SCALING ONLINE OPTIMAL TRANSPORT MINIBATCHES FOR FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

A central challenge in scaling flow models is maintaining stability in both training and integration. In flow matching, this stability depends critically on how source samples are paired with data. Independent random pairing introduces substantial stochasticity, motivating minibatch optimal transport (OT) to construct online source-target couplings that reduce this randomness (Tong et al., 2024). However, as applications extend to larger and higher-dimensional datasets, small minibatches can limit these benefits, while scaling the coupling to substantially larger minibatches at every iteration is computationally prohibitive. Moreover, how OT coupling affects the full stochastic gradient and integration remains theoretically unresolved. In this work, we develop a scalable online entropic OT approximation that enables substantially larger source-target couplings while preserving the benefits of OT. We close the theoretical gap for both unregularized and entropic OT by proving strict reductions in full stochastic-gradient variance and integration error relative to independent pairing, and further characterize how these improvements strengthen as the online coupling minibatch grows. Experiments on single-cell and image-flow training validate these predictions, showing that larger online couplings consistently improve training stability and final generation performance.

1 INTRODUCTION

Much of the recent progress in generative AI has been driven by scale. Autoregressive language models benefit from a standardized learning objective: next-token prediction under a softmax distribution. Once this objective is fixed, much of the challenge of scaling shifts to systems and optimization. Flow models introduce an additional dimension: their supervision is constructed along a generative path from noise to data, and the resulting regression targets depend on how source noise samples are coupled with target data samples. Thus, scaling flow models requires consideration not only of how the model is optimized, but also of how its training targets are constructed.

The choice of such source-target coupling is therefore part of the learning problem itself. In flow matching, each training example pairs a source sample from the noise distribution with a target data sample and regresses the velocity along the resulting interpolation path. Under the standard independent coupling, these pairings are random. Consequently, nearby intermediate states can arise from different endpoint pairs and receive substantially different velocity targets, increasing ambiguity in the regression problem and the variance of stochastic gradients. A natural way to reduce this randomness is to determine the source-target correspondence during training.

Optimal transport (OT) then provides a natural solution by constructing a coupling between source and target samples within each minibatch during training. Such online coupling minimizes the transport cost between the two empirical distributions, yielding more coherent source-target assignments that induce straighter flows and reduce the ambiguity of independent pairing (Pooladian et al., 2023; Tong et al., 2024). However, as flow matching is applied to larger and more heterogeneous datasets, such as high-dimensional single-cell trajectories and large-scale image datasets, a fixed small coupling minibatch can become increasingly restrictive, where each minibatch only exposes OT to a limited subset of the source and target distributions, constraining the candidate correspondences available at each iteration. As the underlying data distribution grows in diversity and complexity, it therefore becomes natural to scale the OT minibatch accordingly.

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.

Still, scaling the minibatch size is computationally challenging, where dense pairwise interactions in standard OT restrict practical minibatches to hundreds of samples. Nevertheless, recent advances in scalable entropic OT make substantially larger contexts accessible. In particular, FlashSinkhorn (Ye et al., 2026) implements Sinkhorn iterations with FlashAttention-style HBM kernels without materializing the full coupling matrix in high-bandwidth memory. To this end, we propose a simple, tractable approximation by replacing online OT coupling with scalable entropic OT, enabling coupling contexts to grow from hundreds to thousands of samples during training.

The price of this tractability is that entropic regularization changes the coupling itself. The regularization strength ε controls this deviation from standard OT, while the minibatch size k controls the scale of the online coupling. Understanding whether the stability benefits of OT persist therefore requires characterizing the joint effects of regularization and coupling scale. Moreover, this question remains unresolved even for standard unregularized minibatch OT. Existing analyses motivate OT through straighter paths and reduced conditional uncertainty (Pooladian et al., 2023; Boïté et al., 2026; Zhang et al., 2026; Chambolle & Hertrich, 2026), but do not provide a general strict characterization of its effect on the full stochastic-gradient variance or the induced numerical integration error.

By making larger online couplings feasible, we determine whether their stability benefits survive the approximation required for scaling. Our contributions can be summarized below:

Contributions. We provide a unified and principled characterization of how online OT coupling affects optimization and numerical integration. On the optimization side, we characterize the full stochastic-gradient variance by (i) proving that population quadratic OT strictly reduces this variance relative to independent coupling, (ii) showing that unregularized minibatch-OT variance reduces as the minibatch size k grows, and (iii) extending same guarantees to entropic-regularized couplings. On the integration side, we establish parallel strict reductions in Euler integration error and characterize how these guarantees improve with scaled minibatch size. These results provide a better theoretical understanding for using and scaling entropic OT as a replacement to the standard online OT couplings.

Related works. OT-based conditional flow matching (OT-CFM) improves both training and generation through better source-target pairings (Tong et al., 2024; Pooladian et al., 2023). Given the difficulty to scale OT-CFM, recent works scale the minibatch context by separating pairing from training: Zhang et al. (2026) precomputed million-scale couplings offline, which requires buffering enough pairs to cover the entire training run before optimization begins; semidiscrete approximation methods offer better efficiency by reusing fitted dataset-level potentials, but need a separate potential-optimization stage (Mousavi-Hosseini et al., 2026) or face non-convergence when scaled to large unconditional datasets (Kong et al., 2026). Therefore, we return to the original challenge, by finding an affordable and stable method for scaling the minibatch that flow matching recomputes from fresh samples at every training iteration.

The existing theory has so far addressed separate pieces. Tong et al. (2024) connect population OT to dynamic OT, while Pooladian et al. (2023) study conditional gradient covariance and asymptotic BatchOT straightness. Finite-batch theory remains limited: Zhang et al. (2026) characterize uncertainty in the resulting OT coupling, while Boïté et al. (2026) analyze its integration benefit only in a restricted two-atom model; more recent work also studies full parameter-gradient variance in restricted Gaussian models (Reu et al., 2025) and integration error through the geometry of a prescribed velocity field (Tao & Choi, 2026). Thus, no prior result establishes that replacing independent pairing with minibatch OT strictly reduces full stochastic-gradient variance while inducing a general integration advantage; we defer further related work to Appendix A.

2 Preliminaries

We introduce flow matching (FM) and coupling definitions used throughout the later analyses.

Continuous normalizing flows. Let $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$ denote the source (noise) and target (image-data) distributions, respectively. A time-dependent vector field $v_\theta : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, parameterized by $\theta \in \mathbb{R}^p$, assigns an instantaneous velocity to each time-state pair (t, x), thereby defining a flow from the source to the target. Specifically, a continuous normalizing flow (CNF) (Lipman et al., 2023) transports source samples through the flow map

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t(x) = v_\theta(t, \phi_t(x)), \qquad \phi_0(x) = x,$$

with the goal that $(\phi_1)_{\#}q_0 \approx q_1$. Flow matching trains this vector field from paths induced by a coupling of the source and target endpoints.

Coupling-based FM. Let π be a coupling of q_0 and q_1. Each FM training tuple is generated first by sampling $(X_0, X_1) \sim \pi$ and then independently sampling $t \sim \mathcal{U}([0,1])$. Unless stated otherwise, we take all expectations $\mathbb{E}_\pi[\cdot]$ and covariances $\mathrm{Cov}_\pi(\cdot)$ over the randomness induced by the sampling procedure associated with coupling π. For simplicity, we denote the complete stochastic FM training tuple by $\xi^\pi = (t, X_0, X_1)$ and its associated variables, including intermediate interpolation point X_t^π and pathwise velocity target U^π, by

$$X_t^\pi = (1-t)X_0 + tX_1, \qquad U^\pi = X_1 - X_0.$$

At a fixed model input (t, X_t^π), different endpoint pairs may produce different velocity targets. Given one sampled FM training tuple ξ^π, its loss is

$$\ell_\pi(\theta; \xi^\pi) = \|v_\theta(t, X_t^\pi) - U^\pi\|_2^2.$$

Under this squared loss, the optimal deterministic velocity prediction at input (t, x) is the conditional mean, defined by $u^\pi(t, X_t^\pi) := \mathbb{E}_\pi[U^\pi \mid t, X_t^\pi]$. The conditional covariance, $\mathrm{Cov}_\pi(U^\pi \mid t, X_t^\pi)$, then measures how much the possible velocity targets still vary at that input; the population FM objective further averages $\ell_\pi(\theta; \xi^\pi)$ over all possible training tuples,

$$\mathcal{L}_\pi(\theta) := \mathbb{E}_\pi[\ell_\pi(\theta; \xi^\pi)].$$

Let $J_\theta(t,x) := \nabla_\theta v_\theta(t,x) \in \mathbb{R}^{d\times p}$, the gradient over $\ell_\pi(\theta; \xi^\pi)$ can be expressed as

$$G_\pi(\theta) := \nabla_\theta \ell_\pi(\theta; \xi^\pi) = 2J_\theta(t, X_t^\pi)^\top \big(v_\theta(t, X_t^\pi) - U^\pi\big).$$

Whenever $\mathbb{E}_\pi[\|G_\pi(\theta)\|_2^2] < \infty$, we can derive the total variance of the per-sample stochastic gradient induced by sampling FM training tuples under coupling π:

$$\sigma^2_{\mathrm{grad},\pi}(\theta) := \mathrm{Tr}\{\mathrm{Cov}_\pi(G_\pi(\theta))\} = \mathbb{E}_\pi\left[\|G_\pi(\theta) - \mathbb{E}_\pi G_\pi(\theta)\|_2^2\right].$$

Here, $\mathrm{Cov}_\pi(G_\pi(\theta)) \in \mathbb{R}^{p\times p}$ is the covariance matrix of the per-sample stochastic gradient, whereas $\sigma^2_{\mathrm{grad},\pi}(\theta)$ is its trace, measuring the total sample-to-sample gradient variance across FM training tuples under coupling π. The law of total covariance further gives us the following decomposition,

$$\sigma^2_{\mathrm{grad},\pi}(\theta) = \mathbb{E}_\pi[\mathrm{Tr}\{\mathrm{Cov}_\pi(G_\pi(\theta) \mid t, X_t^\pi)\}] + \mathrm{Tr}\{\mathrm{Cov}_\pi(\mathbb{E}_\pi[G_\pi(\theta) \mid t, X_t^\pi])\}. \tag{1}$$

The first term is the gradient noise from target ambiguity at a fixed model input, and the second term is the variation of the conditional mean gradient across model inputs.

3 Scalable OT for Flow Matching

We introduce three online coupling setups between q_0 and q_1: π_{ind}, π^*, and $\hat{\pi}_k$. Under the independent coupling π_{ind}, training pairs $X_0 \sim q_0$ and $X_1 \sim q_1$ are sampled independently; in contrast, the optimal quadratic-OT coupling $\pi^\star$ minimizes $\mathbb{E}_\pi[\|X_0 - X_1\|_2^2]$ over all couplings π between q_0 and q_1. Moreover, if $q_0 \ll \mathcal{L}^d$, Brenier's theorem (Brenier, 1991) states that this coupling is unique and induced by a deterministic transport map: $\pi^\star = (\mathrm{id}, T)_{\#}q_0$, where $T = \nabla\varphi$ is q_0-almost everywhere for some convex function φ. In practice, online OT pairing is computed between random source and target minibatches of size k, producing a random empirical plan $\widehat{\pi}_k$.

We denote our method as EOT-CFM, which uses FlashSinkhorn (Ye et al., 2026) to couple fresh source and target contexts at every training iteration. The minibatch size k counts samples per target context, and the regression batch size $\mathcal{B} \le k$ counts pairs used per model update. Given independent samples $x_i \sim q_0$ and $y_j \sim q_1$, the cost implementation is $C_{ij} = \|x_i - y_j\|_2^2/(2d)$, which rescales the cost by the data dimension without changing the unregularized matching. With uniform mass $1/k$ per sample, we solve the following entropic transportation plan:

$$\widehat{\pi}_{k,\varepsilon} = \underset{\substack{P\ge 0,\ P\mathbf{1}=\mathbf{1}/k \\ P^\top\mathbf{1}=\mathbf{1}/k}}{\arg\min} \left\{\langle \mathbf{C}, P\rangle + \varepsilon\,\mathrm{KL}\big(P \big\| \mathbf{1}\mathbf{1}^\top/k^2\big)\right\}. \tag{2}$$

The regularization $\varepsilon > 0$ balances between the cost and entropy, and the iteration count L controls the numerical approximation. Without materializing the full transport plan or invoking a separate JAX process, FlashSinkhorn solves the minibatch coupling directly within the PyTorch training loop using tiled cost evaluations and dual potentials. Specifically, the exact row conditional is $\mathbb{P}(J = j \,|\, I = i) = k \cdot \widehat{\pi}_{k,\varepsilon,ij}$, which normalizes probabilities from the approximate target potential $\widetilde{g}$. Algorithm 1 then constructs the interpolation points $X_{t_b,b}$ and velocity targets U_b defined in §2.

Algorithm 1 Continuous normalizing flow with online entropic couplings (EOT-CFM)

Require: Minibatch size k, batch size $\mathcal{B}$, regularization $\varepsilon > 0$, Sinkhorn iterations L, training steps T.

for $t = 1, \dots, T$ **do**
  Sample noise and data from source $x_i \sim q_0$ and target $y_j \sim q_1$ with minibatch size k
  Approximate the target dual potential $\widetilde{g}$ of Eq. (2) with L FlashSinkhorn iterations
  Select $\mathcal{B}$ distinct source indices $I_1, \dots, I_{\mathcal{B}}$ uniformly
  for $b = 1, \dots, \mathcal{B}$ **do**
    Compute $(\widetilde{r}_{I_b,j})_{j=1}^{k} = \mathrm{softmax}\left(\left(\frac{\widetilde{g}_j - C_{I_b j}}{\varepsilon}\right)_{j=1}^{k}\right)$
    Sample $J_b \sim \mathrm{Cat}\big((\widetilde{r}_{I_b,j})_{j=1}^{k}\big)$ and $t_b \sim \mathcal{U}([0,1])$
  end for
  Compute $\widehat{\mathcal{L}}(\theta) = \frac{1}{\mathcal{B}d} \sum_{b=1}^{\mathcal{B}} \|v_\theta(t_b, X_{t_b,b}) - U_b\|_2^2$
  Update θ with the optimizer using $\nabla_\theta \widehat{\mathcal{L}}(\theta)$
end for

4 Main Results

4.1 On stochastic gradient variance

We first study the benefit of scaling OT minibatch size through its effect on stochastic-gradient variance. The closest existing result we are aware of is Pooladian et al. (2023, Lemma 3.2), which analyzes the conditional gradient variance of minibatch OT at model input (t, x), corresponding to the fixed-input component in Eq. (1). However, it remains a gap to prove that OT coupling strictly reduces the stochastic-gradient variance relative to independent coupling.

Therefore, in §4.1.1, we begin by establishing a strict gradient-variance reduction by filling the gap between optimal coupling $\pi^\star$ and independent coupling π_{ind}. In §4.1.2, we extend this result to entropic-OT couplings. Finally, in §4.1.3, we establish that scaling to larger minibatch size could further reduce gradient variance for both unregularized and entropic OT couplings.

4.1.1 Gradient variance reduction: standard unregularized OT

We first state two standard assumptions used in theoretical analyses of flow matching and neural differential equations: realizability of the population target field and global optimality of the population objective (Islam et al., 2026; Khan et al., 2026; Baheri, 2026). Since the population regression target depends on the coupling π, we note that these assumptions are imposed coupling-wise; all the detailed math derivations and proofs for the later theoretical results are deferred to Appendix B.

Assumption 4.1 (Realizability)**.** For every coupling π under comparison, the model class $\mathcal{V} = \{v_\theta : \theta \in \mathbb{R}^p\}$ contains the coupling-induced conditional mean velocity, i.e., there exists $\theta_\pi \in \mathbb{R}^p$ such that $v_{\theta_\pi}(t, X_t^\pi) = u^\pi(t, X_t^\pi)$ almost surely.

Assumption 4.2 (Optimality)**.** For every coupling π under comparison, $\theta_\pi^\star$ is a global minimizer of $\mathcal{L}_\pi$, the map $\theta \mapsto v_\theta(t, X_t^\pi)$ is differentiable at $\theta_\pi^\star$ almost surely, and $\mathbb{E}_\pi[\|G_\pi(\theta_\pi^\star)\|_2^2] < \infty$.

With this setup, we first derive the stochastic gradient variance at the population optimum.

Proposition 4.3. *Under* Assumptions 4.1 *and* 4.2*, we have*

$$\sigma^2_{\text{grad},\pi}(\theta_\pi^\star) = 4\,\mathbb{E}_\pi\left[\mathrm{Tr}\big\{J_{\theta_\pi^\star}(t, X_t^\pi)^\top \mathrm{Cov}_\pi(U^\pi \mid t, X_t^\pi) J_{\theta_\pi^\star}(t, X_t^\pi)\big\}\right]. \tag{3}$$

Thus, at the population optimum, conditional target covariance is mapped into parameter-gradient space by the model Jacobian. Its average magnitude also satisfies

$$\mathbb{E}_\pi\big[\mathrm{Tr}\{\mathrm{Cov}_\pi(U^\pi \mid t, X_t^\pi)\}\big] = \mathbb{E}_\pi\left[\|U^\pi - u^\pi(t, X_t^\pi)\|_2^2\right].$$

To obtain an explicit positive lower bound for independent coupling, we directly follow from the Assumption (A2') in Pooladian et al. (2023), using the following technical sufficient condition:

Assumption 4.4. There exist a compact convex set $\Omega \subset \mathbb{R}^d$ with nonempty interior, $\alpha \in (0, 1]$, $\gamma > 0$, and α-Hölder functions $f_i : \Omega \to [\gamma, \infty)$ such that, for $i \in \{0, 1\}$ and every Borel $A \subseteq \mathbb{R}^d$,

$$q_i(A) = \int_{A \cap \Omega} f_i(x)\,\mathrm{d}x, \qquad \mathrm{supp}(q_i) = \Omega.$$

Finally, combining the preceding variance identity with this common-support condition, we establish a strict gradient-variance gap between standard unregularized OT and independent coupling.

Theorem 4.5. *Let $\pi^\star = (\mathrm{id}, T)_{\#} q_0$ be the unregularized OT coupling and π_{ind} be the independent coupling under the quadratic cost. Suppose* Assumptions 4.1, 4.2, *and* 4.4 *hold for both couplings and define $\omega_d = \mathcal{L}^d(B(0,1))$, then, for every $\bar{x} \in \Omega^\circ$ and $r > 0$ such that $B(\bar{x}, 2r) \subset \Omega$, we have*

$$\sigma^2_{\mathrm{grad},\pi_{\mathrm{ind}}}(\theta^\star_{\pi_{\mathrm{ind}}}) - \sigma^2_{\mathrm{grad},\pi^\star}(\theta^\star_{\pi^\star}) \geq \frac{4d\gamma^2\omega_d^2}{d+2}\, r^{2d+2} > 0. \tag{4}$$

4.1.2 Gradient-variance reduction: entropic OT

We next study the influence of entropic regularization on gradient variance and introduce following setup inherited from previous minibatch flow matching analysis.

Definition 4.6 (Uniformly bounded parameter Jacobians)**.** Following the maximum-Jacobian setup in the flow-matching gradient-variance bound of Pooladian et al. (2023, Lemma 3.2), a coupling family Π has uniformly bounded parameter Jacobians if there exists $M < \infty$ such that

$$\max_{t\in[0,1],\, x\in\mathbb{R}^d} \|J_{\theta^\star_\pi}(t,x)\|^2_{\mathrm{op}} \leq M, \qquad \forall\, \pi \in \Pi. \tag{5}$$

Remark 4.7. *We apply* Definition 4.6 *to obtain a common Jacobian bound M for all couplings considered in later theorems, allowing us to translate target-uncertainty bounds into full stochastic-gradient variance comparisons. We also note that such Jacobian bound control is widely used in neural-network optimization (Oymak & Soltanolkotabi, 2019, Assumption 2) and recent flow-matching theory analysis (Zeng & Yan, 2025, Proposition 3).*

We use π_ε to denote the population entropic-OT coupling between q_0 and q_1:

$$\pi_\varepsilon \in \operatorname*{arg\,min}_{\pi\in\Pi(q_0,q_1)} \left\{ \frac{1}{2d}\, \mathbb{E}_\pi[\|X_0 - X_1\|_2^2] + \varepsilon\, \mathrm{KL}(\pi\|\pi_{\mathrm{ind}}) \right\}. \tag{6}$$

To compare its variance to that under independent coupling, we first relate its irreducible flow matching regression error to its quadratic transport cost, where W_2 denotes the 2-Wasserstein distance:

Proposition 4.8. *Under* Assumptions 4.1 *and* 4.2, *every coupling π satisfies*

$$0 \leq \mathcal{L}_\pi(\theta^\star_\pi) = \mathbb{E}_\pi[\mathrm{Tr}\{\mathrm{Cov}_\pi(U^\pi \mid t, X_t^\pi)\}] \leq \mathbb{E}_\pi[\|X_1 - X_0\|_2^2] - W_2^2(q_0,q_1). \tag{7}$$

Thus, the population loss at optimality is exactly the average target uncertainty remaining at a fixed model input. As an intermediate step, we establish Lemma 4.9 to show that sufficiently small entropic regularization strictly reduces this uncertainty comparing to independent coupling.

