Expected Batch Optimal Transport Plans and Consequences for Flow Matching

Samuel Boïté[*]
DMA, ENS Paris

Julie Delon
DMA, ENS Paris
Université Paris Cité

Kimia Nadjahi
DI, ENS Paris, CNRS

Abstract

Solving optimal transport (OT) on random minibatches is a common surrogate for exact OT in large-scale learning. In flow matching (FM), this surrogate is used to obtain OT-like couplings that can straighten probability paths and reduce numerical integration cost. Yet, the population-level coupling induced by repeated minibatch OT remains only partially understood. We formalize this coupling as the expected batch OT plan $\overline{\pi}_k$, obtained by averaging empirical OT plans over independent minibatches of size k. We then establish its large-batch consistency and, in the semidiscrete case relevant to generative modeling, derive rates for both the transport-cost bias and the convergence of $\overline{\pi}_k$ to the OT plan. For FM, this yields a population coupling whose induced velocity field is regular enough to define a unique flow from the source to the discrete target. We finally quantify how OT batch size interacts with numerical integration in a tractable two-atom model and in synthetic and image experiments.

1 Introduction

Optimal transport (OT) provides a geometrically meaningful way to compare probability measures and has become a useful principle for designing machine learning methods, with successful applications in generative modeling, among others. In large-scale settings, exactly solving the OT problem is too costly, both computationally and statistically. This has motivated several tractable OT-based alternatives (Peyré and Cuturi, 2019). In particular, minibatch OT was introduced by Fatras et al. (2020) to avoid solving a single large OT problem between two empirical measures: instead, one subsamples smaller batches from the two discrete measures, solves the corresponding OT problems, and averages the resulting solutions. This strategy was then applied to generative adversarial network training, color transfer, and domain adaptation (Fatras et al., 2021a,b).

More recently, minibatch OT has attracted renewed attention through flow matching (FM), a class of generative models that learns a velocity field transporting a simple source distribution μ to a target data distribution ν (Liu et al., 2023; Lipman et al., 2023; Albergo et al., 2024). The FM objective depends on a coupling between source and target, that is, a joint distribution with marginals μ and ν, which determines both the probability path and the associated velocity field. Different choices of coupling therefore lead to different tradeoffs between training and inference. The independent coupling $\mu \otimes \nu$ is simple to implement but may lead to costly inference, whereas the OT plan between μ and ν would yield straight trajectories and allow efficient inference, but is intractable. Minibatch OT offers a practical compromise: it can produce straighter flows than the independent coupling, while being more affordable than exact OT (Pooladian et al., 2023a; Tong et al., 2024).

Many theoretical questions on minibatch OT remain open. In particular, the bias of minibatch OT relative to exact OT has not been precisely quantified, either for the transport cost or plan.

[*]Corresponding author: samuel.boite@ens.fr.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Then, regarding its use in FM, it is not clear whether the velocity field induced by minibatch OT defines a well-posed flow, since a coupling is not necessarily rectifiable (Hertrich et al., 2025, Proposition 13(iii)). Finally, recent benchmarks suggest that the benefits of minibatch OT-FM may only appear at much larger batch sizes (Zhang et al., 2026), while semidiscrete OT can also improve FM performance (Mousavi-Hosseini et al., 2026). Since these strategies make training more expensive, the practical tradeoff remains unclear: should one favor larger transport batches during training, or a larger inference budget at sampling time?

Contributions.

1. In Section 3, we study the expected batch optimal transport plan $\overline{\pi}_k$, obtained by averaging empirical OT plans over random batches of size k, complementing the results of Pooladian et al. (2023a). This plan converges to $\pi^\star$ when the OT plan is unique. In the semidiscrete setting relevant to FM, we further quantify this convergence: the cost bias $\mathbb{E}[W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k)] - W_2^2(\mu,\, \nu)$ decays as $O(k^{-1})$, yielding the plan-convergence rate $W_2^2(\overline{\pi}_k,\, \pi^\star) = O(k^{-1/2})$.
2. In Section 4, we show that when used in FM, the expected batch OT plan induces a well-posed flow: we prove that the associated velocity field is locally Lipschitz, generates a unique flow on $[0, 1)$, and admits a terminal map transporting the Gaussian source μ to the discrete target ν. To our knowledge, this is the first rectifiability result for a minibatch OT coupling in FM, and it provides a rigorous interpretation of this strategy beyond its use as a training heuristic. We give convergence rates for the terminal map's transport cost and cells.
3. We then quantify the tradeoff between OT batch size k and inference budget (number of function evaluations, NFE) through theoretical results and a benchmark covering a wider range of batch sizes and NFE than in related work. In a tractable model where ν is supported on two atoms, we prove that increasing the NFE reduces the integration error much faster than increasing k (Section 4.4). This reveals an asymmetry between these two computational resources, which we also observe in synthetic Gaussian-to-discrete experiments and image benchmarks: increasing k improves performance mainly in the low-NFE regime, while the benefit vanishes at larger NFE. Altogether, these results clarify when larger OT batches can effectively trade training-time computation for inference-time savings.[2]

2 Background on optimal transport and flow matching

We recall the background on OT and FM used in our work. We refer to Peyré and Cuturi (2019) for a comprehensive introduction to OT and to Pierret et al. (2026b) for a recent survey on FM.

2.1 Reminders on optimal transport

OT is a mathematical theory that studies how to transport one probability distribution to another optimally, according to a given cost function. Let $\mu,\, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be probability measures with finite second moments. Denote by $\Pi(\mu,\, \nu)$ the set of couplings between μ and ν, that is, the set of $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ whose marginals are μ and ν. For the quadratic cost, the OT problem is

$$W_2^2(\mu,\, \nu) = \inf_{\pi \in \Pi(\mu,\, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\pi(x,\, y). \tag{1}$$

Any minimizer $\pi^\star$ of (1) is called an OT plan. When μ is absolutely continuous, the OT plan is unique and takes the form $\pi^\star = (\mathrm{id},\, T^\star)_\sharp \mu$, where $\sharp$ denotes the push-forward operator ($T_\sharp \mu$ is the distribution of $T(X)$ where $X \sim \mu$), and where $T^\star : \mathbb{R}^d \to \mathbb{R}^d$ is called the OT map. When $\mu = \frac{1}{k}\sum_{i=1}^k \delta_{x_i}$ and $\nu = \frac{1}{k}\sum_{i=1}^k \delta_{y_i}$ are discrete measures supported on k points of $\mathbb{R}^d$, (1) reduces to finding an optimal permutation $\sigma^\star \in \mathfrak{S}_k$ minimizing the total cost

$$W_2^2(\mu,\, \nu) = \min_{\sigma \in \mathfrak{S}_k} \frac{1}{k} \sum_{i=1}^k \|x_i - y_{\sigma(i)}\|^2.$$

[2] Code is available at `https://github.com/samuelbx/expected-batch-ot`.

2.2 Reminders on flow matching

As in OT, the goal of FM is to transport one probability distribution to another. Let π be a coupling between μ and ν and draw $(X_0,\, X_1) \sim \pi$. For $t \in [0,\, 1]$, consider the linear interpolation $X_t = (1-t)X_0 + tX_1$ and denote its law by ρ_t^π. Then $(\rho_t^\pi)_{t\in[0,\,1]}$ is a probability path connecting $\rho_0^\pi = \mu$ and $\rho_1^\pi = \nu$. We define the associated target velocity field by $u_t^\pi(x) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$. Under regularity assumptions, the pair $(\rho^\pi,\, u^\pi)$ satisfies the continuity equation, and samples from ρ_t^π can be generated by drawing $x(0) \sim \mu$ and integrating the ODE $\dot{x}(s) = u_s^\pi(x(s))$, on $s \in [0,\, t]$.

In practice, u_t^π is generally not available in closed form, so one usually trains a time-dependent model v_t^θ by regressing onto the conditional target $X_1 - X_0$ under $(X_0,\, X_1) \sim \pi$ and $t \sim \mathrm{Unif}[0,\, 1]$. Once the velocity has been trained, the ODE is integrated with a discrete scheme such as Euler or Runge–Kutta. The number of velocity-function evaluations used by the scheme is called the NFE.

The choice of the coupling π is central in FM, as it determines the entire probability path $(\rho_t^\pi)_{t\in[0,\,1]}$ and corresponding target velocity field u_t^π. Choosing the OT plan $\pi^\star$, though generally intractable, would in principle yield straight paths and allow transport in a single integration step. Indeed, when μ is absolutely continuous and $T^\star$ is the OT map between μ and ν, the initial velocity is given by $u_0^{\pi^\star} = T^\star - \mathrm{id}$, so the OT map is recovered exactly by one Euler step from $t = 0$ to $t = 1$.

3 Expected batch optimal transport plan: consistency and rates

We formalize the expected batch OT plan and study its convergence to the OT plan, assuming it is unique. Given $\mu,\, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, minibatch OT consists in solving OT on random batches of size k drawn from μ and ν. This is especially useful for applications where solving the exact OT problem is intractable, and where we can use the resulting matched pairs during training. For $k \in \mathbb{N}^*$, let $\mathbf{x}_k = (x_1, \ldots, x_k)$ and $\mathbf{y}_k = (y_1, \ldots, y_k)$ be the source and target batches, respectively, with $x_i,\, y_i \in \mathbb{R}^d$. We denote the induced empirical measures by $\widehat{\mu}_{\mathbf{x}_k} = \frac{1}{k}\sum_{i=1}^k \delta_{x_i}$ and $\widehat{\nu}_{\mathbf{y}_k} = \frac{1}{k}\sum_{j=1}^k \delta_{y_j}$. Given random batches $(\mathbf{X}_k,\, \mathbf{Y}_k) \sim \mu^{\otimes k} \otimes \nu^{\otimes k}$, we write $\widehat{\mu}_k = \widehat{\mu}_{\mathbf{X}_k}$ and $\widehat{\nu}_k = \widehat{\nu}_{\mathbf{Y}_k}$.

3.1 Definition of the expected batch optimal transport plan

Minibatch OT induces a coupling $\overline{\pi}_k$ between μ and ν by averaging random empirical OT solutions. To define this coupling unambiguously, we assume that for each $k \in \mathbb{N}^*$, we have access to a measurable solver whose output $(\mathbf{x}_k,\, \mathbf{y}_k) \mapsto \widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k}$ is supported on an optimal permutation between $\widehat{\mu}_{\mathbf{x}_k}$ and $\widehat{\nu}_{\mathbf{y}_k}$. For random batches, we write $\widehat{\pi}_k = \widehat{\pi}_{\mathbf{X}_k,\,\mathbf{Y}_k}$. The expected batch OT plan is then obtained by averaging these solver outputs over random batches.

Definition 3.1 (Expected batch OT plan)**.** Fix the measurable solver introduced above. For $k \in \mathbb{N}^*$, the expected batch optimal transport plan $\overline{\pi}_k$ is the unique probability measure on $\mathbb{R}^d \times \mathbb{R}^d$ such that, for every bounded Borel function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

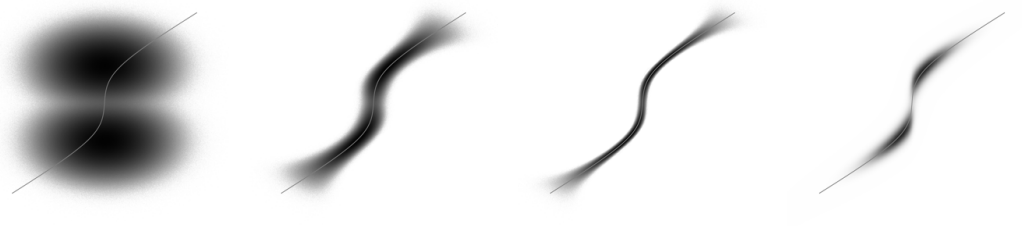
$$\int_{\mathbb{R}^d\times\mathbb{R}^d} g(x,\, y)\, \mathrm{d}\overline{\pi}_k(x,\, y) = \int_{(\mathbb{R}^d)^k\times(\mathbb{R}^d)^k} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} g\, \mathrm{d}\widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k} \right) \mathrm{d}(\mu^{\otimes k} \otimes \nu^{\otimes k})(\mathbf{x}_k,\, \mathbf{y}_k)\,.$$

Hence, $\overline{\pi}_k$ is the law of the pair obtained as follows: first draw $(\mathbf{X}_k,\, \mathbf{Y}_k) \sim \mu^{\otimes k} \otimes \nu^{\otimes k}$, then solve OT between $\widehat{\mu}_k$ and $\widehat{\nu}_k$, and finally select one of the k matched pairs uniformly at random. This construction specializes, in the fully discrete case, to the averaged minibatch transport matrix of Fatras et al. (2021b). It also matches the marginalized minibatch coupling $q^{(k)}$ used in batch OT-FM by Pooladian et al. (2023a).

We show in Appendix A.1 that this construction indeed yields a well-defined probability measure, that $\overline{\pi}_k \in \Pi(\mu,\, \nu)$, and $\overline{\pi}_1 = \mu \otimes \nu$. When one marginal is absolutely continuous, solver choices agree almost surely under random batches, and hence induce the same $\overline{\pi}_k$.

3.2 Asymptotic consistency and rates for the cost

We first study the expected batch OT plan through the transport cost it induces. Since $\overline{\pi}_k \in \Pi(\mu,\, \nu)$, this cost is greater than the optimal value $W_2^2(\mu,\, \nu)$. To quantify the excess cost, we first show that the cost induced by $\overline{\pi}_k$ equals the expected value of $W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k)$ over random batches.



(a) $k = 1$ (independent) (b) $k = 100$ (c) $k = 1\,000$ (d) $\varepsilon = 0.15$ (entropic)

Figure 1: Transport plans between $\mathcal{N}(0, 1)$ and a mixture of two symmetric Gaussians. As k grows, $\overline{\pi}_k$ approaches the OT plan $\pi^\star$ (shown as a line) (a–c). (d) shows the entropically regularized plan.

Proposition 3.2. *For any $k \in \mathbb{N}^*$, $\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\overline{\pi}_k(x, y) = \mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)]$.*

The identity above connects our analysis with classical questions in statistical OT, such as sample complexity of empirical Wasserstein costs. In contrast with most sample-complexity results, which control $\mathbb{E}[|W_2^2(\widehat{\mu}_k, \widehat{\nu}_k) - W_2^2(\mu, \nu)|]$, we study directly the signed bias $\mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)] - W_2^2(\mu, \nu)$. This distinction yields a faster $O(k^{-1})$ rate in the nondegenerate, semidiscrete setting.

Proposition 3.3 (Asymptotic consistency and rates for expected batch OT cost)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$.*

1. *The sequence $(\mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)])_{k \in \mathbb{N}^*}$ is nonincreasing.*
   If $\mu, \nu \in \mathcal{P}_q(\mathbb{R}^d)$ for some $q > 2$, then $\lim_{k \to \infty} \mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)] = W_2^2(\mu, \nu)$.
2. *If $\mu = \mathcal{N}(0, I_d)$ and ν has finite support, then $\mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)] - W_2^2(\mu, \nu) = O(k^{-1})$.*

Thanks to the semidiscrete structure, our rates avoid the exponential dependence on d that appears in the fully discrete case (Fournier and Guillin, 2015). One can show an $O(k^{-1/2})$ rate for $\mathbb{E}[|W_2^2(\widehat{\mu}_k, \widehat{\nu}_k) - W_2^2(\mu, \nu)|]$ using existing results from (Hundrieser et al., 2024; Staudt and Hundrieser, 2025), but the sharper $O(k^{-1})$ rate exploits the signed expectation. This result holds beyond Gaussian μ under the more general Assumptions A or B in Genans et al. (2025).

3.3 Convergence and rates for the plan

We now establish the convergence of $(\overline{\pi}_k)_{k \in \mathbb{N}^*}$ in the space $(\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d), W_2)$. From now on, we assume that μ is absolutely continuous with respect to Lebesgue measure, so that the OT plan $\pi^\star$ between μ and ν is unique.

Before stating our main result, we provide an illustration in Figure 1, which suggests that $\overline{\pi}_k$ approaches $\pi^\star$ as k increases. We also compare $\overline{\pi}_k$ with an entropically regularized plan. Qualitatively, the minibatch strategy induces a form of regularization distinct from entropic regularization. A precise characterization of this effect is outside the scope of the article and left for future work.

The next proposition proves the convergence to $\pi^\star$ as well as quantitative bounds on $W_2^2(\overline{\pi}_k, \pi^\star)$ in the semidiscrete setting, which is relevant for flow matching. The detailed proof is in Appendix A.4.

Proposition 3.4 (Convergence and rates for the expected batch OT plan)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ with μ absolutely continuous with respect to the Lebesgue measure. Then, $(\overline{\pi}_k)_{k \in \mathbb{N}^*}$ converges weakly, and therefore in W_2, to $\pi^\star$: for any bounded continuous function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\lim_{k \to +\infty} \int g \, \mathrm{d}\overline{\pi}_k = \int g \, \mathrm{d}\pi^\star$. Moreover,*

1. *For any $k \geqslant 1$, $W_2^2(\overline{\pi}_k, \pi^\star) \geqslant \frac{1}{2}\big((\mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)])^{\frac{1}{2}} - W_2(\mu, \nu)\big)^2$.*
2. *If $\mu = \mathcal{N}(0, I_d)$ and ν has finite support, then there exists $A_\nu > 0$ such that*

$$W_2^2(\overline{\pi}_k, \pi^\star) \leqslant A_\nu (\mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)] - W_2^2(\mu, \nu))^{1/2} = O(k^{-1/2}).$$

The above result shows that if the expected batch OT cost is far above $W_2^2(\mu, \nu)$, then $\overline{\pi}_k$ is necessarily far from $\pi^\star$ in Wasserstein distance. However, the converse fails in general. More precisely, we show that for any $\pi \in \Pi(\mu, \nu)$, there is no universal modulus that controls $W_2^2(\pi, \pi^\star)$ in terms of the cost bias (Appendix A.4.4). Plan convergence thus requires additional hypotheses on μ, ν. We use the W_2

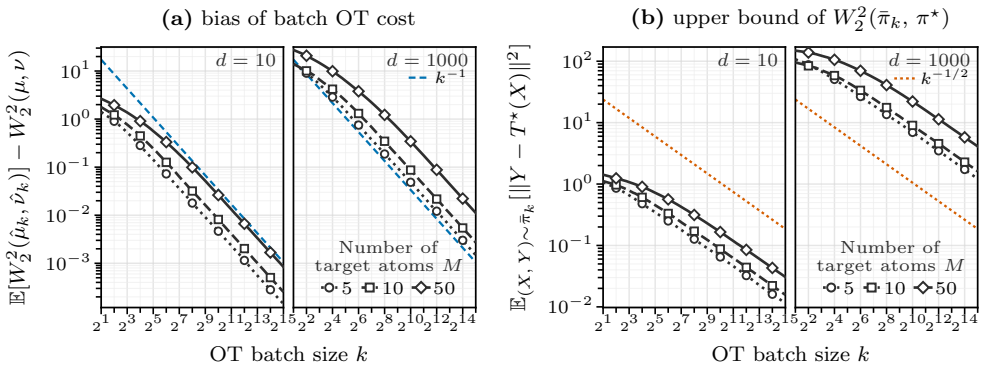


Figure 2: Numerical rates in Gaussian-to-discrete experiments. The expected batch OT cost bias decays as $O(k^{-1})$ (left). The quantity $\mathbb{E}_{(X,Y)\sim\overline{\pi}_k}[\|Y - T^\star(X)\|^2]$, which upper bounds $W_2^2(\overline{\pi}_k, \pi^\star)$, decays as $O(k^{-1/2})$ (right). Standard errors are below plotting resolution.

distance to compare plans because it is consistent with the geometry of the underlying OT problem. Moreover, all couplings in $\Pi(\mu, \nu)$ have the same second moment, so the weak convergence upgrades to W_2 convergence, and it is natural to consider the corresponding rate.

3.4 Numerical illustration of convergence rates

We illustrate the convergence rates of the expected batch OT cost and plan in Gaussian-to-discrete settings. In Figure 2, we consider $\mu = \mathcal{N}(0, I_d)$ and ν supported on M distinct atoms, drawn uniformly from $[0, 1]^d$ and then fixed, for $d \in \{10, 1000\}$ and $M \in \{5, 10, 50\}$.

We evaluate the cost bias and observe a decay in $O(k^{-1})$, illustrating Proposition 3.3. We also observe an $O(k^{-1/2})$ decay for a computable upper bound on $W_2^2(\overline{\pi}_k, \pi^\star)$. The finite-sample regime appears slower than the asymptotic regime. Directly estimating $W_2^2(\overline{\pi}_k, \pi^\star)$ is both computationally and statistically challenging, since it requires discretizing both $\overline{\pi}_k$ and $\pi^\star$ and solving a discrete OT problem on $\mathbb{R}^d \times \mathbb{R}^d$. Instead, we estimate $\mathbb{E}_{(X,Y)\sim\overline{\pi}_k}\left[\|Y - T^\star(X)\|^2\right]$, where $T^\star$ is the OT map from μ to ν. This quantity upper bounds $W_2^2(\overline{\pi}_k, \pi^\star)$: if $(X, Y) \sim \overline{\pi}_k$, then $((X, Y), (X, T^\star(X)))$ is a coupling of $\overline{\pi}_k$ and $\pi^\star$.

