0.452	0.229
ARI range	0.13–0.75	0.88–1.00	0.99–1.00	0.71–0.94	0.38–0.70	0.43–0.91

M DIMENSIONALITY-REDUCTION EMBEDDINGS

DistR treats clustering and dimensionality reduction as two readings of one Gromov–Wasserstein coupling (Van Assel et al., 2025). BGW has the same two readings: the row-wise maximum of π is the partition of Table 1, and the unrestricted prototype relation D' is an embedding kernel on the same Bregman barycenter map equation 6. That D' update is the full-matrix map of Theorem 1; the clustering restriction is Corollary 2. This appendix records DistR's official neighbor embedding, then probes the kernel family of that joint P, to test how much of the embedding is carried by the low-dimensional kernel.

Swiss roll and S-curve have ambient dimension three, so the panels are visualizations of an already three-dimensional manifold. Digits and breast cancer have ambient dimensions 64 and 30. DistR uses the official neighbor-embedding DR: symmetric entropic affinity on the input, Student kernel on the embedding, and KL matching. Figure 8 places those embeddings next to PCA, t-SNE, and UMAP.

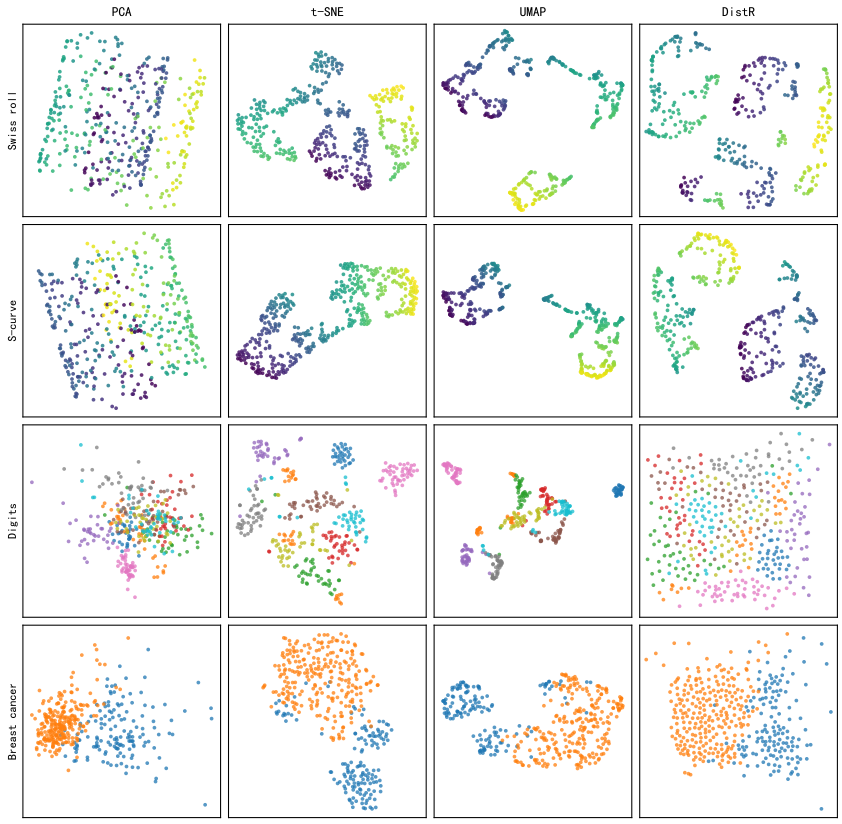

Figure 8: DistR's official n-point embedding beside PCA, t-SNE, and UMAP (digits silhouette 0.054). Swiss roll and S-curve are ambient-dimension-three visualizations; digits and breast cancer reduce from 64 and 30.