Lemma 4.9. *Suppose* Assumptions 4.1 *and* 4.2 *hold for π_{ind}, $\pi^\star$, and every π_ε under comparison, and assume $\mathcal{L}_{\pi_{\mathrm{ind}}}(\theta^\star_{\pi_{\mathrm{ind}}}) > 0$, which holds whenever q_0 has a strictly positive Lebesgue density and q_1 is not a Dirac mass, including the standard choice $q_0 = \mathcal{N}(0, I_d)$. Then, for any $\varepsilon_0 > 0$ that satisfies*

$$\mathbb{E}_{\pi_{\varepsilon_0}}[\|X_1 - X_0\|_2^2] - W_2^2(q_0,q_1) < \mathcal{L}_{\pi_{\mathrm{ind}}}(\theta^\star_{\pi_{\mathrm{ind}}}), \tag{8}$$

every $0 < \varepsilon \leq \varepsilon_0$ gives us Eq. (9), and such ε_0 is guaranteed to exist by Remark 4.10,

$$\mathcal{L}_{\pi^\star}(\theta^\star_{\pi^\star}) < \mathcal{L}_{\pi_\varepsilon}(\theta^\star_{\pi_\varepsilon}) < \mathcal{L}_{\pi_{\mathrm{ind}}}(\theta^\star_{\pi_{\mathrm{ind}}}). \tag{9}$$

Remark 4.10. *We note that such ε_0 always exists. For any $\widetilde{\pi} \in \Pi(q_0,q_1)$ with $0 < \mathrm{KL}(\widetilde{\pi}\|\pi_{\mathrm{ind}}) < \infty$ and $\mathbb{E}_{\widetilde{\pi}}[\|X_1 - X_0\|_2^2] - W_2^2(q_0,q_1) < \mathcal{L}_{\pi_{\mathrm{ind}}}(\theta^\star_{\pi_{\mathrm{ind}}})$, Eq. (9) holds whenever*

$$0 < \varepsilon < \frac{\mathcal{L}_{\pi_{\mathrm{ind}}}(\theta^\star_{\pi_{\mathrm{ind}}}) + W_2^2(q_0,q_1) - \mathbb{E}_{\widetilde{\pi}}[\|X_1 - X_0\|_2^2]}{2d\, \mathrm{KL}(\widetilde{\pi}\|\pi_{\mathrm{ind}})}. \tag{10}$$

A coupling $\widetilde{\pi}$ satisfying these conditions always exists; we defer its proof to Appendix B.5.

Finally, we transform Lemma 4.9 into the stochastic-gradient variance reduction.

Theorem 4.11. *Suppose the conditions of* Lemma 4.9 *hold. Then, every $0 < \varepsilon \leq \varepsilon_0$ satisfies that*

$$0 < \sigma^2_{\mathrm{grad},\pi_\varepsilon}(\theta^\star_{\pi_\varepsilon}) \leq 4\{\mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] - W_2^2(q_0,q_1)\}\, \mathrm{ess\,sup}\, \|J_{\theta^\star_{\pi_\varepsilon}}(t, X_t^{\pi_\varepsilon})\|^2_{\mathrm{op}}. \tag{11}$$

Moreover, under Definition 4.6 *over all sufficiently small ε, there exists $\varepsilon_1 \in (0, \varepsilon_0]$ such that every $0 < \varepsilon \leq \varepsilon_1$ satisfies*

$$\sigma^2_{\mathrm{grad},\pi^\star}(\theta^\star_{\pi^\star}) < \sigma^2_{\mathrm{grad},\pi_\varepsilon}(\theta^\star_{\pi_\varepsilon}) < \sigma^2_{\mathrm{grad},\pi_{\mathrm{ind}}}(\theta^\star_{\pi_{\mathrm{ind}}}). \tag{12}$$

4.1.3 GRADIENT-VARIANCE REDUCTION: LARGER MINIBATCH

We now move to characterize how the gradient variance changes when minibatch size is scaled. Similar to the previous proof structure, we first derive the result upon standard unregularized OT and then extend to the results to entropic OT. Our math idea follows from by connecting the cost residue in Eq. (13) to the velocity uncertainty and then bridging this uncertainty to the gradient variance.

Unregularized OT. For each minibatch, we independently draw k source endpoints from q_0 and k target endpoints from q_1 and denote the distribution of this selected pair by π_k, with $\pi_1 = \pi_{\text{ind}}$. Since π_k is a coupling between q_0 and q_1, its quadratic transport cost cannot be smaller than the cost of optimal coupling π^*, i.e., $W_2^2(q_0, q_1)$. Therefore, we can represent this gap as the excess cost residue Δ_k:

$$\Delta_k := \mathbb{E}_{\pi_k}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1). \tag{13}$$

Clearly, $\Delta_k \geq 0$, and we apply this cost residual in an interesting way by directly connecting it to the velocity-target uncertainty that drives stochastic gradient variance:

Proposition 4.12. *Given $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$, then without any assumptions, for every $k \geq 1$, we have*

$$\mathbb{E}_{\pi_k}[\mathrm{Tr}\{\mathrm{Cov}_{\pi_k}(U^{\pi_k} \mid t, X_t^{\pi_k})\}] \leq \Delta_k \xrightarrow{k\to\infty} 0. \tag{14}$$

Indeed, Proposition 4.12 plays an important role by bridging the gap between the velocity-target uncertainty and OT minibatch, which allows us to further connect it to stochastic gradient variance:

Theorem 4.13. *Suppose that $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$ and* Assumptions 4.1 *and* 4.2 *hold for every π_k. Then,*

$$0 \leq \sigma^2_{\text{grad},\pi_k}(\theta^\star_{\pi_k}) \leq 4\Delta_k \operatorname{ess\,sup} \|J_{\theta^\star_{\pi_k}}(t, X_t^{\pi_k})\|_{\text{op}}^2. \tag{15}$$

Given the setup in Definition 4.6*, the right-hand side of Eq.* (15) *converges to zero, so for any fixed k_0 with $\sigma^2_{\text{grad},\pi_{k_0}}(\theta^\star_{\pi_{k_0}}) > 0$, Eq.* (16) *strictly holds for all sufficiently large k:*

$$\sigma^2_{\text{grad},\pi_k}(\theta^\star_{\pi_k}) < \sigma^2_{\text{grad},\pi_{k_0}}(\theta^\star_{\pi_{k_0}}). \tag{16}$$

Entropic OT. For each $k \geq 1$ and $\varepsilon > 0$, we denote $\pi_{k,\varepsilon}$ as the distribution of the endpoint pair generated by the entropic minibatch procedure in §3 and $\widehat{\pi}_{k,\varepsilon}$ as the random empirical plan within minibatches. Similarly, $\pi_{1,\varepsilon} = \pi_{\text{ind}}$. For simplicity, we abbreviate the average target uncertainty in Proposition 4.12 by $\mathsf{R}(\pi) = \mathbb{E}_\pi[\mathrm{Tr}\{\mathrm{Cov}_\pi(U^\pi \mid t, X_t^\pi)\}]$. Shown in Proposition 4.14, we first characterize the effect of increased k using the similar cost residue as in Eq. (13):

Proposition 4.14. *Suppose $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$. For every fixed $\varepsilon > 0$, $\pi_{k,\varepsilon} \to \pi_\varepsilon$ in W_2, and moreover,*

$$\limsup_{k\to\infty} \mathsf{R}(\pi_{k,\varepsilon}) \leq \mathsf{R}(\pi_\varepsilon) \leq \mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1). \tag{17}$$

Note that $\pi_{k,\varepsilon}$ and π_{ind} are mutually absolutely continuous for every finite k, so $\mathsf{R}(\pi_{k,\varepsilon}) = 0$ if and only if $\mathsf{R}(\pi_{\text{ind}}) = 0$.

Increasing k at a fixed ε indeed approaches the population entropic coupling, but this convergence alone cannot give the complete reduction result, which also depends on the coupling-specific model Jacobian. So, we next establish the gradient variance reduction under larger minibatch size:

Theorem 4.15. *Fix a finite baseline $k_0 \geq 1$ and consider that $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathsf{R}(\pi_{\text{ind}}) > 0$, and* Assumptions 4.1 *and* 4.2 *hold for every $\pi_{k,\varepsilon}$ under comparison. Then, for every $k \geq 1$ and $\varepsilon > 0$, we have*

$$4\mathsf{R}(\pi_{k,\varepsilon}) \leq \sigma^2_{\text{grad},\pi_{k,\varepsilon}}(\theta^\star_{\pi_{k,\varepsilon}}) \leq 4\mathsf{R}(\pi_{k,\varepsilon}) \operatorname{ess\,sup} \|J_{\theta^\star_{\pi_{k,\varepsilon}}}(t, X_t^{\pi_{k,\varepsilon}})\|_{\text{op}}^2. \tag{18}$$

Under Definition 4.6 *over all sufficiently large k and sufficiently small ε, there exists $\bar{\varepsilon} > 0$ such that, for every fixed $0 < \varepsilon \leq \bar{\varepsilon}$, there is a finite batch threshold $K > k_0$ for which every $k \geq K$ satisfies*

$$\sigma^2_{\text{grad},\pi_{k,\varepsilon}}(\theta^\star_{\pi_{k,\varepsilon}}) < \sigma^2_{\text{grad},\pi_{k_0,\varepsilon}}(\theta^\star_{\pi_{k_0,\varepsilon}}). \tag{19}$$

4.2 ON INTEGRATION-ERROR REDUCTION

Beyond optimization, OT couplings induce straighter probability paths, but how this geometry translates into numerical benefits remains unclear. We study integration error at the distribution level by comparing the output of Euler discretization with the target distribution.

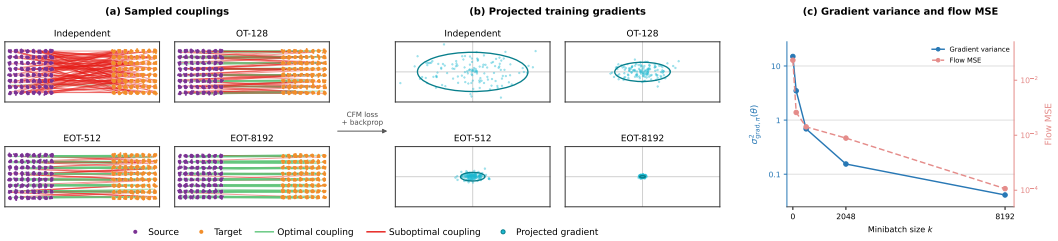

Figure 1: Results on a synthetic multimodal task over independent random coupling, OT-CFM using the minibatch size of 128, and EOT-CFM with contexts up to 8192. (a) Source-target pairings, with green and red lines indicating optimal and suboptimal couplings, respectively. Larger minibatches produce more optimal assignments for subsequent flow training. (b) Stochastic gradients evaluated at a trained flow model with regression batch size 128 and projected onto common two-dimensional axes; ellipses summarize their spread. (c) Full gradient variance (left axis) and final flow-generation MSE (right axis) averaged over three training seeds, scaling along the OT minibatch size, where $k = 0$ denotes independent coupling. Scaling the minibatch size consistently reduces both gradient variance and MSE error, yielding more stable training and better flow generation.

Definition 4.16 (One-step distributional Euler error)**.** For a coupling $\pi \in \Pi(q_0, q_1)$, one Euler step from $t = 0$ produces $X_0 + \mathbb{E}_\pi[X_1 - X_0 \mid X_0] = \mathbb{E}_\pi[X_1 \mid X_0]$. Its distributional integration error is

$$\mathrm{err}(\pi) := W_2^2\big((\mathbb{E}_\pi[X_1 \mid X_0 = \cdot])_{\#} q_0, q_1\big)\,, \tag{20}$$

i.e., the Wasserstein distance between the one-step output distribution and the target distribution q_1.

Shown in Proposition 4.17, we first establish the relationship of distributional integration error across optimal, entropic, and independent couplings:

Proposition 4.17. *Suppose $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$, $q_0 \ll \mathcal{L}^d$, and q_1 is not a Dirac mass. Let $\pi^\star = (\mathrm{id}, T)_{\#} q_0$ be the population quadratic-OT coupling, let π_k be the unregularized OT plan at batch size k, with $\pi_1 = \pi_{\mathrm{ind}}$, and for every fixed $\varepsilon > 0$, let π_ε be the population entropic coupling. Then,*

$$\mathrm{err}(\pi^*) < \mathrm{err}(\pi_\varepsilon) < \mathrm{err}(\pi_{\mathrm{ind}}).$$

Finally, based on Proposition 4.17, we establish the error reduction over larger k by connecting the expected empirical EOT value in Eq. (21) to the integration error.

$$S_{k,\varepsilon} := \mathbb{E}\big[\langle \mathbf{C}, \widehat{\pi}_{k,\varepsilon}\rangle + \varepsilon\, \mathrm{KL}\big(\widehat{\pi}_{k,\varepsilon} \big\| \mathbf{1}\mathbf{1}^\top / k^2\big)\big]\,. \tag{21}$$

Theorem 4.18. *Under the conditions of* Proposition 4.17*, for every fixed $\varepsilon > 0$ and $k \geq 2$, we have*

$$\mathrm{err}(\pi_{k,\varepsilon}) \leq \mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|_2^2] - \frac{d^2 (S_{1,\varepsilon} - S_{k,\varepsilon})^2}{\mathbb{E}[\|X_0 - \mathbb{E}[X_0]\|_2^2]} < \mathrm{err}(\pi_{\mathrm{ind}}). \tag{22}$$

The upper bound in Eq. (22) is also non-increasing in k, and $\mathrm{err}(\pi_{k,\varepsilon}) \xrightarrow{k\to\infty} \mathrm{err}(\pi_\varepsilon)$. Let $\varepsilon_k > 0$ be any sequence with $\varepsilon_k \to 0$, for every fixed finite baseline k_0, there exists $k_1 > k_0$ such that all $k > k_1$ satisfies

$$\mathrm{err}(\pi_{k,\varepsilon_k}) < \mathrm{err}(\pi_{k_0,\varepsilon_k}). \tag{23}$$

Figure 3 illustrates the empirical benefit of Eq. (23) through a qualitative comparison of generated samples at the same NFE; we defer further explanations to §5.

5 Experiments

We study three classes of experiments to show the empirical benefits of scaling minibatch size: (i) first, we consider classical 2D synthetic datasets, where the effects of minibatch scaling can be clearly visualized and interpreted; (ii) we then work on single-cell interpolation flows; apart from the traditional medium-dimensional dataset used in OT-CFM, we include a large-scale, high-dimensional dataset to demonstrate the benefits of scaled minibatches; (iii) finally, we evaluate image flow training on CIFAR-100 (Krizhevsky, 2009) and ImageNet-64×64 (Deng et al., 2009).

Table 1: Results on EB and PBMC (mean $\pm$ std over 5 seeds), with performance metrics (W_1, Energy) and gradient variance. Larger contexts improve prediction quality and reduce gradient variance.

EB			PBMC			
Method	Exact W_1 ($\downarrow$)	$\sigma^2_{\text{grad},\pi}(\theta)$ ($\downarrow$)	**Method**	Exact W_1 ($\downarrow$)	CellFlow Energy ($\downarrow$)	$\sigma^2_{\text{grad},\pi}(\theta)$ ($\downarrow$)
OT-CFM-128	0.756 ± 0.027	0.247 ± 0.047	OT-CFM-128	2.612 ± 0.021	0.693 ± 0.102	1.183 ± 0.343
EOT-CFM-512	0.746 ± 0.021	0.101 ± 0.017	EOT-CFM-1024	2.531 ± 0.017	0.435 ± 0.115	0.697 ± 0.096
EOT-CFM-1024	0.719 ± 0.017	0.080 ± 0.012	EOT-CFM-4096	2.505 ± 0.028	0.385 ± 0.161	0.515 ± 0.035
EOT-CFM-1536	0.714 ± 0.016	0.071 ± 0.011	EOT-CFM-8192	2.432 ± 0.015	0.329 ± 0.046	0.384 ± 0.037

2D-couplings. We consider two synthetic problems: a symmetric problem with 64 Gaussian modes arranged on an 8×8 grid, and a double-ring problem with 8 modes on each ring. In both problems, the source and target distributions are separated by a horizontal translation, and the ground-truth coupling matches points with the same mode label. We compare independent coupling, OT-CFM with a 128×128 minibatch, and EOT with context sizes $n \in \{512, 2048, 8192\}$, using squared Euclidean cost and training a three-layer MLP for 6,000 updates with batch size 128 with AdamW.

Results are shown in Figure 1. Specifically, larger OT contexts produce increasingly coherent source-target pairings, as shown by the reduction in red suboptimal crossing connections and the increase in green optimal coupling connections. This improvement directly translates into lower SGD variance, with tighter projected-gradient distributions that shows scaling the OT context provides cleaner pairings and substantially better training stability. More 2D examples are provided in Appendix C.

High-dimensional single-cell interpolation. We evaluate on both the embryoid body (EB) benchmark (Moon et al., 2019) used in Tong et al. (2024) and a larger, higher-dimensional PBMC cytokine-response dataset (Oesinghaus et al., 2025), processed by CellFlow (Klein et al., 2025) for flow model training. For EB, it represents 16,819 cells from five time points (the smallest point has 1900 cells) using five whitened principal components and independently split each time point into $80\%/10\%/10\%$ training, validation, and test sets. We leave out each intermediate time point $t \in \{1, 2, 3\}$, train a three-layer MLP of width 64 on the remaining transitions for at most 1,000 epochs.

For PBMC, we use 90,452 PBS control cells and 20,471 IFN-β-stimulated cells from Donor 1, represent each cell using 100 whitened principal components, and apply the same data split. We train a three-layer MLP of width 256 for 100,000 steps. Both experiments use AdamW with learning rate 10^{-3}, weight decay 10^{-5}, and a regression batch size of 128, and generate cells using 100 Euler steps. For EB, we report exact W_1 averaged over three held-out time points and five seeds. For PBMC, we compare 2,048 generated and held-out target cells and report exact assignment W_1, CellFlow energy, and gradient variance, averaged over five seeds. Results are presented in Table 1. Since the smallest time point from EB contains around 1900 cells, the maximum minibatch cannot be scaled as PBMC. For both dataset, larger minibatch leads to both better empirical performance and more stable training, where the improvement consistently gets better with larger minibatch size. With the result deferred to Appendix C, we also use the high dimensional PBMC dataset to study the ablation over ε, moderate thousands-level context already gives a clear effect of scaling and remains robust to the entropic regularization intensity.

Image generation. We evaluate on CIFAR-100 and ImageNet-64×64. For CIFAR, all methods follow the same U-Net architecture and optimization hyperparameters from OT-CFM's CIFAR training script (Tong et al., 2024). We evaluate EMA checkpoints using fixed-step and adaptive Dopri5 integration, and FIDs are computed with CleanFID with 50,000 generated samples.

For ImageNet training, we follow the unconditional ADM-style U-Net configuration of Pooladian et al. (2023). We train for 500,000 updates with a global batch size of 1,024 across $8\times$NVIDIA H100 GPUs and a constant learning rate of 10^{-4}. Similar to CIFAR-100, all coupling methods share the same architecture and training configuration, and we generate 50,000 samples from the EMA checkpoint and report FID against the canonical ADM ImageNet-64 reference statistics. Evaluation results are shown in Table 2, and we further present in-training gradient variance and qualitative image generation results in Figure 2. Experiments on both single-cell and image flow matching show that larger coupling minibatches generally reduce gradient variance and improve performance, consistent with our entropic analysis in §4.1.3. Figure 3 further shows that the gains in generation quality are particularly better at low NFEs. Together, these findings suggest that scaling the entropic coupling context benefits both optimization and sampling, combining lower gradient noise during training with better generation quality under limited integration budgets.

Table 2: FIDs across different minibatch size and NFEs on CIFAR and ImageNet dataset.

Minibatch context	**CIFAR**-100 NFE 25	NFE 50	NFE 100	NFE 1000	Adaptive NFE	Adaptive FID
OT-CFM-128	12.85	10.49	9.20	8.00	142.40	7.89
EOT-CFM-2048	12.77	10.28	8.77	7.79	147.16	7.71
EOT-CFM-4096	12.05	9.80	8.59	7.46	147.80	7.36
EOT-CFM-8192	11.85	9.63	8.44	7.34	146.86	7.25

Minibatch context	**ImageNet**-64×64 NFE 25	NFE 50	NFE 100	NFE 1000	Adaptive NFE	Adaptive FID
OT-CFM-128	16.95	14.88	13.98	13.33	183.91	13.26
EOT-CFM-6144	16.82	14.69	13.80	13.24	188.17	13.20
EOT-CFM-8192	16.44	14.57	13.73	13.12	187.32	13.01
EOT-CFM-10240	16.36	14.47	13.59	13.05	187.19	12.94

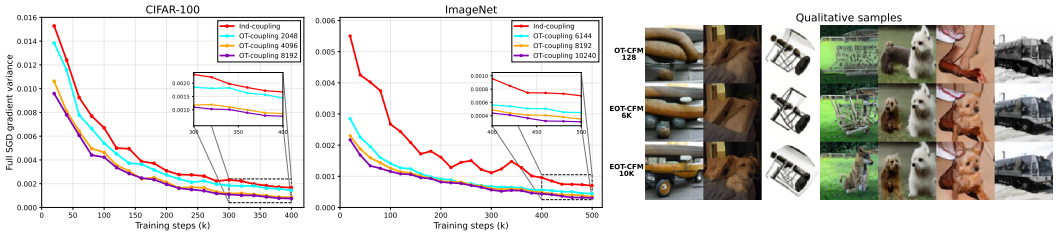

Figure 2: Full SGD gradient variance and qualitative samples. Variance during training on CIFAR-100 (**left**) and ImageNet (**middle**) under different minibatch size, with insets highlighting late-stage behavior. Samples from OT-CFM and EOT-CFM from 128 to 10K, generated at NFE 100. (**right**).