4 Flow matching under expected batch optimal transport

In this section, we study the effect of minibatch OT on the dynamics of flow matching. We focus on the Gaussian-to-discrete setting, where the source μ is a standard Gaussian, and the target ν is supported on pairwise distinct atoms $v_1, \ldots, v_M$. We first show that the velocity field of FM, when derived from the expected batch OT plan $\overline{\pi}_k$, is regular enough to define a well-posed ODE flow on $[0, 1)$. We then study the extent to which increasing the batch size k can serve as a substitute for finer numerical integration, thereby complementing the empirical studies of Pooladian et al. (2023a); Tong et al. (2024); Zhang et al. (2026). The theoretical statements in this section concern the exact target velocity induced by the coupling. The image experiments then separately study learned neural approximations.

4.1 Posterior representation of the velocity field

In this Gaussian-to-discrete setting, the velocity field admits a closed-form expression, as studied for instance by Bertrand et al. (2025, Proposition 1). We prove the following result in Appendix B.1.

Lemma 4.1. *Assume $\mu = \mathcal{N}(0,\, I_d)$ and $\nu = \sum_{j=1}^{M} q_j \delta_{v_j}$, with $q_j > 0$ and pairwise distinct atoms $v_j \in \mathbb{R}^d$. Let $\pi \in \Pi(\mu,\, \nu)$ and for $t \in [0, 1)$, let $X_t = (1-t)X_0 + tX_1$, $(X_0, X_1) \sim \pi$. Then,*

$$u_t^\pi(x) = \frac{m_t^\pi(x) - x}{1-t},\ m_t^\pi(x) = \mathbb{E}_\pi[X_1 \mid X_t = x] = \sum_{j=1}^{M} v_j \mathbb{P}_{(X_0,\, X_1)\sim\pi}(X_1 = v_j \mid X_t = x). \quad (2)$$

Moreover, if f_π denotes the density of π with respect to the product of the Lebesgue measure on $\mathbb{R}^d$ and the counting measure on $\{v_1, \dots, v_M\}$, then, for ρ_t^π-a.e. x (equivalently, whenever the denominator below is positive),

$$\mathbb{P}_{(X_0,X_1)\sim\pi}(X_1 = v_j \mid X_t = x) = f_\pi\left(\frac{x - tv_j}{1-t}, v_j\right) \Bigg/ \sum_{\ell=1}^{M} f_\pi\left(\frac{x - tv_\ell}{1-t}, v_\ell\right). \quad (3)$$

The velocity field points from x to the posterior mean $m_t^\pi(x)$, which always lies in the convex hull of the target atoms. Observe that if we write $\overline{a}_j^\pi(x_0) = \mathbb{P}_{(X_0,\, X_1)\sim\pi}(X_1 = v_j \mid X_0 = x_0)$, then $f_\pi(x_0,\, v_j) = \varphi(x_0)\overline{a}_j^\pi(x_0)$, where φ is the standard Gaussian density (Appendix B.2). Thus the velocity field is also completely determined by the assignment probabilities $\overline{a}_j^\pi$ between X_0 and X_1. In the specific case where $\pi = \mu \otimes \nu$ and ν is uniform, the assignment probabilities in (3) form a softmax with logits $-\frac{\|x - tv_j\|^2}{2(1-t)^2}$, $j \in \{1, \dots, M\}$.

4.2 Smoothness and rectifiability of the expected batch OT plan

When using batch OT for FM, a natural question is whether the velocity field generated by $\overline{\pi}_k$ defines a well-posed flow. The following result, proven in Appendix B.2, confirms that this is the case.

Proposition 4.2 (Rectifiability of the expected batch OT plan)**.** *Fix $k \in \mathbb{N}^*$ and consider the Gaussian-to-discrete setting. The expected batch OT plan $\overline{\pi}_k$ is rectifiable in the following sense.*

1. (Local Lipschitzness of $u_t^{\overline{\pi}_k}$). *For any compact $K \subseteq \mathbb{R}^d$, the map $(t,\, x) \in [0,\, 1) \times K \mapsto u_t^{\overline{\pi}_k}(x)$ is jointly continuous and for any $\tau \in (0,\, 1)$, it is locally Lipschitz in x uniformly in $t \in [0,\, \tau]$.*

2. (Pre-terminal flow). *For any $x_0 \in \mathbb{R}^d$, the ODE $\dot{x}(t) = u_t^{\overline{\pi}_k}(x(t))$, $x(0) = x_0$, admits a unique solution on $[0,\, 1)$, denoted by $t \mapsto \phi_t^{\overline{\pi}_k}(x_0)$. Moreover, for any $t \in [0,\, 1)$, the map $x_0 \mapsto \phi_t^{\overline{\pi}_k}(x_0)$ is continuous and $(\phi_t^{\overline{\pi}_k})_\sharp \mu = \rho_t^{\overline{\pi}_k}$.*

3. (Terminal flow). *There exists a Borel map $\phi^{\overline{\pi}_k} : \mathbb{R}^d \to \{v_1, \dots, v_M\}$ such that $\phi^{\overline{\pi}_k}(x_0) = \lim_{t\uparrow 1} \phi_t^{\overline{\pi}_k}(x_0)$ for μ-a.e. x_0, and $(\phi^{\overline{\pi}_k})_\sharp \mu = \nu$.*

The terminal flow map $\phi^{\overline{\pi}_k}$ partitions $\mathbb{R}^d$ into regions $\Gamma_{k,\,j} = (\phi^{\overline{\pi}_k})^{-1}(\{v_j\})$, $j \in \{1, \dots, M\}$, according to the final atom reached when starting from $x_0 \in \mathbb{R}^d$. As k grows, these regions approach the Laguerre cells $C_j^\star = (T^\star)^{-1}(\{v_j\})$ of the semidiscrete OT map (Peyré and Cuturi, 2019, Section 5) between μ and ν (see Figure 3). The next result, shown in Appendix B.3, transfers the rates of Sections 3.2 and 3.3 to the terminal flow map $\phi^{\overline{\pi}_k}$. Here $A \triangle B = (A \setminus B) \cup (B \setminus A)$ denotes the symmetric difference.

Corollary 4.3 (Rates for the terminal flow map)**.** *In the Gaussian-to-discrete setting,*

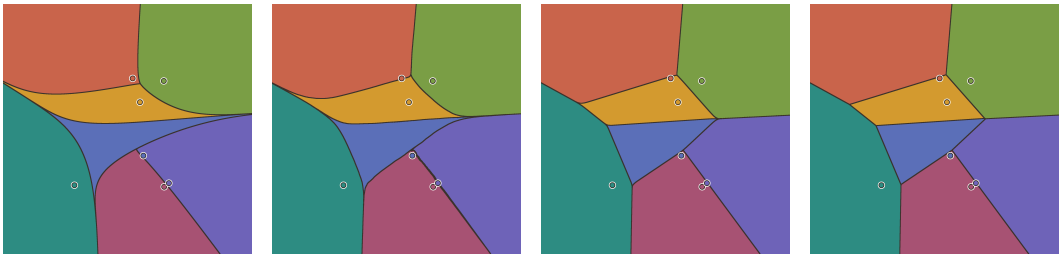
1. (Transport cost) $0 \leqslant \int_{\mathbb{R}^d} \|\phi^{\overline{\pi}_k}(x) - x\|^2 \,\mathrm{d}\mu(x) - W_2^2(\mu,\, \nu) = O(k^{-1})$.

2. (FM cells) $\mu(\{x \in \mathbb{R}^d,\ \phi^{\overline{\pi}_k}(x) \neq T^\star(x)\}) = O(k^{-1/2})$, *therefore* $\mu(\Gamma_{k,\,j} \triangle C_j^\star) = O(k^{-1/2})$.

The transport cost of $\phi^{\overline{\pi}_k}$ can be nonmonotonic in k, as highlighted in the numerical counterexample of Appendix B.3.1. Moreover, the convergence of the FM cells is in μ-measure and is not uniform, as shown in Appendix B.3.2 using the four-point example by Pierret et al. (2026a).

4.3 Posterior concentration

The concentration in time of the posterior weights $\mathbb{P}_{(X_0,X_1)\sim\pi}(X_1 = v_j \mid X_t = x)$ determines how quickly the posterior mean $m_t^\pi(x)$ converges to a single specific atom, according to (2). Figure 4



(a) $k = 1$ (b) $k = 25$ (c) $k = 1\,000$ (d) OT plan $\pi^\star$

Figure 3: The terminal map $\phi^{\overline{\pi}_k}$ induced by the flow partitions $\mathbb{R}^2$ into cells that approach the semidiscrete OT Laguerre partition. The case $k = 1$ corresponds to the independent coupling. In this example, $\mu = \mathcal{N}(0,\, I_2)$ and ν is uniformly supported on 7 fixed atoms, shown as colored points.

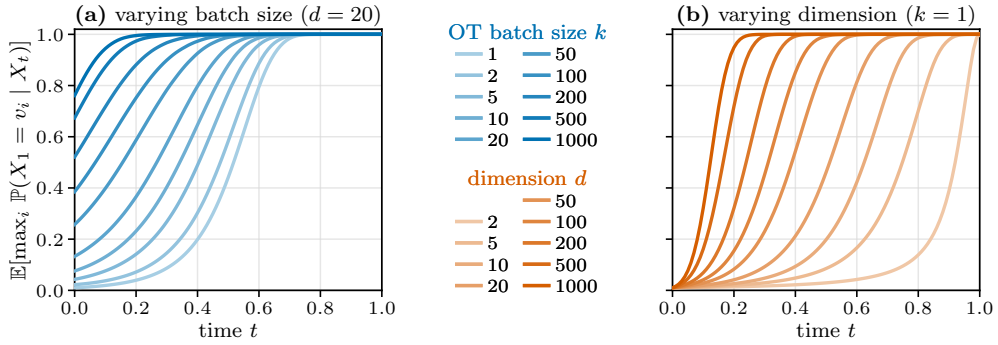


Figure 4: Posterior concentration increases with OT batch size k (left), and with ambient dimension d when $k = 1$ (right) in a Gaussian-to-discrete setting with $M = 100$ atoms. Standard errors over sampled trajectories are below plotting resolution. The batch-size experiment uses a fixed cloud of 100 atoms drawn uniformly in $[-1,\, 1]^{20}$; the dimension experiment uses one such cloud for each d.

reports $\mathbb{E}\left[\max_{1 \leqslant i \leqslant M} \mathbb{P}_{(X_0, X_1) \sim \overline{\pi}_k}(X_1 = v_i \mid X_t)\right]$ when k increases (left panel) for a fixed dimension $d = 20$, or when d increases (right panel) for the independent coupling $k = 1$. Both panels show stronger concentration over time as k or d increases. For $k = 1$, the increase in concentration with d was already observed by Bertrand et al. (2025); we turn this observation into a quantitative theoretical statement in Appendix B.4. The increase with k is harder to prove, since the OT batch plan is not available in closed form, but we hope to provide a theoretical analysis in future work.

4.4 Numerical integration

We now turn to the interaction between the OT batch size k and the numerical integration budget, measured by the number of function evaluations (NFE). For any $\pi \in \Pi(\mu,\, \nu)$, define the explicit Euler iterates with $n \in \mathbb{N}^*$ uniform time steps $t_j = j/n$ by

$$f^{\pi}_{0,\,n} = \mathrm{id} \quad \text{and} \quad f^{\pi}_{j+1,\,n} = (\mathrm{id} + \tfrac{1}{n} u^{\pi}_{t_j}) \circ f^{\pi}_{j,\,n} \quad \text{for } j \in \{0, \dots, n-1\}. \tag{4}$$

We denote the final Euler map by $f^{\pi}_n = f^{\pi}_{n,\,n}$. For $n,\, k \in \mathbb{N}^*$, we quantify the expected Euler integration error of the target velocity $u^{\overline{\pi}_k}$ by $\mathscr{E}_{n,\,k} = \mathbb{E}_{X_0 \sim \mu}\left[\left\| f^{\overline{\pi}_k}_n(X_0) - \phi^{\overline{\pi}_k}(X_0) \right\|\right]$.

Can the OT batch size efficiently compensate for fewer function evaluations? To isolate the effect of the two parameters on $\mathscr{E}_{n,\,k}$, we consider a tractable model and compare the asymptotics along the two extreme regimes $k = 1$ and $n = 1$. This simple setting already reveals a striking difference in how $\mathscr{E}_{n,\,k}$ may decay with respect to n versus k.

Proposition 4.4 (Asymptotic comparison between OT batch size and NFE)**.** *Let* $\mu = \mathcal{N}(0,\, 1)$ *and* $\nu = \mathrm{Unif}\{-1,\, 1\}$. *Then* $\log \mathscr{E}_{n,\,1} \sim -\frac{2\pi}{\sqrt{3}} n^{\frac{2}{3}}$ *as* $n \to \infty$ *and* $\log \mathscr{E}_{1,\,k} \sim -\frac{1}{2} \log k$ *as* $k \to \infty$.

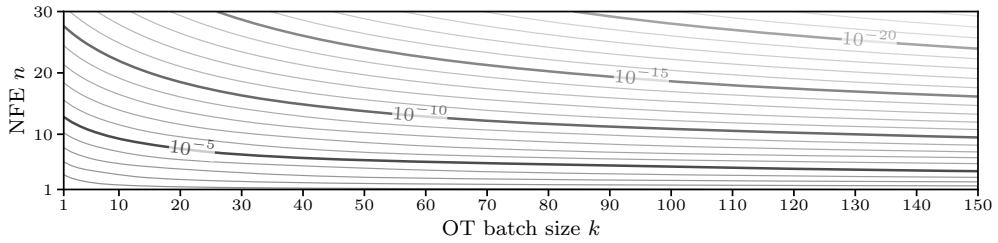


Figure 5: Level sets of the Euler integration error $\mathscr{E}_{n,\,k}$ in the two-atom model of Proposition 4.4.

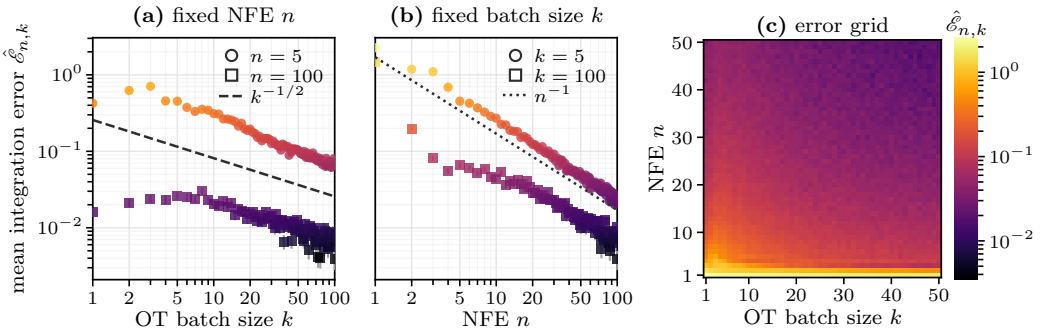


Figure 6: In a Gaussian-to-discrete setting with 100 atoms in dimension 20, the numerical integration error is consistent with a $1/n$ scaling in the number of Euler steps n and a $1/\sqrt{k}$ scaling in the OT batch size k. Error bars show ± 1 standard error over sampled initial conditions.

Increasing the number of Euler steps thus causes the integration error to decay as $\exp(-c\,n^{2/3})$, whereas increasing the OT batch size only yields a $k^{-1/2}$-type decay. In this specific setting, the influence of the NFE is considerably more pronounced, and this asymmetry remains clear even outside the asymptotic regime. Figure 5 displays the level sets of $\mathscr{E}_{n,\,k}$ in the $(n,\,k)$ plane. It shows for example that to compensate for a reduction in NFE from 25 to 10, one needs to increase OT batch size k from 1 to 90, well beyond the two-atom support size.

We observe a similar qualitative asymmetry in higher-dimensional semidiscrete examples with larger finite support. Although we were not able to prove it formally, we conjecture that in this general case, $\mathscr{E}_{n,\,k}$ decays as $1/n$ and $1/\sqrt{k}$. As a result, halving the expected integration error would require either doubling the NFE, and hence roughly doubling inference time, or multiplying the OT batch size by 4, thereby multiplying the OT solve time during training by $4^3 = 64$. This highlights the difficulty of using larger OT batches as a surrogate for additional inference steps. Figure 6 shows the (Monte Carlo) average integration error $\hat{\mathscr{E}}_{n,\,k}$ as a function of n and k, for $\nu = \mathrm{Unif}\{v_1, \ldots, v_{100}\}$ with $\{v_1, \ldots, v_{100}\} \subseteq [-1,\,1]^{20}$. For each pair $(n,\,k)$, we estimate the numerical integration error of the expected batch OT flow by comparing a coarse n-step Euler discretization with a much finer reference discretization. The error also appears to be non-monotonic in k for small values, reflecting the fact that the monotonicity of the expected transport cost (Proposition 3.3) does not necessarily imply monotonicity of the induced numerical error.

4.5 Image dataset experiments

We now evaluate expected batch OT on two image datasets: CIFAR-10 (Krizhevsky, 2009) and SVHN (Netzer et al., 2011). These experiments probe the learned neural velocity field, rather than the exact target velocity studied above.

For each dataset, we define ν as the uniform empirical measure on the flattened images after normalization to $[-1,\,1]^{3072}$. We keep the optimization minibatch size fixed and vary only the OT batch size $k \in \{1,\,2,\,4, \ldots, 8192\}$. For each value of k, we train a separate flow model using exact

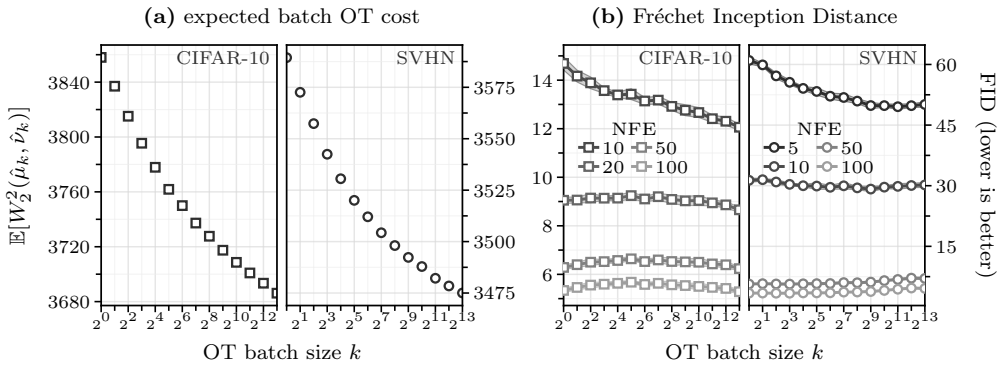


Figure 7: **Image experiment**. The expected batch OT cost decreases steadily up to $k = 8192$ (left). Increasing k improves FID mainly in the low-NFE regime, while the effect weakens or reverses at large NFE (right). Error bands show ± 1 empirical standard deviation over training seeds.

OT couplings computed on batches of size k. We then generate samples with explicit Euler using a prescribed NFE, and evaluate them with the Fréchet Inception Distance (FID, Heusel et al. (2017)). We stop at $k = 8192$ because larger exact OT batches exceeded our computational budget. For each dataset and each k, we report FID over six training seeds.

Figure 7 (left) shows that $\mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)]$ decreases steadily on both datasets, with no visible saturation up to $k = 8192$. By Proposition 3.4(1), the expected batch OT plan should not be regarded as effectively converged. Figure 7 (right) shows that increasing k improves FID mainly when the integration scheme is coarse ($\mathrm{NFE} = 10$). At high NFE, the effect is weaker and can reverse: on SVHN, large k degrades the FID at $\mathrm{NFE} = 50$ and 100. This effect can also be seen qualitatively on generated samples: see Figures 14 to 17 in Appendix C.

We recall that the effect of k and NFE occurs at different stages: increasing k affects training time, while NFE determines the cost of sampling at inference. The improvement at low NFE in Figure 7 therefore suggests a possible training-inference tradeoff: larger OT batches can improve low-NFE sampling, at the price of additional training computation. On our hardware, exact OT batches up to $k = 1024$ add about 3% training overhead, whereas $k = 8192$ increases training time by 55% (Table 2; see Appendix C.2.4 for details).

Our experiment complements previous empirical studies. Pooladian et al. (2023a); Tong et al. (2024) fix the training batch size B and compare independent pairing with minibatch OT only in the case $k = B$; thus, they do not explore a grid of OT batch sizes k. Zhang et al. (2026) study much larger batch couplings with entropic OT, instead of exact batch OT. Moreover, these works do not track the expected batch OT cost $\mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)]$ as a function of k. A comparison between exact batch OT and entropic or sliced alternatives is left for future work.