Figure 9 replaces only the low-dimensional kernel on the t-SNE joint P. The four kernels are Gaussian $\exp(-d^2/2)$, Laplace $\exp(-d)$, Student-t with $\nu = 1$, and Student-t with $\nu = 0.5$, which span light to heavy tails and so isolate the tail behaviour of the low-dimensional kernel. Each column is an n-point embedding of P; DistR remains the last column of Figure 8.

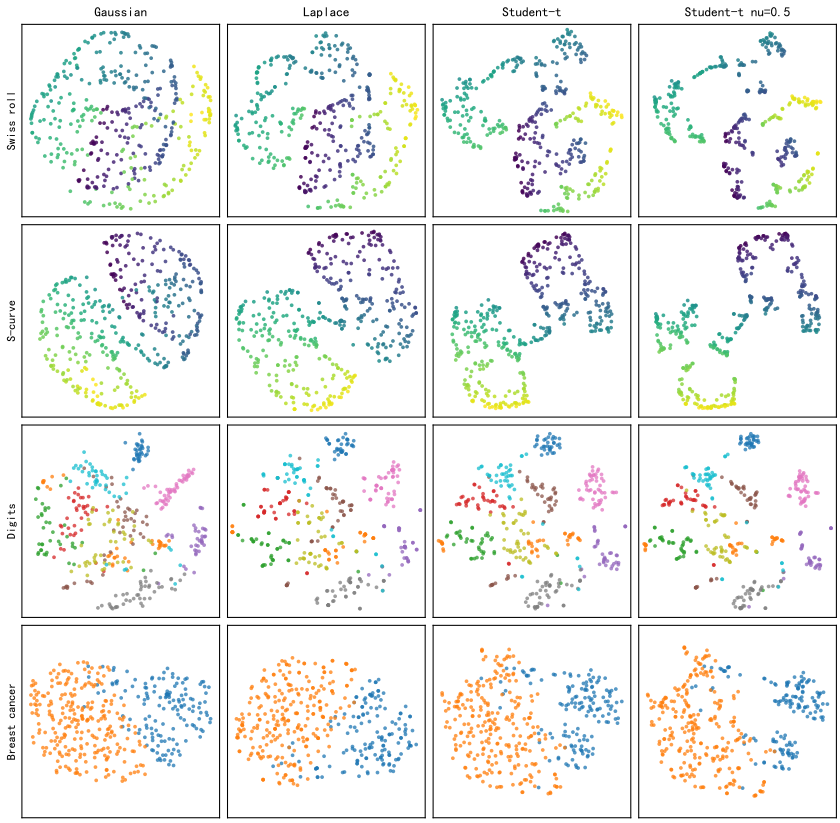

Figure 9: Heavier tails raise digits silhouette from DistR 0.054 to Student-t $\nu = 0.5$ 0.362 on a shared t-SNE joint P; breast cancer stays in 0.369–0.423. Columns: Gaussian, Laplace, $\nu = 1$, $\nu = 0.5$.

On digits the silhouette is 0.054 (DistR), 0.247 (Gaussian), 0.299 (Laplace), 0.324 ($\nu = 1$), and 0.362 ($\nu = 0.5$). On breast cancer the same scores are 0.369, 0.423, 0.408, 0.398, and 0.411. Heavier tails increase class separation on digits, from Gaussian 0.247 to $\nu = 0.5$ 0.362, whereas on breast cancer the four kernels stay within 0.398–0.423 and the ordering does not hold (Gaussian is highest at 0.423), so the tail effect is dataset-specific.

N ARTIFACT INVENTORY

The supplementary artifact stores the immutable machine-readable records underlying the appendix tables and figures, together with regeneration commands. Graph files are not redistributed, and numerical records remain unchanged during anonymization. Wiki(dir) Table 5 reports the replayed endpoint AMI 0.321; the TPE record stored selection-time AMI 0.323 because that trial omitted its precomputed GWL/SpecGWL initial transport. Dataset terms follow the SNAP and SpecGWL sources in Table 4; the SNAP library BSD page is a software license, distinct from those dataset terms.