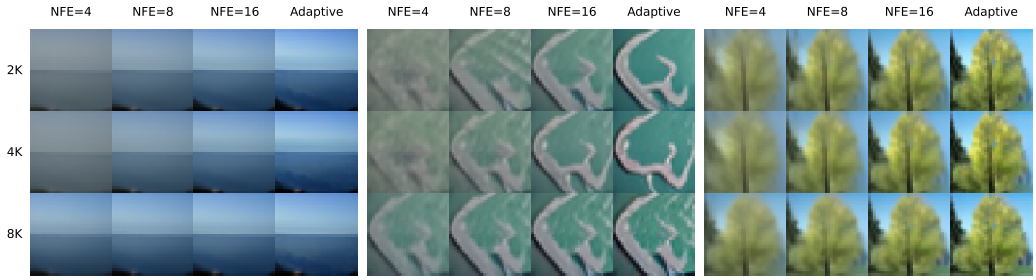

Figure 3: Generated images from initial noise. Rows correspond to minibatch sizes of 2K, 4K, and 8K, while columns compare Euler sampling at NFE 4, 8, and 16 with adaptive Dopri5. Smaller coupling minibatches tend to produce darker, less vivid images, whereas larger minibatches yield brighter samples with more vibrant colors, particularly at low NFEs.

Thanks to its bandwidth-efficient design and compatibility with PyTorch/Triton over HBM GPUs, FlashSinkhorn integrates naturally into efficient FM training. Increasing the minibatch size incurs only a modest increase in peak GPU memory. On ImageNet, the standard OT-CFM baseline reaches a peak memory usage of 39.96 GB, while FlashSinkhorn with a coupling size of 10K uses only 41.08 GB. For runtime, even at the largest context 10K, full OT coupling and pair sampling wall time account for less than 10% of the total time per gradient optimization step. The measured coupling latency is approximately 0.03 seconds at the 4K context and around 0.08 seconds at the 8K context. In addition to the benefit brought by entropic regularization, FlashAttention-style engineering also offers efficient GPU kernels, the native PyTorch integration removes the need for PyTorch-JAX tensor interchange and coordination between separate execution and memory-management runtimes, keeping coupling computation and pair sampling within the flow-matching optimization loop.

6 Conclusion

In this work, we study online coupling as a fundamental scaling axis for flow matching and fill the gap by establishing a strict reduction in full stochastic-gradient variance from independent pairing to population OT, characterize how larger online minibatch couplings recover this advantage, and extend the analysis to entropically regularized OT together with parallel guarantees for integration error. These results motivate a scalable approximation based on large online entropic couplings. Across both cell flow and image flow experiments, scaling the online coupling consistently improves training stability and generation quality. In the end, our results provide both a principled theoretical framework for rigorously analyzing the concrete effect brought by the more informative couplings and simple yet practical foundation for scaling flow matching.

AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing the writing and delivery of the paper content. We have not used generative AI tools for other source of disclosures listed by ICLR. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *NeurIPS*, 2017.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *ICML*, 2017.

Arip Asadulaev, Aleksandr Korotin, Vage Egiazarian, Petr Mokrov, and Evgeny Burnaev. Neural optimal transport with general cost functionals. In *ICLR*, 2024.

Ali Baheri. Geometry-grounded flow matching on compact manifolds. In *ICLR 2026 Workshop on Geometry-grounded Representation Learning and Generative Modeling*, 2026.

Mathias Beiglböck, Marcel Nutz, and Nizar Touzi. Complete duality for martingale optimal transport on the line. 2017.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Samuel Boïté, Julie Delon, and Kimia Nadjahi. Expected batch optimal transport plans and consequences for flow matching. *arXiv preprint arXiv:2605.12174*, 2026.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. In *Communications on Pure and Applied Mathematics*, 1991.

Sergio Calvo-Ordonez, Matthieu Meunier, Alvaro Cartea, Christoph Reisinger, Yarin Gal, and Jose Miguel Hernandez-Lobato. Weighted conditional flow matching. *arXiv preprint arXiv:2507.22270*, 2025.

Antonin Chambolle and Johannes Hertrich. Limit points of reflow with minibatch optimal transport. *arXiv preprint arXiv:2608.07042*, 2026.

Peter Chen, Yue Xie, and Qingpeng Zhang. SICNN: Sparsity-induced input convex neural network for optimal transport. In *Optimization for Machine Learning*, 2024.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Stephan Eckstein and Marcel Nutz. Quantitative stability of regularized optimal transport and convergence of sinkhorn's algorithm. *SIAM Journal on Mathematical Analysis*, 54(6):5922–5948, 2022.

Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Scalable computations of wasserstein barycenter via input convex neural networks. In *ICML*, 2021.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. In *AISTATS*, 2020.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. POT: Python optimal transport. In *JMLR*, 2021.

Alberto González-Sanz, Marcel Nutz, and Austin J Stromme. Finite-sample bounds for regularized optimal transport. *arXiv preprint arXiv:2606.25947*, 2026.

Mohammad Mohaiminul Islam, Thijs P Kuipers, Sharvaree Vadgama, Coen de Vente, Afsana Khan, Clara I Sánchez, and Erik J Bekkers. Longitudinal flow matching for trajectory modeling. In *AISTATS*, 2026.

Fairoz Nower Khan, Nabuat Zaman Nahim, Ruiquan Huang, Haibo Yang, and Peizhong Ju. Flow matching for offline reinforcement learning with discrete actions. *arXiv preprint arXiv:2602.06138*, 2026.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025.

Lingkai Kong, Molei Tao, Yang Liu, Bryan Wang, Jinmiao Fu, Chien-Chih Wang, and Huidong Liu. Alignflow: Improving flow-based generative models with semi-discrete optimal transport. In *ICLR*, 2026.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. In *NeurIPS*, 2024.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. In *ICLR*, 2021a.

Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *ICLR*, 2021b.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *ICLR*, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Tech Report*, 2009.

Tianyi Lin, Nhat Ho, and Michael Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *ICML*, 2019.

Tianyi Lin, Nhat Ho, Marco Cuturi, and Michael I Jordan. On the complexity of approximating multimarginal optimal transport. In *JMLR*, 2022.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Elon Litman. Scaled-dot-product attention as one-sided entropic optimal transport. *arXiv preprint arXiv:2508.08369*, 2025.

Elon Litman and Gabe Guo. You need better attention priors. In *ICML*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *ICML*, 2020.

Hugo Malamut and Maxime Sylvestre. Convergence rates of the regularized optimal transport: Disentangling suboptimality and entropy. In *SIAM Journal on Mathematical Analysis*, 2025.

Kevin R Moon, David Van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 2019.

Gilles Mordant. The entropic optimal (self-) transport problem: Limit distributions for decreasing regularization with application to score function estimation. *arXiv preprint arXiv:2412.12007*, 2024.

Alireza Mousavi-Hosseini, Stephen Zhang, Michal Klein, and Marco Cuturi. Flow matching with semidiscrete couplings. In *ICLR*, 2026.

Marcel Nutz and Johannes Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 2022.

Lukas Oesinghaus, Sören Becker, Larsen Vornholz, Efthymia Papalexi, Joey Pangallo, Amir Ali Moinfar, Jenni Liu, Alyssa La Fleur, Maiia Shulman, Simone Marrujo, et al. A single-cell cytokine dictionary of human peripheral blood. *BioRXiv*, 2025.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *ICML*, 2019.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *ICML*, 2023.

Teodora Reu, Sixtine Dromigny, Michael Bronstein, and Francisco Vargas. Gradient variance reveals failure modes in flow-based generative models. In *NeurIPS*, 2025.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, 2015.

Philippe Rigollet and Austin J Stromme. On the sample complexity of entropic optimal transport. *The Annals of Statistics*, 2025.

Chenxi Tao and Seung-Kyum Choi. On the role of strain and vorticity in numerical integration error for flow matching. *arXiv preprint arXiv:2605.06680*, 2026.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. In *TMLR*, 2024.

Jonathan Weed. An explicit analysis of the entropic penalty in linear programming. In *COLT*, 2018.

Felix X-F Ye, Xingjie Li, An Yu, Ming-Ching Chang, Linsong Chu, and Davis Wertheimer. Flashsinkhorn: Io-aware entropic optimal transport on gpu. In *ICML*, 2026.

Weili Zeng and Yichao Yan. Flow matching in the low-noise regime: Pathologies and a contrastive remedy. *arXiv preprint arXiv:2509.20952*, 2025.

Stephen Y. Zhang, Alireza Mousavi-Hosseini, Michal Klein, and Marco Cuturi. On fitting flow models with large sinkhorn couplings. In *TMLR*, 2026.

A Further Related Works

We walk through more related works and discuss some related theoretical results from both the FM and OT perspectives:

Flow matching and minibatch couplings. Normalizing flows and neural ODEs provide the modeling background for flow matching (Rezende & Mohamed, 2015; Chen et al., 2018), while flow matching and rectified flow introduce regression-based objectives for learning continuous transport dynamics (Lipman et al., 2023; Liu et al., 2023). More broadly, minibatch Wasserstein objectives have been studied as stochastic surrogates for population OT, including their asymptotic and gradient properties (Fatras et al., 2020); these results concern the minibatch transport objective itself rather than using its random plan as flow-matching supervision. Tong et al. (2024) introduce OT-CFM and connect the population OT coupling to dynamic OT, whereas Pooladian et al. (2023) analyze gradient covariance conditional on a fixed model input and establish asymptotic BatchOT straightness and transport-cost convergence.

Very recent work characterizes limit points of Reflow combined with fixed-size minibatch OT through cyclical monotonicity and straightness properties (Chambolle & Hertrich, 2026); we note that this concerns repeated Reflow rather than the optimization or integration behavior of a single online minibatch-CFM objective.

Scaling OT couplings for flow matching. Zhang et al. (2026) scales Sinkhorn couplings to millions of samples through an offline construction, where paired samples are precomputed independently of model training and served through an augmented data loader. We note that their finite-batch lower bound motivates increasing coupling size, but it does not establish a strict OT-versus-independent comparison or a full stochastic-gradient result. Boïté et al. (2026) characterize the expected exact-OT batch plan and its transport-cost convergence, while their integration result is restricted to an exact velocity field in a one-dimensional two-atom model, and they also show that when the population OT plan is unique, $\widehat{\pi}_k$ converges weakly to $\pi^\star$, i.e., $\widehat{\pi}_k \xrightarrow[k\to\infty]{w} \pi^\star$. However, we note that this coupling-level convergence does not imply any level of convergence for stochastic gradient variance.

Semidiscrete methods instead construct dataset-level noise-data alignments, avoiding a fresh discrete OT solve at every iteration (Mousavi-Hosseini et al., 2026; Kong et al., 2026). Calvo-Ordonez et al. (2025) replace minibatch OT with Gibbs-weighted independent pairs, avoiding an explicit OT solve but potentially altering the marginals, so larger minibatches provide more weighted pairs rather than scaling a larger candidate pool for optimizing source-target assignments. Our setting instead retains online couplings recomputed from fresh minibatches, with clear scaling benefits merely to thousand-level of minibatch size.

Scalable optimal transport. Brenier's characterization motivates representing quadratic-cost OT maps as gradients of convex potentials (Brenier, 1991), for which input-convex neural networks provide a convenient parameterization (Amos et al., 2017). This approach has been used to learn OT maps, plans, and Wasserstein barycenters (Makkuva et al., 2020; Korotin et al., 2021a; Fan et al., 2021; Korotin et al., 2021b; 2023; Chen et al., 2024; Asadulaev et al., 2024). Within flow matching, optimal flow matching constrains the velocity parameterization to recover the population quadratic-OT displacement directly (Kornilov et al., 2024). These methods learn a population-level map or plan through model structure or a separate neural optimization problem, rather than analyzing the random coupling recomputed from each flow-training minibatch.

Optimization and integration theory. Reu et al. (2025) analyze the variance of the full parameter gradient in restricted Gaussian transport models under specific analytic velocity parameterizations, and show that low gradient variance can also coexist with undesirable deterministic pairings. We note that these results do not associate independent pairing with random online minibatch OT in a general model class. For numerical integration, Tao & Choi (2026) decompose the velocity Jacobian into strain and vorticity and establish favorable Euler properties for exact OT displacement fields. This analysis concerns the geometry of a given velocity field rather than an integration-error ordering induced by finite minibatch couplings. Thus, neither line characterizes how full-gradient variance and integration error jointly change with coupling size and entropic regularization.

Computational entropic OT. Entropic regularization makes discrete OT amenable to Sinkhorn matrix scaling (Cuturi, 2013). Subsequent work establishes approximation-complexity guarantees for

Sinkhorn, Greenkhorn, accelerated variants, and multimarginal extensions (Altschuler et al., 2017; Lin et al., 2019; 2022), while POT provides widely used implementations (Flamary et al., 2021). FlashSinkhorn targets GPU efficiency by fusing Sinkhorn updates and avoiding materialization of the full coupling matrix in high-bandwidth memory (Ye et al., 2026), along with prior results with scaled dotproduct attention for onesided EOT problem (Litman, 2025) and later extension with improved learning prior (Litman & Guo, 2026).

Statistical and limiting theory of entropic OT. For finite linear programs, Weed (2018) give explicit non-asymptotic guarantees for convergence of entropically penalized solutions toward unregularized optima. At the population level, EOT potentials, plans, and costs converge toward their unregularized counterparts as $\varepsilon \to 0$ (Nutz & Wiesel, 2022; Malamut & Sylvestre, 2025). Statistical results establish parametric sample-complexity bounds at fixed regularization (Rigollet & Stromme, 2025), limit distributions when regularization decreases with sample size (Mordant, 2024), and finite-sample bias and variance bounds with explicit dependence on regularization and intrinsic dimension (González-Sanz et al., 2026).

B Missing Proofs

We first clarify a measurability point used throughout the proofs. Since $[0,1]\times\mathbb{R}^d$ and $\mathbb{R}^d$ are standard Borel spaces, a regular conditional law of U^π given (t, X_t^π) exists. We fix jointly measurable versions of its conditional mean and covariance. Any two such versions agree almost surely, so their values on null sets do not affect the results below.

B.1 Proposition 4.3

Proof. For any θ with $\mathcal{L}_\pi(\theta) < \infty$, conditioning on (t, X_t^π) and expanding the squared loss gives

$$\mathcal{L}_\pi(\theta) = \mathbb{E}_\pi\left[\|v_\theta(t, X_t^\pi) - u^\pi(t, X_t^\pi)\|_2^2\right] + \mathbb{E}_\pi\left[\|U^\pi - u^\pi(t, X_t^\pi)\|_2^2\right].$$

The second term is independent of θ. By realizability, the first term can attain zero. Therefore, global optimality implies $v_{\theta_\pi^\star}(t, X_t^\pi) = u^\pi(t, X_t^\pi)$ almost surely, and substituting this into the per-sample gradient yields

$$G_\pi(\theta_\pi^\star) = 2J_{\theta_\pi^\star}(t, X_t^\pi)^\top (u^\pi(t, X_t^\pi) - U^\pi).$$

So, we have

$$\mathbb{E}_\pi[G_\pi(\theta_\pi^\star) \mid t, X_t^\pi] = 0.$$

Thus, the conditional-mean term in Eq. (1) vanishes. Conditional on (t, X_t^π), the Jacobian is fixed, so the conditional gradient covariance is $4J_{\theta_\pi^\star}^\top \mathrm{Cov}_\pi(U^\pi \mid t, X_t^\pi) J_{\theta_\pi^\star}$, with every Jacobian evaluated at (t, X_t^π). Taking its expected trace would give us Eq. (3). □

B.2 Theorem 4.5

Before proving Theorem 4.5, we derive two supporting facts. First, we define the essential lower and upper model sensitivities at the coupling-specific population optimum by m_π and M_π, respectively:

$$m_\pi := \operatorname{ess\,inf} \lambda_{\min}(J_{\theta_\pi^\star}(t, X_t^\pi) J_{\theta_\pi^\star}(t, X_t^\pi)^\top), \qquad M_\pi := \operatorname{ess\,sup} \|J_{\theta_\pi^\star}(t, X_t^\pi)\|_{\mathrm{op}}^2.$$

Here, the essential extrema are taken under the distribution of (t, X_t^π), and together, they can yield the following variance sandwich bound:

Proposition B.1 (Variance sandwich)**.** *Under* Assumptions 4.1 *and* 4.2*, every coupling π under comparison satisfies*

$$4m_\pi \mathbb{E}_\pi\left[\|U^\pi - u^\pi(t, X_t^\pi)\|_2^2\right] \le \sigma_{\mathrm{grad},\pi}^2(\theta_\pi^\star), \tag{24}$$

and, if $M_\pi < \infty$, we have

$$\sigma_{\mathrm{grad},\pi}^2(\theta_\pi^\star) \le 4M_\pi \mathbb{E}_\pi\left[\|U^\pi - u^\pi(t, X_t^\pi)\|_2^2\right]. \tag{25}$$

Proof. For notation simplicity, we denote $C = \mathrm{Cov}_\pi(U^\pi \mid t, X_t^\pi) \succeq 0$ and $A = J_{\theta_\pi^\star} J_{\theta_\pi^\star}^\top \succeq 0$ (with the Jacobian evaluated at (t, X_t^π)). Cyclicity of the trace gives $\mathrm{Tr}(J_{\theta_\pi^\star}^\top C J_{\theta_\pi^\star}) = \mathrm{Tr}(CA)$, and

$$\lambda_{\min}(A)\,\mathrm{Tr}(C) \le \mathrm{Tr}(CA) \le \|A\|_{\mathrm{op}}\,\mathrm{Tr}(C) = \|J_{\theta_\pi^\star}\|_{\mathrm{op}}^2\,\mathrm{Tr}(C).$$

By the definition of the essential infimum, $\lambda_{\min}(A) \geq m_\pi$ almost surely. Applying the lower trace inequality pointwise and taking expectations in Eq. (3), we have

$$\sigma^2_{\text{grad},\pi}(\theta^\star_\pi) = 4\mathbb{E}_\pi[\mathrm{Tr}(CA)] \geq 4\mathbb{E}_\pi[m_\pi \,\mathrm{Tr}(C)] = 4m_\pi \mathbb{E}_\pi[\mathrm{Tr}(C)].$$

Similarly, when $M_\pi < \infty$, the definition of the essential supremum implies that

$$\|A\|_{\text{op}} = \|J_{\theta^\star_\pi}(t, X^\pi_t)\|^2_{\text{op}} \leq M_\pi \qquad \text{almost surely.}$$

Applying the upper trace inequality pointwise and taking expectations yields that

$$\sigma^2_{\text{grad},\pi}(\theta^\star_\pi) = 4\mathbb{E}_\pi[\mathrm{Tr}(CA)] \leq 4\mathbb{E}_\pi[M_\pi \,\mathrm{Tr}(C)] = 4M_\pi \mathbb{E}_\pi[\mathrm{Tr}(C)].$$

Finally, since $u^\pi(t, X^\pi_t) = \mathbb{E}_\pi[U^\pi \mid t, X^\pi_t]$, the definition of conditional covariance and the tower property gives us:

$$\begin{aligned}\mathbb{E}_\pi[\mathrm{Tr}(C)] &= \mathbb{E}_\pi\left[\mathbb{E}_\pi\left[\|U^\pi - u^\pi(t, X^\pi_t)\|_2^2 \,\middle|\, t, X^\pi_t\right]\right] \\ &= \mathbb{E}_\pi\left[\|U^\pi - u^\pi(t, X^\pi_t)\|_2^2\right].\end{aligned}$$

Substituting this identity into the preceding lower and upper bounds proves Eqs. (24) and (25), respectively. We also provide some further discussions over Jacobian norm and essential extrema in Remark B.2. □

Remark B.2. *At each (t, x), $\lambda_{\min}(J_{\theta^\star_\pi}(t, x)J_{\theta^\star_\pi}(t, x)^\top)$ and $\|J_{\theta^\star_\pi}(t, x)\|^2_{\text{op}}$ measure the weakest and strongest model sensitivities across output directions. Accordingly, m_π and M_π are their essential lower and upper bounds under the distribution of (t, X^π_t) (also note that the essential extrema are suitable here, because* Assumption 4.2 *requires differentiability only almost everywhere and are therefore unaffected by arbitrary values on null sets). We also note that the operator norm in M_π gives a sharper bound than the Frobenius norm, while playing the same role as the maximum Jacobian factor in Pooladian et al. (2023, Lemma 3.2).*

Second, we show the nondegeneracy supplied by an affine output layer with trainable bias.

Fact B.3. Over standard OT-CFM's MLP structure of width-d output layer trainable bias, let $\theta = (\eta, b)$, where $b \in \mathbb{R}^d$, one always have $v_{\eta,b}(t, x) = \bar{v}_\eta(t, x) + b$. We note that is explicitly shown in Tong et al. (2024, Appendix E.2) and its code implementation[1].

With Fact B.3, we then establish Lemma B.4:

Lemma B.4. *Under* Assumptions 4.1 *and* 4.2*, we have $J_{\theta^\star_\pi}(t, x)J_{\theta^\star_\pi}(t, x)^\top \succeq I_d$ almost everywhere, so $m_\pi \geq 1$ and $\sigma^2_{\text{grad},\pi}(\theta^\star_\pi) \geq 4\mathbb{E}_\pi\left[\|U^\pi - u^\pi(t, X^\pi_t)\|_2^2\right]$.*

Proof. Since $\nabla_b v_{\eta,b}(t, x) = I_d$, the full parameter Jacobian at the coupling-specific optimum is

$$J_{\theta^\star_\pi}(t, x) = \begin{bmatrix}\nabla_\eta \bar{v}_{\eta^\star_\pi}(t, x) & I_d\end{bmatrix}, \quad J_{\theta^\star_\pi}(t, x)J_{\theta^\star_\pi}(t, x)^\top = \nabla_\eta \bar{v}_{\eta^\star_\pi}(t, x)\nabla_\eta \bar{v}_{\eta^\star_\pi}(t, x)^\top + I_d \succeq I_d.$$

The conclusion follows from Eq. (24). □

We now prove the theorem:

Proof of Theorem 4.5. We first consider population OT. Choose a Borel representative of the Brenier map and a Borel set D of full q_0-measure on which $T = \nabla\varphi$. Monotonicity of gradients of convex functions gives us

$$\langle T(x) - T(y), x - y\rangle \geq 0, \qquad x, y \in D.$$

For $t < 1$, we write $F_t(x) = (1 - t)x + tT(x)$. If $x \neq y$, then we have

$$\langle F_t(x) - F_t(y), x - y\rangle \geq (1 - t)\|x - y\|_2^2 > 0,$$

so F_t is injective on D.