5 Conclusion

We studied the expected batch OT plan as a population-level object underlying minibatch OT for flow matching, complementing Fatras et al. (2021b); Pooladian et al. (2023a). In the semidiscrete setting, we established quantitative convergence rates and proved that the induced velocity field is regular enough to define a well-posed flow. Taken together, these results put minibatch OT on firmer theoretical ground and sharpen its practical interpretation: increasing the OT batch size can meaningfully improve low-NFE generation, but it does not eliminate the need for accurate inference-time integration and the gains must be weighed against additional training-time OT cost.

Limitations and future directions.

- Although the expected batch OT plan can be formulated in full generality, most of our quantitative results are obtained in the semidiscrete setting, which is the one directly relevant to generative

modeling. Extending comparable convergence rates and structural properties to more general source-target configurations would further clarify the scope of the construction.

- Our theoretical analysis concerns the true velocity field induced by the coupling, whereas practical FM relies on an approximation learned by a neural network. Understanding how our results transfer to learned velocities is therefore a central next step, even if analyzing the exact target dynamics is a necessary starting point.
- We obtained a theoretical comparison between OT batch size and numerical integration in a tractable, one-dimensional two-atom model using the Euler method. Deriving general bounds for the behavior of $\mathscr{E}_{n,\,k}$ as a function of both n and k, and for other numerical schemes, appears much more challenging and remains open.
- Expected batch OT is only one possible coupling surrogate between independent pairing and exact OT. Similar questions about rectifiability and interaction with numerical integration could be addressed for other couplings used in practice, such as entropic and sliced variants.

Acknowledgements

This research was supported, in part, by the Agence nationale de la recherche (ANR) under the France 2030 programme, through the PEPR IA projects PDE-AI (ANR-23-PEIA-0004) and MacLeOD (ANR-25-PEIA-0005). The authors declare no competing interests.

References

Albergo, M. S., Goldstein, M., Boffi, N. M., Ranganath, R., and Vanden-Eijnden, E. (2024). Stochastic interpolants with data-dependent couplings. In *ICML*, pages 921–937.

Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser Basel, 1 edition.

Bertrand, Q., Gagneux, A., Massias, M., and Emonet, R. (2025). On the closed-form of flow matching: Generalization does not arise from target stochasticity. In *NeurIPS*.

Bobkov, S. and Ledoux, M. (2019). One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 261(1259):1–126.

Boucheron, S. and Massart, P. (2011). A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150(3–4):405–433.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. In *NeurIPS*.

Dodson, N., Gao, X., Wang, Q., Wang, Y., and Wan, Z. (2026). Two calm ends and the wild middle: A geometric picture of memorization in diffusion models. In *ICML*, pages 25643–25694.

Fatras, K., Séjourné, T., Courty, N., and Flamary, R. (2021a). Unbalanced minibatch optimal transport; applications to domain adaptation. In *ICML*, pages 3186–3197.

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. (2020). Learning with minibatch Wasserstein: asymptotic and gradient properties. In *AISTATS*, pages 2131–2141.

Fatras, K., Zine, Y., Majewski, S., Flamary, R., Gribonval, R., and Courty, N. (2021b). Minibatch optimal transport distances; analysis and applications. arXiv:2101.01792v1.

Flamary, R., Courty, N., Gramfort, A., et al. (2021). POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3–4):707–738.

Genans, F., Godichon-Baggioni, A., Vialard, F.-X., and Wintenberger, O. (2025). Stochastic optimization in semi-discrete optimal transport: Convergence analysis and minimax rate. In *NeurIPS*.

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Hertrich, J., Chambolle, A., and Delon, J. (2025). On the relation between rectified flows and optimal transport. In *NeurIPS*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*.

Hundrieser, S., Staudt, T., and Munk, A. (2024). Empirical optimal transport between different measures adapts to lower complexity. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 60(2):824–846.

Klatt, M., Munk, A., and Zemel, Y. (2022). Limit laws for empirical optimal solutions in random linear programs. *Annals of Operations Research*, 315(1):251–278.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow matching for generative modeling. In *ICLR*.

Liu, X., Gong, C., and Liu, Q. (2023). Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*.

Mousavi-Hosseini, A., Zhang, S., Klein, M., and Cuturi, M. (2026). Flow matching with semidiscrete couplings. In *ICLR*.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Parmar, G., Zhang, R., and Zhu, J.-Y. (2022). On aliased resizing and surprising subtleties in GAN evaluation. In *CVPR*, pages 11400–11410.

Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607.

Pierret, E., Hertrich, J., Hurault, S., and Delon, J. (2026a). Tessellations of semi-discrete flow matching. In *NeurIPS*. To appear. arXiv:2605.07513v1.

Pierret, E., Tosel, V., Delon, J., and Newson, A. (2026b). Flow matching for applied mathematicians. HAL: hal-05538982v2.

Poli, M., Massaroli, S., Yamashita, A., Asama, H., Park, J., and Ermon, S. (2021). TorchDyn: Implicit models and neural numerical methods in PyTorch. Software package.

Pooladian, A.-A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., and Chen, R. T. Q. (2023a). Multisample flow matching: Straightening flows with minibatch couplings. In *ICML*, pages 28100–28127.

Pooladian, A.-A., Divol, V., and Niles-Weed, J. (2023b). Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *ICML*, pages 28128–28150.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser Cham, 1 edition.

Staudt, T. and Hundrieser, S. (2025). Convergence of empirical optimal transport in unbounded settings. *Bernoulli*, 31(3):1929–1954.

Tong, A., Fatras, K., Malkin, N., et al. (2024). Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*.

TorchVision maintainers and contributors (2024). TorchVision: PyTorch's computer vision library. Software release 0.20.0.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, 1 edition.

Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3):261–272.

Wan, Z., Wang, Q., Mishne, G., and Wang, Y. (2025). Elucidating flow matching ODE dynamics via data geometry and denoisers. In *ICML*, pages 62020–62083.

Zhang, S. Y., Mousavi-Hosseini, A., Klein, M., and Cuturi, M. (2026). On fitting flow models with large Sinkhorn couplings. *Transactions on Machine Learning Research*.

A Proofs for Section 3

This appendix follows the structure of Section 3. We first justify the construction of $\overline{\pi}_k$ and prove its elementary properties (Appendix A.1). We then analyze its transport cost through the empirical OT cost $\mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)]$ (Appendices A.2 and A.3). The last part proves convergence of $\overline{\pi}_k$ as a plan and gives the cost-to-plan estimates used in Proposition 3.4 (Appendix A.4).

Throughout, if $\mathcal{Y} \subseteq \mathbb{R}^d$ is a finite set containing at least two points, we define

$$\mathrm{sep}(\mathcal{Y}) = \min_{\substack{y,\, y' \in \mathcal{Y} \\ y \neq y'}} \|y - y'\|, \quad \text{and} \quad \mathrm{diam}(\mathcal{Y}) = \max_{y, y' \in \mathcal{Y}} \|y - y'\|.$$

A.1 Construction and elementary properties (Section 3.1)

We detail the construction of the expected batch OT plan $\overline{\pi}_k$ and record its elementary properties.

A.1.1 Empirical OT solver

An empirical OT problem may have several solutions, and the solver used in the implementation of batch OT typically selects one. To model this, we fix a measurable empirical OT solver, and define the corresponding expected batch OT plan as the average of these empirical OT plans over random batches. We then show that the solver-dependence disappears when μ or ν is absolutely continuous.

For $k \in \mathbb{N}^*$, write $E_k = (\mathbb{R}^d)^k \times (\mathbb{R}^d)^k$, $F = \mathbb{R}^d \times \mathbb{R}^d$ and $\alpha_k = \mu^{\otimes k} \otimes \nu^{\otimes k}$. For $(\mathbf{x}_k,\, \mathbf{y}_k) \in E_k$, recall that $\widehat{\mu}_{\mathbf{x}_k} = \frac{1}{k}\sum_{i=1}^k \delta_{x_i}$ and $\widehat{\nu}_{\mathbf{y}_k} = \frac{1}{k}\sum_{j=1}^k \delta_{y_j}$. We consider a measurable solver

$$\begin{aligned} \widehat{\sigma}_k \colon \quad E_k &\longrightarrow \mathfrak{S}_k, \\ (\mathbf{x}_k,\, \mathbf{y}_k) &\longmapsto \widehat{\sigma}_{\mathbf{x}_k,\, \mathbf{y}_k}, \end{aligned}$$

such that, for every $(\mathbf{x}_k,\, \mathbf{y}_k) \in E_k$,

$$\widehat{\sigma}_{\mathbf{x}_k,\, \mathbf{y}_k} \in \operatorname*{argmin}_{\sigma \in \mathfrak{S}_k} C_\sigma(\mathbf{x}_k,\, \mathbf{y}_k), \qquad C_\sigma(\mathbf{x}_k,\, \mathbf{y}_k) = \frac{1}{k}\sum_{i=1}^k \|x_i - y_{\sigma(i)}\|^2.$$

Let $\widehat{\pi}_{\mathbf{x}_k,\, \mathbf{y}_k} = \frac{1}{k}\sum_{i=1}^k \delta_{(x_i,\, y_{\widehat{\sigma}_{\mathbf{x}_k,\, \mathbf{y}_k}(i)})}$ be the corresponding OT plan between $\widehat{\mu}_{\mathbf{x}_k}$ and $\widehat{\nu}_{\mathbf{y}_k}$.

Existence of a measurable solver. Fix a total order $\preceq_k$ on the finite set $\mathfrak{S}_k$, and choose the $\preceq_k$-smallest minimizer $\widehat{\sigma}_k^{\preceq}$. This map is Borel because each C_σ is continuous, and for every $\sigma \in \mathfrak{S}_k$,

$$\{\widehat{\sigma}_k^{\preceq} = \sigma\} = \bigcap_{\sigma' \prec_k \sigma} \{C_\sigma < C_{\sigma'}\} \cap \bigcap_{\sigma' \succeq_k \sigma} \{C_\sigma \leqslant C_{\sigma'}\}.$$

Uniqueness for absolutely continuous measures. Generic uniqueness of empirical OT plans when both random supports are absolutely continuous is proved in Klatt et al. (2022, Lemma 6.5 and Proposition 6.6). The same genericity argument gives the one-sided version needed here. Suppose that μ is absolutely continuous and condition on $\mathbf{y}_k = (y_1, \ldots, y_k)$. For two permutations $\sigma, \tau \in \mathfrak{S}_k$, a tie in quadratic cost is equivalent to

$$C_\sigma(\mathbf{x}_k,\, \mathbf{y}_k) - C_\tau(\mathbf{x}_k,\, \mathbf{y}_k) = -\frac{2}{k}\sum_{i=1}^k \langle x_i,\, y_{\sigma(i)} - y_{\tau(i)} \rangle = 0.$$

If the two permutations induce distinct empirical plans, this is a nontrivial hyperplane in $(\mathbb{R}^d)^k$. It therefore has $\mu^{\otimes k}$-measure zero. A finite union over pairs $(\sigma,\, \tau)$ shows that, almost surely, all optimal permutations induce the same empirical plan. Since every coupling between the labelled uniform empirical measures is a convex combination of permutation plans, the empirical OT plan is almost surely unique as a probability measure. The case where ν is absolutely continuous is symmetric.

A.1.2 Well-posedness of the expected batch OT plan

We can now define the expected batch OT plan using the measurable OT solver introduced above.

Justification for Definition 3.1. The expected batch OT plan is defined by averaging the solver outputs $\widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k}$ over random batches. We show that this is well-posed. For $z = (\mathbf{x}_k,\,\mathbf{y}_k) \in E_k$, let $\eta_z = \widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k} \in \mathcal{P}(F)$. We first show that $(\mathbf{x}_k,\,\mathbf{y}_k) \mapsto \widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k}$ is a Borel family of probability measures. For Borel $A \subseteq F$,

$$\widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k}(A) = \frac{1}{k}\sum_{i=1}^{k}\sum_{\sigma\in\mathfrak{S}_k} \mathbb{1}_A\big(x_i,\,y_{\sigma(i)}\big)\,\mathbb{1}_{\{\sigma=\widehat{\sigma}_{\mathbf{x}_k,\,\mathbf{y}_k}\}}.$$

Since $\widehat{\sigma}_k$ is Borel by assumption, the right-hand side is a Borel function of $(\mathbf{x}_k,\,\mathbf{y}_k)$. Therefore, by (Ambrosio et al., 2005, Section 5.3), there exists a unique probability measure $\widetilde{\eta}_k \in \mathcal{P}(E_k \times F)$ such that, for every bounded Borel map $f : E_k \times F \to \mathbb{R}$,

$$\int_{E_k\times F} f(z,\,z')\,\mathrm{d}\widetilde{\eta}_k(z,\,z') = \int_{E_k}\left(\int_F f(z,\,z')\,\mathrm{d}\eta_z(z')\right)\mathrm{d}(\mu^{\otimes k}\otimes\nu^{\otimes k})(z).$$

Let $\mathrm{pr}^F : E_k \times F \to F$ be the projection onto the second factor, and define $\overline{\pi}_k = \mathrm{pr}^F_\sharp\,\widetilde{\eta}_k$. For bounded Borel $g : F \to \mathbb{R}$, applying the change-of-variables formula for push-forward measures and then the identity above, we obtain

$$\begin{aligned}
\int_F g(z')\,\mathrm{d}\overline{\pi}_k(z') &= \int_{E_k\times F} g(z')\,\mathrm{d}\widetilde{\eta}_k(z,\,z') \\
&= \int_{E_k}\left(\int_F g(z')\,\mathrm{d}\eta_z(z')\right)\mathrm{d}(\mu^{\otimes k}\otimes\nu^{\otimes k})(z) \\
&= \int_{(\mathbb{R}^d)^k\times(\mathbb{R}^d)^k}\left(\int_{\mathbb{R}^d\times\mathbb{R}^d} g\,\mathrm{d}\widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k}\right)\mathrm{d}(\mu^{\otimes k}\otimes\nu^{\otimes k})(\mathbf{x}_k,\,\mathbf{y}_k).
\end{aligned}$$

We now show that $\overline{\pi}_k$ is the law of a random pair produced by the minibatch OT procedure. If we draw $(\mathbf{X}_k,\,\mathbf{Y}_k) \sim \mu^{\otimes k}\otimes\nu^{\otimes k}$ and let $I \sim \mathrm{Unif}\{1,\dots,k\}$ be independent of $(\mathbf{X}_k,\,\mathbf{Y}_k)$, then $(X_I,\,Y_{\widehat{\sigma}_{\mathbf{X}_k,\,\mathbf{Y}_k}(I)}) \sim \overline{\pi}_k$. Indeed, for every bounded Borel function $g : \mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}$, conditioning on I gives

$$\begin{aligned}
\mathbb{E}[g(X_I,\,Y_{\widehat{\sigma}_{\mathbf{X}_k,\,\mathbf{Y}_k}(I)})] &= \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} g\left(X_i,\,Y_{\widehat{\sigma}_{\mathbf{X}_k,\,\mathbf{Y}_k}(i)}\right)\right] \\
&= \mathbb{E}\left[\int_{\mathbb{R}^d\times\mathbb{R}^d} g(x,\,y)\,\mathrm{d}\widehat{\pi}_{\mathbf{X}_k,\,\mathbf{Y}_k}(x,\,y)\right] \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} g(x,\,y)\,\mathrm{d}\overline{\pi}_k(x,\,y).
\end{aligned}$$

□

A.1.3 Elementary properties of the expected batch OT plan

We show that $\overline{\pi}_k$ is an admissible coupling between μ and ν (Pooladian et al., 2023a, Lemma 4.1). Indeed, let $f : \mathbb{R}^d \to \mathbb{R}$ be bounded and Borel, and let $\mathrm{pr}^1 : \mathbb{R}^d\times\mathbb{R}^d \to \mathbb{R}^d$ be defined by

$\mathrm{pr}^1(x,\, y) = x$. Then

$$\begin{aligned}
\int_{\mathbb{R}^d} f(x)\,\mathrm{d}(\mathrm{pr}^1_\sharp\,\overline{\pi}_k)(x) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} f(x)\,\mathrm{d}\overline{\pi}_k(x,\, y) \\
&= \int_{(\mathbb{R}^d)^k\times(\mathbb{R}^d)^k} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} f(x)\,\mathrm{d}\widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k}(x,\, y)\right)\,\mathrm{d}\alpha_k(\mathbf{x}_k,\, \mathbf{y}_k) \\
&= \int_{(\mathbb{R}^d)^k\times(\mathbb{R}^d)^k} \left(\int_{\mathbb{R}^d} f(x)\,\mathrm{d}\widehat{\mu}_{\mathbf{x}_k}(x)\right)\,\mathrm{d}\alpha_k(\mathbf{x}_k,\, \mathbf{y}_k) \qquad (5) \\
&= \int_{(\mathbb{R}^d)^k} \frac{1}{k}\sum_{i=1}^k f(x_i)\,\mathrm{d}\mu^{\otimes k}(\mathbf{x}_k) \\
&= \int_{\mathbb{R}^d} f(x)\,\mathrm{d}\mu(x).
\end{aligned}$$

Equation (5) follows from the fact that $\widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k}$ is an OT plan between $\widehat{\mu}_{\mathbf{x}_k}$ and $\widehat{\nu}_{\mathbf{y}_k}$, thus its first marginal is $\widehat{\mu}_{\mathbf{x}_k}$. We conclude that $\mathrm{pr}^1_\sharp\,\overline{\pi}_k = \mu$. The same argument can be adapted to show that $\mathrm{pr}^2_\sharp\,\overline{\pi}_k = \nu$, with $\mathrm{pr}^2(x,\, y) = y$. Therefore, $\overline{\pi}_k \in \Pi(\mu,\, \nu)$.

We now show that $\overline{\pi}_1 = \mu\otimes\nu$. By definition, $\widehat{\pi}_{\mathbf{x}_1,\,\mathbf{y}_1} = \delta_{(x_1,\,y_1)}$. For every Borel set $A \subseteq \mathbb{R}^d\times\mathbb{R}^d$, using $g = \mathbb{1}_A$ in Definition 3.1,

$$\begin{aligned}
\overline{\pi}_1(A) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathbb{1}_A(x,\, y)\,\mathrm{d}\overline{\pi}_1(x,\, y) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \mathbb{1}_A\,\mathrm{d}\widehat{\pi}_{\mathbf{x}_1,\,\mathbf{y}_1}\right)\,\mathrm{d}(\mu\otimes\nu)(\mathbf{x}_1,\, \mathbf{y}_1) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \widehat{\pi}_{\mathbf{x}_1,\,\mathbf{y}_1}(A)\,\mathrm{d}(\mu\otimes\nu)(x_1,\, y_1) \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \mathbb{1}_A(x_1,\, y_1)\,\mathrm{d}(\mu\otimes\nu)(x_1,\, y_1) \\
&= (\mu\otimes\nu)(A).
\end{aligned}$$

Therefore, $\overline{\pi}_1 = \mu\otimes\nu$.

A.2 Properties of the expected batch OT cost (Proposition 3.2 and Proposition 3.3(1))

We prove the equivalent formulation of the expected batch OT cost (Proposition 3.2). We then establish that this cost is nonincreasing in k and converges to the population OT cost (Proposition 3.3(1)).

Proof of Proposition 3.2. Let $k \in \mathbb{N}^*$. For $n \in \mathbb{N}^*$, define the truncated cost function $g_n : (x, y) \mapsto \min(\|x-y\|^2,\, n)$ for $(x, y) \in \mathbb{R}^d\times\mathbb{R}^d$. Applying Definition 3.1 with g_n gives,

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} g_n\,\mathrm{d}\overline{\pi}_k = \int_{(\mathbb{R}^d)^k\times(\mathbb{R}^d)^k} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} g_n\,\mathrm{d}\widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k}\right)\,\mathrm{d}(\mu^{\otimes k}\otimes\nu^{\otimes k})(\mathbf{x}_k,\, \mathbf{y}_k). \qquad (6)$$

Since $g_{n+1} \geqslant g_n \geqslant 0$ and $\lim_{n\to+\infty} g_n(x,\, y) = \|x-y\|^2$, by the monotone convergence theorem,

$$\lim_{n\to+\infty} \int_{\mathbb{R}^d\times\mathbb{R}^d} g_n\,\mathrm{d}\overline{\pi}_k = \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2\,\mathrm{d}\overline{\pi}_k. \qquad (7)$$

Moreover, for any $(\mathbf{x}_k,\, \mathbf{y}_k) \in (\mathbb{R}^d)^k\times(\mathbb{R}^d)^k$,

$$\begin{aligned}
\lim_{n\to+\infty} \int_{\mathbb{R}^d\times\mathbb{R}^d} g_n\,\mathrm{d}\widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k} &= \lim_{n\to+\infty} \frac{1}{k}\sum_{i=1}^k g_n(x_i,\, y_{\widehat{\sigma}_{\mathbf{x}_k,\,\mathbf{y}_k}(i)}) \\
&= \frac{1}{k}\sum_{i=1}^k \|x_i - y_{\widehat{\sigma}_{\mathbf{x}_k,\,\mathbf{y}_k}(i)}\|^2 = W_2^2(\widehat{\mu}_{\mathbf{x}_k},\, \widehat{\nu}_{\mathbf{y}_k}).
\end{aligned}$$

Since $(\int_{\mathbb{R}^d\times\mathbb{R}^d} g_n \,\mathrm{d}\widehat{\pi}_{\mathbf{x}_k,\,\mathbf{y}_k})_{n\in\mathbb{N}^*}$ is nondecreasing, we apply the monotone convergence theorem once more on the right-hand side of (6). Combining with (7), we then obtain

$$\begin{aligned}\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \,\mathrm{d}\overline{\pi}_k(x,\,y) &= \int_{(\mathbb{R}^d)^k\times(\mathbb{R}^d)^k} W_2^2(\widehat{\mu}_{\mathbf{x}_k},\,\widehat{\nu}_{\mathbf{y}_k})\,\mathrm{d}(\mu^{\otimes k}\otimes\nu^{\otimes k})(\mathbf{x}_k,\,\mathbf{y}_k)\\ &= \mathbb{E}_{(\mathbf{X}_k,\,\mathbf{Y}_k)\sim\mu^{\otimes k}\otimes\nu^{\otimes k}}[W_2^2(\widehat{\mu}_k,\,\widehat{\nu}_k)].\end{aligned}$$

□

Proof of Proposition 3.3(1). Monotonicity. We show that $\mathbb{E}[W_2^2(\widehat{\mu}_{k+1},\,\widehat{\nu}_{k+1})] \leqslant \mathbb{E}[W_2^2(\widehat{\mu}_k,\,\widehat{\nu}_k)]$ for any $k\in\mathbb{N}^*$. This result was first stated in Pooladian et al. (2023a, Theorem D.8) with an incomplete proof.