It's clear that Borel map $\Phi(t, x) = (t, F_t(x))$ from $[0, 1) \times D$ into $[0, 1) \times \mathbb{R}^d$ is injective. By the Lusin-Souslin theorem, its image is Borel and its inverse is Borel as well. Extending the inverse by a fixed point off its image yields a Borel map Ψ such that, almost surely, we have

$$X_0 = \mathrm{pr}_2[\Psi(t, X^{\pi^\star}_t)].$$

[1] See code at commit `cacd4dc`, where it uses `torch.nn.Linear` with `bias=True` by default

Since $t=1$ has probability zero, this recovers X_0 from $(t, X_t^{\pi^\star})$ almost surely, so $U^{\pi^\star} = T(X_0) - X_0$ is measurable with respect to $\sigma(t, X_t^{\pi^\star})$ (and also $\mathrm{Cov}_{\pi^\star}(U^{\pi^\star} \mid t, X_t^{\pi^\star}) = 0$ almost surely).

Assumptions 4.1 and 4.2 then give $v_{\theta^\star_{\pi^\star}}(t, X_t^{\pi^\star}) = u^{\pi^\star}(t, X_t^{\pi^\star}) = U^{\pi^\star}$ almost surely. Therefore, every per-sample gradient vanishes almost surely and $\sigma^2_{\mathrm{grad},\pi^\star}(\theta^\star_{\pi^\star}) = 0$.

We next consider independent coupling and first extend f_0, f_1 by zero outside Ω. For fixed t, the map

$$(x_0, x_1) \mapsto (x, u) = ((1-t)x_0 + tx_1, x_1 - x_0)$$

has inverse $x_0 = x - tu$, $x_1 = x + (1-t)u$, and absolute Jacobian determinant one. Therefore, (X_t, U) has joint density $h_t(x, u) = f_0(x - tu) f_1(x + (1-t)u)$. For notation clarity, we define

$$\rho_t(x) = \int h_t(x, u)\,\mathrm{d}u, \qquad m_t(x) = \rho_t(x)^{-1} \int u\, h_t(x, u)\,\mathrm{d}u,$$

and, when $\rho_t(x) > 0$, we can derive that

$$\mathrm{Cov}_{\pi_{\mathrm{ind}}}(U^{\pi_{\mathrm{ind}}} \mid t, X_t^{\pi_{\mathrm{ind}}} = x) = \rho_t(x)^{-1} \int (u - m_t(x))(u - m_t(x))^\top h_t(x, u)\,\mathrm{d}u,$$

which sets $m_t(x) = 0$ when $\rho_t(x) = 0$. Thus, ρ_t is the density of X_t, while note that m_t and above are versions of its conditional mean velocity and covariance, respectively.

Next, we fix $\bar{x} \in \Omega^\circ$ and $r > 0$ with $B(\bar{x}, 2r) \subset \Omega$.

For $x \in B(\bar{x}, r)$, $u \in B(0, r)$, and $t \in [0, 1]$, it's clear that both $x - tu$ and $x + (1-t)u$ lie in Ω. Therefore, we have

$$h_t(x, u) \geq \gamma^2, \qquad \rho_t(x) \geq \gamma^2 \omega_d r^d > 0.$$

For every $a \in \mathbb{R}^d$ and such (t, x), we can further derive

$$\begin{aligned}
\rho_t(x) a^\top \mathrm{Cov}_{\pi_{\mathrm{ind}}}(U^{\pi_{\mathrm{ind}}} \mid t, X_t^{\pi_{\mathrm{ind}}} = x) a &\geq \gamma^2 \int_{B(0,r)} (a^\top u - a^\top m_t(x))^2\,\mathrm{d}u \\
&\geq \frac{\gamma^2 \omega_d r^{d+2}}{d+2} \|a\|_2^2,
\end{aligned}$$

where the second step follows because the cross term integrates to zero by symmetry and $\int_{B(0,r)} uu^\top\,\mathrm{d}u = \omega_d r^{d+2} I_d/(d+2)$. Therefore, we can finally derive

$$\rho_t(x)\,\mathrm{Cov}_{\pi_{\mathrm{ind}}}(U^{\pi_{\mathrm{ind}}} \mid t, X_t^{\pi_{\mathrm{ind}}} = x) \succeq \frac{\gamma^2 \omega_d r^{d+2}}{d+2} I_d.$$

The joint law of (t, X_t) has density $\rho_t(x)\,\mathrm{d}t\,\mathrm{d}x$. Taking the trace and integrating over $t \in [0, 1]$ and $x \in B(\bar{x}, r)$ eventually yields

$$\mathbb{E}_{\pi_{\mathrm{ind}}}\left[\|U^{\pi_{\mathrm{ind}}} - u^{\pi_{\mathrm{ind}}}(t, X_t^{\pi_{\mathrm{ind}}})\|_2^2\right] \geq \frac{d\gamma^2 \omega_d^2}{d+2} r^{2d+2}.$$

By Lemma B.4, we have

$$\sigma^2_{\mathrm{grad},\pi_{\mathrm{ind}}}(\theta^\star_{\pi_{\mathrm{ind}}}) \geq \frac{4d\gamma^2 \omega_d^2}{d+2} r^{2d+2} > 0.$$

Subtracting the zero population-OT variance, we can conclude Eq. (4). □

B.3 Proposition 4.8

Proof. By Assumptions 4.1 and 4.2, $v_{\theta^\star_\pi}(t, X_t^\pi) = u^\pi(t, X_t^\pi)$ almost surely. Therefore, the loss at optimality equals the conditional target variance, and conditional orthogonality gives us that

$$\mathbb{E}_\pi[\|U^\pi\|_2^2] = \mathbb{E}_\pi[\|u^\pi(t, X_t^\pi)\|_2^2] + \mathcal{L}_\pi(\theta^\star_\pi). \tag{26}$$

For every smooth compactly supported test function ϕ, differentiation along the linear interpolation and conditioning on (t, X_t^π) at each fixed t, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\phi(X_t^\pi)] = \mathbb{E}[\nabla\phi(X_t^\pi)^\top U^\pi] = \mathbb{E}[\nabla\phi(X_t^\pi)^\top u^\pi(t, X_t^\pi)].$$

Thus, together with u^π, the marginal path induced by π satisfies the continuity equation and is admissible in the Benamou–Brenier formulation (Benamou & Brenier, 2000). So, its integrated kinetic energy is at least $W_2^2(q_0, q_1)$, i.e.,

$$\mathbb{E}_\pi[\|u^\pi(t, X_t^\pi)\|_2^2] \geq W_2^2(q_0, q_1).$$

Since $\mathbb{E}_\pi[\|U^\pi\|_2^2] = \mathbb{E}_\pi[\|X_1 - X_0\|_2^2]$, combining this inequality with Eq. (26) proves Eq. (7). For $\pi^\star$, recall its transport cost is exactly $W_2^2(q_0, q_1)$, so the same bound also gives $\mathcal{L}_{\pi^\star}(\theta^\star_{\pi^\star}) = 0$. □

B.4 Lemma 4.9

Proof. Let $0 < \varepsilon_1 < \varepsilon_2$. Since $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$, independent coupling is a finite-objective competitor, so both quantities are finite. We use π_{ε_2} as a competitor for the ε_1-problem and π_{ε_1} as a competitor for the ε_2-problem. Their optimality gives us that

$$\begin{aligned}
\frac{\mathbb{E}_{\pi_{\varepsilon_1}}[\|X_1 - X_0\|_2^2]}{2d} + \varepsilon_1 \,\mathrm{KL}(\pi_{\varepsilon_1} \mid \pi_{\text{ind}}) &\leq \frac{\mathbb{E}_{\pi_{\varepsilon_2}}[\|X_1 - X_0\|_2^2]}{2d} + \varepsilon_1 \,\mathrm{KL}(\pi_{\varepsilon_2} \mid \pi_{\text{ind}}), \\
\frac{\mathbb{E}_{\pi_{\varepsilon_2}}[\|X_1 - X_0\|_2^2]}{2d} + \varepsilon_2 \,\mathrm{KL}(\pi_{\varepsilon_2} \mid \pi_{\text{ind}}) &\leq \frac{\mathbb{E}_{\pi_{\varepsilon_1}}[\|X_1 - X_0\|_2^2]}{2d} + \varepsilon_2 \,\mathrm{KL}(\pi_{\varepsilon_1} \mid \pi_{\text{ind}}).
\end{aligned}$$

Adding these inequalities and canceling the transport costs yields

$$(\varepsilon_2 - \varepsilon_1)\,(\mathrm{KL}(\pi_{\varepsilon_2} \mid \pi_{\text{ind}}) - \mathrm{KL}(\pi_{\varepsilon_1} \mid \pi_{\text{ind}})) \leq 0.$$

Because $\varepsilon_2 > \varepsilon_1$, we have

$$\mathrm{KL}(\pi_{\varepsilon_1} \mid \pi_{\text{ind}}) \geq \mathrm{KL}(\pi_{\varepsilon_2} \mid \pi_{\text{ind}}).$$

Substituting this into the first optimality inequality further gives

$$\mathbb{E}_{\pi_{\varepsilon_1}}[\|X_1 - X_0\|_2^2] - \mathbb{E}_{\pi_{\varepsilon_2}}[\|X_1 - X_0\|_2^2] \leq 2d\varepsilon_1\,(\mathrm{KL}(\pi_{\varepsilon_2} \mid \pi_{\text{ind}}) - \mathrm{KL}(\pi_{\varepsilon_1} \mid \pi_{\text{ind}})) \leq 0.$$

Therefore, the relative entropy is non-increasing in ε, while the transport cost is non-decreasing:

$$\begin{aligned}
\mathrm{KL}(\pi_{\varepsilon_1} \mid \pi_{\text{ind}}) &\geq \mathrm{KL}(\pi_{\varepsilon_2} \mid \pi_{\text{ind}}), \\
\mathbb{E}_{\pi_{\varepsilon_1}}[\|X_1 - X_0\|_2^2] &\leq \mathbb{E}_{\pi_{\varepsilon_2}}[\|X_1 - X_0\|_2^2].
\end{aligned} \tag{27}$$

For every $0 < \varepsilon \leq \varepsilon_0$, we can now apply Eqs. (7), (8) and (27) in order to obtain

$$\begin{aligned}
\mathcal{L}_{\pi_\varepsilon}(\theta^\star_{\pi_\varepsilon}) &\leq \mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) \\
&\leq \mathbb{E}_{\pi_{\varepsilon_0}}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) \\
&< \mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}),
\end{aligned}$$

which proves the strict upper inequality in Eq. (9).

$$\exp\left(\frac{\varphi_\varepsilon(x_0) + \psi_\varepsilon(x_1) - \|x_1 - x_0\|_2^2/(2d)}{\varepsilon}\right) > 0$$

for π_{ind}-almost every (x_0, x_1) (Nutz & Wiesel, 2022). Thus, $\pi_\varepsilon \sim \pi_{\text{ind}}$, meaning that the two couplings have exactly the same null sets. Therefore, we have

$$\mathbb{E}_\pi[U^\pi \mid t, X_t^\pi] = f(t, X_t^\pi) = U^\pi \qquad \text{almost surely,}$$

so the loss is zero.

$$\big(t, x - tu, x + (1-t)u\big).$$

We write $P_\pi = \Phi_\#(\lambda \otimes \pi)$ for the induced law of (t, X_t^π, U^π). For any Borel set A, the definition of pushforward gives

$$\begin{aligned}
P_{\pi_\varepsilon}(A) = 0 \quad &\Longleftrightarrow \quad (\lambda \otimes \pi_\varepsilon)(\Phi^{-1}(A)) = 0 \\
&\Longleftrightarrow \quad (\lambda \otimes \pi_{\text{ind}})(\Phi^{-1}(A)) = 0 \\
&\Longleftrightarrow \quad P_{\pi_{\text{ind}}}(A) = 0.
\end{aligned}$$

Therefore, $P_{\pi_\varepsilon} \sim P_{\pi_{\text{ind}}}$. For any Borel measurable $f : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, its graph

$$A_f := \{(t, x, u) : u = f(t, x)\}$$

is Borel. Applying the preceding equivalence to A_f^c gives

$$P_{\pi_\varepsilon}(A_f) = 1 \quad \Longleftrightarrow \quad P_{\pi_{\text{ind}}}(A_f) = 1.$$

Thus, a measurable function determines U from (t, X_t) almost surely under one law exactly when it does under the other. Combining this with the zero-loss characterization above, we obtain

$$\mathcal{L}_{\pi_\varepsilon}(\theta^\star_{\pi_\varepsilon}) = 0 \quad \Longleftrightarrow \quad \mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}) = 0.$$

Therefore, as $\mathcal{L}_{\pi_\varepsilon}(\theta^\star_{\pi_\varepsilon}) > 0$, with Proposition 4.8 and optimality of $\pi^\star$, we end up with

$$0 \le \mathcal{L}_{\pi^\star}(\theta^\star_{\pi^\star}) \le \mathbb{E}_{\pi^\star}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) = 0.$$

Consequently, $\mathcal{L}_{\pi^\star}(\theta^\star_{\pi^\star}) = 0 < \mathcal{L}_{\pi_\varepsilon}(\theta^\star_{\pi_\varepsilon})$, which proves the strict lower inequality. □

B.5 Remark 4.10

Proof. Because $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$, the normalized quadratic cost satisfies

$$\int \frac{\|x_1 - x_0\|_2^2}{2d} (q_0 \otimes q_1)(dx_0, dx_1) \le \frac{1}{d} \left(\int \|x_0\|_2^2 \, q_0(dx_0) + \int \|x_1\|_2^2 \, q_1(dx_1) \right) < \infty.$$

The cost is also continuous, so vanishing-regularization convergence of entropic OT implies that every weak cluster point of π_ε as $\varepsilon \downarrow 0$ is a quadratic-OT optimizer (Nutz & Wiesel, 2022).

To pass from weak convergence to convergence of the unbounded quadratic costs, we make the uniform-integrability argument explicit. This is the same de la Vallée–Poussin argument used below in Lemma B.9; for similar one usage, also see the proof of Lemma 6.4 in Beiglböck et al. (2017). Since the squared norms under both marginals are integrable, the de la Vallée–Poussin criterion supplies a common nondecreasing, convex function $\Phi : [0, \infty) \to [0, \infty)$ such that

$$\lim_{s \to \infty} \frac{\Phi(s)}{s} = \infty, \qquad \int \Phi(\|x_0\|_2^2) \, q_0(dx_0) + \int \Phi(\|x_1\|_2^2) \, q_1(dx_1) < \infty.$$

Indeed, we can apply the criterion to the equally weighted mixture of the two squared-norm distributions, which has finite first moment. For every pair (x_0, x_1), we have

$$\frac{\|x_1 - x_0\|_2^2}{4} \le \frac{\|x_0\|_2^2 + \|x_1\|_2^2}{2}.$$

Therefore, monotonicity and convexity of Φ give us the following:

$$\Phi\left(\frac{\|x_1 - x_0\|_2^2}{4} \right) \le \frac{1}{2} \Phi(\|x_0\|_2^2) + \frac{1}{2} \Phi(\|x_1\|_2^2).$$

Integrating under any coupling $\pi \in \Pi(q_0, q_1)$, its fixed marginals yield the uniform bound

$$\sup_{\pi \in \Pi(q_0, q_1)} \int \Phi\left(\frac{\|x_1 - x_0\|_2^2}{4} \right) d\pi \le K = \frac{1}{2} \int \Phi(\|x_0\|_2^2) \, q_0(dx_0) + \frac{1}{2} \int \Phi(\|x_1\|_2^2) \, q_1(dx_1) < \infty.$$

For any $a > 0$, superlinearity gives a threshold $M_a < \infty$ such that $s \le a\Phi(\frac{s}{4})$ whenever $s > M_a$. Consequently, for every $M \ge M_a$, we have

$$\sup_{\pi \in \Pi(q_0, q_1)} \int \|x_1 - x_0\|_2^2 \, \mathbf{1}_{\{\|x_1 - x_0\|_2^2 > M\}} \, d\pi \le a \sup_{\pi \in \Pi(q_0, q_1)} \int \Phi\left(\frac{\|x_1 - x_0\|_2^2}{4} \right) d\pi \le aK.$$

Since $a > 0$ is arbitrary, we obtain

$$\tau(M) := \sup_{\pi \in \Pi(q_0, q_1)} \int \|x_1 - x_0\|_2^2 \, \mathbf{1}_{\{\|x_1 - x_0\|_2^2 > M\}} \, d\pi \longrightarrow 0 \qquad \text{as } M \to \infty.$$

Thus, the quadratic costs are uniformly integrable over all couplings with these fixed marginals.

We now use this bound to establish the cost limit. First, the family $\{\pi_\varepsilon\}_{\varepsilon > 0}$ is tight. To see this, for any $\delta > 0$, we choose compact sets $K_0, K_1 \subset \mathbb{R}^d$ such that $q_i(K_i^c) < \delta/2$. The marginal constraints give us:

$$\sup_{\varepsilon > 0} \pi_\varepsilon\big((K_0 \times K_1)^c\big) \le q_0(K_0^c) + q_1(K_1^c) < \delta.$$

Therefore, from any sequence $\varepsilon_n \downarrow 0$, we can extract a subsequence such that

$$\pi_{\varepsilon_{n_j}} \Rightarrow \bar{\pi}.$$

The limit $\bar{\pi}$ still has marginals q_0, q_1, and the vanishing-regularization result above gives us

$$\mathbb{E}_{\bar{\pi}}[\|X_1 - X_0\|_2^2] = W_2^2(q_0, q_1).$$

For fixed $M > 0$, we define

$$c_M := \min\{\|x_1 - x_0\|_2^2, M\}.$$

As c_M is bounded and continuous, weak convergence further gives us

$$\int c_M \, d\pi_{\varepsilon_{n_j}} \longrightarrow \int c_M \, d\bar{\pi}.$$

Moreover, since

$$0 \le \|x_1 - x_0\|_2^2 - c_M \le \|x_1 - x_0\|_2^2 \, \mathbf{1}_{\{\|x_1 - x_0\|_2^2 > M\}},$$

so we have

$$\begin{aligned}\left|\mathbb{E}_{\pi_{\varepsilon_{n_j}}}[\|X_1 - X_0\|_2^2] - \mathbb{E}_{\bar{\pi}}[\|X_1 - X_0\|_2^2]\right| &\le \left|\int c_M \, d\pi_{\varepsilon_{n_j}} - \int c_M \, d\bar{\pi}\right| \\ &\quad + \int (\|x_1 - x_0\|_2^2 - c_M) \, d\pi_{\varepsilon_{n_j}} + \int (\|x_1 - x_0\|_2^2 - c_M) \, d\bar{\pi} \\ &\le \left|\int c_M \, d\pi_{\varepsilon_{n_j}} - \int c_M \, d\bar{\pi}\right| + 2\tau(M).\end{aligned}$$

Taking $j \to \infty$ at fixed M, and then $M \to \infty$, we have

$$\mathbb{E}_{\pi_{\varepsilon_{n_j}}}[\|X_1 - X_0\|_2^2] \longrightarrow \mathbb{E}_{\bar{\pi}}[\|X_1 - X_0\|_2^2] = W_2^2(q_0, q_1).$$

This subsequence argument also gives convergence of the full family. Otherwise, since

$$\mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] \ge W_2^2(q_0, q_1),$$

there would exist $\delta > 0$ and a sequence $\varepsilon_n \downarrow 0$ such that

$$\mathbb{E}_{\pi_{\varepsilon_n}}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) \ge \delta$$

for every n. The preceding claim would produce a subsequence whose costs converge to $W_2^2(q_0, q_1)$, which is a contradiction, so we get

$$\mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] \longrightarrow W_2^2(q_0, q_1) \qquad \text{as } \varepsilon \downarrow 0. \tag{28}$$

Since $\mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}) > 0$, we can now choose $\varepsilon_0 > 0$ sufficiently small that

$$0 \le \mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) < \frac{\mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}})}{2} < \mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}), \qquad 0 < \varepsilon \le \varepsilon_0.$$

In particular, Eq. (8) holds at ε_0.