Let independent batches $\mathbf{X}_{k+1}=(X_1,\dots,X_{k+1})\sim\mu^{\otimes k+1}$ and $\mathbf{Y}_{k+1}=(Y_1,\dots,Y_{k+1})\sim\nu^{\otimes k+1}$. For each $s\in\{1,\dots,k+1\}$, define the leave-one-out batches $\mathbf{X}_{-s}=(X_i)_{i\neq s}$ and $\mathbf{Y}_{-s}=(Y_j)_{j\neq s}$. Denote the corresponding empirical measures by $\widehat{\mu}_{-s}=\widehat{\mu}_{\mathbf{X}_{-s}}$ and $\widehat{\nu}_{-s}=\widehat{\nu}_{\mathbf{Y}_{-s}}$, and let π_{-s} be an OT plan between $\widehat{\mu}_{-s}$ and $\widehat{\nu}_{-s}$. As $\overline{\pi}=\frac{1}{k+1}\sum_{s=1}^{k+1}\pi_{-s}$ averages batch OT plans between discrete measures $\widehat{\mu}_{k+1}$ and $\widehat{\nu}_{k+1}$, we have $\overline{\pi}\in\Pi(\widehat{\mu}_{k+1},\,\widehat{\nu}_{k+1})$ by Fatras et al. (2021b, Lemma 9), but we detail the first marginal for completeness: by linearity of the push-forward operator,

$$\mathrm{pr}^1_\sharp\,\overline{\pi}=\frac{1}{k+1}\sum_{s=1}^{k+1}\mathrm{pr}^1_\sharp\,\pi_{-s}=\frac{1}{k+1}\sum_{s=1}^{k+1}\widehat{\mu}_{-s}=\widehat{\mu}_{k+1}.$$

Therefore, by definition of the Wasserstein distance,

$$\begin{aligned}W_2^2(\widehat{\mu}_{k+1},\,\widehat{\nu}_{k+1}) &\leqslant \int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^2\,\mathrm{d}\overline{\pi}(x,y)\\ &= \frac{1}{k+1}\sum_{s=1}^{k+1}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^2\,\mathrm{d}\pi_{-s}(x,y)\\ &= \frac{1}{k+1}\sum_{s=1}^{k+1}W_2^2(\widehat{\mu}_{-s},\,\widehat{\nu}_{-s}).\end{aligned}\tag{8}$$

Finally, for each s, $(\mathbf{X}_{-s},\,\mathbf{Y}_{-s})\sim\mu^{\otimes k}\otimes\nu^{\otimes k}$, hence $\mathbb{E}[W_2^2(\widehat{\mu}_{-s},\,\widehat{\nu}_{-s})]=\mathbb{E}[W_2^2(\widehat{\mu}_k,\,\widehat{\nu}_k)]$, and taking expectations in (8) gives $\mathbb{E}[W_2^2(\widehat{\mu}_{k+1},\,\widehat{\nu}_{k+1})]\leqslant\mathbb{E}[W_2^2(\widehat{\mu}_k,\,\widehat{\nu}_k)]$.

Consistency. Let $\mu,\,\nu\in\mathcal{P}_q(\mathbb{R}^d)$ with $q>2$. By Appendix A.1.3 and the definition of the Wasserstein distance, $\mathbb{E}[W_2^2(\widehat{\mu}_k,\,\widehat{\nu}_k)]\geqslant W_2^2(\mu,\,\nu)$. On the other hand, by the triangle inequality,

$$W_2(\widehat{\mu}_k,\,\widehat{\nu}_k)\leqslant W_2(\widehat{\mu}_k,\,\mu)+W_2(\mu,\,\nu)+W_2(\widehat{\nu}_k,\,\nu).$$

Therefore, denoting $a_k=W_2(\widehat{\mu}_k,\,\mu)+W_2(\widehat{\nu}_k,\,\nu)$,

$$W_2^2(\widehat{\mu}_k,\,\widehat{\nu}_k)-W_2^2(\mu,\,\nu)\leqslant 2W_2(\mu,\,\nu)a_k+a_k^2.$$

By admissibility of $\overline{\pi}_k$ in Appendix A.1.3 and applying Proposition 3.2, $W_2^2(\mu,\,\nu)\leqslant\mathbb{E}[W_2^2(\widehat{\mu}_k,\,\widehat{\nu}_k)]$, thus

$$0\leqslant\mathbb{E}[W_2^2(\widehat{\mu}_k,\,\widehat{\nu}_k)]-W_2^2(\mu,\,\nu)\leqslant 2W_2(\mu,\,\nu)\mathbb{E}[a_k]+\mathbb{E}[a_k^2].$$

Now, by the Cauchy-Schwarz inequality, $\mathbb{E}[a_k]\leqslant\mathbb{E}[W_2^2(\widehat{\mu}_k,\,\mu)]^{1/2}+\mathbb{E}[W_2^2(\widehat{\nu}_k,\,\nu)]^{1/2}$. Moreover, as $(a+b)^2\leqslant 2(a^2+b^2)$, we have $\mathbb{E}[a_k^2]\leqslant 2\mathbb{E}[W_2^2(\widehat{\mu}_k,\,\mu)]+2\mathbb{E}[W_2^2(\widehat{\nu}_k,\,\nu)]$. Since $\mu,\,\nu\in\mathcal{P}_q(\mathbb{R}^d)$, by Fournier and Guillin (2015, Theorem 1), both terms on the right-hand side converge to zero as $k\to+\infty$. We conclude that $\lim_{k\to+\infty}\mathbb{E}[W_2^2(\widehat{\mu}_k,\,\widehat{\nu}_k)]=W_2^2(\mu,\,\nu)$. □

A.3 Convergence rates for the expected batch OT cost (Proposition 3.3(2))

We now prove the convergence rates in Proposition 3.3(2). We first recall the dual formulation of semidiscrete OT in our setting (Peyré and Cuturi, 2019, Section 5).

Assume that $\nu=\sum_{j=1}^M q_j\delta_{v_j}$, where $q_j>0$ and the atoms $v_1,\dots,v_M\in\mathbb{R}^d$ are pairwise distinct. If $M=1$, the claim is immediate, hence we assume $M\geqslant 2$. Let $(X_1,\dots,X_k)$ and $(Y_1,\dots,Y_k)$ be

independent batches with joint law $\mu^{\otimes k} \otimes \nu^{\otimes k}$. Denote the corresponding empirical measures by $\widehat{\mu}_k = \frac{1}{k}\sum_{i=1}^k \delta_{X_i}$ and $\widehat{\nu}_k = \frac{1}{k}\sum_{i=1}^k \delta_{Y_i}$. Writing $Y_i = v_{J_i}$, the indices $J_1, \dots, J_k$ are independent with $\mathbb{P}(J_i = j) = q_j$, and we can write

$$\widehat{\nu}_k = \sum_{j=1}^M \widehat{q}_{k,j}\delta_{v_j} \quad \text{with} \quad \widehat{q}_{k,j} = \frac{1}{k}\sum_{i=1}^k \mathbb{1}_{\{J_i=j\}}.$$

For every $\lambda = (\lambda_1, \dots, \lambda_M) \in \mathbb{R}^M$, let $\psi_\lambda(x) = \min_{1\leqslant j\leqslant M}\{\|x - v_j\|^2 - \lambda_j\}$, and define

$$H(\lambda) = \int_{\mathbb{R}^d} \psi_\lambda(x)\,\mathrm{d}\mu(x) + \sum_{j=1}^M q_j\lambda_j \quad \text{and} \quad H_k(\lambda) = \int_{\mathbb{R}^d} \psi_\lambda(x)\,\mathrm{d}\widehat{\mu}_k(x) + \sum_{j=1}^M \widehat{q}_{k,j}\lambda_j.$$

Since $H(\lambda + t\mathbf{1}_M) = H(\lambda)$ for any $t \in \mathbb{R}$, where $\mathbf{1}_M = (1, \dots, 1) \in \mathbb{R}^M$, we work on $\Lambda = (\mathbb{R}\,\mathbf{1}_M)^\perp$ to remove such invariance. Then, by the dual formulation of semidiscrete OT (Peyré and Cuturi, 2019, Section 5),

$$W_2^2(\mu,\, \nu) = \sup_{\lambda\in\Lambda} H(\lambda) \quad \text{and} \quad W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k) = \sup_{\lambda\in\Lambda} H_k(\lambda).$$

Proposition A.1. *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\nu = \sum_{i=1}^M q_i\delta_{v_i}$ with $q_i > 0$ and pairwise distinct $v_i \in \mathbb{R}^d$. If Assumption A or B in Genans et al. (2025) holds, then $\mathbb{E}[W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k)] - W_2^2(\mu,\, \nu) = O(k^{-1})$.*

The above Assumption B holds for $\mu = \mathcal{N}(0,\, I_d)$ by Genans et al. (2025, §4), establishing Proposition 3.3(2).

Under Assumption A or B, by Genans et al. (2025, Proposition 4.1), the restriction of H to Λ admits a unique maximizer $\lambda^\star$, H is $\mathcal{C}^2$ in a neighborhood of $\lambda^\star$, and $D^2H(\lambda^\star)$ is negative definite on Λ. We also show that $\lambda \mapsto \psi_\lambda(x)$ is 1-Lipschitz continuous for any $x \in \mathbb{R}^d$.

Lemma A.2. *For any $x \in \mathbb{R}^d$ and $\lambda,\, \lambda' \in \mathbb{R}^M$, $|(\psi_\lambda - \psi_{\lambda'})(x)| \leqslant \|\lambda - \lambda'\|_\infty$.*

Proof. Let $x \in \mathbb{R}^d$ and $\lambda,\, \lambda' \in \mathbb{R}^M$. For any $\ell \in \{1, \dots, M\}$,

$$\psi_\lambda(x) \leqslant \|x - v_\ell\|^2 - \lambda_\ell = \|x - v_\ell\|^2 - \lambda'_\ell + (\lambda'_\ell - \lambda_\ell) \leqslant \|x - v_\ell\|^2 - \lambda'_\ell + \|\lambda - \lambda'\|_\infty.$$

Taking the minimum over ℓ gives $\psi_\lambda(x) - \psi_{\lambda'}(x) \leqslant \|\lambda - \lambda'\|_\infty$. Additionally, by exchanging λ and λ', we obtain $\psi_{\lambda'}(x) - \psi_\lambda(x) \leqslant \|\lambda' - \lambda\|_\infty$, hence the claim. □

To prove Proposition A.1, we reduce the cost bias to a quantity defined on a deterministic compact $K \subseteq \Lambda$.

Lemma A.3. *Under the assumptions of Proposition A.1, there exists a compact $K \subseteq \Lambda$ containing $\lambda^\star$ with $\mathbb{E}[W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k)] - W_2^2(\mu,\, \nu) = \mathbb{E}[Z_k] + o(k^{-1})$, where $Z_k = \sup_{\lambda\in K}\{H_k(\lambda) - H_k(\lambda^\star)\}$.*

Proof. Adapting the proofs of Genans et al. (2025, Lemma 3.1) and Pooladian et al. (2023b, Lemma E.2), we first show that, on an event $\mathcal{G}_k$ of high probability, the empirical semi-dual H_k admits a maximizer in a deterministic compact $K \subseteq \Lambda$. By tightness of μ, choose a compact $K_0 \subseteq \mathbb{R}^d$ such that $\mu(K_0) \geqslant 1 - \frac{1}{8}\min_{1\leqslant j\leqslant M} q_j$, and define the event $\mathcal{G}_k = \mathcal{G}_k^s \cap \mathcal{G}_k^t$, where

$$\mathcal{G}_k^s = \left\{\widehat{\mu}_k(K_0) \geqslant 1 - \frac{1}{4}\min_{1\leqslant j\leqslant M} q_j\right\} \quad \text{and} \quad \mathcal{G}_k^t = \left\{\min_{1\leqslant j\leqslant M} \widehat{q}_{k,\,j} \geqslant \frac{1}{2}\min_{1\leqslant j\leqslant M} q_j\right\}.$$

By Hoeffding's inequality and the union bound, there exist $c,\, C > 0$ such that $\mathbb{P}(\mathcal{G}_k^c) \leqslant Ce^{-ck}$. $(\star)$

For $\lambda \in \mathbb{R}^M$, denote the Laguerre cells $C_j(\lambda) = \{x \in \mathbb{R}^d,\, \psi_\lambda(x) = \|x - v_j\|^2 - \lambda_j\}$. By Peyré and Cuturi (2019, Proposition 2.4), H_k admits a maximizer $\lambda \in \mathbb{R}^M$. Let $\pi = (\pi_{ij})$ be an OT plan between $\widehat{\mu}_k$ and $\widehat{\nu}_k$. The primal-dual optimality condition (Peyré and Cuturi, 2019, Eq. 2.23) implies that if $\pi_{ij} > 0$, then $\psi_\lambda(X_i) + \lambda_j = \|X_i - v_j\|^2$, hence $X_i \in C_j(\lambda)$, and

$$\widehat{q}_{k,\,j} = \sum_{i=1}^k \pi_{ij} = \sum_{i=1}^k \pi_{ij}\,\mathbb{1}_{X_i\in C_j(\lambda)} \leqslant \sum_{i=1}^k \frac{1}{k}\,\mathbb{1}_{X_i\in C_j(\lambda)} = \widehat{\mu}_k(C_j(\lambda)).$$

Applying the union bound to complements,

$\widehat{\mu}_k(C_j(\lambda)\cap K_0) = 1-\widehat{\mu}_k(C_j(\lambda)^c\cup K_0^c) \geqslant 1-(\widehat{\mu}_k(C_j(\lambda)^c)+\widehat{\mu}_k(K_0^c)) = \widehat{\mu}_k(C_j(\lambda))+\widehat{\mu}_k(K_0)-1.$

On $\mathcal{G}_k$, $\widehat{\mu}_k(K_0) \geqslant 1 - \frac{1}{4}\min_j q_j \geqslant 1 - \frac{1}{2}\min_i \widehat{q}_{k,\,i}$. Therefore, we can choose $x_j \in C_j(\lambda) \cap K_0$ for every j, as

$$\widehat{\mu}_k(C_j(\lambda) \cap K_0) \geqslant \widehat{q}_{k,\,j} - \frac{1}{2}\min_{1\leqslant i\leqslant M} \widehat{q}_{k,\,i} \geqslant \frac{1}{2}\min_{1\leqslant i\leqslant M} \widehat{q}_{k,\,i} > 0.$$

Set $C_0 = \max_i \sup_{x\in K_0}\|x - v_i\|^2$. As $x_j \in C_j(\lambda)$, for every i, j, $\lambda_i - \lambda_j \leqslant \|x_j - v_i\|^2 - \|x_j - v_j\|^2 \leqslant C_0$, and taking the maximum over i, j, we get $\max_i \lambda_i - \min_j \lambda_j \leqslant C_0$. Since H_k is invariant along $\mathbb{R}\mathbf{1}_M$, we can assume that $\lambda \in \Lambda$, so that $\min_j \lambda_j \leqslant 0 \leqslant \max_i \lambda_i$, thus $\|\lambda\|_\infty \leqslant C_0$. Therefore, on $\mathcal{G}_k$, H_k admits a maximizer $\widehat{\lambda}_k$ in the compact $K = \{\lambda \in \Lambda,\ \|\lambda\|_\infty \leqslant C_0 + \|\lambda^\star\|_\infty + 1\}$ containing $\lambda^\star$, and defining Z_k as in the statement,

$$(W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k) - W_2^2(\mu,\, \nu))\mathbb{1}_{\mathcal{G}_k} = (Z_k + (H_k - H)(\lambda^\star))\mathbb{1}_{\mathcal{G}_k}.$$

Since $\mathbb{E}[(H_k - H)(\lambda^\star)] = 0$, we have $\mathbb{E}[(H_k - H)(\lambda^\star)\mathbb{1}_{\mathcal{G}_k}] = -\mathbb{E}[(H_k - H)(\lambda^\star)\mathbb{1}_{\mathcal{G}_k^c}]$, thus

$$\begin{aligned}
&\mathbb{E}[W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k)] - W_2^2(\mu,\, \nu)\\
&= \mathbb{E}[(W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k) - W_2^2(\mu,\, \nu))(\mathbb{1}_{\mathcal{G}_k} + \mathbb{1}_{\mathcal{G}_k^c})]\\
&= \mathbb{E}[Z_k\mathbb{1}_{\mathcal{G}_k}] + \mathbb{E}[(H_k - H)(\lambda^\star)\mathbb{1}_{\mathcal{G}_k}] + \mathbb{E}[(W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k) - W_2^2(\mu,\, \nu))\mathbb{1}_{\mathcal{G}_k^c}]\\
&= \mathbb{E}[Z_k] - \underbrace{\mathbb{E}[Z_k\mathbb{1}_{\mathcal{G}_k^c}]}_{\text{(i)}} - \underbrace{\mathbb{E}[(H_k - H)(\lambda^\star)\mathbb{1}_{\mathcal{G}_k^c}]}_{\text{(ii)}} + \underbrace{\mathbb{E}[W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k)\mathbb{1}_{\mathcal{G}_k^c}]}_{\text{(iii)}} - \underbrace{W_2^2(\mu,\, \nu)\mathbb{P}(\mathcal{G}_k^c)}_{\text{(iv)}}.
\end{aligned}$$

We now show that the four last terms are $o(k^{-1})$.

(i) For each $\lambda \in K$, applying the triangle inequality and Lemma A.2,

$$|H_k(\lambda) - H_k(\lambda^\star)| \leqslant \frac{1}{k}\sum_{i=1}^k |(\psi_\lambda - \psi_{\lambda^\star})(X_i)| + \sum_{j=1}^M \widehat{q}_{k,j}|\lambda_j - \lambda_j^\star| \leqslant 2\|\lambda - \lambda^\star\|_\infty.$$

Therefore $0 \leqslant Z_k \leqslant 2\sup_{\lambda\in K}\|\lambda - \lambda^\star\|_\infty < \infty$, and $\mathbb{E}[Z_k\mathbb{1}_{\mathcal{G}_k^c}] = O(\mathbb{P}(\mathcal{G}_k^c)) = o(k^{-1})$.