For the explicit cutoff, fix a competitor $\widetilde{\pi} \in \Pi(q_0, q_1)$ satisfying

$$0 < \mathrm{KL}(\widetilde{\pi} \| \pi_{\text{ind}}) < \infty, \qquad \mathbb{E}_{\widetilde{\pi}}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) < \mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}).$$

Multiplying the objective in Eq. (6) by $2d$, optimality of π_ε against $\widetilde{\pi}$ gives

$$\mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] + 2d\varepsilon \, \mathrm{KL}(\pi_\varepsilon \| \pi_{\text{ind}}) \le \mathbb{E}_{\widetilde{\pi}}[\|X_1 - X_0\|_2^2] + 2d\varepsilon \, \mathrm{KL}(\widetilde{\pi} \| \pi_{\text{ind}}).$$

Since $\mathrm{KL}(\pi_\varepsilon \| \pi_{\text{ind}}) \ge 0$, dropping its contribution and subtracting $W_2^2(q_0, q_1)$ yields

$$\mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) \le \mathbb{E}_{\widetilde{\pi}}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) + 2d\varepsilon \, \mathrm{KL}(\widetilde{\pi} \| \pi_{\text{ind}}).$$

The competitor condition ensures that

$$\Delta := \mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}) + W_2^2(q_0, q_1) - \mathbb{E}_{\widetilde{\pi}}[\|X_1 - X_0\|_2^2] > 0.$$

Thus, whenever $0 < \varepsilon < \frac{\Delta}{2d\,\mathrm{KL}(\widetilde{\pi}\|\pi_{\text{ind}})}$, which is exactly Eq. (10), we have

$$2d\varepsilon\,\mathrm{KL}(\widetilde{\pi}\|\pi_{\text{ind}}) < \Delta.$$

Substituting this into the preceding cost bound gives

$$\begin{aligned}\mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) &\le \mathbb{E}_{\widetilde{\pi}}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) + 2d\varepsilon\,\mathrm{KL}(\widetilde{\pi}\|\pi_{\text{ind}}) \\ &< \mathbb{E}_{\widetilde{\pi}}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) + \Delta \\ &= \mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}).\end{aligned}$$

So, Eq. (8) holds with this ε as the reference ε_0, and Lemma 4.9 gives the claimed strict ordering in Eq. (9).

Finally, we verify that a competitor satisfying both conditions always exists. By Eq. (28), choose a sufficiently small $\eta > 0$ such that

$$\mathbb{E}_{\pi_\eta}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) < \mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}),$$

and set $\widetilde{\pi} = \pi_\eta$. To establish finiteness of its relative entropy, we use π_{ind} as a competitor in the η-problem. Obviously, $\mathrm{KL}(\pi_{\text{ind}}\|\pi_{\text{ind}}) = 0$, and optimality gives us

$$\mathbb{E}_{\pi_\eta}[\|X_1 - X_0\|_2^2] + 2d\eta\,\mathrm{KL}(\pi_\eta\|\pi_{\text{ind}}) \le \mathbb{E}_{\pi_{\text{ind}}}[\|X_1 - X_0\|_2^2] < \infty.$$

Consequently, we have

$$0 \le \mathrm{KL}(\pi_\eta\|\pi_{\text{ind}}) \le \frac{\mathbb{E}_{\pi_{\text{ind}}}[\|X_1 - X_0\|_2^2] - \mathbb{E}_{\pi_\eta}[\|X_1 - X_0\|_2^2]}{2d\eta} \le \frac{\mathbb{E}_{\pi_{\text{ind}}}[\|X_1 - X_0\|_2^2]}{2d\eta} < \infty.$$

We still need to show that

$$\mathrm{KL}(\pi_\eta\|\pi_{\text{ind}}) > 0.$$

Again, we can consider the contradiction, i.e., suppose instead that

$$\mathrm{KL}(\pi_\eta\|\pi_{\text{ind}}) = 0.$$

Since relative entropy vanishes only when the two probability measures agree, this would imply $\pi_\eta = \pi_{\text{ind}}$. Applying Eq. (7) to independent coupling then gives

$$\mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}) \le \mathbb{E}_{\pi_{\text{ind}}}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) = \mathbb{E}_{\pi_\eta}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1) < \mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}),$$

which is a contradiction. Therefore, we can now conclude

$$0 < \mathrm{KL}(\pi_\eta\|\pi_{\text{ind}}) < \infty,$$

and $\widetilde{\pi} = \pi_\eta$ satisfies all the required conditions, so the guarantee we claimed holds true. □

B.6 Theorem 4.11

Before establish the proof, we first justify the sufficient condition for $\mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}) > 0$ stated in Lemma 4.9.

Suppose q_0 has a Lebesgue density $f_0 > 0$ almost everywhere and q_1 is not a Dirac mass. Under π_{ind}, the joint law of (X_0, X_1) is $f_0(x_0)\,dx_0\,q_1(dy)$. Fix $t \in (0, 1)$ and we write $X_t = (1-t)X_0 + tX_1$, then for any Borel sets $A, B \subseteq \mathbb{R}^d$, independence gives us that

$$\mathbb{P}(X_t \in A,\ X_1 \in B) = \int_B \int_{\mathbb{R}^d} \mathbf{1}_A\big((1-t)x_0 + ty\big) f_0(x_0)\,dx_0\,q_1(dy).$$

For each fixed y, we can do the change of variables $x = (1-t)x_0 + ty$, so that $x_0 = (x - ty)/(1-t)$ and $dx_0 = (1-t)^{-d}dx$. Substituting into the preceding integral and applying Tonelli's theorem yields the following

$$\begin{aligned}\mathbb{P}(X_t \in A,\ X_1 \in B) &= \int_B \int_A (1-t)^{-d} f_0\left(\frac{x - ty}{1-t}\right) dx\,q_1(dy) \\ &= \int_A \int_B (1-t)^{-d} f_0\left(\frac{x - ty}{1-t}\right) q_1(dy)\,dx.\end{aligned}$$

By taking $B = \mathbb{R}^d$, we can show that X_t has Lebesgue density

$$\rho_t(x) = (1-t)^{-d} \int_{\mathbb{R}^d} f_0\left(\frac{x - tz}{1-t}\right) q_1(dz).$$

For every x such that $0 < \rho_t(x) < \infty$, we abbreviate the following for clarity:

$$K_t(x, dy) = \frac{(1-t)^{-d} f_0\left(\frac{x-ty}{1-t}\right)}{\rho_t(x)} q_1(dy).$$

The preceding joint-probability identity then becomes $\mathbb{P}(X_t \in A,\ X_1 \in B) = \int_A K_t(x, B)\rho_t(x)\, dx$, so $K_t(x, \cdot)$ is a version of the conditional law of X_1 given $X_t = x$. Consequently, for almost every x under the law of X_t, we have

$$\frac{d\,\mathcal{L}(X_1 \mid X_t = x)}{dq_1}(y) = \frac{f_0\left(\frac{x-ty}{1-t}\right)}{\displaystyle\int_{\mathbb{R}^d} f_0\left(\frac{x - tz}{1-t}\right) q_1(dz)}.$$

The denominator depends only on t and x, so the Radon-Nikodym derivative is proportional to

$$y \mapsto f_0\big(\frac{x - ty}{1-t}\big).$$

We note that the derivative is strictly positive almost everywhere, so the conditional law is equivalent to the non-Dirac law q_1 and therefore has strictly positive covariance trace. Since

$$U^{\pi_{\text{ind}}} = \frac{X_1 - X_t^{\pi_{\text{ind}}}}{1-t},$$

its conditional covariance also has strictly positive trace. By integrating over $(t, X_t^{\pi_{\text{ind}}})$, we can prove $\mathcal{L}_{\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}) > 0$, and the condition holds for $q_0 = \mathcal{N}(0, I_d)$.

We now prove the theorem:

Proof. Following from Fact B.3, we write the affine output layer as $v_{\eta,b}(t,x) = \bar{v}_\eta(t,x) + b$. Since $\nabla_b v_{\eta,b}(t,x) = I_d$, its bias-gradient block at the population optimum is

$$\nabla_b \ell_\pi(\theta^\star_\pi; \xi^\pi) = 2\big(u^\pi(t, X_t^\pi) - U^\pi\big).$$

Its conditional mean at (t, X_t^π) is zero, so the law of total covariance therefore gives us that

$$\mathrm{Tr}\{\mathrm{Cov}_\pi(\nabla_b \ell_\pi(\theta^\star_\pi; \xi^\pi))\} = 4\mathcal{L}_\pi(\theta^\star_\pi). \tag{29}$$

For the complete gradient, Eq. (3) and the positive-semidefinite trace bound gives us that

$$\sigma^2_{\text{grad},\pi_\varepsilon}(\theta^\star_{\pi_\varepsilon}) \le 4\mathcal{L}_{\pi_\varepsilon}(\theta^\star_{\pi_\varepsilon})\, \operatorname{ess\,sup} \|J_{\theta^\star_{\pi_\varepsilon}}(t, X_t^{\pi_\varepsilon})\|^2_{\text{op}}.$$

Combining this inequality with Eq. (7) proves the upper bound in Eq. (11). For the strict lower bound, Eq. (29) and Lemma 4.9 show that the bias-gradient block has strictly positive variance for every finite ε.

Since the total gradient-variance trace is at least the trace of this principal block, we can show that $\sigma^2_{\text{grad},\pi_\varepsilon}(\theta^\star_{\pi_\varepsilon}) > 0$. At the same time, population OT has zero gradient variance by Eqs. (3) and (7). Therefore, whenever the right-hand side of Eq. (11) is smaller than the independent-coupling variance, the strict chain in Eq. (12) follows.

Finally, the independent-coupling variance is strictly positive by Eq. (29). The theorem's uniform condition means that there exist $\varepsilon_J > 0$ and $M < \infty$ such that $\{\pi_\varepsilon : 0 < \varepsilon \le \varepsilon_J\}$ satisfies Definition 4.6 with bound M. Thus, for every $0 < \varepsilon \le \varepsilon_J$, we have

$$\operatorname{ess\,sup} \|J_{\theta^\star_{\pi_\varepsilon}}(t, X_t^{\pi_\varepsilon})\|^2_{\text{op}} \le \max_{t\in[0,1],\, x\in\mathbb{R}^d} \|J_{\theta^\star_{\pi_\varepsilon}}(t,x)\|^2_{\text{op}} \le M.$$

We denote $D_\varepsilon := \mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1)$, Eq. (28) then gives us that $4MD_\varepsilon \to 0$ as $\varepsilon \downarrow 0$. Therefore, there exists $\varepsilon_{\text{g}} > 0$ such that

$$4MD_\varepsilon < \sigma^2_{\text{grad},\pi_{\text{ind}}}(\theta^\star_{\pi_{\text{ind}}}) \qquad \text{for every } 0 < \varepsilon \le \varepsilon_{\text{g}}.$$

By setting $\varepsilon_1 = \min\{\varepsilon_0, \varepsilon_J, \varepsilon_{\text{g}}\} > 0$, then for every $0 < \varepsilon \le \varepsilon_1$, the right-hand side of Eq. (11) is smaller than the independent-coupling variance, which concludes Eq. (12). □

B.7 PROPOSITION 4.12

Proof. We take $\widehat{q}_{0,k}$ and $\widehat{q}_{1,k}$ to denote the empirical distributions of the two endpoint batches used to construct π_k. Note that conditional on these batches, the OT solution is a permutation and the selected index is uniform.

Therefore, the selected source and target each remain uniformly distributed over their respective batches; after averaging over the batches, we have $\pi_k \in \Pi(q_0, q_1)$. Moreover, one can derive that

$$\mathbb{E}_{\pi_k}[\|X_1 - X_0\|_2^2] = \mathbb{E}[W_2^2(\widehat{q}_{0,k}, \widehat{q}_{1,k})].$$

To establish this, we first show that, for any $q \in \mathcal{P}_2(\mathbb{R}^d)$ along with its empirical distribution $\widehat{q}_k$, $\mathbb{E}[W_2^2(\widehat{q}_k, q)]$ will converge to zero.

For any $R > 0$, let P_R be the metric projection onto the closed ball $\overline{B}(0, R)$, and let $q^R = (P_R)_{\#}q$ and $\widehat{q}_k^R = (P_R)_{\#}\widehat{q}_k$, coupling each sample with its projection gives us that

$$\mathbb{E}[W_2^2(\widehat{q}_k, \widehat{q}_k^R)] \le a_R := \mathbb{E}_q[\|X - P_R(X)\|_2^2], \qquad W_2^2(q, q^R) \le a_R.$$

For fixed $R > 0$, the empirical measure $\widehat{q}_k^R$ converges to q^R almost surely in W_2. Both measures are supported on $\overline{B}(0, R)$, whose diameter is $2R$, so $W_2^2(\widehat{q}_k^R, q^R) \le 4R^2$. The almost-sure convergence of $\widehat{q}_k^R$ to q^R in W_2 further gives

$$W_2^2(\widehat{q}_k^R, q^R) \longrightarrow 0 \qquad \text{almost surely.}$$

Since both measures are supported on the radius-R ball, we also have $0 \le W_2^2(\widehat{q}_k^R, q^R) \le 4R^2$. This deterministic bound is integrable and independent of k, so dominated convergence gives

$$\lim_{k\to\infty} \mathbb{E}[W_2^2(\widehat{q}_k^R, q^R)] = \mathbb{E}\left[\lim_{k\to\infty} W_2^2(\widehat{q}_k^R, q^R)\right] = 0.$$

Next, the triangle inequality gives

$$W_2(\widehat{q}_k, q) \le W_2(\widehat{q}_k, \widehat{q}_k^R) + W_2(\widehat{q}_k^R, q^R) + W_2(q^R, q).$$

First using $(a + b + c)^2 \le 3(a^2 + b^2 + c^2)$ and then simply taking expectations, we have

$$\begin{aligned}\mathbb{E}[W_2^2(\widehat{q}_k, q)] &\le 3\mathbb{E}[W_2^2(\widehat{q}_k, \widehat{q}_k^R)] + 3\mathbb{E}[W_2^2(\widehat{q}_k^R, q^R)] + 3W_2^2(q^R, q) \\ &\le 3a_R + 3\mathbb{E}[W_2^2(\widehat{q}_k^R, q^R)] + 3a_R \\ &= 6a_R + 3\mathbb{E}[W_2^2(\widehat{q}_k^R, q^R)],\end{aligned}$$

where the second inequality uses the two projection bounds established above. Taking the limit superior as $k \to \infty$, while keeping R fixed, therefore gives

$$\limsup_{k\to\infty} \mathbb{E}[W_2^2(\widehat{q}_k, q)] \le 6a_R + 3 \lim_{k\to\infty} \mathbb{E}[W_2^2(\widehat{q}_k^R, q^R)] = 6a_R.$$

Since $q \in \mathcal{P}_2(\mathbb{R}^d)$, $a_R \to 0$ as $R \to \infty$, which proves our claim. Applying it to both marginals yields

$$|W_2(\widehat{q}_{0,k}, \widehat{q}_{1,k}) - W_2(q_0, q_1)| \le W_2(\widehat{q}_{0,k}, q_0) + W_2(\widehat{q}_{1,k}, q_1),$$

The right-hand side converges to 0 in L^2, so $W_2(\widehat{q}_{0,k}, \widehat{q}_{1,k}) \to W_2(q_0, q_1)$ in L^2. Its square therefore converges in L^1, which proves $\Delta_k \to 0$ in Eq. (13). Since π_k is a coupling of q_0 and q_1, its cost is at least $W_2^2(q_0, q_1)$, so $\Delta_k \ge 0$.

For any coupling π, conditional orthogonality gives us that

$$\mathbb{E}_\pi[\|U^\pi\|_2^2] = \mathbb{E}_\pi[\|u^\pi(t, X_t^\pi)\|_2^2] + \mathbb{E}_\pi[\mathrm{Tr}\{\mathrm{Cov}_\pi(U^\pi \mid t, X_t^\pi)\}].$$

As earlier shown in the proof of Proposition 4.8, the interpolation induced by π, together with u^π, is admissible in the Benamou–Brenier formulation. Therefore, $\mathbb{E}_\pi[\|u^\pi(t, X_t^\pi)\|_2^2] \ge W_2^2(q_0, q_1)$.

Applying the preceding kinetic-energy bound to π_k, we have

$$\mathbb{E}_{\pi_k}\left[\|u^{\pi_k}(t, X_t^{\pi_k})\|_2^2\right] \ge W_2^2(q_0, q_1).$$

We can then rearrange the conditional-orthogonality identity and use $U^{\pi_k} = X_1 - X_0$ to get

$$\begin{aligned}
0 &\le \mathbb{E}_{\pi_k}[\mathrm{Tr}\{\mathrm{Cov}_{\pi_k}(U^{\pi_k} \mid t, X_t^{\pi_k})\}] \\
&= \mathbb{E}_{\pi_k}\left[\|U^{\pi_k} - u^{\pi_k}(t, X_t^{\pi_k})\|_2^2\right] && (\because \text{conditional-variance identity}) \\
&= \mathbb{E}_{\pi_k}\left[\|U^{\pi_k}\|_2^2\right] - \mathbb{E}_{\pi_k}\left[\|u^{\pi_k}(t, X_t^{\pi_k})\|_2^2\right] && (\because \text{conditional orthogonality}) \\
&\le \mathbb{E}_{\pi_k}\left[\|U^{\pi_k}\|_2^2\right] - W_2^2(q_0, q_1) && (\because \text{mean-velocity energy lower bound}) \\
&= \mathbb{E}_{\pi_k}\left[\|X_1 - X_0\|_2^2\right] - W_2^2(q_0, q_1) = \Delta_k.
\end{aligned}$$

This proves the upper bound in Eq. (14). Since we have already established that $\Delta_k \to 0$, the squeeze theorem further gives

$$\mathbb{E}_{\pi_k}[\mathrm{Tr}\{\mathrm{Cov}_{\pi_k}(U^{\pi_k} \mid t, X_t^{\pi_k})\}] \longrightarrow 0 \qquad \text{as } k \to \infty.$$

So, we prove $\Delta_k \to 0$. □

B.8 THEOREM 4.13

Before moving into the proof of Theorem 4.13. We first provide the following supporting lemma, which shows that every finite-batch coupling has zero target uncertainty exactly when independent coupling does.

Lemma B.5. *For every finite $k \ge 1$, π_k and π_{ind} are mutually absolutely continuous, so it has*

$$\mathbb{E}_{\pi_k}[\mathrm{Tr}\{\mathrm{Cov}_{\pi_k}(U^{\pi_k} \mid t, X_t^{\pi_k})\}] = 0 \quad \Longleftrightarrow \quad \mathbb{E}_{\pi_{\text{ind}}}[\mathrm{Tr}\{\mathrm{Cov}_{\pi_{\text{ind}}}(U^{\pi_{\text{ind}}} \mid t, X_t^{\pi_{\text{ind}}})\}] = 0.$$

If $q_0 = f_0 \mathcal{L}^d$ with $f_0 > 0$ Lebesgue-almost everywhere and q_1 is not a Dirac mass, both quantities are strictly positive.

Proof. We write the two endpoint batches as $(X_1, \dots, X_k)$ and $(Y_1, \dots, Y_k)$, and let I be uniform on $\{1, \dots, k\}$, independently of both batches, along with τ be the permutation returned by a fixed measurable tie-breaking rule for empirical OT. Then, $(X_I, Y_{\tau(I)}) \sim \pi_k$.

For any Borel set $E \subset \mathbb{R}^d \times \mathbb{R}^d$, if $(q_0 \otimes q_1)(E) = 0$, all finitely many candidate pairs (X_i, Y_j) avoid E almost surely, so $\pi_k(E) = 0$. Conversely, we denote $E_x = \{y : (x, y) \in E\}$, and the event that every candidate target paired with X_I lies in E_{X_I} forces the selected pair to lie in E. Therefore, we get

$$\pi_k(E) \ge \mathbb{P}((X_I, Y_j) \in E \text{ for every } j) = \int q_1(E_x)^k \, q_0(\mathrm{d}x).$$

If $(q_0 \otimes q_1)(E) > 0$, then $q_1(E_x) > 0$ on a set of positive q_0-measure, so the final integral is strictly positive. Thus, one have $\pi_k \sim \pi_{\text{ind}}$.

Next, as we discussed before, the map $(t, x_0, x_1) \mapsto (t, (1-t)x_0 + tx_1, x_1 - x_0)$ is a Borel bijection with inverse $x_0 = x - tu$, $x_1 = x + (1-t)u$. Therefore, equivalent endpoint couplings induce equivalent laws of (t, X_t, U).

As the covariance trace is nonnegative, its expectation is zero exactly when $\mathrm{Cov}(U \mid t, X_t) = 0$ almost surely, or equivalently when U is almost surely a measurable function of (t, X_t). This graph property is preserved under equivalent laws, proving the stated equivalence.

Finally, using the stated density conditions and fix $t \in (0, 1)$. Under independent coupling, for almost every x under the law of X_t, the conditional law of X_1 given $X_t = x$ has Radon–Nikodym derivative with respect to q_1 proportional to

$$y \longmapsto f_0\left(\frac{x - ty}{1-t}\right).$$

We claim that this derivative is strictly positive for q_1-almost every y. To prove it, we set $t \in (0, 1)$ and let

$$N := \{z \in \mathbb{R}^d : f_0(z) = 0\}.$$

The assumption $f_0 > 0$ Lebesgue-almost everywhere gives $\mathcal{L}^d(N) = 0$. For every fixed $y \in \mathbb{R}^d$, the change of variables $x = (1-t)z + ty$ gives

$$\int_{\mathbb{R}^d} \mathbf{1}_N\left(\frac{x - ty}{1-t}\right) dx = (1-t)^d \int_{\mathbb{R}^d} \mathbf{1}_N(z)\, dz = (1-t)^d \mathcal{L}^d(N) = 0.$$

Thus, for each fixed y, the numerator $f_0((x-ty)/(1-t))$ is positive for Lebesgue-almost every x. We now interchange the order of integration. Since the indicator is nonnegative, Tonelli's theorem gives us

$$\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\mathbf{1}_N\left(\frac{x-ty}{1-t}\right)\,q_1(dy)\right]dx=\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}\mathbf{1}_N\left(\frac{x-ty}{1-t}\right)\,dx\right]q_1(dy)=0.$$

The inner integral on the left is nonnegative, so it must vanish for Lebesgue-almost every x. Equivalently, we have

$$q_1\left(\left\{y: f_0\left(\frac{x-ty}{1-t}\right)=0\right\}\right)=0 \qquad \text{for Lebesgue-almost every } x.$$

Therefore, for Lebesgue-almost every x, we have

$$f_0\left(\frac{x-ty}{1-t}\right)>0 \qquad \text{for } q_1\text{-almost every } y.$$

We still remain for considering the normalizing factor in the conditional density. Under independent coupling, X_t has Lebesgue density

$$\rho_t(x)=(1-t)^{-d}\int_{\mathbb{R}^d}f_0\left(\frac{x-ty}{1-t}\right)\,q_1(dy).$$

The preceding positivity gives $\rho_t(x)>0$ for Lebesgue-almost every x. Moreover, we can show that

$$\int_{\mathbb{R}^d}\rho_t(x)\,dx=\int_{\mathbb{R}^d}\left[(1-t)^{-d}\int_{\mathbb{R}^d}f_0\left(\frac{x-ty}{1-t}\right)\,dx\right]q_1(dy)=\int_{\mathbb{R}^d}\left[\int_{\mathbb{R}^d}f_0(z)\,dz\right]q_1(dy)=1,$$

so $\rho_t(x)<\infty$ for Lebesgue-almost every x. Since the law of X_t is absolutely continuous with respect to Lebesgue measure, all these properties also hold for almost every x under the law of X_t. For such x, the conditional Radon–Nikodym derivative therefore satisfies

$$\frac{d\mathcal{L}(X_1\mid X_t=x)}{dq_1}(y)=\frac{(1-t)^{-d}f_0\left(\frac{x-ty}{1-t}\right)}{\rho_t(x)}>0 \qquad \text{for } q_1\text{-almost every } y.$$

Thus, for almost every x under the law of X_t, the conditional law of X_1 given $X_t=x$ is equivalent to q_1, which concludes our claim. Furthermore, the conditional law is equivalent to the non-Dirac distribution q_1 and has strictly positive covariance trace.