(ii) Since $|\psi_{\lambda^\star}(x)| \leqslant C(1 + \|x\|^2)$ and $(H_k - H)(\lambda^\star) = \int_{\mathbb{R}^d} \psi_{\lambda^\star}\,\mathrm{d}(\widehat{\mu}_k - \mu) + \sum_{j=1}^M (\widehat{q}_{k,j} - q_j)\lambda_j^\star$,

$$|H_k - H|(\lambda^\star) \leqslant \frac{1}{k}\sum_{\ell=1}^k |\psi_{\lambda^\star}(X_\ell)| + \int_{\mathbb{R}^d} |\psi_{\lambda^\star}(x)|\,\mathrm{d}\mu(x) + C \leqslant C\left(1 + \frac{1}{k}\sum_{\ell=1}^k \|X_\ell\|^2\right).$$

Recall that $\mathcal{G}_k = \mathcal{G}_k^s \cap \mathcal{G}_k^t$. As $\mathcal{G}_k^t$ is independent of $X_1, \ldots, X_k$, we have

$$\mathbb{E}\left[\left(1 + \frac{1}{k}\sum_{\ell=1}^k \|X_\ell\|^2\right)\mathbb{1}_{(\mathcal{G}_k^t)^c}\right] = \mathbb{E}_{X\sim\mu}[1 + \|X\|^2]\mathbb{P}((\mathcal{G}_k^t)^c) = o(k^{-1}).$$

Moreover, $\mathcal{G}_k^s = \{\frac{1}{k}\sum_{\ell=1}^k \mathbb{1}_{X_\ell\notin K_0} \leqslant \frac{1}{4}\min_j q_j\}$, and $\mathbb{E}[\mathbb{1}_{X_\ell\notin K_0}] \leqslant \frac{1}{8}\min_j q_j$ by choice of K_0. Set $B_m = \mathbb{1}_{\{X_m\notin K_0\}}$. Chernoff's inequality applied conditionally to the variables $(B_m)_{m\neq\ell}$ gives, uniformly in X_ℓ, $\mathbb{P}((\mathcal{G}_k^s)^c \mid X_\ell) \leqslant Ce^{-ck}$. Therefore

$$\begin{aligned}
\mathbb{E}\left[\left(1 + \frac{1}{k}\sum_{\ell=1}^k \|X_\ell\|^2\right)\mathbb{1}_{(\mathcal{G}_k^s)^c}\right] &= \frac{1}{k}\sum_{\ell=1}^k \mathbb{E}[(1 + \|X_\ell\|^2)\mathbb{P}((\mathcal{G}_k^s)^c \mid X_\ell)]\\
&\leqslant Ce^{-ck}\mathbb{E}_{X\sim\mu}[1 + \|X\|^2] = o(k^{-1}).
\end{aligned}$$

Therefore $\mathbb{E}[|H_k - H|(\lambda^\star)\mathbb{1}_{\mathcal{G}_k^c}] = o(k^{-1})$.

(iii) Using the coupling $\pi = \frac{1}{k}\sum_{\ell=1}^k \delta_{(X_\ell,\, Y_\ell)}$, we get that $W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k) \leqslant \frac{1}{k}\sum_{\ell=1}^k \|X_\ell - Y_\ell\|^2$, where $\|X_\ell - Y_\ell\|^2 \leqslant 2\|X_\ell\|^2 + 2\max_j\|v_j\|^2$. By the same calculation as above, $\mathbb{E}[W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k)\mathbb{1}_{\mathcal{G}_k^c}] = o(k^{-1})$.

(iv) By the Hoeffding bound $(\star)$, $W_2^2(\mu,\,\nu)\mathbb{P}(\mathcal{G}_k^c) = o(k^{-1})$.

□

We then control $\mathbb{E}[Z_k]$ by combining a peeling argument with bounds on suprema of localized empirical processes, of the type developed in van der Vaart and Wellner (1996) or Boucheron and Massart (2011, Theorem 7). We can conveniently rewrite H_k and H as single empirical averages. Let $P = \mu \otimes \sum_{j=1}^M q_j \delta_j$ on $\mathcal{Z} = \mathbb{R}^d \times \{1, \dots, M\}$. Let $\zeta_i = (X_i,\, J_i) \sim P$ and $P_k = \frac{1}{k}\sum_{i=1}^k \delta_{\zeta_i}$. For $\lambda \in \mathbb{R}^M$, define $h_\lambda(x,\, j) = \psi_\lambda(x) + \lambda_j$ and $g_\lambda = h_\lambda - h_{\lambda^\star}$. Then $g_{\lambda^\star} = 0$, and

$$H(\lambda) = \int_{\mathcal{Z}} h_\lambda \,\mathrm{d}P, \qquad H_k(\lambda) = \int_{\mathcal{Z}} h_\lambda \,\mathrm{d}P_k,$$

$$H(\lambda) - H(\lambda^\star) = \int_{\mathcal{Z}} g_\lambda \,\mathrm{d}P, \qquad H_k(\lambda) - H_k(\lambda^\star) = \int_{\mathcal{Z}} g_\lambda \,\mathrm{d}P_k.$$

Therefore, we can decompose the quantity defining Z_k as

$$H_k(\lambda) - H_k(\lambda^\star) = H(\lambda) - H(\lambda^\star) + \int_{\mathcal{Z}} g_\lambda \,\mathrm{d}(P_k - P). \tag{9}$$

To control the first term in (9), we use the fact that, by Genans et al. (2025, Lemma 4.2 and Remark 4.3), under the assumptions of Proposition A.1, there exists $\kappa > 0$ such that, for every $\lambda \in K$,

$$H(\lambda) - H(\lambda^\star) \leqslant -\kappa \|\lambda - \lambda^\star\|_\infty^2. \tag{10}$$

To control the second term in (9), we bound $\int_{\mathcal{Z}} g_\lambda \,\mathrm{d}(P_k - P)$ locally. For $a \geqslant 0$, define

$$S_k(a) = \sup_{\substack{\lambda \in \Lambda \\ \|\lambda - \lambda^\star\|_\infty \leqslant a}} \left| \int_{\mathcal{Z}} g_\lambda \,\mathrm{d}(P_k - P) \right|.$$

Lemma A.4. *There exist $C > 0$, depending only on M, and $c > 0$ such that, for every $a,\, t > 0$,*

$$\forall k \in \mathbb{N}^*, \quad \mathbb{E}[S_k(a)] \leqslant C\frac{a}{\sqrt{k}} \quad \textit{and} \quad \mathbb{P}\left(S_k(a) - \mathbb{E}[S_k(a)] \geqslant t\right) \leqslant \exp\left(-c\,\frac{kt^2}{a^2}\right).$$

Proof. By definition of g_λ, for any $z = (x, j) \in \mathcal{Z}$, $g_\lambda(x,\, j) - g_{\lambda'}(x,\, j) = \psi_\lambda(x) - \psi_{\lambda'}(x) + \lambda_j - \lambda'_j$. By Lemma A.2, $|g_\lambda(x,\, j) - g_{\lambda'}(x,\, j)| \leqslant 2\|\lambda - \lambda'\|_\infty$. Hence, choosing $\lambda' = \lambda^\star$ and taking the sup over $z = (x, j) \in \mathcal{Z}$ yields $\|g_\lambda\|_\infty \leqslant 2\|\lambda - \lambda^\star\|_\infty$. In particular, if $\|\lambda - \lambda^\star\|_\infty \leqslant a$ for $a > 0$, then $\|g_\lambda\|_\infty \leqslant 2a$.

For the first inequality, let $\mathscr{G}_a = \{g_\lambda : \lambda \in \Lambda,\ \|\lambda - \lambda^\star\|_\infty \leqslant a\}$. The bound above gives that the class of functions $\mathscr{G}_a$ is uniformly bounded by the envelope $z \mapsto 2a$. By van der Vaart and Wellner (1996, §2.14) (see also Pooladian et al. (2023b, Lemma H.3)),

$$\mathbb{E}[S_k(a)] \leqslant \frac{C_1}{\sqrt{k}} \int_0^{C_2 a} \sqrt{\log(2N(r, \mathscr{G}_a, \|\cdot\|_\infty))}\,\mathrm{d}r,$$

where $C_1,\, C_2 > 0$ and $N(r,\, \mathscr{G}_a,\, \|\cdot\|_\infty)$ denotes the r-covering number in $\mathscr{G}_a$ with respect to the sup norm. Since $\lambda \mapsto g_\lambda$ is 2-Lipschitz uniformly in z by the bound above, for every $r > 0$,

$$N(r, \mathscr{G}_a, \|\cdot\|_\infty) \leqslant N(r/2,\, B_\infty(\lambda^\star,\, a), \|\cdot\|_\infty).$$

Since $N(r/2,\, B_\infty(\lambda^\star,\, a), \|\cdot\|_\infty) \leqslant (1 + 4a/r)^M$,

$$\begin{aligned}
\mathbb{E}[S_k(a)] &\leqslant \frac{C_1}{\sqrt{k}} \int_0^{C_2 a} \sqrt{\log(2(1 + 4a/r)^M)}\,\mathrm{d}r \\
&\leqslant C\frac{a}{\sqrt{k}} \left(\sqrt{\log 2} + \int_0^{C_2} \sqrt{\log(1 + 4/s)}\,\mathrm{d}s \right),
\end{aligned}$$

where the last line follows from $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$ for $x,\, y \geqslant 0$, and the change of variable $r = as$. Since $\int_0^{C_2} \sqrt{\log(1+4/s)}\,\mathrm{d}s < +\infty$, this concludes the first inequality.

We now prove the second inequality. By definition, we can write $S_k(a) = F(\zeta_1, \dots, \zeta_k)$. If one coordinate ζ_i is replaced by an arbitrary ζ_i', then for any λ satisfying $\|\lambda - \lambda^\star\|_\infty \leqslant a$,

$$\left| \frac{1}{k} g_\lambda(\zeta_i) - \frac{1}{k} g_\lambda(\zeta_i') \right| \leqslant \frac{1}{k} \left(|g_\lambda(\zeta_i)| + |g_\lambda(\zeta_i')| \right) \leqslant \frac{4a}{k},$$

by the bound above. Taking the supremum over λ gives that F satisfies the bounded-differences property with constant $4a/k$. The final result follows from applying McDiarmid's inequality. □

We now have everything to control $\mathbb{E}[Z_k]$ and prove Proposition A.1 using Lemma A.3.

Proof of Proposition A.1. By Lemma A.3, we only need to prove that $\mathbb{E}[Z_k] = O(k^{-1})$. Let $D_K = \sup_{\lambda \in K} \|\lambda - \lambda^\star\|_\infty < \infty$. For $\ell \geqslant 0$, set $a_\ell = \frac{A 2^\ell}{\sqrt{k}}$, where $A \geqslant 1$ will be chosen large enough as explained below. Let L_k be the smallest integer such that $a_{L_k} \geqslant D_K$. Then, any $\lambda \in K$ either satisfies $\|\lambda - \lambda^\star\|_\infty \leqslant a_0$, or belongs to one of the sets $a_\ell < \|\lambda - \lambda^\star\|_\infty \leqslant a_{\ell+1}$ with $0 \leqslant \ell < L_k$. By (10), if $\|\lambda - \lambda^\star\|_\infty \leqslant a_0$, then $H_k(\lambda) - H_k(\lambda^\star) \leqslant S_k(a_0)$. On the set $a_\ell < \|\lambda - \lambda^\star\|_\infty \leqslant a_{\ell+1}$, the same decomposition gives

$$H_k(\lambda) - H_k(\lambda^\star) \leqslant -\kappa a_\ell^2 + S_k(a_{\ell+1}).$$

Taking the supremum over K, we obtain

$$Z_k \leqslant \max\Big(S_k(a_0),\, \max_{0 \leqslant \ell < L_k} \{ S_k(a_{\ell+1}) - \kappa a_\ell^2 \} \Big).$$

Since $S_k(a_0) \geqslant 0$, denoting $x_+ = \max(x, 0)$, this implies

$$Z_k \leqslant S_k(a_0) + \sum_{\ell=0}^{L_k - 1} (S_k(a_{\ell+1}) - \kappa a_\ell^2)_+. \tag{11}$$

The first term is already of the right order, because Lemma A.4 gives

$$\mathbb{E}[S_k(a_0)] \leqslant C \frac{a_0}{\sqrt{k}} = C \frac{A}{k} \quad \text{and} \quad \mathbb{E}[S_k(a_{\ell+1})] \leqslant C \frac{a_{\ell+1}}{\sqrt{k}} \leqslant C \frac{A 2^\ell}{k}.$$

On the other hand, $\kappa a_\ell^2 = \kappa \frac{A^2 4^\ell}{k}$. Choosing A large enough, depending only on the constants in the above equation, ensures that for every $\ell \geqslant 0$, $\mathbb{E}[S_k(a_{\ell+1})] \leqslant \frac{\kappa}{2} a_\ell^2$. By Lemma A.4, for every $t > 0$,

$$\begin{aligned} \mathbb{P}(S_k(a_{\ell+1}) - \kappa a_\ell^2 > t) &\leqslant \mathbb{P}\left(S_k(a_{\ell+1}) - \mathbb{E}[S_k(a_{\ell+1})] > \frac{\kappa}{2} a_\ell^2 + t \right) \\ &\leqslant \exp\left(-c \frac{k(\frac{\kappa}{2} a_\ell^2 + t)^2}{a_{\ell+1}^2} \right). \end{aligned}$$

Using $\mathbb{E}[X_+] = \int_0^\infty \mathbb{P}(X > t)\,\mathrm{d}t$, followed by the change of variables $u = \frac{\sqrt{k}}{a_{\ell+1}} \left(\frac{\kappa}{2} a_\ell^2 + t \right)$, and the elementary Gaussian tail bound $\int_x^\infty e^{-cu^2}\mathrm{d}u \leqslant C e^{-cx^2/2}$, we get

$$\mathbb{E}[(S_k(a_{\ell+1}) - \kappa a_\ell^2)_+] \leqslant C \frac{a_{\ell+1}}{\sqrt{k}} \exp\left(-c \frac{k a_\ell^4}{a_{\ell+1}^2} \right).$$

Since $a_{\ell+1} = 2a_\ell$, $\frac{k a_\ell^4}{a_{\ell+1}^2} = \frac{k a_\ell^2}{4} = \frac{A^2 4^\ell}{4}$ and $\frac{a_{\ell+1}}{\sqrt{k}} = \frac{2A2^\ell}{k}$. Thus

$$\mathbb{E}[(S_k(a_{\ell+1}) - \kappa a_\ell^2)_+] \leqslant C \frac{A 2^\ell}{k} \exp(-c A^2 4^\ell).$$

Summing over $\ell \in \{0, \dots, L_k - 1\}$ and extending the finite sum to an infinite one,

$$\sum_{\ell=0}^{L_k - 1} \mathbb{E}[(S_k(a_{\ell+1}) - \kappa a_\ell^2)_+] \leqslant \frac{CA}{k} \sum_{\ell=0}^{\infty} 2^\ell \exp(-c A^2 4^\ell) \leqslant \frac{C}{k},$$

because A is fixed and the series is finite. Taking expectations in (11), we conclude that $\mathbb{E}[Z_k] \leqslant \frac{C}{k}$. Substituting this into Lemma A.3 proves the claim. □

A.4 Convergence of the expected plan and comparison with costs (Proposition 3.4)

In this section, we prove the plan convergence results of Proposition 3.4 and record the cost-to-plan estimates used to compare convergence of couplings with convergence of their transport costs.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space carrying two independent i.i.d. sequences $(X_i)_{i\in\mathbb{N}^*}$ and $(Y_j)_{j\in\mathbb{N}^*}$ with respective laws μ and ν. For each $k \in \mathbb{N}^*$, set $\mathbf{X}_k = (X_1, \dots, X_k)$ and $\mathbf{Y}_k = (Y_1, \dots, Y_k)$, and write $\widehat{\mu}_k = \widehat{\mu}_{\mathbf{X}_k}$, $\widehat{\nu}_k = \widehat{\nu}_{\mathbf{Y}_k}$ and $\widehat{\pi}_k = \widehat{\pi}_{\mathbf{X}_k, \mathbf{Y}_k}$. We assume, as in Proposition 3.4, that μ is absolutely continuous, so that the OT plan between μ and ν is unique. We denote it by $\pi^\star$.

A.4.1 Weak convergence of the expected batch OT plan

Proof of weak convergence in Proposition 3.4. We prove weak convergence of the expected batch OT plan $\overline{\pi}_k$ to the unique OT plan $\pi^\star$ by first showing that $\widehat{\pi}_k \rightharpoonup \pi^\star$ almost surely, and then passing to the expected plan by dominated convergence.

Convergence of the empiricals (cf. Pooladian et al. (2023a, Proposition D.1)). As $k \to \infty$, $\widehat{\pi}_k \rightharpoonup \pi^\star$ almost surely. Indeed, Bobkov and Ledoux (2019, Theorem 2.13) give, almost surely, $W_2(\widehat{\mu}_k,\, \mu) \to 0$ and $W_2(\widehat{\nu}_k,\, \nu) \to 0$. Fix an outcome ω for which both convergences hold. By Santambrogio (2015, Theorem 5.11), $\widehat{\mu}_k \rightharpoonup \mu$ and $\widehat{\nu}_k \rightharpoonup \nu$. Moreover, the triangle inequality gives

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \,\mathrm{d}\widehat{\pi}_k = W_2^2(\widehat{\mu}_k,\, \widehat{\nu}_k) \leqslant \left(W_2(\widehat{\mu}_k,\, \mu) + W_2(\mu,\, \nu) + W_2(\nu,\, \widehat{\nu}_k)\right)^2,$$

so the sequence $(\widehat{\pi}_k)$ has uniformly bounded quadratic cost. Since each $\widehat{\pi}_k$ is an OT plan between $\widehat{\mu}_k$ and $\widehat{\nu}_k$, by Ambrosio et al. (2005, Proposition 7.1.3), $(\widehat{\pi}_k)$ is relatively compact in $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, and every weak limit point is an OT plan between μ and ν. As this OT plan is unique, every weak limit point equals $\pi^\star$, thus $\widehat{\pi}_k \rightharpoonup \pi^\star$. Since this holds on an event of probability one, the conclusion follows.

Convergence of the expected plan. We now show that as $k \to \infty$, $\overline{\pi}_k \rightharpoonup \pi^\star$. Let g be a bounded continuous map on $\mathbb{R}^d \times \mathbb{R}^d$. Since $(\mathbf{X}_k,\, \mathbf{Y}_k) \sim \mu^{\otimes k} \otimes \nu^{\otimes k}$, the transfer formula gives, for every bounded Borel map f on $(\mathbb{R}^d)^k \times (\mathbb{R}^d)^k$,

$$\int_\Omega f(\mathbf{X}_k(\omega),\, \mathbf{Y}_k(\omega))\,\mathrm{d}\mathbb{P}(\omega) = \int_{(\mathbb{R}^d)^k\times(\mathbb{R}^d)^k} f(\mathbf{x}_k,\, \mathbf{y}_k)\,\mathrm{d}(\mu^{\otimes k}\otimes\nu^{\otimes k})(\mathbf{x}_k,\, \mathbf{y}_k).$$

Hence, by Definition 3.1, using the bounded Borel map $f : (\mathbf{x}_k,\, \mathbf{y}_k) \mapsto \int_{\mathbb{R}^d\times\mathbb{R}^d} g\,\mathrm{d}\widehat{\pi}_{\mathbf{x}_k,\, \mathbf{y}_k}$,

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} g\,\mathrm{d}\overline{\pi}_k = \int_\Omega \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} g\,\mathrm{d}\widehat{\pi}_k(\omega)\right)\mathrm{d}\mathbb{P}(\omega).$$

By the convergence of empiricals, for $\mathbb{P}$-almost every ω, $\widehat{\pi}_k(\omega) \rightharpoonup \pi^\star$ as $k \to +\infty$, thus

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} g\,\mathrm{d}\widehat{\pi}_k(\omega) \xrightarrow[k\to+\infty]{} \int_{\mathbb{R}^d\times\mathbb{R}^d} g\,\mathrm{d}\pi^\star.$$

Moreover,

$$\left|\int_{\mathbb{R}^d\times\mathbb{R}^d} g\,\mathrm{d}\widehat{\pi}_k(\omega)\right| \leqslant \|g\|_\infty.$$

Hence the dominated convergence theorem yields

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} g\,\mathrm{d}\overline{\pi}_k \xrightarrow[k\to+\infty]{} \int_{\mathbb{R}^d\times\mathbb{R}^d} g\,\mathrm{d}\pi^\star.$$

Since all measures $\overline{\pi}_k$ and $\pi^\star$ have marginals μ and ν, their second moments on $\mathbb{R}^d \times \mathbb{R}^d$ are equal. Thus, by Santambrogio (2015, Theorem 5.11), the weak convergence just proved is also convergence in W_2. □

A.4.2 Lower bound on plan distance (Proposition 3.4(1))

We now prove a lower bound that converts excess quadratic transport cost into a lower bound on the distance to the OT plan $\pi^\star$. Specifically, for every $\pi \in \Pi(\mu,\, \nu)$, we show that

$$W_2(\pi,\, \pi^\star) \geqslant \frac{1}{\sqrt{2}}\left(\left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2\,\mathrm{d}\pi(x,\, y)\right)^{\frac{1}{2}} - W_2(\mu,\, \nu)\right).$$

Proof of Proposition 3.4(1). Let $S(x,\, y) = y - x$. The map S is $\sqrt{2}$-Lipschitz, since

$$\|S(x,\, y) - S(x',\, y')\| \leqslant \|x - x'\| + \|y - y'\| \leqslant \sqrt{2}\|(x,\, y) - (x',\, y')\|_{\mathbb{R}^d \times \mathbb{R}^d},$$

where $\|(x,\, y) - (x',\, y')\|^2_{\mathbb{R}^d \times \mathbb{R}^d} = \|x - x'\|^2 + \|y - y'\|^2$.