Since $U=(X_1-X_t)/(1-t)$, the conditional covariance of U also has strictly positive trace. Integrating first over X_t and then over $t\in(0,1)$ proves positivity for independent coupling and therefore, by the preceding equivalence, for every $k<\infty$. □

We now formally prove Theorem 4.13:

Proof of Theorem 4.13. At the coupling-specific population optimum, Eq. (3) and the positive-semidefinite trace bound used in Proposition B.1 give

$$0\le\sigma^2_{\mathrm{grad},\pi_k}(\theta^\star_{\pi_k})\le 4\,\operatorname{ess\,sup}\|J_{\theta^\star_{\pi_k}}(t,X_t^{\pi_k})\|^2_{\mathrm{op}}\,\mathbb{E}_{\pi_k}[\mathrm{Tr}\{\mathrm{Cov}_{\pi_k}(U^{\pi_k}\mid t,X_t^{\pi_k})\}].$$

Applying Eq. (14) proves the upper bound in Eq. (15). The uniform condition means that there exist $K_J\in\mathbb{N}$ and $M<\infty$ such that, for every $k\ge K_J$, the following essential optima control exists:

$$\operatorname{ess\,sup}\|J_{\theta^\star_{\pi_k}}(t,X_t^{\pi_k})\|^2_{\mathrm{op}}\le\max_{t\in[0,1],\,x\in\mathbb{R}^d}\|J_{\theta^\star_{\pi_k}}(t,x)\|^2_{\mathrm{op}}\le M.$$

Since $\Delta_k\to 0$, the right-hand side of Eq. (15) is at most $4M\Delta_k\to 0$ for $k\ge K_J$. The squeeze theorem therefore gives $\sigma^2_{\mathrm{grad},\pi_k}(\theta^\star_{\pi_k})\to 0$, so Eq. (16) holds for all sufficiently large k whenever the fixed baseline variance is positive.

For completeness, we also connect the per-sample statement to an update averaged over all matched pairs.

We denote $\{G_{k,1},\dots,G_{k,k}\}$ as the gradients generated at $\theta^\star_{\pi_k}$ by the k matched pairs in one empirical OT batch, with an independent uniform time sampled for each pair, and let I be an independent

uniform index. Then, $G_{k,I}$ has the same law as $G_{\pi_k}(\theta^\star_{\pi_k})$. The proof of Proposition 4.3 gives $\mathbb{E}[G_{k,I}] = 0$, so the averaged update also has mean zero. Jensen's inequality further gives us the following bound:

$$\mathrm{Tr}\Big\{\mathrm{Cov}\Big(k^{-1}\textstyle\sum_{i=1}^k G_{k,i}\Big)\Big\} = \mathbb{E}\left[\left\|k^{-1}\textstyle\sum_{i=1}^k G_{k,i}\right\|_2^2\right] \le \mathbb{E}[\|G_{k,I}\|_2^2] = \sigma^2_{\mathrm{grad},\pi_k}(\theta^\star_{\pi_k}).$$

Thus, the averaged matched-batch update also has variance converging to zero under the same vanishing-product condition. □

B.9 Proposition 4.14

To ensure rigor, we first establish several lemmas required to pass from convergence of entropic couplings to convergence of their target uncertainty.

Lemma B.6. *If $\pi_n, \pi \in \Pi(q_0, q_1)$, with $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$, and $\pi_n \to \pi$ in W_2, then we have*

$$\limsup_{n\to\infty} \mathsf{R}(\pi_n) \le \mathsf{R}(\pi). \tag{30}$$

Proof. We write P_n and P as the laws of $((t, X_t), U)$ induced by π_n and π, respectively. Then, we can couple π_n and π optimally and use the same independent t for both couplings.

Since $X_t = (1-t)X_0 + tX_1$ and $U = X_1 - X_0$ are linear in the endpoints, $P_n \to P$ in W_2. The conditional-mean characterization of squared-error regression gives

$$\mathsf{R}(\pi) = \inf_f \mathbb{E}_P[\|U - f(t, X_t)\|_2^2],$$

where the infimum may be restricted to bounded continuous f, which are dense in the relevant L^2 space. For every such fixed f, the integrand is continuous with quadratic growth, so its expectation converges under W_2 convergence. Therefore, we end up with

$$\limsup_{n\to\infty} \mathsf{R}(\pi_n) \le \mathbb{E}_P[\|U - f(t, X_t)\|_2^2].$$

Taking the infimum over f proves Eq. (30). □

For the convergence of empirical entropic plans, we first try to conclude the following local value bound. Before that, for notation simplicity, given probability measure μ and ν, we define

$$\mathsf{S}_\varepsilon(\mu, \nu) := \inf_{\rho \in \Pi(\mu,\nu)} \left\{ \frac{1}{2d}\mathbb{E}_\rho[\|X_1 - X_0\|_2^2] + \varepsilon\,\mathrm{KL}(\rho\|\mu\otimes\nu) \right\}.$$

Lemma B.7. *Let $\widehat{q}_{0,k}$ and $\widehat{q}_{1,k}$ be the two empirical endpoint distributions, set $a_k = W_2(\widehat{q}_{0,k}, q_0) + W_2(\widehat{q}_{1,k}, q_1)$, and let $C_{\mathrm{ind}} = \mathbb{E}_{\pi_{\mathrm{ind}}}[\|X_1 - X_0\|_2^2]$. Then, for every $\varepsilon > 0$, we have*

$$\mathsf{S}_\varepsilon(\widehat{q}_{0,k}, \widehat{q}_{1,k}) \le \mathsf{S}_\varepsilon(q_0, q_1) + \tfrac{1}{2d}(2\sqrt{C_{\mathrm{ind}}}\,a_k + a_k^2). \tag{31}$$

Proof. Let ρ_ε minimize $\mathsf{S}_\varepsilon(q_0, q_1)$, and transport each of its marginals to the corresponding empirical marginal through optimal W_2 kernels. Applying the product of these kernels to ρ_ε produces a coupling $\rho'_\varepsilon \in \Pi(\widehat{q}_{0,k}, \widehat{q}_{1,k})$.

We claim that $\mathrm{KL}(\rho'_\varepsilon\|\widehat{q}_{0,k}\otimes\widehat{q}_{1,k}) \le \mathrm{KL}(\rho_\varepsilon\|q_0\otimes q_1)$. To prove it, we fix the sampled endpoint batches and let $K_i(x_i, d\widehat{x}_i)$ be the transport kernel from q_i to $\widehat{q}_{i,k}$, for $i \in \{0,1\}$. Thus, for every Borel set $A_i \subseteq \mathbb{R}^d$, we have $\int K_i(x_i, A_i)\, q_i(dx_i) = \widehat{q}_{i,k}(A_i)$.

We denote $x = (x_0, x_1)$ and $\widehat{x} = (\widehat{x}_0, \widehat{x}_1)$, and the product kernel is $K_0(x_0, d\widehat{x}_0)K_1(x_1, d\widehat{x}_1)$. For any probability measure μ on the endpoint space, the law obtained by applying this kernel to μ assigns to each Borel set A with $\int \left(\int_A K_0(x_0, d\widehat{x}_0)K_1(x_1, d\widehat{x}_1)\right)\mu(dx)$. So, we then have

$$\rho'_\varepsilon(A) = \int \left(\int_A K_0(x_0, d\widehat{x}_0)K_1(x_1, d\widehat{x}_1)\right)\rho_\varepsilon(dx).$$

Applying the same kernel to the independent reference measure gives, for every Borel rectangle $A_0 \times A_1$, we can do the following derivation:

$$\begin{aligned}\int \left(\int_{A_0 \times A_1} K_0(x_0, d\widehat{x}_0) K_1(x_1, d\widehat{x}_1) \right) (q_0 \otimes q_1)(dx) &= \int K_0(x_0, A_0) K_1(x_1, A_1)\, q_0(dx_0) q_1(dx_1) \\ &= \left(\int K_0(x_0, A_0)\, q_0(dx_0) \right) \left(\int K_1(x_1, A_1)\, q_1(dx_1) \right) \\ &= \widehat{q}_{0,k}(A_0) \widehat{q}_{1,k}(A_1) \\ &= (\widehat{q}_{0,k} \otimes \widehat{q}_{1,k})(A_0 \times A_1).\end{aligned}$$

Since these rectangles determine the probability measure, for every Borel set A, we have

$$\int \left(\int_A K_0(x_0, d\widehat{x}_0) K_1(x_1, d\widehat{x}_1) \right) (q_0 \otimes q_1)(dx) = (\widehat{q}_{0,k} \otimes \widehat{q}_{1,k})(A).$$

We next derive the entropy inequality directly. Since independent coupling is a finite-objective competitor and $\varepsilon > 0$, optimality gives $\mathrm{KL}(\rho_\varepsilon \mid q_0 \otimes q_1) < \infty$. Indeed, since $\rho_\varepsilon \ll q_0 \otimes q_1$, the density ratio $\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(x)$ is then well defined.

Now, we consider the joint reference law $(q_0 \otimes q_1)(dx) K_0(x_0, d\widehat{x}_0) K_1(x_1, d\widehat{x}_1)$. Under this law, the original endpoint pair X has law $q_0 \otimes q_1$, while the transformed endpoint pair $\widehat{X}$ has law $\widehat{q}_{0,k} \otimes \widehat{q}_{1,k}$. We note that all the expectations below are taken under this joint reference law.

For any Borel set A in the transformed endpoint space, the definition of ρ_ε and the conditional-expectation identity give

$$\begin{aligned}\rho'_\varepsilon(A) &= \int \left(\int_A K_0(x_0, d\widehat{x}_0) K_1(x_1, d\widehat{x}_1) \right) \rho_\varepsilon(dx) \\ &= \int \frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(x) \left(\int_A K_0(x_0, d\widehat{x}_0) K_1(x_1, d\widehat{x}_1) \right) (q_0 \otimes q_1)(dx) \\ &= \mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X) \mathbf{1}_{\{\widehat{X} \in A\}} \right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X) \,\middle|\, \widehat{X} \right] \mathbf{1}_{\{\widehat{X} \in A\}} \right] \\ &= \int_A \mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X) \,\middle|\, \widehat{X} = \widehat{x} \right] (\widehat{q}_{0,k} \otimes \widehat{q}_{1,k})(d\widehat{x}).\end{aligned}$$

Therefore, the transformed density ratio is

$$\frac{d\rho'_\varepsilon}{d(\widehat{q}_{0,k} \otimes \widehat{q}_{1,k})}(\widehat{x}) = \mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X) \,\middle|\, \widehat{X} = \widehat{x} \right]$$

for $(\widehat{q}_{0,k} \otimes \widehat{q}_{1,k})$-almost every $\widehat{x}$.

Finally, the function $s \mapsto s \log s$ (with $0 \log 0 = 0$) is convex on $[0, \infty)$. Conditional Jensen's inequality therefore gives

$$\begin{aligned}&\mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X) \,\middle|\, \widehat{X} \right] \log\left(\mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X) \,\middle|\, \widehat{X} \right] \right) \\ &\qquad \le \mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X) \log\left(\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X) \right) \,\middle|\, \widehat{X} \right].\end{aligned}$$

Taking expectations and using the two marginal laws of the joint reference law, we get

$$\begin{aligned}
\mathrm{KL}\big(\rho'_\varepsilon \mid \widehat{q}_{0,k} \otimes \widehat{q}_{1,k}\big) &= \int \frac{d\rho'_\varepsilon}{d(\widehat{q}_{0,k} \otimes \widehat{q}_{1,k})}(\widehat{x}) \log\left(\frac{d\rho'_\varepsilon}{d(\widehat{q}_{0,k} \otimes \widehat{q}_{1,k})}(\widehat{x})\right) (\widehat{q}_{0,k} \otimes \widehat{q}_{1,k})(d\widehat{x}) \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X)\middle|\widehat{X}\right] \log\left(\mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X)\middle|\widehat{X}\right]\right)\right] \\
&\le \mathbb{E}\left[\mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X) \log\left(\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X)\right)\middle|\widehat{X}\right]\right] \\
&= \mathbb{E}\left[\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X) \log\left(\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(X)\right)\right] \\
&= \int \frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(x) \log\left(\frac{d\rho_\varepsilon}{d(q_0 \otimes q_1)}(x)\right) (q_0 \otimes q_1)(dx) \\
&= \mathrm{KL}(\rho_\varepsilon \mid q_0 \otimes q_1).
\end{aligned}$$

So our claim is proved, and by using Minkowski's inequality, we have

$$\sqrt{\mathbb{E}_{\rho'_\varepsilon}[\|X_1 - X_0\|_2^2]} \le \sqrt{\mathbb{E}_{\rho_\varepsilon}[\|X_1 - X_0\|_2^2]} + a_k.$$

Since independent coupling is feasible and has zero relative entropy, the raw transport cost of ρ_ε is at most C_{ind}. Substituting these bounds into the empirical objective, we can then conclude Eq. (31). □

Lemma B.8. *Considering two sampled endpoint batches $(X_1,\dots,X_k)$ and $(Y_1,\dots,Y_k)$, then for any uniform-marginal matrix $P = (p_{ij})_{i,j=1}^k \in \mathbb{R}_+^{k\times k}$ satisfying $P\mathbf{1} = P^\top\mathbf{1} = \mathbf{1}/k$, we write $\Gamma_P = \sum_{i,j=1}^k p_{ij}\,\delta_{(X_i,Y_j)} \in \Pi(\widehat{q}_{0,k}, \widehat{q}_{1,k})$ as its pushforward to endpoint space. Then, the index-level problem in Eq. (2) and the measure-level problem defining $\mathsf{S}_\varepsilon(\widehat{q}_{0,k}, \widehat{q}_{1,k})$ have the same minimum.*

Proof. Let $\{x^1,\dots,x^A\}$ and $\{y^1,\dots,y^B\}$ be the distinct values in the two batches. We write

$$I_a := \{i : X_i = x^a\}, \qquad J_b := \{j : Y_j = y^b\}, \qquad m_a := |I_a|, \qquad n_b := |J_b|.$$

For a feasible matrix P, we define its location-level block masses by

$$\overline{P}_{ab} := \sum_{i\in I_a}\sum_{j\in J_b} p_{ij}.$$

Then, $\overline{P}$ is a coupling between the location masses $(m_a/k)_{a=1}^A$ and $(n_b/k)_{b=1}^B$, and

$$\Gamma_P = \sum_{a=1}^A \sum_{b=1}^B \overline{P}_{ab}\,\delta_{(x^a,y^b)}.$$

Whenever $\overline{P}_{ab} > 0$, we set $r_{ij}^{ab} := p_{ij}/\overline{P}_{ab}$ on $I_a \times J_b$; when $\overline{P}_{ab} = 0$, one can choose r^{ab} arbitrarily. The relative-entropy chain rule gives us that

$$\mathrm{KL}\big(P\big\|\mathbf{1}\mathbf{1}^\top/k^2\big) = \mathrm{KL}\Big(\overline{P}\Big\|\big(m_a n_b/k^2\big)_{a,b}\Big) + \sum_{a,b}\overline{P}_{ab}\,\mathrm{KL}\big(r^{ab}\big\|\mathrm{Unif}(I_a\times J_b)\big)\,. \tag{32}$$

So, clearly, the first term in Eq. (32) is exactly $\mathrm{KL}(\Gamma_P\|\widehat{q}_{0,k} \otimes \widehat{q}_{1,k})$, while the second term is nonnegative. Moreover, the quadratic cost is constant on every block $I_a \times J_b$, so we have

$$\langle \mathbf{C}, P\rangle = \frac{1}{2d}\int \|x-y\|_2^2\,\Gamma_P(\mathrm{d}x,\mathrm{d}y).$$

Conversely, every location-level coupling $\overline{P}$ has the feasible block-uniform lift

$$p_{ij} := \frac{\overline{P}_{ab}}{m_a n_b}, \qquad i\in I_a,\ j\in J_b.$$

Indeed, for $i \in I_a$, we have

$$\sum_j p_{ij} = \frac{1}{m_a}\sum_b \overline{P}_{ab} = \frac{1}{k},$$

and the column constraints follow identically. For this lift every conditional entropy term in Eq. (32) vanishes. Thus, minimizing the index-level objective first over lifts of a fixed endpoint coupling, and then over endpoint couplings, gives exactly the measure-level entropic problem. The block-uniform lift is the unique entropy-minimizing lift of each endpoint coupling.

Since relative entropy is strictly convex on the finite location-level coupling polytope, the measure-level optimizer is unique, and its block-uniform lift is the unique index-level optimizer. □

With these lemmas, we can finally now prove Proposition 4.14:

Proof of Proposition 4.14. We denote $\widehat{q}_{0,k}$ and $\widehat{q}_{1,k}$ as the empirical marginals and set $\widehat{\gamma}_{k,\varepsilon} = \Gamma_{\widehat{\pi}_{k,\varepsilon}}$. By Lemma B.8, conditional on the endpoint batches, $\widehat{\gamma}_{k,\varepsilon}$ is exactly the unique minimizer of $\mathsf{S}_\varepsilon(\widehat{q}_{0,k}, \widehat{q}_{1,k})$. The empirical marginals converge almost surely to q_0 and q_1 in W_2.

Fix $\varepsilon > 0$, we can work on an event of probability on which both empirical W_2 convergences hold. We claim that the sequence $\{\widehat{\gamma}_{k,\varepsilon}\}_k$ is tight as its two marginals are tight.

Let γ be the weak limit of an arbitrary convergent subsequence $\{\widehat{\gamma}_{k_j,\varepsilon}\}_{j\geq 1}$, its marginals are q_0 and q_1. Joint lower semicontinuity of the quadratic cost and relative entropy along this subsequence gives us

$$\frac{1}{2d}\mathbb{E}_\gamma\|X_1 - X_0\|_2^2 + \varepsilon\,\mathrm{KL}(\gamma\|q_0 \otimes q_1) \leq \liminf_{j\to\infty} \mathsf{S}_\varepsilon(\widehat{q}_{0,k_j}, \widehat{q}_{1,k_j}).$$

For the entropy term, the joint lower semicontinuity follows directly from its variational representation

$$\mathrm{KL}(\mu\|\nu) = \sup_{f\in C_b}\left\{\int f\,\mathrm{d}\mu - \log\int e^f\,\mathrm{d}\nu\right\},$$

which is because $\widehat{q}_{0,k} \otimes \widehat{q}_{1,k} \Rightarrow q_0 \otimes q_1$. On the other hand, Lemma B.7 and $a_k \to 0$ (almost surely) give

$$\limsup_{k\to\infty} \mathsf{S}_\varepsilon(\widehat{q}_{0,k}, \widehat{q}_{1,k}) \leq \mathsf{S}_\varepsilon(q_0, q_1). \tag{33}$$

The full-sequence bound in Eq. (33) also holds along $\{k_j\}$. Therefore, we have

$$\frac{1}{2d}\mathbb{E}_\gamma\|X_1 - X_0\|_2^2 + \varepsilon\,\mathrm{KL}(\gamma\|q_0\otimes q_1) \leq \liminf_{j\to\infty} \mathsf{S}_\varepsilon(\widehat{q}_{0,k_j}, \widehat{q}_{1,k_j}) \leq \limsup_{j\to\infty} \mathsf{S}_\varepsilon(\widehat{q}_{0,k_j}, \widehat{q}_{1,k_j}) \leq \mathsf{S}_\varepsilon(q_0, q_1),$$

so γ minimizes the population objective in Eq. (6). We note the objective has a unique minimizer, i.e., independent coupling has finite objective, and relative entropy is strictly convex on the finite-objective domain, so $\gamma = \pi_\varepsilon$. Since every weak cluster point is the same, we end up with

$$\widehat{\gamma}_{k,\varepsilon} \Rightarrow \pi_\varepsilon \qquad \text{almost surely.}$$

Furthermore, we have

$$\int (\|x\|_2^2 + \|y\|_2^2)\,\widehat{\gamma}_{k,\varepsilon}(\mathrm{d}x, \mathrm{d}y) = \int \|x\|_2^2\,\widehat{q}_{0,k}(\mathrm{d}x) + \int \|y\|_2^2\,\widehat{q}_{1,k}(\mathrm{d}y),$$

which converges almost surely to the corresponding second moment under $q_0 \otimes q_1$. Thus, the almost-sure convergence is in W_2. Conditional on the endpoint batches, $\widehat{\gamma}_{k,\varepsilon}$ is precisely the law of the selected endpoint pair. Therefore, for every bounded continuous g, we have

$$\int g\,\mathrm{d}\pi_{k,\varepsilon} = \mathbb{E}\left[\int g\,\mathrm{d}\widehat{\gamma}_{k,\varepsilon}\right] \longrightarrow \int g\,\mathrm{d}\pi_\varepsilon$$

by bounded convergence, so $\pi_{k,\varepsilon} \Rightarrow \pi_\varepsilon$. Every expected coupling $\pi_{k,\varepsilon}$ has marginals q_0, q_1, and therefore has the same product-space second moment as π_ε. It follows that $\pi_{k,\varepsilon} \to \pi_\varepsilon$ in W_2.

For any coupling π, the conditional orthogonality and Benamou–Brenier argument in the proof of Proposition 4.8 gives

$$0 \leq \mathsf{R}(\pi) \leq \mathbb{E}_\pi[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1). \tag{34}$$

Applying Lemma B.6 to the established W_2 convergence yields $\limsup_{k\to\infty} \mathsf{R}(\pi_{k,\varepsilon}) \leq \mathsf{R}(\pi_\varepsilon)$, and applying Eq. (34) to π_ε proves Eq. (17).

We still need to prove the finite-batch equivalence. Let $E \subseteq \mathbb{R}^d \times \mathbb{R}^d$ be Borel, we write $E_x = \{y : (x, y) \in E\}$. If $\pi_{\text{ind}}(E) = 0$, every one of the finitely many candidate source-target pairs avoids E almost surely, so $\pi_{k,\varepsilon}(E) = 0$. Conversely, if $\pi_{\text{ind}}(E) > 0$, then

$$\mathbb{P}(Y_j \in E_{X_I} \text{ for every } j) = \int q_1(E_x)^k \, q_0(\mathrm{d}x) > 0,$$

and this event forces the selected pair to belong to E, so $\pi_{k,\varepsilon} \sim \pi_{\text{ind}}$. The Borel bijection

$$(t, x_0, x_1) \mapsto \big(t, (1-t)x_0 + tx_1, x_1 - x_0\big)$$

preserves equivalence of the induced laws. Finally, $\mathsf{R}(\pi) = 0$ exactly when U^π is almost surely a measurable function of (t, X_t^π), and this graph property is invariant under equivalent laws. We now prove the proposition. □

B.10 THEOREM 4.15

We first establish Lemma B.9 before prove Theorem 4.15.

Lemma B.9. *Suppose $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$, and we choose any $k \in \mathbb{N}$. Conditional on two sampled endpoint batches, let $\mathcal{O}_k$ be the face of uniform-marginal matrices minimizing the empirical quadratic cost in Eq. (2), and we define*

$$\widehat{\pi}_{k,0}^{\mathrm{ME}} := \operatorname*{arg\,min}_{P \in \mathcal{O}_k} \mathrm{KL}\big(P \big\| \mathbf{1}\mathbf{1}^\top / k^2\big)\,. \tag{35}$$

Let $\pi_{k,0}^{\mathrm{ME}}$ be the averaged selected-pair coupling induced by this plan. Then, it gives that

$$\big\|\pi_{k,\varepsilon} - \pi_{k,0}^{\mathrm{ME}}\big\|_{\mathrm{TV}} \longrightarrow 0, \qquad \mathsf{R}(\pi_{k,\varepsilon}) \longrightarrow \mathsf{R}(\pi_{k,0}^{\mathrm{ME}}) \tag{36}$$

as $\varepsilon \downarrow 0$. Moreover, we have

$$\pi_{k,0}^{\mathrm{ME}} \sim \pi_{\text{ind}}, \qquad \mathsf{R}(\pi_{k,0}^{\mathrm{ME}}) = 0 \iff \mathsf{R}(\pi_{\text{ind}}) = 0. \tag{37}$$

Proof. Let

$$\mathcal{T}_k := \left\{ P \in \mathbb{R}_+^{k \times k} : P\mathbf{1} = \frac{1}{k}\mathbf{1},\ P^\top \mathbf{1} = \frac{1}{k}\mathbf{1} \right\}$$

be the empirical transport poly-tope. Conditional on the sampled batches, we write

$$C(P) := \langle \mathbf{C}, P \rangle, \qquad H(P) := \mathrm{KL}\big(P \big\| \mathbf{1}\mathbf{1}^\top / k^2\big)\,, \qquad C_\star := \min_{P \in \mathcal{T}_k} C(P).$$

We note that $\mathcal{T}_k$ is both compact and convex, and the optimal face $\mathcal{O}_k = \{P \in \mathcal{T}_k : C(P) = C_\star\}$ is nonempty, compact, and convex. Because the reference matrix has strictly positive entries, H is continuous and strictly convex on $\mathcal{T}_k$. Therefore, Eq. (35) has a unique solution.

We denote $\widehat{\pi}_{k,\varepsilon}$ as the unique empirical entropic optimizer, and its optimality against $\widehat{\pi}_{k,0}^{\mathrm{ME}}$ gives

$$C(\widehat{\pi}_{k,\varepsilon}) + \varepsilon H(\widehat{\pi}_{k,\varepsilon}) \le C_\star + \varepsilon H(\widehat{\pi}_{k,0}^{\mathrm{ME}}).$$

Since $C(\widehat{\pi}_{k,\varepsilon}) \ge C_\star$ and $H \ge 0$, we can show that

$$0 \le C(\widehat{\pi}_{k,\varepsilon}) - C_\star \le \varepsilon H(\widehat{\pi}_{k,0}^{\mathrm{ME}}), \qquad H(\widehat{\pi}_{k,\varepsilon}) \le H(\widehat{\pi}_{k,0}^{\mathrm{ME}}).$$

Therefore, every cluster point of $\widehat{\pi}_{k,\varepsilon}$ as $\varepsilon \downarrow 0$ lies in $\mathcal{O}_k$ and has entropy no larger than that of $\widehat{\pi}_{k,0}^{\mathrm{ME}}$.

The uniqueness in Eq. (35) forces every cluster point to equal $\widehat{\pi}_{k,0}^{\mathrm{ME}}$. Compactness then gives convergence of the full sequence in the finite-dimensional polytope, equivalently in entrywise ℓ_1 and total variation.

For every positive ε, the unique optimizer is a Borel function of the sampled cost matrix by the measurable maximum theorem. Taking, for example, $\varepsilon_n = 1/n$, the pointwise convergence just proved shows that $\widehat{\pi}_{k,0}^{\mathrm{ME}}$ is the pointwise limit of measurable optimizers and is therefore measurable as well.

For a matrix $P \in \mathcal{T}_k$, we denote $\Gamma_P = \sum_{i,j=1}^k P_{ij}\,\delta_{(X_i,Y_j)}$ as its selected-pair endpoint law conditional on the batches. Pushforward contracts total variation, so we have

$$\|\Gamma_{\widehat{\pi}_{k,\varepsilon}} - \Gamma_{\widehat{\pi}^{\mathrm{ME}}_{k,0}}\|_{\mathrm{TV}} \le \frac{1}{2}\|\widehat{\pi}_{k,\varepsilon} - \widehat{\pi}^{\mathrm{ME}}_{k,0}\|_1 \longrightarrow 0$$

for every realized pair of batches. Convexity of total variation and dominated convergence now yield

$$\|\pi_{k,\varepsilon} - \pi^{\mathrm{ME}}_{k,0}\|_{\mathrm{TV}} \le \mathbb{E}\Big[\|\Gamma_{\widehat{\pi}_{k,\varepsilon}} - \Gamma_{\widehat{\pi}^{\mathrm{ME}}_{k,0}}\|_{\mathrm{TV}}\Big] \longrightarrow 0.$$

This proves the first convergence in Eq. (36).

We next prove the continuity of R needed for the second convergence. More generally, let $\rho_n, \rho \in \Pi(q_0, q_1)$ satisfy $\|\rho_n - \rho\|_{\mathrm{TV}} \to 0$, and let P_n, P be the corresponding laws of $((t, X_t), U)$, where $U = X_1 - X_0$. Appending the same independent $t \sim \mathcal{U}([0,1])$ preserves total-variation convergence, while measurable pushforward contracts it. Thus, we have $\|P_n - P\|_{\mathrm{TV}} \to 0$.

Let $T_L : \mathbb{R}^d \to \overline{B}(0, L)$ denote metric projection, we set $\mathsf{R}_L(\rho') = \inf_f \mathbb{E}_{\rho'}\big[\|T_L(U) - f(t, X_t)\|_2^2\big]$. The infimum may be restricted to predictors with values in $\overline{B}(0, L)$, because projection onto this ball cannot increase the distance to $T_L(U)$, so the resulting losses are bounded by $4L^2$.

Given $\|\mu - \nu\|_{\mathrm{TV}} = \sup_A |\mu(A) - \nu(A)|$, comparison with an arbitrarily accurate minimizer under the other law gives us that

$$|\mathsf{R}_L(\rho_n) - \mathsf{R}_L(\rho)| \le 8L^2\|P_n - P\|_{\mathrm{TV}} \longrightarrow 0. \tag{38}$$

The final missing point is that we need to remove the truncation uniformly. We write $A = \|X_0\|_2^2$ and $B = \|X_1\|_2^2$; because their two fixed marginal laws are integrable, the de la Vallée–Poussin criterion supplies a nondecreasing, convex, superlinear function Φ such that

$$\mathbb{E}_{q_0}[\Phi(A)] + \mathbb{E}_{q_1}[\Phi(B)] < \infty.$$

For every $\rho' \in \Pi(q_0, q_1)$, we have

$$\frac{\|X_1 - X_0\|_2^2}{4} \le \frac{A + B}{2},$$

and monotonicity and convexity give the following bound

$$\sup_{\rho' \in \Pi(q_0,q_1)} \mathbb{E}_{\rho'}\left[\Phi\left(\frac{\|X_1 - X_0\|_2^2}{4}\right)\right] \le \frac{1}{2}\mathbb{E}_{q_0}[\Phi(A)] + \frac{1}{2}\mathbb{E}_{q_1}[\Phi(B)] < \infty.$$

Therefore, $\{\|X_1 - X_0\|_2^2 : \rho' \in \Pi(q_0, q_1)\}$ is uniformly integrable, and we end up with

$$e_L = \sup_{\rho' \in \Pi(q_0,q_1)} \mathbb{E}_{\rho'}\big[\|U - T_L(U)\|_2^2\big] \longrightarrow 0. \tag{39}$$

For any coupling ρ', the square root of its Bayes risk is the ℓ_2-distance from U to the closed subspace of (t, X_t)-measurable vector-valued functions. Since the distance to a closed set is 1-Lipschitz, we have

$$\left|\sqrt{\mathsf{R}(\rho')} - \sqrt{\mathsf{R}_L(\rho')}\right| \le \big(\mathbb{E}_{\rho'}\|U - T_L(U)\|_2^2\big)^{1/2} \le e_L^{1/2}. \tag{40}$$

Combining Eqs. (38) to (40), first taking $n \to \infty$ at fixed L and then taking $L \to \infty$, we get $\mathsf{R}(\rho_n) \to \mathsf{R}(\rho)$. Applying this fact to the endpoint couplings above proves the second convergence in Eq. (36).

Finally, consider $E \subseteq \mathbb{R}^d \times \mathbb{R}^d$ be Borel with $E_x = \{y : (x, y) \in E\}$. If $(q_0 \otimes q_1)(E) = 0$, all k^2 sampled candidate pairs avoid E almost surely, which gives $\pi^{\mathrm{ME}}_{k,0}(E) = 0$. Conversely, if $(q_0 \otimes q_1)(E) > 0$, we just consider I be any uniformly selected source index, which gives us

$$\mathbb{P}(Y_j \in E_{X_I} \text{ for every } j) = \int q_1(E_x)^k\, q_0(\mathrm{d}x) > 0.$$

On this event, whichever target index is selected from row I, the selected pair still lies in E. Therefore, $\pi^{\mathrm{ME}}_{k,0}(E) > 0$, proving $\pi^{\mathrm{ME}}_{k,0} \sim \pi_{\mathrm{ind}}$. Again, recall that the map

$$(t, x_0, x_1) \longmapsto \big(t, (1-t)x_0 + tx_1, x_1 - x_0\big)$$

is a Borel bijection, with inverse $(t, x, u) \mapsto (t, x - tu, x + (1-t)u)$. Therefore, equivalent endpoint laws induce equivalent laws of (t, X_t, U). Since

$$\mathsf{R}(\rho) = 0 \quad \Longleftrightarrow \quad U \text{ is almost surely a measurable function of } (t, X_t),$$

so this zero-risk graph property is invariant under equivalence, and we can then conclude Eq. (37). □

Before establishing the proof for Theorem 4.15, we define $D_\varepsilon = \mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] - W_2^2(q_0, q_1)$. The theorem's joint uniform condition means that there exist $K_J \in \mathbb{N}$, $\varepsilon_J > 0$, and $M < \infty$ such that

$$\operatorname{ess\,sup} \|J_{\theta^\star_{\pi_{k,\varepsilon}}}(t, X_t^{\pi_{k,\varepsilon}})\|_{\mathrm{op}}^2 \le \max_{t\in[0,1],\, x\in\mathbb{R}^d} \|J_{\theta^\star_{\pi_{k,\varepsilon}}}(t,x)\|_{\mathrm{op}}^2 \le M, \quad k \ge K_J, \quad 0 < \varepsilon \le \varepsilon_J. \tag{41}$$

Moreover, Eq. (28) gives $D_\varepsilon \to 0$ as $\varepsilon \downarrow 0$. Now, we can start the proof:

Proof of Theorem 4.15. At a realizable coupling-specific population optimum, Eq. (3) gives the exact gradient-variance identity. The trainable additive output bias implies $J_{\theta^\star_{\pi_{k,\varepsilon}}} J_{\theta^\star_{\pi_{k,\varepsilon}}}^\top \succeq I_d$ by Fact B.3, with both Jacobians evaluated at $(t, X_t^{\pi_{k,\varepsilon}})$. Applying the positive-semidefinite trace bounds used in Proposition B.1 proves both inequalities in Eq. (18).

We next obtain a positive lower bound for the prescribed finite baseline as $\varepsilon \downarrow 0$ and set $r_0 = \mathsf{R}(\pi^{\mathrm{ME}}_{k_0,0})$. By Lemma B.9 and the assumption $\mathsf{R}(\pi_{\mathrm{ind}}) > 0$, we have

$$r_0 > 0, \qquad \mathsf{R}(\pi_{k_0,\varepsilon}) \longrightarrow r_0 \quad \text{as } \varepsilon \downarrow 0.$$

So, there exists $\varepsilon_{\mathrm{b}} > 0$ such that

$$\mathsf{R}(\pi_{k_0,\varepsilon}) > \frac{r_0}{2} \qquad \text{for every } 0 < \varepsilon \le \varepsilon_{\mathrm{b}}.$$

By the lower inequality in Eq. (18), this gives

$$\sigma^2_{\mathrm{grad},\pi_{k_0,\varepsilon}}(\theta^\star_{\pi_{k_0,\varepsilon}}) > 2r_0 \qquad \text{for every } 0 < \varepsilon \le \varepsilon_{\mathrm{b}}. \tag{42}$$

Now fix $0 < \varepsilon \le \varepsilon_J$. Proposition 4.14 and Eq. (41) imply

$$\limsup_{k\to\infty} \mathsf{R}(\pi_{k,\varepsilon}) \le D_\varepsilon, \qquad \limsup_{k\to\infty} \operatorname{ess\,sup} \|J_{\theta^\star_{\pi_{k,\varepsilon}}}(t, X_t^{\pi_{k,\varepsilon}})\|_{\mathrm{op}}^2 \le M.$$

Note that both sequences are nonnegative and the two right-hand sides are finite. Therefore, applying Eq. (18) and the elementary limsup product inequality gives

$$\limsup_{k\to\infty} \sigma^2_{\mathrm{grad},\pi_{k,\varepsilon}}(\theta^\star_{\pi_{k,\varepsilon}}) \le 4MD_\varepsilon. \tag{43}$$

Since $D_\varepsilon \to 0$, there exists $\varepsilon_{\mathrm{s}} > 0$ such that

$$4MD_\varepsilon < r_0 \qquad \text{for every } 0 < \varepsilon \le \varepsilon_{\mathrm{s}}. \tag{44}$$

We set $\bar\varepsilon = \min\{\varepsilon_J, \varepsilon_{\mathrm{b}}, \varepsilon_{\mathrm{s}}\} > 0$. For every fixed $0 < \varepsilon \le \bar\varepsilon$, Eqs. (42) to (44) give

$$\limsup_{k\to\infty} \sigma^2_{\mathrm{grad},\pi_{k,\varepsilon}}(\theta^\star_{\pi_{k,\varepsilon}}) < r_0 < 2r_0 < \sigma^2_{\mathrm{grad},\pi_{k_0,\varepsilon}}(\theta^\star_{\pi_{k_0,\varepsilon}}).$$

By the limit of superior, we have a finite threshold after which the strict upper comparison holds. Increasing that threshold if necessary, choose a finite $K > \max\{k_0, K_J\}$, which may depend on k_0 and ε; then, Eq. (19) holds for every $k \ge K$.

Now, it remains for us only to verify strict positivity. For every finite k and $\varepsilon > 0$, Proposition 4.14 gives $\pi_{k,\varepsilon} \sim \pi_{\mathrm{ind}}$. As in Lemma B.9, equivalence preserves the graph property characterizing zero target uncertainty. Therefore, $\mathsf{R}(\pi_{k,\varepsilon}) > 0$, and the lower inequality in Eq. (18) gives $\sigma^2_{\mathrm{grad},\pi_{k,\varepsilon}} > 0$. This completes the proof, and we can conclude the variance reduction. □

B.10.1 Proposition 4.17

Before establishing the proof, we need to prove two supporting lemmas. We first prove the following one-step coupling bound:

Lemma B.10. *Let $q_0, q_1 \in \mathcal{P}_2(\mathbb{R}^d)$, let $\pi \in \Pi(q_0, q_1)$, and we denote $m_\pi(x) = \mathbb{E}_\pi[X_1 \mid X_0 = x]$. Then, we have*

$$0 \le \mathrm{err}(\pi) \le \mathbb{E}_\pi[\|X_1 - m_\pi(X_0)\|_2^2] = \mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|_2^2] - \mathbb{E}[\|m_\pi(X_0) - \mathbb{E}[X_1]\|_2^2]. \tag{45}$$

Moreover, for every measurable $g \in L^2(q_0; \mathbb{R}^d)$, the following reduction holds true:

$$\mathrm{err}(\pi) \le \mathbb{E}_\pi[\|X_1 - m_\pi(X_0)\|_2^2] \le \mathbb{E}_\pi[\|X_1 - g(X_0)\|_2^2]. \tag{46}$$

If $\pi \sim q_0 \otimes q_1$ and q_1 is not a Dirac mass, then $\mathrm{err}(\pi) > 0$.

Proof. The proof is quite straightforward. Specifically, the pair $(m_\pi(X_0), X_1)$ couples $(m_\pi)_{\#}q_0$ with q_1, which gives the first upper bound. Conditional orthogonality gives the equality in Eq. (45) and the predictor inequality in Eq. (46).

Now suppose $\pi \sim q_0 \otimes q_1$. For q_0-almost every x, the conditional law $\pi(\mathrm{d}y \mid X_0 = x)$ is equivalent to the non-Dirac law q_1, so its conditional variance is strictly positive. If $\mathrm{err}(\pi) = 0$, then $(m_\pi)_{\#}q_0 = q_1$, while conditional orthogonality gives us that

$$\mathbb{E}[\|X_1\|_2^2] = \mathbb{E}[\|m_\pi(X_0)\|_2^2] + \mathbb{E}_\pi[\|X_1 - m_\pi(X_0)\|_2^2] > \mathbb{E}[\|X_1\|_2^2],$$

which is a contradiction, so this proves the lemma. □

We next prove the convergence to Monge coupling:

Lemma B.11. *Let $T \in L^2(q_0; \mathbb{R}^d)$ and let $\gamma_n \in \Pi(q_0, q_1)$ satisfy $\gamma_n \to (\mathrm{id}, T)_{\#}q_0$ in W_2. Then,*

$$\mathbb{E}_{\gamma_n}[\|X_1 - T(X_0)\|_2^2] \longrightarrow 0, \qquad \mathrm{err}(\gamma_n) \longrightarrow 0. \tag{47}$$

Proof. Fix $g \in C_b(\mathbb{R}^d; \mathbb{R}^d)$. Since $(x, y) \mapsto \|y - g(x)\|_2^2$ is continuous with quadratic growth, W_2 convergence and the common first marginal give

$$\limsup_{n\to\infty} \mathbb{E}_{\gamma_n}[\|X_1 - T(X_0)\|_2^2] \le 4\|g - T\|^2_{L^2(q_0)}.$$

Bounded continuous functions are dense in $L^2(q_0)$, so the first limit follows. The second follows from Eq. (46) with $g = T$. □

Given these two supporting lemmas, we can now prove Proposition 4.17:

Proof of Proposition 4.17. For population OT, $X_1 = T(X_0)$ almost surely, so the one-step output has law q_1 and $\mathrm{err}(\pi^\star) = 0$. Under independent coupling, $\mathbb{E}[X_1 \mid X_0] = \mathbb{E}[X_1]$, so we can show that

$$\mathrm{err}(\pi_{\text{ind}}) = \mathbb{E}[\|X_1 - \mathbb{E}[X_1]\|_2^2] > 0.$$

We next prove the two population-entropic claims. First, we write objective as

$$\mathcal{J}_\varepsilon(\pi) = \frac{1}{2d}\mathbb{E}_\pi[\|X_1 - X_0\|_2^2] + \varepsilon\,\mathrm{KL}(\pi\|\pi_{\text{ind}}).$$

Since π_{ind} has finite objective, π_ε is absolutely continuous with respect to π_{ind}, and there exists a Radon–Nikodym density f such that $\mathrm{d}\pi_\varepsilon = f\,\mathrm{d}\pi_{\text{ind}}$.