Let $\gamma \in \Pi(S_\sharp \pi,\, S_\sharp \pi^\star)$, and let $(D,\, D^\star) \sim \gamma$. By the reverse triangle inequality for the L_2 norm,

$$\left(\mathbb{E}[\|D - D^\star\|^2]\right)^{\frac{1}{2}} \geqslant \left|\left(\mathbb{E}[\|D\|^2]\right)^{\frac{1}{2}} - \left(\mathbb{E}[\|D^\star\|^2]\right)^{\frac{1}{2}}\right|.$$

Taking the infimum over $\Pi(S_\sharp \pi,\, S_\sharp \pi^\star)$ yields

$$\begin{aligned} W_2(S_\sharp \pi,\, S_\sharp \pi^\star) &\geqslant \left|\left(\mathbb{E}_{D \sim S_\sharp \pi}[\|D\|^2]\right)^{\frac{1}{2}} - \left(\mathbb{E}_{D^\star \sim S_\sharp \pi^\star}[\|D^\star\|^2]\right)^{\frac{1}{2}}\right| \\ &= \left|\left(\mathbb{E}_{(X,\, Y) \sim \pi}[\|S(X,\, Y)\|^2]\right)^{\frac{1}{2}} - \left(\mathbb{E}_{(X,\, Y) \sim \pi^\star}[\|S(X,\, Y)\|^2]\right)^{\frac{1}{2}}\right| \\ &= \left|\left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\pi(x,\, y)\right)^{\frac{1}{2}} - W_2(\mu,\, \nu)\right|. \end{aligned}$$

Since $\pi^\star$ is optimal,

$$\left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\pi(x,\, y)\right)^{\frac{1}{2}} \geqslant W_2(\mu,\, \nu),$$

and hence

$$W_2(S_\sharp \pi,\, S_\sharp \pi^\star) \geqslant \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\pi(x,\, y)\right)^{\frac{1}{2}} - W_2(\mu,\, \nu). \tag{12}$$

For any $\gamma \in \Pi(\pi,\, \pi^\star)$, the push-forward $(S \times S)_\sharp \gamma$ belongs to $\Pi(S_\sharp \pi,\, S_\sharp \pi^\star)$. The Lipschitz bound on S gives

$$W_2^2(S_\sharp \pi,\, S_\sharp \pi^\star) \leqslant \mathbb{E}_{(Q,\, Q^\star) \sim \gamma}[\|S(Q) - S(Q^\star)\|^2] \leqslant 2\mathbb{E}_{(Q,\, Q^\star) \sim \gamma}[\|Q - Q^\star\|^2_{\mathbb{R}^d \times \mathbb{R}^d}].$$

Taking the infimum over $\Pi(\pi,\, \pi^\star)$ gives $W_2(S_\sharp \pi,\, S_\sharp \pi^\star) \leqslant \sqrt{2}\, W_2(\pi,\, \pi^\star)$. Combining this with (12) yields the result. □

A.4.3 Upper bound on plan distance (Proposition 3.4(2))

We prove the converse estimate to the previous lower bound in the Gaussian-to-discrete setting: excess quadratic transport cost controls the W_2-distance to the semidiscrete OT plan. Assume $\mu = \mathcal{N}(0,\, I_d)$, $\nu = \sum_{j=1}^M q_j \delta_{v_j}$ with $q_j > 0$ and pairwise distinct atoms, and denote $\pi^\star \in \Pi(\mu,\, \nu)$ the unique OT plan. We show that there exists $A_\nu > 0$ such that for any $\pi \in \Pi(\mu,\, \nu)$,

$$W_2^2(\pi,\, \pi^\star) \leqslant A_\nu \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\pi(x,\, y) - W_2^2(\mu,\, \nu)\right)^{\frac{1}{2}}.$$

We use the following formula for the distance from a point to a hyperplane:

Lemma A.5. *Let $n_H \in \mathbb{R}^d \backslash \{0\}$, $b \in \mathbb{R}$ and define, for $x \in \mathbb{R}^d$, $h(x) = \langle n_H,\, x\rangle - b$. Consider the hyperplane $H = \{x \in \mathbb{R}^d \,:\, h(x) = 0\}$. Then* $\mathrm{dist}(x,\, H) = \frac{|h(x)|}{\|n_H\|}$.

Proof. The orthogonal projection of x onto H is $x - \frac{h(x)}{\|n_H\|^2} n_H$, hence the distance is $\frac{|h(x)|}{\|n_H\|}$. □

Proof of Proposition 3.4(2). We prove that $W_2^2(\pi,\, \pi^\star) \leqslant A_\nu \,\mathrm{Exc}(\pi)^{\frac{1}{2}}$, where we denote

$$\mathrm{Exc}(\pi) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\pi(x,\, y) - W_2^2(\mu,\, \nu).$$

Write $\nu = \sum_{j=1}^M q_j \delta_{v_j}$ and $\mathcal{Y} = \{v_1, \ldots, v_M\}$. If $M = 1$, the claim is immediate, hence we assume $M \geqslant 2$.

Reminders on semidiscrete OT. We refer the reader to Peyré and Cuturi (2019, Section 5) for standard results on semidiscrete OT. Because μ is absolutely continuous and ν is finitely supported, there exists a unique OT map $T^\star : \mathbb{R}^d \to \mathcal{Y}$ between μ and ν. Specifically, there exist dual weights $\lambda_1^\star, \dots, \lambda_M^\star \in \mathbb{R}$ such that if we define $P_j^\star(x) = \|x - v_j\|^2 - \lambda_j^\star$ and the Laguerre cells

$$C_j = \left\{ x \in \mathbb{R}^d,\ P_j^\star(x) \leqslant \min_{1 \leqslant \ell \leqslant M} P_\ell^\star(x) \right\},$$

then $T^\star(x) = v_j$ for $x \in C_j$ away from the interfaces; on the μ-null interfaces, we use an arbitrary tie-breaking. For $j \neq \ell$, the interface between C_j and C_ℓ is contained in the affine hyperplane $H_{j\ell} = \{x \in \mathbb{R}^d : P_j^\star(x) = P_\ell^\star(x)\}$. We write $\Sigma^\star = \cup_{1 \leqslant j < \ell \leqslant M} H_{j\ell}$, which contains the decision boundary of $T^\star$.

Reduction to misclassification. Now let $(X,\, Y) \sim \pi$ and define $Y^\star = T^\star(X)$, so that $(X,\, Y^\star) \sim \pi^\star$. Using this as a coupling between π and $\pi^\star$, we obtain

$$W_2^2(\pi,\, \pi^\star) \leqslant \mathbb{E}[\|(X,\, Y) - (X,\, Y^\star)\|^2_{\mathbb{R}^d \times \mathbb{R}^d}] = \mathbb{E}[\|Y - Y^\star\|^2].$$

Since Y and $Y^\star$ both take values in the finite set $\mathcal{Y}$, $\|Y - Y^\star\|^2 \leqslant \operatorname{diam}(\mathcal{Y})^2\, \mathbb{1}_{\{Y \neq Y^\star\}}$, hence

$$W_2^2(\pi,\, \pi^\star) \leqslant \operatorname{diam}(\mathcal{Y})^2\, \mathbb{P}(Y \neq Y^\star). \tag{13}$$

Excess cost is the expected dual slack. By the dual form of semidiscrete OT,

$$W_2^2(\mu,\, \nu) = \int_{\mathbb{R}^d} \min_{1 \leqslant m \leqslant M} P_m^\star(x)\, \mathrm{d}\mu(x) + \sum_{m=1}^M q_m \lambda_m^\star.$$

For every j, define $\lambda^\star(v_j) = \lambda_j^\star$, and set $\Delta(x,\, y) = \|x - y\|^2 - (\min_m P_m^\star(x) + \lambda^\star(y))$. By dual feasibility, $\Delta(x,\, y) \geqslant 0$ for all $(x,\, y)$. Moreover, because π has marginals μ and ν,

$$\mathbb{E}[\Delta(X,\, Y)] = \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2\, \mathrm{d}\pi(x,\, y) - W_2^2(\mu,\, \nu) = \operatorname{Exc}(\pi). \tag{14}$$

Excess cost controls misclassification away from the boundary. Now fix $\delta > 0$ and define the δ-neighborhood of the decision boundary by $\Sigma_\delta^\star = \{x \in \mathbb{R}^d,\ \operatorname{dist}(x,\, \Sigma^\star) \leqslant \delta\}$. Take $x \in C_j \backslash \Sigma_\delta^\star$, let $\ell \neq j$ and consider $h_{j\ell} = P_\ell^\star - P_j^\star$. This is an affine function of gradient $\nabla h_{j\ell} = 2(v_j - v_\ell)$, and $\|\nabla h_{j\ell}\| = 2\|v_j - v_\ell\| \geqslant 2\operatorname{sep}(\mathcal{Y})$. As $H_{j\ell} = h_{j\ell}^{-1}(\{0\})$, by Lemma A.5,

$$|h_{j\ell}(x)| = \|\nabla h_{j\ell}\| \operatorname{dist}(x,\, H_{j\ell}).$$

Because $x \notin \Sigma_\delta^\star$, we have $\operatorname{dist}(x,\, H_{j\ell}) > \delta$, hence

$$|h_{j\ell}(x)| \geqslant 2\operatorname{sep}(\mathcal{Y})\, \delta.$$

But $x \in C_j$, so $P_j^\star(x) \leqslant P_\ell^\star(x)$, hence $h_{j\ell}(x) \geqslant 0$, and

$$P_\ell^\star(x) - P_j^\star(x) \geqslant 2\operatorname{sep}(\mathcal{Y})\, \delta. \tag{15}$$

Now, since $x \in C_j$, $P_j^\star(x) = \min_{1 \leqslant m \leqslant M} P_m^\star(x)$, thus for $y = v_\ell$,

$$\Delta(x,\, v_\ell) = (\|x - v_\ell\|^2 - \lambda_\ell^\star) - \min_{1 \leqslant m \leqslant M} P_m^\star(x) = P_\ell^\star(x) - P_j^\star(x).$$

Using (15), we obtain $\Delta(x,\, v_\ell) \geqslant 2\operatorname{sep}(\mathcal{Y})\, \delta$,

$$\Delta(X,\, Y)\, \mathbb{1}_{\{X \notin \Sigma_\delta^\star,\, Y \neq Y^\star\}} \geqslant 2\operatorname{sep}(\mathcal{Y})\, \delta\, \mathbb{1}_{\{X \notin \Sigma_\delta^\star,\, Y \neq Y^\star\}}.$$

Since $\Delta \geqslant 0$ everywhere, (14) implies $\operatorname{Exc}(\pi) \geqslant 2\operatorname{sep}(\mathcal{Y})\, \delta\, \mathbb{P}(X \notin \Sigma_\delta^\star,\, Y \neq Y^\star)$, hence

$$\mathbb{P}(X \notin \Sigma_\delta^\star,\, Y \neq Y^\star) \leqslant \frac{\operatorname{Exc}(\pi)}{2\operatorname{sep}(\mathcal{Y})\, \delta}. \tag{16}$$

Gaussian anti-concentration near the boundary. Since $\Sigma^\star$ is the union of at most $\frac{M(M-1)}{2}$ affine hyperplanes, it suffices to bound the Gaussian mass of a δ-tube around one hyperplane $H = \{x \in \mathbb{R}^d : \langle n, x\rangle = b\}$ for $\|n\| = 1$.

By Lemma A.5, $\operatorname{dist}(x, H) = |\langle n, x\rangle - b|$. If $X \sim \mathcal{N}(0, I_d)$, then $\langle n, X\rangle \sim \mathcal{N}(0, 1)$, and

$$\mathbb{P}(\operatorname{dist}(X, H) \leqslant \delta) = \mathbb{P}_{Z\sim\mathcal{N}(0,1)}(|Z - b| \leqslant \delta).$$

The standard Gaussian density is bounded by $(2\pi)^{-\frac{1}{2}}$, so

$$\mathbb{P}_{Z\sim\mathcal{N}(0,1)}(|Z - b| \leqslant \delta) \leqslant \frac{2\delta}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}\delta.$$

By the union bound,

$$\mathbb{P}(X \in \Sigma^\star_\delta) \leqslant \alpha\delta \quad \text{with} \quad \alpha = \frac{M(M-1)}{2}\sqrt{\frac{2}{\pi}}. \tag{17}$$

Conclusion. Combining (16) and (17), we obtain

$$\begin{aligned}\mathbb{P}(Y \neq Y^\star) &\leqslant \mathbb{P}(X \in \Sigma^\star_\delta) + \mathbb{P}(X \notin \Sigma^\star_\delta,\, Y \neq Y^\star)\\ &\leqslant \alpha\delta + \frac{\operatorname{Exc}(\pi)}{2\operatorname{sep}(\mathcal{Y})\,\delta}.\end{aligned}$$

Minimizing the right-hand side over $\delta > 0$,

$$\mathbb{P}(Y \neq Y^\star) \leqslant 2\sqrt{\frac{\alpha}{2\operatorname{sep}(\mathcal{Y})}}\operatorname{Exc}(\pi)^{\frac{1}{2}}. \tag{18}$$

Finally, substituting (18) into (13), $W_2^2(\pi, \pi^\star) \leqslant A_\nu \operatorname{Exc}(\pi)^{\frac{1}{2}}$ with $A_\nu = 2\operatorname{diam}(\mathcal{Y})^2\left(\frac{M(M-1)}{2\sqrt{2\pi}\operatorname{sep}(\mathcal{Y})}\right)^{\frac{1}{2}}$. Taking $\pi = \overline{\pi}_k$ and using Propositions 3.2 and 3.3,

$$W_2^2(\overline{\pi}_k, \pi^\star) \leqslant A_\nu\left(\mathbb{E}[W_2^2(\widehat{\mu}_k, \widehat{\nu}_k)] - W_2^2(\mu, \nu)\right)^{1/2} \leqslant Ck^{-1/2}.$$

□

A.4.4 No universal cost-to-plan modulus

We show that there is no universal modulus converting excess transport cost into W_2-closeness to the OT plan, even when the OT plan is unique.

Lemma A.6. *There is no function ω defined for $r \geqslant 0$, with $\omega(r) \geqslant 0$ and $\omega(r) \xrightarrow[r\to 0]{} 0$, such that for all pairs $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^2)$ whose OT plan $\pi^\star$ is unique, and for every $\pi \in \Pi(\mu, \nu)$,*

$$W_2(\pi, \pi^\star) \leqslant \omega\left(\int_{\mathbb{R}^2\times\mathbb{R}^2}\|x - y\|^2\,\mathrm{d}\pi(x, y) - W_2^2(\mu, \nu)\right).$$

Proof. For $\varepsilon > 0$, let $K_\varepsilon = [-\varepsilon, \varepsilon] \times [-1, 1]$, $\mu_\varepsilon = \operatorname{Unif}(K_\varepsilon)$ and $\nu = \operatorname{Unif}\{(-1, 0), (1, 0)\}$. The OT map $S^\star_\varepsilon$ between μ_ε and ν is unique by Brenier's theorem, as μ_ε is absolutely continuous (Santambrogio, 2015, Theorem 1.22). Moreover, for every $(x_1, x_2) \in K_\varepsilon$, we have $S^\star_\varepsilon(x_1, x_2) = (\operatorname{sign}(x_1), 0)$, as it is optimal to send (x_1, x_2) to $(-1, 0)$ if $x_1 < 0$, and to $(1, 0)$ otherwise. Define the map $S_\varepsilon : (x_1, x_2) \mapsto (\operatorname{sign}(x_2), 0)$. The values of S_ε and $S^\star_\varepsilon$ do not matter on the μ_ε-negligible sets $\mathbb{R} \times \{0\}$ and $\{0\} \times \mathbb{R}$. Denote by $\pi_\varepsilon = (\mathrm{id}, S_\varepsilon)_\sharp\mu_\varepsilon$ and $\pi^\star_\varepsilon = (\mathrm{id}, S^\star_\varepsilon)_\sharp\mu_\varepsilon$ the corresponding transport plans. Then

$$\delta_\varepsilon = \int_{\mathbb{R}^2\times\mathbb{R}^2}\|x - y\|^2\,\mathrm{d}\pi_\varepsilon(x, y) - W_2^2(\mu_\varepsilon, \nu) = \frac{1}{4\varepsilon}\int_{K_\varepsilon}\left(\|x - S_\varepsilon(x)\|^2 - \|x - S^\star_\varepsilon(x)\|^2\right)\mathrm{d}x.$$

Given $x = (x_1, x_2) \in \mathbb{R}^2$, we have $S_\varepsilon(x) \neq S^\star_\varepsilon(x)$ if and only if x_1 and x_2 have opposite signs, that is, $x_1x_2 < 0$. For instance, if $x_1 < 0$ and $x_2 > 0$,

$$\|x - S_\varepsilon(x)\|^2 - \|x - S^\star_\varepsilon(x)\|^2 = \left((x_1 - 1)^2 + x_2^2\right) - \left((x_1 + 1)^2 + x_2^2\right) = 4|x_1|,$$

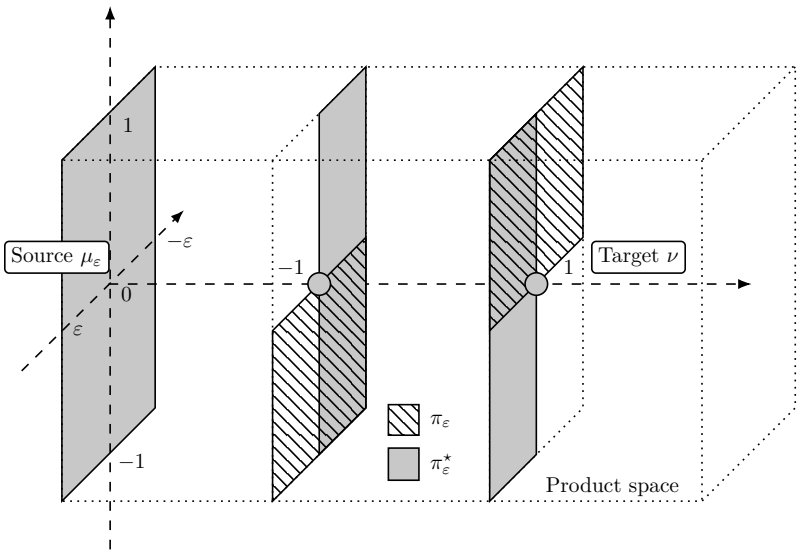


Figure 8: Couplings π_ε and $\pi_\varepsilon^\star$ in $\Pi(\mu_\varepsilon, \nu)$.

which also holds if $x_1 > 0$ and $x_2 < 0$ by symmetry. Therefore,

$$\delta_\varepsilon = \frac{1}{4\varepsilon}\left(-\int_{-\varepsilon}^{0}\int_{0}^{1} 4x_1 \,\mathrm{d}x_2 \,\mathrm{d}x_1 + \int_{0}^{\varepsilon}\int_{-1}^{0} 4x_1 \,\mathrm{d}x_2 \,\mathrm{d}x_1\right) = \frac{2}{\varepsilon}\int_0^\varepsilon x_1 \,\mathrm{d}x_1 = \varepsilon.$$

Consider the map $\Psi : (x, y) \in \mathbb{R}^2 \times \mathbb{R}^2 \mapsto x_2 y_1 \in \mathbb{R}$. On $K_\varepsilon \times \{(-1, 0), (1, 0)\}$, since $a + b \leqslant \sqrt{2(a^2 + b^2)}$,

$$\begin{aligned}|\Psi(x, y) - \Psi(x', y')| &= |x_2 y_1 - x_2' y_1'| \leqslant |x_2 - x_2'| + |y_1 - y_1'| \\ &\leqslant \sqrt{2}\sqrt{\|x - x'\|^2 + \|y - y'\|^2} = \sqrt{2}\|(x, y) - (x', y')\|_{\mathbb{R}^2 \times \mathbb{R}^2}.\end{aligned}$$

Thus Ψ is $\sqrt{2}$-Lipschitz, and as in the proof of Appendix A.4.2,

$$W_2(\pi_\varepsilon, \pi_\varepsilon^\star) \geqslant \frac{1}{\sqrt{2}} W_2(\Psi_\sharp \pi_\varepsilon, \Psi_\sharp \pi_\varepsilon^\star).$$

Let $(X, Y) \sim \pi_\varepsilon$. Then $Y_1 = \operatorname{sign}(X_2)$, hence $\Psi(X, Y) = X_2 \operatorname{sign}(X_2) = |X_2| \sim \mathrm{Unif}[0, 1]$, and $\Psi_\sharp \pi_\varepsilon = \mathrm{Unif}[0, 1]$, with quantile function $F^-_{\Psi_\sharp \pi_\varepsilon}(u) = u$. Likewise, let $(X, Y) \sim \pi_\varepsilon^\star$. Then $Y_1 = \operatorname{sign}(X_1)$ is independent of $X_2 \sim \mathrm{Unif}[-1, 1]$, so $\Psi(X, Y) = X_2 Y_1 \sim \mathrm{Unif}[-1, 1]$, hence $\Psi_\sharp \pi_\varepsilon^\star = \mathrm{Unif}[-1, 1]$, with quantile function $F^-_{\Psi_\sharp \pi_\varepsilon^\star}(u) = -1 + 2u$.