We claim that $f > 0$ almost everywhere, which can be proved via contradiction, i.e., otherwise if $A = \{f = 0\}$ had positive π_{ind}-mass, the feasible mixture $\pi_s = (1 - s)\pi_\varepsilon + s\pi_{\text{ind}}$ would satisfy, with $h(r) = r \log r$, that

$$\frac{\mathcal{J}_\varepsilon(\pi_s) - \mathcal{J}_\varepsilon(\pi_\varepsilon)}{s} \le \frac{\mathbb{E}_{\pi_{\text{ind}}}[\|X_1 - X_0\|_2^2] - \mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2]}{2d} + \varepsilon\{\pi_{\text{ind}}(A)\log s - \mathrm{KL}(\pi_\varepsilon\|\pi_{\text{ind}})\},$$

which tends to $-\infty$ as $s \downarrow 0$ and contradicts the optimality, so $\pi_\varepsilon \sim \pi_{\text{ind}}$, and Lemma B.10 gives $\mathrm{err}(\pi_\varepsilon) > 0$.

We next compare π_ε strictly with independence. We choose w such that $\langle w, X_1 - \mathbb{E}[X_1]\rangle$ is not a constant and set $b(X_1) = \tanh\langle w, X_1 - \mathbb{E}[X_1]\rangle - \mathbb{E}[\tanh\langle w, X_1 - \mathbb{E}[X_1]\rangle]$, and it's clear that $v \coloneqq \mathbb{E}[(X_1 - \mathbb{E}[X_1])b(X_1)] \neq 0$. Since $q_0 \ll \mathcal{L}^d$, the variable $\langle v, X_0 - \mathbb{E}[X_0]\rangle$ is not a constant.

We define $a(X_0)$ by applying and centering $\tanh$ in the same way, then rescaling it so that both $\|a\|_\infty, \|b\|_\infty \le 1$. For $|s| < 1$, the density $1 + s, a(x)b(y)$ defines a coupling $\rho_s \in \Pi(q_0, q_1)$, and its entropy derivative at $s = 0$ vanishes, whereas

$$\frac{\mathrm{d}}{\mathrm{d}s}\frac{1}{2d}\mathbb{E}_{\rho_s}[\|X_1 - X_0\|_2^2]\Big|_{s=0} = -\frac{1}{d}\,\langle \mathbb{E}[(X_0 - \mathbb{E}[X_0])a(X_0)], v\rangle < 0.$$

Therefore, $\mathcal{J}_\varepsilon(\pi_\varepsilon) \le \mathcal{J}_\varepsilon(\rho_s) < \mathcal{J}_\varepsilon(\pi_{\text{ind}})$ for all sufficiently small $s > 0$, and non-negativity of relative entropy gives us that

$$\mathbb{E}_{\pi_\varepsilon}[\|X_1 - X_0\|_2^2] < \mathbb{E}_{\pi_{\text{ind}}}[\|X_1 - X_0\|_2^2].$$

Expanding the two costs shows that $\mathbb{E}_{\pi_\varepsilon}[X_1 \mid X_0]$ is not almost surely equal to $\mathbb{E}[X_1]$. The strict upper comparison $\mathrm{err}(\pi_\varepsilon) < \mathrm{err}(\pi_{\mathrm{ind}})$ now follows from Eq. (45).

Finally, choosing any minibatch batch size k, we denote $(Z_1, \dots, Z_k)$ and $(Y_1, \dots, Y_k)$ as the endpoint batches, let I be any uniform source index, and let τ be the empirical matching. If $(q_0 \otimes q_1)(E) = 0$, every candidate pair (Z_i, Y_j) avoids E almost surely, so $\pi_k(E) = 0$. Conversely, if $(q_0 \otimes q_1)(E) > 0$ and $E_z = \{y : (z, y) \in E\}$, then

$$\pi_k(E) \geq \mathbb{P}\big(Y_j \in E_{Z_I} \text{ for every } j\big) = \int q_1(E_z)^k q_0(\mathrm{d}z) > 0.$$

Therefore, $\pi_k \sim q_0 \otimes q_1$, and Lemma B.10 gives $\mathrm{err}(\pi_k) > 0$.

Now, recall that Proposition 4.12 already proves the convergence of π_k to $W_2^2(q_0, q_1)$. Also, the tightness, lower semicontinuity, and uniqueness of the Brenier coupling imply $\pi_k \Rightarrow \pi^\star$. Since all couplings have the same product-space second moment, this convergence holds in W_2. Lemma B.11 therefore gives

$$\mathrm{err}(\pi_k) \leq \mathbb{E}_{\pi_k}[\|X_1 - T(X_0)\|_2^2] \longrightarrow 0.$$

Since every fixed finite baseline has positive error, we can conclude the strict error reduction. □

B.11 THEOREM 4.18

Proof. For simplicity, we abbreviate the following notations: $\mu_i = \mathbb{E}[X_i]$, $V_i = \mathbb{E}[\|X_i - \mu_i\|_2^2]$, and $C(\pi) = \mathbb{E}_\pi[\|X_1 - X_0\|_2^2]$ with $C_{\mathrm{ind}} = C(\pi_{\mathrm{ind}})$. Since $q_0 \ll \mathcal{L}^d$ and q_0 is non-Dirac, we have $V_0 > 0$. For any coupling π, expanding the two quadratic costs gives

$$C_{\mathrm{ind}} - C(\pi) = 2\mathbb{E}_\pi[\langle X_0 - \mu_0, \mathbb{E}_\pi[X_1 \mid X_0] - \mu_1 \rangle].$$

Together with Cauchy–Schwarz and Lemma B.10, we have

$$\mathrm{err}(\pi) \leq V_1 - \frac{\{C_{\mathrm{ind}} - C(\pi)\}^2}{4V_0}. \tag{48}$$

To apply this inequality, we consider $\mathcal{B}$ as two sampled endpoint batches. Under the actual sampling law $\mathbb{P}_{\mathrm{ot}}$, the conditional distribution of the selected indices (I, J) given $\mathcal{B}$ is $\widehat{\pi}_{k,\varepsilon}$. We can define a reference law $\mathbb{P}_{\mathrm{ref}}$ using the same batches but conditionally independent uniform indices, so $(I, J) \mid \mathcal{B} \sim \mathbf{1}\mathbf{1}^\top/k^2$. The map $(\mathcal{B}, I, J) \mapsto (X_I, Y_J)$ pushes $\mathbb{P}_{\mathrm{ot}}$ forward to $\pi_{k,\varepsilon}$ and $\mathbb{P}_{\mathrm{ref}}$ forward to π_{ind}, so the conditional KL chain rule, along with the same data processing proof structure from Lemma B.7, give us that

$$\mathrm{KL}(\pi_{k,\varepsilon} \| \pi_{\mathrm{ind}}) \leq \mathrm{KL}(\mathbb{P}_{\mathrm{ot}} \| \mathbb{P}_{\mathrm{ref}}) = \mathbb{E}\big[\mathrm{KL}\big(\widehat{\pi}_{k,\varepsilon} \big\| \mathbf{1}\mathbf{1}^\top/k^2\big)\big]. \tag{49}$$

Moreover, selecting (I, J) according to $\widehat{\pi}_{k,\varepsilon}$ gives us that

$$\frac{C(\pi_{k,\varepsilon})}{2d} = \mathbb{E}[\langle \mathbf{C}, \widehat{\pi}_{k,\varepsilon} \rangle]. \tag{50}$$

Adding Eq. (49) to Eq. (50) and recalling the definition of $S_{k,\varepsilon}$, we have

$$\frac{C(\pi_{k,\varepsilon})}{2d} + \varepsilon\, \mathrm{KL}(\pi_{k,\varepsilon} \| \pi_{\mathrm{ind}}) \leq S_{k,\varepsilon}.$$

Since $S_{1,\varepsilon} = C_{\mathrm{ind}}/(2d)$, we obtain

$$C_{\mathrm{ind}} - C(\pi_{k,\varepsilon}) \geq 2d(S_{1,\varepsilon} - S_{k,\varepsilon}).$$

Substitution into Eq. (48) proves the middle inequality in Eq. (22). Proposition 4.14 gives $\pi_{k,\varepsilon} \sim \pi_{\mathrm{ind}}$ at every finite k, so strict positivity follows from Lemma B.10, and we finish proving the first statement in the theorem.

We now prove the second statement using the idea of induction, which the upper bound is nonincreasing and strict for $k \geq 2$.

First, we set $n = k + 1$ and fix two realized n-point batches. For each source index a and target index b, remove those two points, let P^{ab} solve the resulting $k \times k$ problem, embed it into an $n \times n$ matrix

Q^{ab} by inserting a zero row and column, and set $\overline{P} = n^{-2}\sum_{a,b} Q^{ab}$. We note that this matrix has uniform n-point marginals. If

$$M(a,b,i,j) = n^{-2}Q^{ab}_{ij}, \qquad R(a,b,i,j) = \frac{\mathbf{1}_{\{i\neq a,\, j\neq b\}}}{n^2k^2},$$

then their (i,j)-marginals are respectively $\overline{P}$ and the uniform $n\times n$ matrix. Data processing gives

$$\mathrm{KL}(\overline{P}\|\mathbf{1}\mathbf{1}^\top/n^2) \le \frac{1}{n^2}\sum_{a,b}\mathrm{KL}(P^{ab}\|\mathbf{1}\mathbf{1}^\top/k^2).$$

The same averaging identity holds for the cost, so taking expectations proves $S_{k+1,\varepsilon} \le S_{k,\varepsilon}$.

We now prove the strictness part, and it's enough for us to start with $k=2$.

Let $U_2 = \mathbf{1}\mathbf{1}^\top/4$ and $D = \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1\end{smallmatrix}\right)$. The feasible perturbation $U_2 + sD$ has zero entropy derivative at $s=0$, whereas its cost derivative is

$$-\frac{1}{d}\langle Z_1 - Z_2, Y_1 - Y_2\rangle.$$

Since q_1 is non-Dirac, $Y_1 - Y_2 \neq 0$ with positive probability; conditional on any such value, absolute continuity of q_0 makes the displayed inner product nonzero almost surely. Therefore, the empirical optimum is strictly better than U_2 with positive probability.

Also, the nonnegative gap is integrable, so $S_{2,\varepsilon} < S_{1,\varepsilon}$, and monotonicity extends this strict inequality to every $k \ge 2$. This proves both the strict final inequality and monotonicity of the upper bound in Eq. (22).

We now proceed to prove the final Eq. (23). Similar to before, we write the population value as

$$S_\varepsilon := \inf_{\pi\in\Pi(q_0,q_1)}\left\{\frac{C(\pi)}{2d} + \varepsilon\,\mathrm{KL}(\pi\|\pi_{\text{ind}})\right\}.$$

The index-to-measure reduction and uniform shadow bound in Lemmas B.8 and B.7 give the following, respectively,

$$0 \le S_{k,\varepsilon} - S_\varepsilon \le e_k := \frac{2\sqrt{C_{\text{ind}}}\,\mathbb{E}[a_k] + \mathbb{E}[a_k^2]}{2d}, \qquad a_k = W_2(\widehat{q}_{0,k}, q_0) + W_2(\widehat{q}_{1,k}, q_1). \tag{51}$$

For every $q \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathbb{E}[W_2^2(\widehat{q}_k, q)] \to 0$ (i.e., project onto a fixed ball, use bounded-support empirical W_2 convergence and dominated convergence, and then send the radius to infinity), so $e_k \to 0$.

We set $C_\star = W_2^2(q_0,q_1)$ and try to first show $S_\varepsilon \to C_\star/(2d)$. First, we note that the optimal coupling $\pi^\star$ may have infinite relative entropy with respect to π_{ind}, so we do the following math technique by approximating it through finite-entropy couplings.

Specifically, we first partition both endpoint spaces into increasingly fine finite partitions and set $p_{AB} = \pi^\star(A\times B)$, and we set ρ_n by assigning mass p_{AB} to each block $A\times B$ according to the product measure $q_0(\mathrm{d}x \mid A)q_1(\mathrm{d}y \mid B)$. Since $\sum_B p_{AB} = q_0(A)$ and $\sum_A p_{AB} = q_1(B)$, ρ_n has marginals q_0 and q_1. Moreover, we have

$$\mathrm{KL}(\rho_n\|\pi_{\text{ind}}) = \sum_{A,B} p_{AB}\log\frac{p_{AB}}{q_0(A)q_1(B)} < \infty.$$

As the partitions become finer, ρ_n converges weakly to $\pi^\star$. Since all ρ_n have the same marginals as $\pi^\star$, their product-space second moments also agree, so $\rho_n \to \pi^\star$ in W_2. Therefore, $C(\rho_n)\to C_\star$, and

$$\frac{C_\star}{2d} \le S_\varepsilon \le \frac{C(\rho_n)}{2d} + \varepsilon\,\mathrm{KL}(\rho_n\|\pi_{\text{ind}}).$$

To prove the claim, we simply need to take $\varepsilon \downarrow 0$ and $n\to\infty$. Now, let $\varepsilon_k \to 0$. The raw selected-pair cost, i.e., Eq. (51), and the preceding population limit give

$$\frac{C_\star}{2d} \le \frac{C(\pi_{k,\varepsilon_k})}{2d} \le S_{k,\varepsilon_k} \le S_{\varepsilon_k} + e_k \longrightarrow \frac{C_\star}{2d}.$$

Tightness and uniqueness of the Brenier coupling imply $\pi_{k,\varepsilon_k} \Rightarrow \pi^\star$, and the common product-space second moment upgrades this to W_2 convergence. Lemma B.11 therefore gives $\mathrm{err}(\pi_{k,\varepsilon_k}) \to 0$, with no rate condition relating k and ε_k.

To derive Eq. (23), we still need to use the following continuity fact: let $\rho_n, \rho \in \Pi(q_0, q_1)$ satisfy $\|\rho_n - \rho\|_{\mathrm{TV}} \to 0$, and we can apply the technique by disintegrating them as $q_0(\mathrm{d}x)K_n(x, \mathrm{d}y)$ and $q_0(\mathrm{d}x)K(x, \mathrm{d}y)$. Their common first marginal gives

$$\|\rho_n - \rho\|_{\mathrm{TV}} = \int \|K_n(x,\cdot) - K(x,\cdot)\|_{\mathrm{TV}} q_0(\mathrm{d}x).$$

For metric projection P_L onto the radius-L ball, the corresponding truncated conditional means satisfy

$$\int \left\| \int P_L(y) K_n(x, \mathrm{d}y) - \int P_L(y) K(x, \mathrm{d}y) \right\|_2^2 q_0(\mathrm{d}x) \le 4L^2 \|\rho_n - \rho\|_{\mathrm{TV}}.$$

Conditional Jensen bounds each truncation tail by $\int \|y - P_L(y)\|_2^2 q_1(\mathrm{d}y)$, and we can take $n \to \infty$ and then $L \to \infty$, which proves convergence of the conditional means in $L^2(q_0)$.

Coupling their pushforward laws through $X_0 \sim q_0$ gives $\mathrm{err}(\rho_n) \to \mathrm{err}(\rho)$, and we now derive the final ingredient. Fix $\varepsilon > 0$, and recall the population objective

$$J_\varepsilon(\rho) := \frac{C(\rho)}{2d} + \varepsilon\, \mathrm{KL}(\rho \| \pi_{\mathrm{ind}}), \qquad S_\varepsilon = J_\varepsilon(\pi_\varepsilon).$$

Since $\pi_{k,\varepsilon} \in \Pi(q_0, q_1)$, population optimality and the preceding bounds on the averaged coupling and empirical value give

$$S_\varepsilon \le J_\varepsilon(\pi_{k,\varepsilon}) \le S_{k,\varepsilon} \le S_\varepsilon + e_k.$$

Therefore, the population objective gap satisfies

$$0 \le J_\varepsilon(\pi_{k,\varepsilon}) - J_\varepsilon(\pi_\varepsilon) \le e_k \longrightarrow 0.$$

To translate this objective convergence into convergence of the couplings, we apply the entropy Pythagorean inequality (see Lemma 4.4 in Eckstein & Nutz (2022)) to the rescaled objective $J_\varepsilon/\varepsilon$. This gives

$$\mathrm{KL}(\pi_{k,\varepsilon} \| \pi_\varepsilon) \le \frac{J_\varepsilon(\pi_{k,\varepsilon}) - J_\varepsilon(\pi_\varepsilon)}{\varepsilon} \le \frac{e_k}{\varepsilon}.$$

Using the convention $\|\mu - \nu\|_{\mathrm{TV}} = \sup_A |\mu(A) - \nu(A)|$, Pinsker's inequality then yields

$$\|\pi_{k,\varepsilon} - \pi_\varepsilon\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\,\mathrm{KL}(\pi_{k,\varepsilon} \| \pi_\varepsilon)} \le \sqrt{\frac{e_k}{2\varepsilon}} \longrightarrow 0,$$

where the limit follows because $\varepsilon > 0$ is fixed and $e_k \to 0$. Both couplings have the same marginals q_0, q_1, so the total-variation continuity established above applies and gives

$$\mathrm{err}(\pi_{k,\varepsilon}) \longrightarrow \mathrm{err}(\pi_\varepsilon) \qquad \text{as } k \to \infty.$$

This proves the fixed-regularization convergence claimed in the theorem.

For any fixed finite k_0, Lemma B.9 gives $\pi_{k_0,\varepsilon} \to \pi^{\mathrm{ME}}_{k_0,0}$ in total variation and $\pi^{\mathrm{ME}}_{k_0,0} \sim \pi_{\mathrm{ind}}$. With the continuity, along with Lemma B.10, we have

$$\mathrm{err}(\pi_{k_0,\varepsilon}) \longrightarrow \mathrm{err}(\pi^{\mathrm{ME}}_{k_0,0}) > 0,$$

i.e., the candidate error tends to zero, so Eq. (23) holds for all sufficiently large k, and we can finally conclude the theorem. □

C Further Empirical Results

Additional 2D examples. We provide additional experimental results and details. First, we present additional 2D synthetic experiment on rings in Figure 4, using the same training settings as the multimodal block synthetic experiment in Figure 1. The results show a similar conclusion, where larger minibatch sizes consistently lead to more stable training and lower flow-generation error.

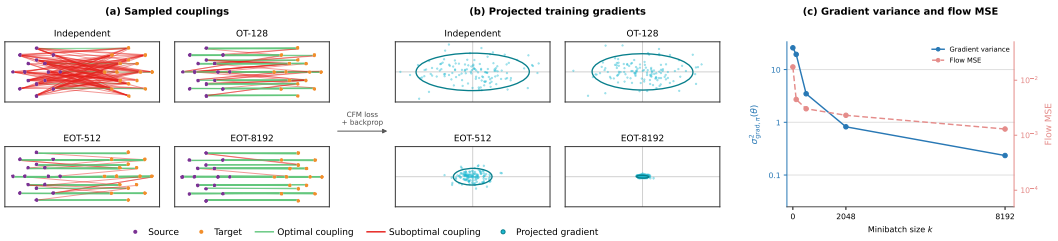

Figure 4: Results on a synthetic ring task over independent random coupling, OT-CFM using the minibatch size of 128, and EOT-CFM with contexts up to 8192. From left to right, each panel corresponds the flow couplings, the projected training gradients, and the final gradient variance and generation quality under different context size. All the training setup follows from Figure 1.

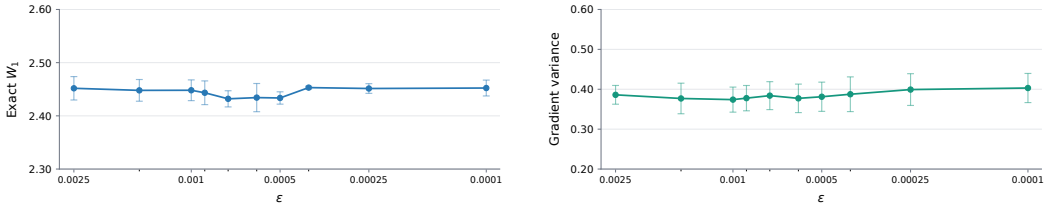

Figure 5: Ablation over ε on the PBMC dataset with an 8K entropic OT context, with the exact W_1 error of generated samples (left) and training gradient variance (right) shown. We sweep ε from $\{0.0025, 0.0015, 0.001, 0.0009, 0.00075, 0.0006, 0.0005, 0.0004, 0.00025, 0.0001\}$.

Ablation on ε. We also ablate ε on the PBMC dataset, whose size lets us reach substantially larger minibatch contexts than the EB dataset. Offline pipelines that scale to million-level contexts must drive ε to very low levels to keep the coupling sharp, which makes its tuning delicate. Online couplings instead pair each freshly sampled source point within the current minibatch, so that a moderate context of a few thousand points already yields a clear scaling effect while remaining robust to the choice of ε.

As shown in Figure 5, we sweep ε from 0.0001 to 0.0025 at 8K minibatch entropic OT context. Following the exactly same training configuration detailed in §5, each ε trial is running over 5 seeds, with the mean performance and its standard deviation shown. Performance peaks around $\varepsilon = 0.001$ and varies little across the swept range. For EB training, we use a constant $\varepsilon = 0.01$ across all context sizes, since the largest context can only be scaled to 1.5K. For image flow training, where the context can be scaled similarly to PBMC, the optimal ε at the largest minibatch context is 0.002 for CIFAR and 0.001 for ImageNet, consistent with the 8K-context ablation.