By Peyré and Cuturi (2019, Remark 2.30), in dimension 1, the 2-Wasserstein distance is the L^2 distance between quantile functions:

$$W_2^2(\Psi_\sharp \pi_\varepsilon, \Psi_\sharp \pi_\varepsilon^\star) = \int_0^1 (\underbrace{F^-_{\Psi_\sharp \pi_\varepsilon}(u)}_{u} - \underbrace{F^-_{\Psi_\sharp \pi_\varepsilon^\star}(u)}_{-1+2u})^2 \,\mathrm{d}u = \int_0^1 (u - 1)^2 \,\mathrm{d}u = \frac{1}{3}.$$

We built a family of instances (μ_ε, ν) with unique OT plans $\pi_\varepsilon^\star$ and couplings $\pi_\varepsilon \in \Pi(\mu_\varepsilon, \nu)$ such that $\int_{\mathbb{R}^2 \times \mathbb{R}^2} \|x - y\|^2 \,\mathrm{d}\pi_\varepsilon(x, y) - W_2^2(\mu_\varepsilon, \nu) = \delta_\varepsilon \downarrow 0$ but $W_2(\pi_\varepsilon, \pi_\varepsilon^\star) \geqslant 1/\sqrt{6}$, ruling out the existence of any such universal function ω. □

A.4.5 One-dimensional expected batch OT plan rate

We show that in one dimension the expected batch OT plan converges to the monotone OT plan $\pi^\star$ at rate $O(k^{-1})$ in squared W_2 distance, provided $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ and one quantile map is Lipschitz.

Lemma A.7. *If μ and ν have quantile functions F_μ^- and F_ν^-, and one of the quantiles is L-Lipschitz,*

$$W_2^2(\overline{\pi}_k, \pi^\star) \leqslant \frac{L^2}{3(k + 1)}.$$

Proof. We give the argument for the case where F_ν^- is L-Lipschitz on $[0,\,1]$: the case of F_μ^- is symmetric.

Let $(U_1, \dots, U_k, V_1, \dots, V_k) \sim \mathrm{Unif}([0,\,1])^{\otimes 2k}$ and write the corresponding order statistics

$$U_{(1)} \leqslant \cdots \leqslant U_{(k)} \quad \text{and} \quad V_{(1)} \leqslant \cdots \leqslant V_{(k)}.$$

Let also $I \sim \mathrm{Unif}(\{1, \dots, k\})$ be independent of everything else.

In one dimension, the OT plan between two empirical measures with equal weights matches the i-th order statistic of one sample with the i-th order statistic of the other. Hence $(X,\,Y) = (F_\mu^-(U_{(I)}),\,F_\nu^-(V_{(I)}))$ has law $\overline{\pi}_k$.

Now define $Y^\star = F_\nu^-(U_{(I)})$. Since I is independent, $U_{(I)} \sim \mathrm{Unif}([0,\,1])$. Therefore $(X,\,Y^\star) = (F_\mu^-(U_{(I)}),\,F_\nu^-(U_{(I)}))$ has law $\pi^\star$. Thus $((X,\,Y),\,(X,\,Y^\star))$ is a coupling of $\overline{\pi}_k$ and $\pi^\star$, and

$$W_2^2(\overline{\pi}_k,\,\pi^\star) \leqslant \mathbb{E}[\|(X,\,Y) - (X,\,Y^\star)\|_{\mathbb{R}^2}^2] = \mathbb{E}[|Y - Y^\star|^2].$$

Since F_ν^- is L-Lipschitz, $|Y - Y^\star| \leqslant L\,|V_{(I)} - U_{(I)}|$, and $W_2^2(\overline{\pi}_k,\,\pi^\star) \leqslant L^2\,\mathbb{E}[|U_{(I)} - V_{(I)}|^2]$.

Conditioning on I gives

$$\mathbb{E}[|U_{(I)} - V_{(I)}|^2] = \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}[|U_{(i)} - V_{(i)}|^2].$$

For each i, the random variables $U_{(i)}$ and $V_{(i)}$ are independent and identically distributed, so

$$\mathbb{E}[|U_{(i)} - V_{(i)}|^2] = 2\ \mathrm{Var}(U_{(i)}).$$

As $\mathrm{Var}(U_{(i)}) = \frac{i(k+1-i)}{(k+1)^2(k+2)}$, and using $\sum_{i=1}^{k} i(k+1-i) = \frac{k(k+1)(k+2)}{6}$, we obtain

$$\mathbb{E}[|U_{(I)} - V_{(I)}|^2] = \frac{1}{3(k+1)}.$$

Substituting this into the previous bound yields $W_2^2(\overline{\pi}_k,\,\pi^\star) \leqslant \frac{L^2}{3(k+1)}$. □

B Proofs for Section 4

This appendix follows the structure of Section 4. We first prove the posterior formulas for Gaussian-to-discrete FM velocities (Appendix B.1). We then show that the expected batch OT plan is rectifiable by establishing regularity of its induced velocity field and constructing the terminal flow map (Appendix B.2). We next prove the posterior-concentration and one-step Euler estimates used to interpret the independent coupling (Appendix B.4). Finally, we analyze the one-dimensional two-atom model and prove the asymptotic comparison between OT batch size and NFE (Appendix B.5).

Notation for the proofs. In the main text of Section 4, we write $(X_0,\,X_1,\,X_t)$, with $X_t = (1-t)X_0 + tX_1$. In this appendix only, we write $(X,\,Y,\,Z_t)$ for the same objects, with $X = X_0$, $Y = X_1$, and $Z_t = X_t = (1-t)X + tY$. Accordingly, a point z in the support of Z_t corresponds to a point x in the support of X_t in the main text whenever x is used there as the current time-t state.

B.1 Conditional mean and posterior formulas (Lemma 4.1)

Proof of Lemma 4.1. Let $(X,\,Y) \sim \pi$ and $Z_t = (1-t)X + tY$. Since $Y - X = \frac{Y - Z_t}{1-t}$, taking conditional expectations given $Z_t = z$ yields

$$u_t^\pi(z) = \mathbb{E}[Y - X \mid Z_t = z] = \frac{m_t^\pi(z) - z}{1-t}.$$

The finite-support expression for m_t^π follows by conditioning on the values of Y. The posterior formula for π is Bayes' formula, at points where the density of Z_t is positive, applied after the change of variables $z = (1-t)x + tv_j$. □

B.2 Rectifiability of the expected batch OT plan (Proposition 4.2)

This section proves rectifiability of the expected batch OT plan by first expressing its FM velocity through averaged assignment probabilities, then proving the regularity needed for the pre-terminal ODE flow, and finally showing that the flow converges to a discrete terminal map. If the target has a single atom, the claim is immediate, hence we assume $M \geqslant 2$. Let $k \in \mathbb{N}^*$, $\mu = \mathcal{N}(0,\, I_d)$ and $\nu = \sum_{j=1}^{M} q_j \delta_{v_j}$ with $M \geqslant 2$ and $q_j > 0$, where the atoms $v_1, \dots, v_M$ are pairwise distinct. Set $q_{\min} = \min_{1 \leqslant j \leqslant M} q_j$. Let $\mathcal{Y} = \{v_1, \dots, v_M\}$ denote the support of ν.

B.2.1 Local Lipschitz regularity of the batch OT target velocity

Draw $\mathbf{X}_k = (X_1, \dots, X_k) \sim \mu^{\otimes k}$ and $\mathbf{Y}_k = (Y_1, \dots, Y_k) \sim \nu^{\otimes k}$ independently, and let $I \sim \mathrm{Unif}\{1, \dots, k\}$ be independent of everything else. Set $\widetilde{X} = X_I$ and $\widetilde{Y} = Y_{\widehat{\sigma}_{\mathbf{X}_k,\, \mathbf{Y}_k}(I)}$, with $\widehat{\sigma}_{\mathbf{X}_k,\, \mathbf{Y}_k}$ as in Appendix A.1.1. By Definition 3.1, $(\widetilde{X},\, \widetilde{Y}) \sim \overline{\pi}_k$.

To prove that the velocity field $u_t^{\overline{\pi}_k}$ is Lipschitz, we first express it in Lemma B.1 in terms of the assignment probabilities $\bar{a}_j(x) = \mathbb{P}(\widetilde{Y} = v_j \mid \widetilde{X} = x)$. We then show in Lemma B.2 that the functions $\bar{a}_j$ are globally Lipschitz and bounded from below, which then yields Proposition 4.2(1).

For each $i \in \{1, \dots, k\}$ and $x \in \mathbb{R}^d$, let $\widehat{Y}^{(i)}(x)$ be the target atom matched to the i-th source position when X_i is replaced by x, while all other source points and all target points are kept fixed:

$$\widehat{Y}^{(i)}(x) = Y_{\widehat{\sigma}_k(X_1, \dots, X_{i-1},\, x,\, X_{i+1}, \dots, X_k,\, \mathbf{Y}_k)(i)}.$$

For $j \in \{1, \dots, M\}$, set $a_j^{(i)}(x) = \mathbb{P}(\widehat{Y}^{(i)}(x) = v_j)$. Since I is uniform, $\overline{a}_j(x)$ is obtained by averaging over the possible positions i: $\overline{a}_j(x) = \frac{1}{k} \sum_{i=1}^{k} a_j^{(i)}(x)$.

Lemma B.1. *Let φ denote the standard Gaussian density. For $t \in [0,\, 1)$ and $z \in \mathbb{R}^d$, write*

$$x_{t,j}(z) = \frac{z - t v_j}{1 - t} \quad \textit{and} \quad w_j(t,\, z) = (1-t)^{-d} \varphi(x_{t,j}(z)) \overline{a}_j(x_{t,j}(z)).$$

Then the density of $\overline{\pi}_k$ with respect to the Lebesgue measure in x and the counting measure on $\{v_1, \dots, v_M\}$ is $f_{\overline{\pi}_k}(x,\, v_j) = \varphi(x) \overline{a}_j(x)$, and

$$m_t^{\overline{\pi}_k}(z) = \sum_{j=1}^{M} v_j \, \frac{w_j(t,\, z)}{\sum_{\ell=1}^{M} w_\ell(t,\, z)}, \quad u_t^{\overline{\pi}_k}(z) = \sum_{j=1}^{M} \frac{v_j - z}{1 - t} \, \frac{w_j(t,\, z)}{\sum_{\ell=1}^{M} w_\ell(t,\, z)}.$$

Proof. Let g be bounded and Borel. Conditioning on the random index I gives

$$\mathbb{E}[g(\widetilde{X},\, \widetilde{Y})] = \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}\big[g(X_i, \widehat{Y}^{(i)}(X_i))\big].$$

Now condition on $X_i = x$. Since $X_i \sim \mathcal{N}(0,\, I_d)$ is independent of $(X_r)_{r \neq i}$ and $\mathbf{Y}_k$,

$$\mathbb{E}\big[g(X_i,\, \widehat{Y}^{(i)}(X_i))\big] = \sum_{j=1}^{M} \int_{\mathbb{R}^d} g(x,\, v_j)\, \varphi(x)\, a_j^{(i)}(x)\, \mathrm{d}x.$$

Averaging over i yields

$$\mathbb{E}[g(\widetilde{X},\, \widetilde{Y})] = \sum_{j=1}^{M} \int_{\mathbb{R}^d} g(x,\, v_j)\, \varphi(x)\, \overline{a}_j(x)\, \mathrm{d}x.$$

Thus $(\widetilde{X},\, \widetilde{Y})$ has density $(x,\, v_j) \mapsto \varphi(x) \overline{a}_j(x)$.

Denote $Z_t = (1-t)\widetilde{X} + t\widetilde{Y}$. The change of variables $z = (1-t)x + t v_j$ shows that $(Z_t,\, \widetilde{Y})$ has density $(z,\, v_j) \mapsto w_j(t,\, z)$. Hence the density of Z_t is $\sum_{\ell=1}^{M} w_\ell(t,\, z)$, and Bayes' formula gives

$$\mathbb{P}(\widetilde{Y} = v_j \mid Z_t = z) = \frac{w_j(t,\, z)}{\sum_{\ell=1}^{M} w_\ell(t,\, z)}.$$

The formulas for $m_t^{\overline{\pi}_k}(z) = \mathbb{E}[\widetilde{Y} \mid Z_t = z]$ and $u_t^{\overline{\pi}_k}$ follow. □

Lemma B.2. *For every $j \in \{1, \dots, M\}$, the map $\overline{a}_j : \mathbb{R}^d \to [0,\, 1]$ is globally Lipschitz:*

$$\forall x,\, x' \in \mathbb{R}^d, \quad |\overline{a}_j(x) - \overline{a}_j(x')| \leqslant (k!)^2 \frac{\operatorname{diam}(\mathcal{Y})}{\operatorname{sep}(\mathcal{Y})\sqrt{2\pi}} \|x - x'\|.$$

Moreover, $\overline{a}_j(x) \geqslant q_j^k$ for every $x \in \mathbb{R}^d$.

Proof. If $k = 1$, then $a_j^{(1)}(x) = q_j$ for every x, so the Lipschitz bound and lower bound are immediate. We therefore assume $k \geqslant 2$. Fix $i \in \{1, \dots, k\}$ and $x,\, x' \in \mathbb{R}^d$. Since $a_j^{(i)}(x) = \mathbb{P}(\widehat{Y}^{(i)}(x) = v_j)$, we have

$$\begin{aligned} |a_j^{(i)}(x) - a_j^{(i)}(x')| &= |\mathbb{E}[\mathbb{1}_{\{\widehat{Y}^{(i)}(x) = v_j\}} - \mathbb{1}_{\{\widehat{Y}^{(i)}(x') = v_j\}}]| \\ &\leqslant \mathbb{E}[|\mathbb{1}_{\{\widehat{Y}^{(i)}(x) = v_j\}} - \mathbb{1}_{\{\widehat{Y}^{(i)}(x') = v_j\}}|] \\ &\leqslant \mathbb{P}(\widehat{Y}^{(i)}(x) \neq \widehat{Y}^{(i)}(x')). \end{aligned}$$

It remains to bound this probability by a constant times $\|x - x'\|$. For each permutation $\sigma \in \mathfrak{S}_k$ and each $z \in \mathbb{R}^d$, define the cost of transporting $(X_1,\, \dots,\, X_{i-1},\, z,\, X_{i+1},\, \dots,\, X_k)$ to $\mathbf{Y}_k$ using σ:

$$C_\sigma^{(i)}(z) = \frac{1}{k}\left(\|z - Y_{\sigma(i)}\|^2 + \sum_{r \neq i} \|X_r - Y_{\sigma(r)}\|^2 \right).$$

By definition, $\widehat{Y}^{(i)}(z)$ is equal to $Y_{\sigma(i)}$, where σ is the permutation selected by the solver, minimizing $\sigma \mapsto C_\sigma^{(i)}(z)$. Let $E_{\sigma,\,\sigma'}$ be the event that the solver selects σ at x and σ' at x', and that $\widehat{Y}^{(i)}(x) \neq \widehat{Y}^{(i)}(x')$. Then, by the union bound,

$$\mathbb{P}(\widehat{Y}^{(i)}(x) \neq \widehat{Y}^{(i)}(x')) \leqslant \sum_{\sigma, \sigma' \in \mathfrak{S}_k} \mathbb{P}(E_{\sigma,\sigma'}).$$

Fix a pair $\sigma,\, \sigma' \in \mathfrak{S}_k$. Then

$$C_\sigma^{(i)}(z) - C_{\sigma'}^{(i)}(z) = \frac{2}{k}\langle Y_{\sigma'(i)} - Y_{\sigma(i)},\, z\rangle - U_{\sigma,\,\sigma'}^{(i)}, \quad \text{with} \quad U_{\sigma,\,\sigma'}^{(i)} = \frac{2}{k}\sum_{r \neq i} \langle Y_{\sigma(r)} - Y_{\sigma'(r)},\, X_r\rangle.$$

The target-norm terms cancel in this difference because σ and σ' are permutations of the same target batch. On the event $E_{\sigma,\,\sigma'}$, we have $C_\sigma^{(i)}(x) - C_{\sigma'}^{(i)}(x) \leqslant 0$ and $C_\sigma^{(i)}(x') - C_{\sigma'}^{(i)}(x') \geqslant 0$, which implies that the random variable $U_{\sigma,\,\sigma'}^{(i)}$ belongs to the interval

$$I_{\sigma,\,\sigma'}^{(i)}(x,\, x') = \left[\frac{2}{k}\langle Y_{\sigma'(i)} - Y_{\sigma(i)},\, x\rangle,\, \frac{2}{k}\langle Y_{\sigma'(i)} - Y_{\sigma(i)},\, x'\rangle \right].$$

Since $E_{\sigma,\sigma'}$ implies that the labels at position i are different, we have

$$E_{\sigma,\,\sigma'} \subseteq \{Y_{\sigma(i)} \neq Y_{\sigma'(i)}\} \cap \{U_{\sigma,\,\sigma'}^{(i)} \in I_{\sigma,\,\sigma'}^{(i)}(x,\, x')\}.$$

Therefore, conditionally on $\mathbf{Y}_k$,

$$\mathbb{P}(E_{\sigma,\,\sigma'} \mid \mathbf{Y}_k) \leqslant \mathbb{1}_{\{Y_{\sigma(i)} \neq Y_{\sigma'(i)}\}}\, \mathbb{P}(U_{\sigma,\,\sigma'}^{(i)} \in I_{\sigma,\,\sigma'}^{(i)}(x,\, x') \mid \mathbf{Y}_k).$$

Since $\sigma,\, \sigma'$ are fixed, conditionally on $\mathbf{Y}_k$, the variable $U_{\sigma,\,\sigma'}^{(i)} \mid \mathbf{Y}_k$ is centered and Gaussian, and

$$\operatorname{Var}(U_{\sigma,\,\sigma'}^{(i)} \mid \mathbf{Y}_k) = \left(\frac{2}{k}\right)^2 \sum_{r \neq i} \|Y_{\sigma(r)} - Y_{\sigma'(r)}\|^2.$$

Moreover, on the event $\{Y_{\sigma(i)} \neq Y_{\sigma'(i)}\}$, as the families $(Y_{\sigma(r)})_{r \neq i}$ and $(Y_{\sigma'(r)})_{r \neq i}$ cannot coincide,

$$\operatorname{Var}(U_{\sigma,\,\sigma'}^{(i)} \mid \mathbf{Y}_k) \geqslant \left(\frac{2\, \operatorname{sep}(\mathcal{Y})}{k}\right)^2.$$

Let Φ denote the standard Gaussian cumulative distribution function. The function Φ is $1/\sqrt{2\pi}$-Lipschitz. It follows that on the event $\{Y_{\sigma(i)} \neq Y_{\sigma'(i)}\}$,

$$\mathbb{P}(U^{(i)}_{\sigma,\,\sigma'} \in I^{(i)}_{\sigma,\,\sigma'}(x,\,x') \mid \mathbf{Y}_k) \leqslant \frac{|I^{(i)}_{\sigma,\,\sigma'}(x,\,x')|}{\sqrt{2\pi}\,\sqrt{\mathrm{Var}(U^{(i)}_{\sigma,\,\sigma'} \mid \mathbf{Y}_k)}} \leqslant \frac{\mathrm{diam}(\mathcal{Y})}{\mathrm{sep}(\mathcal{Y})\sqrt{2\pi}}\,\|x - x'\|.$$

Hence

$$\mathbb{P}(E_{\sigma,\sigma'} \mid \mathbf{Y}_k) \leqslant \mathbb{1}_{\{Y_{\sigma(i)} \neq Y_{\sigma'(i)}\}} \frac{\mathrm{diam}(\mathcal{Y})}{\mathrm{sep}(\mathcal{Y})\sqrt{2\pi}} \|x - x'\| \leqslant \frac{\mathrm{diam}(\mathcal{Y})}{\mathrm{sep}(\mathcal{Y})\sqrt{2\pi}} \|x - x'\|.$$

Since the right-hand side does not depend on $\mathbf{Y}_k$, the same upper bound holds for $\mathbb{P}(E_{\sigma,\sigma'})$, and

$$\mathbb{P}(\widehat{Y}^{(i)}(x) \neq \widehat{Y}^{(i)}(x')) \leqslant (k!)^2 \frac{\mathrm{diam}(\mathcal{Y})}{\mathrm{sep}(\mathcal{Y})\sqrt{2\pi}}\,\|x - x'\|.$$

Combining this with the first inequality of the proof shows that $a^{(i)}_j$ is Lipschitz, and the same holds for $\overline{a}_j$ by averaging over i. Finally, the event $\{Y_1 = \cdots = Y_k = v_j\}$ has probability q_j^k, and on this event one has $\widehat{Y}^{(i)}(x) = v_j$ for every i and every x. Thus $a^{(i)}_j(x) \geqslant q_j^k$, and therefore $\overline{a}_j(x) \geqslant q_j^k$. □

Proof of Proposition 4.2(1). As in Hertrich et al. (2025, proof of Theorem 14), we derive the regularity of the FM velocity from the regularity of the conditional posterior weights, which in our setting are the $\overline{a}_j$ controlled in Lemma B.2. Let $\tau \in (0,\,1)$ and let $K \subseteq \mathbb{R}^d$ be compact. We show that there exists a constant $L > 0$ such that, for every $t \in [0,\,\tau]$, the vector field $u_t^{\overline{\pi}_k}$ is L-Lipschitz on K. Moreover, we show that $(t,\,z) \in [0,\,1) \times K \mapsto u_t^{\overline{\pi}_k}(z)$ is continuous.

First, observe that by Lemma B.2, each $\overline{a}_j$ is globally Lipschitz on $\mathbb{R}^d$. It is also bounded by 1 as a probability. The standard Gaussian φ is also globally bounded and globally Lipschitz on $\mathbb{R}^d$.

As products of globally Lipschitz and bounded functions, the functions $\varphi\overline{a}_j$ are globally Lipschitz and bounded. Since $(t,\,z) \mapsto x_{t,j}(z)$ is also Lipschitz in z with a constant uniform in $t \in [0,\,\tau]$, it follows that each function $w_j(t,\,z) = (1-t)^{-d}\varphi(x_{t,j}(z))\overline{a}_j(x_{t,j}(z))$ is also bounded and globally Lipschitz, with Lipschitz constant uniform in $t \in [0,\,\tau]$. All of these functions are also continuous in $(t,\,z)$, as products and compositions of continuous functions. Hence the sum $S(t,\,z) = \sum_{\ell=1}^M w_\ell(t,\,z)$ is also continuous on $[0,\,\tau] \times \mathbb{R}^d$, globally Lipschitz in z uniformly in $t \in [0,\,\tau]$.

Now, let K be a compact set. Since $(t,\,z) \mapsto x_{t,j}(z)$ is continuous on the compact $[0,\,\tau] \times K$, the set

$$K' = \{x_{t,j}(z) :\, t \in [0,\,\tau],\; z \in K,\; 1 \leqslant j \leqslant M\}$$

is also compact. The Gaussian φ is bounded from below by a positive constant $c_\varphi > 0$ on K' and $\overline{a}_j$ is bounded from below everywhere by $(\min_j q_j)^k$. It follows that for every $(t,\,z) \in [0,\,\tau] \times K$, $w_j(t,\,z) \geqslant (1-t)^{-d} c_\varphi (\min_j q_j)^k > 0$, and the sum $S(t,\,z)$ is also bounded from below on $[0,\,\tau] \times K$ by a positive constant. Therefore each ratio w_j/S is continuous on $[0,\,\tau] \times K$, and Lipschitz in z on K with a constant uniform in $t \in [0,\,\tau]$. Since, by Lemma B.1,

$$m_t^{\overline{\pi}_k}(z) = \sum_{j=1}^M v_j\,\frac{w_j(t,\,z)}{S(t,\,z)} \text{ and } u_t^{\overline{\pi}_k}(z) = \frac{m_t^{\overline{\pi}_k}(z) - z}{1-t}$$

the same result holds for $(t,\,z) \mapsto m_t^{\overline{\pi}_k}(z)$ and $(t,\,z) \mapsto u_t^{\overline{\pi}_k}(z)$. □

Absolute continuity alone is insufficient for the regularity of the target velocity $u_t^{\overline{\pi}_k}$: when μ has disconnected support, the velocity can even be discontinuous, as the following example shows.

Example B.3. In dimension 1, define $S_+ = (\frac{1}{2},\,1)$ and $S_- = -S_+$. Fix $p \in (0,\,1)$, let μ be the probability measure with density $f = 2p\mathbb{1}_{S_+} + 2(1-p)\mathbb{1}_{S_-}$, and $\nu = \mathrm{Unif}\{-1,\,1\}$. Then for every $z_0 \in (0,\,1/7)$, setting $t_0 = \frac{1+2z_0}{3}$, the map $(t,\,z) \mapsto u_t^{\overline{\pi}_1}(z)$ is discontinuous at $(t_0,\,z_0)$, even when restricted to either coordinate direction (see Figure 9).

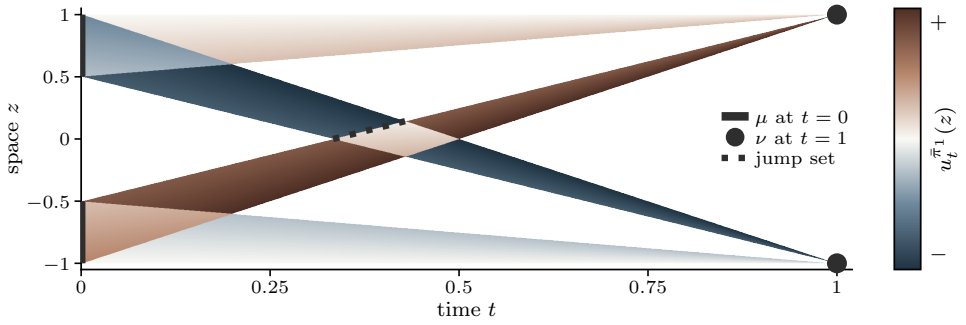


Figure 9: The velocity field $u_t^{\overline{\pi}_1}$ described in Example B.3 exhibits jump discontinuities.

B.2.2 Pre-terminal flow

Proof of Proposition 4.2(2). Denote $u_t = u_t^{\overline{\pi}_k}$, $\rho_t = \rho_t^{\overline{\pi}_k}$, and $R_{\mathcal{Y}} = \max_{1 \leqslant j \leqslant M} \|v_j\|$. The proof relies on combining Pierret et al. (2026b, Theorem 2.7) and Pierret et al. (2026b, Proposition 3.2). The main difference is that the velocity field $u_t^{\overline{\pi}_k}$ is locally Lipschitz with a constant that may blow up as $t \to 1$. For this reason, we work on $[0,\, \tau]$ for some $\tau < 1$, and since τ is arbitrary, this gives the conclusion on $[0,\, 1)$.

Pierret et al. (2026b, Proposition 3.2) ensures that the pair $(\rho^{\overline{\pi}_k},\, u^{\overline{\pi}_k})$ satisfies the continuity equation, since μ and ν have finite second moments. Moreover, for every $t \in [0,\, \tau]$ and every $z \in \mathbb{R}^d$, since $m_t^{\overline{\pi}_k}(z) \in \mathrm{conv}(\mathcal{Y})$, we have $\|m_t^{\overline{\pi}_k}(z)\| \leqslant R_{\mathcal{Y}}$. Thus by Lemma 4.1,

$$\|u_t(z)\| = \frac{\|m_t^{\overline{\pi}_k}(z) - z\|}{1-t} \leqslant \frac{R_{\mathcal{Y}} + \|z\|}{1-\tau}. \tag{19}$$

Hence, for every compact $K \subseteq \mathbb{R}^d$, $\sup_{z \in K} \int_0^\tau \|u_t(z)\|\, \mathrm{d}t < +\infty$. Moreover, the Lipschitz constant L of Proposition 4.2(1) is integrable on $[0,\, \tau]$. Therefore, applying Pierret et al. (2026b, Theorem 2.7) on $[0,\, \tau]$ gives the existence and uniqueness of $\phi_t^{\overline{\pi}_k}$, and the property $(\phi_t^{\overline{\pi}_k})_\sharp \mu = \rho_t$ on $[0,\, \tau]$. The continuity of $x \mapsto \phi_t^{\overline{\pi}_k}(x)$ for every fixed $t \in [0,\, \tau]$ is a consequence of the inequality (19) and the Cauchy–Lipschitz theorem (Pierret et al., 2026b, Section 2.2). □

B.2.3 Terminal flow

We adapt the analysis of Wan et al. (2025) to the expected batch OT setting. Fix $\delta \in (0,\, \mathrm{sep}(\mathcal{Y})/8)$. The proof is divided into four steps.

1. *Uniform concentration near atoms.* We show that the posterior mean $m_t^{\overline{\pi}_k}(z)$ is exponentially close to v_j, uniformly for $z \in B(v_j, 2\delta)$ and t close enough to 1.
2. *Trapping near an atom.* We show that once a trajectory enters $B(v_j, \delta)$ sufficiently close to terminal time, it remains in $B(v_j, 2\delta)$ and converges to v_j.
3. *Almost-sure entry.* We show that for μ-almost every initial point, the trajectory enters one of the balls $B(v_j, \delta)$ along a sequence of times $t_n \uparrow 1$, and therefore converges to the corresponding atom.
4. *Terminal map and push-forward.* We define the terminal map as the almost-sure limit of the flow and prove that it pushes μ forward to ν.

Step 1: Uniform concentration near atoms. Let $h = 1 - t$. By Lemma B.1,

$$m_t^{\overline{\pi}_k}(z) = \sum_{i=1}^M v_i \frac{w_i(t,\, z)}{\sum_{\ell=1}^M w_\ell(t,\, z)} \quad \text{and} \quad w_i(t,\, z) = h^{-d} \varphi\left(\frac{z - (1-h)v_i}{h}\right) \overline{a}_i\left(\frac{z-(1-h)v_i}{h}\right).$$

Fix $z \in B(v_j,\, 2\delta)$ and $i \neq j$. Since $\delta < \mathrm{sep}(\mathcal{Y})/8$, for h small enough,

$$\|z - (1-h)v_i\|^2 - \|z - (1-h)v_j\|^2 \geqslant \frac{\mathrm{sep}(\mathcal{Y})^2}{4}. \tag{20}$$

Using $\overline{a}_i \leqslant 1$ and $\overline{a}_j \geqslant q_{\min}^k$ from Lemma B.2, it follows that

$$\frac{w_i(t,\, z)}{w_j(t,\, z)} \leqslant q_{\min}^{-k} \exp\!\left(-\frac{\operatorname{sep}(\mathcal{Y})^2}{8h^2}\right).$$

Summing over $i \neq j$ yields, as $\sum_{\ell=1}^M w_\ell(t,\, z) \geqslant w_j(t,\, z)$,

$$1 - \frac{w_j(t,\, z)}{\sum_{\ell=1}^M w_\ell(t,\, z)} \leqslant (M-1) q_{\min}^{-k} \exp\!\left(-\frac{\operatorname{sep}(\mathcal{Y})^2}{8(1-t)^2}\right).$$

Hence

$$\|m_t^{\overline{\pi}_k}(z) - v_j\| \leqslant \operatorname{diam}(\mathcal{Y}) \sum_{i \neq j} \frac{w_i(t,\, z)}{\sum_{\ell=1}^M w_\ell(t,\, z)} \leqslant \operatorname{diam}(\mathcal{Y})(M-1) q_{\min}^{-k} \exp\!\left(-\frac{\operatorname{sep}(\mathcal{Y})^2}{8(1-t)^2}\right).$$

Thus, choosing $t_\delta \in (0,1)$ close enough to 1 so that (20) holds for all $t \in [t_\delta, 1)$, all $z \in B(v_j, 2\delta)$, all j, and all $i \neq j$, we obtain

$$\|m_t^{\overline{\pi}_k}(z) - v_j\| \leqslant A \exp\left(-\frac{a}{(1-t)^2}\right),$$

for every j, every $t \in [t_\delta, 1)$, and every $z \in B(v_j, 2\delta)$, with

$$A = \operatorname{diam}(\mathcal{Y})(M-1) q_{\min}^{-k}, \qquad a = \frac{\operatorname{sep}(\mathcal{Y})^2}{8}.$$

Step 2: Trapping near an atom. Increase t_δ, if necessary, so that

$$\int_{t_\delta}^1 \frac{A}{1-s} \exp\left(-\frac{a}{(1-s)^2}\right) \mathrm{d}s < \delta.$$

This preserves the conclusion of Step 1, since increasing t_δ only restricts the range of times under consideration.

Fix $x \in \mathbb{R}^d$, $j \in \{1, \ldots, M\}$ and $t_* \in [t_\delta,\, 1)$ such that $\|\phi_{t_*}^{\overline{\pi}_k}(x) - v_j\| < \delta$. For $t \in [t_*,\, 1)$, set $z_t = \phi_t^{\overline{\pi}_k}(x)$ and $y_t = z_t - v_j$. With $\xi_t = m_t^{\overline{\pi}_k}(z_t) - v_j$, the ODE and the definition of ξ_t give

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{y_t}{1-t}\right) = \frac{\xi_t}{(1-t)^2}.$$

Hence, for every $t \in [t_*,\, 1)$,

$$y_t = \frac{1-t}{1-t_*} y_{t_*} + (1-t) \int_{t_*}^t \frac{\xi_s}{(1-s)^2} \,\mathrm{d}s. \tag{21}$$

Let T be the supremum of the times $r \in [t_*,\, 1)$ such that $\|y_t\| < 2\delta$ for every $t \in [t_*,\, r]$. By continuity and the assumption at time t_*, $T > t_*$. For every $t < T$, the trajectory remains in $B(v_j,\, 2\delta)$ on $[t_*,\, t]$, hence

$$\forall s \in [t_*,\, t], \quad \|\xi_s\| \leqslant A \exp\left(-\frac{a}{(1-s)^2}\right). \tag{22}$$

Using (21) and the inequality $1 - t \leqslant 1 - s$ for $s \leqslant t$, we obtain

$$\begin{aligned}
\|y_t\| &\leqslant \frac{1-t}{1-t_*}\|y_{t_*}\| + (1-t)\int_{t_*}^t \frac{\|\xi_s\|}{(1-s)^2}\,\mathrm{d}s \\
&< \delta + \int_{t_*}^t \frac{A}{1-s} \exp\!\left(-\frac{a}{(1-s)^2}\right) \mathrm{d}s \\
&< 2\delta.
\end{aligned}$$

Thus $\|y_t\| < 2\delta$ for every $t \in [t_*,\, T)$. Letting $t \uparrow T$ in the preceding estimate gives $\|y_T\| < 2\delta$ if $T < 1$. Hence, by continuity, there exists $\varepsilon \in (0,\, 1-T)$ such that $\|y_t\| < 2\delta$ for all $t \in [T,\, T+\varepsilon]$, contradicting the definition of T as a supremum. Therefore $T = 1$.

It remains to prove convergence. Since $T = 1$, (22) holds on $[t_*, 1)$. The first term in (21) converges to zero as $t \uparrow 1$. For the second term, fix $\varepsilon > 0$ and choose $T_0 \in [t_*, 1)$ such that $\int_{T_0}^1 \frac{A}{1-s} \exp\Big(-\frac{a}{(1-s)^2}\Big)\, \mathrm{d}s < \varepsilon$. Then, for $t \geqslant T_0$,

$$(1-t)\int_{t_*}^{t} \frac{\|\xi_s\|}{(1-s)^2}\,\mathrm{d}s \leqslant (1-t)\int_{t_*}^{T_0} \frac{\|\xi_s\|}{(1-s)^2}\,\mathrm{d}s + \int_{T_0}^{t} \frac{A}{1-s}\exp\Big(-\frac{a}{(1-s)^2}\Big)\,\mathrm{d}s.$$

The first term tends to zero as $t \uparrow 1$, while the second term is bounded by ε. Since $\varepsilon > 0$ was arbitrary, the integral term in the representation of y_t also tends to zero. Hence $y_t \to 0$, that is, $\phi_t^{\overline{\pi}^k}(x) \to v_j$ as $t \uparrow 1$.

Step 3: Almost-sure entrance. Let $\mathcal{Y}_\delta = \{z \in \mathbb{R}^d : \mathrm{dist}(z, \mathcal{Y}) < \delta\}$. Since $Z_t = (1-t)X + tY \to Y \in \mathcal{Y}$ almost surely as $t \uparrow 1$, we have $\rho_t^{\overline{\pi}^k}(\mathcal{Y}_\delta^c) \to 0$. Choose a sequence $t_n \uparrow 1$ with $t_n \geqslant t_\delta$ and $\rho_{t_n}^{\overline{\pi}^k}(\mathcal{Y}_\delta^c) \leqslant 2^{-n}$. By Proposition 4.2(2), $(\phi_{t_n}^{\overline{\pi}^k})_\sharp \mu = \rho_{t_n}^{\overline{\pi}^k}$. Hence

$$\sum_{n\geqslant 1} \mu\Big(\{x : \phi_{t_n}^{\overline{\pi}^k}(x) \notin \mathcal{Y}_\delta\}\Big) = \sum_{n\geqslant 1} \rho_{t_n}^{\overline{\pi}^k}(\mathcal{Y}_\delta^c) < \infty.$$

By the Borel–Cantelli lemma, for μ-almost every x there exists $N(x)$ such that $\phi_{t_n}^{\overline{\pi}^k}(x) \in \mathcal{Y}_\delta$ for all $n \geqslant N(x)$. Fix such an x, and choose an index $n_0 \geqslant N(x)$. Since $\delta < \mathrm{sep}(\mathcal{Y})/8$, the balls $B(v_1, \delta), \ldots, B(v_M, \delta)$ are pairwise disjoint. Thus there is a unique $j(x) \in \{1, \ldots, M\}$ such that $\phi_{t_{n_0}}^{\overline{\pi}^k}(x) \in B(v_{j(x)}, \delta)$. Applying step 2 with $t_* = t_{n_0}$ and $j = j(x)$ yields $\phi_t^{\overline{\pi}^k}(x) \to v_{j(x)}$ as $t \uparrow 1$. Consequently, $\phi_{t_n}^{\overline{\pi}^k}(x) \in B(v_{j(x)}, \delta)$ for all sufficiently large n.

Step 4: Terminal map and push-forward. By step 3, for μ-almost every x the limit $\phi^{\overline{\pi}^k}(x) = \lim_{t\uparrow 1} \phi_t^{\overline{\pi}^k}(x)$ exists and belongs to $\mathcal{Y}$. Choose a sequence $t_n \uparrow 1$ as in step 3. Since each $\phi_{t_n}^{\overline{\pi}^k}$ is continuous by Proposition 4.2(2), the set

$$E = \{x \in \mathbb{R}^d \,:\, \lim_{n\to\infty} \phi_{t_n}^{\overline{\pi}^k}(x) \text{ exists and belongs to } \mathcal{Y}\}$$

is Borel, and the limit map is Borel on E. Since $\mu(E) = 1$, extending it by the fixed value v_1 on E^c yields a Borel map $\phi^{\overline{\pi}^k} : \mathbb{R}^d \to \mathcal{Y}$. Let f be bounded and continuous. By dominated convergence,

$$\int_{\mathbb{R}^d} f(\phi^{\overline{\pi}^k}(x))\,\mathrm{d}\mu(x) = \lim_{n\to\infty} \int_{\mathbb{R}^d} f(\phi_{t_n}^{\overline{\pi}^k}(x))\,\mathrm{d}\mu(x).$$

Using Proposition 4.2(2),

$$\int_{\mathbb{R}^d} f(\phi_{t_n}^{\overline{\pi}^k}(x))\,\mathrm{d}\mu(x) = \int_{\mathbb{R}^d} f(z)\,\mathrm{d}\rho_{t_n}^{\overline{\pi}^k}(z).$$

Since $\rho_t^{\overline{\pi}^k} \rightharpoonup \nu$ as $t \uparrow 1$, it follows that

$$\int_{\mathbb{R}^d} f(\phi^{\overline{\pi}^k}(x))\,\mathrm{d}\mu(x) = \int_{\mathbb{R}^d} f(y)\,\mathrm{d}\nu(y).$$

Therefore $(\phi^{\overline{\pi}^k})_\sharp \mu = \nu$. This proves Proposition 4.2(3).

B.3 Rates for the terminal flow map (Corollary 4.3)

Proof of Corollary 4.3. By Proposition 4.2, $(\phi^{\overline{\pi}^k})_\sharp \mu = \nu$, thus $W_2^2(\mu, \nu) \leqslant \int_{\mathbb{R}^d} \|\phi^{\overline{\pi}^k}(x) - x\|^2\,\mathrm{d}\mu(x)$. We upper bound the previous quantity using Liu et al. (2023, Theorem D.5): we derive their argument in our present setting for completeness. Fix $\tau \in (0, 1)$. By the Cauchy–Schwarz inequality,

$$\|\phi_\tau^{\overline{\pi}^k}(x) - x\|^2 \leqslant \tau \int_0^\tau \|u_t^{\overline{\pi}^k}(\phi_t^{\overline{\pi}^k}(x))\|^2\,\mathrm{d}t.$$

Since $(\phi_t^{\overline{\pi}_k})_\sharp \mu = \rho_t^{\overline{\pi}_k}$, integrating over x yields

$$\int_{\mathbb{R}^d} \|\phi_\tau^{\overline{\pi}_k}(x) - x\|^2 \,\mathrm{d}\mu(x) \leqslant \tau \int_0^\tau \int_{\mathbb{R}^d} \|u_t^{\overline{\pi}_k}(z)\|^2 \,\mathrm{d}\rho_t^{\overline{\pi}_k}(z)\,\mathrm{d}t.$$

For each $t < 1$, Jensen's inequality applied to $u_t^{\overline{\pi}_k}(Z_t) = \mathbb{E}_{(X,\,Y)\sim\overline{\pi}_k}[Y - X \mid Z_t]$ gives

$$\int_{\mathbb{R}^d} \|u_t^{\overline{\pi}_k}(z)\|^2 \,\mathrm{d}\rho_t^{\overline{\pi}_k}(z) \leqslant \int_{\mathbb{R}^d\times\mathbb{R}^d} \|y - x\|^2 \,\mathrm{d}\overline{\pi}_k(x,\,y).$$

Consequently,

$$\int_{\mathbb{R}^d} \|\phi_\tau^{\overline{\pi}_k}(x) - x\|^2 \,\mathrm{d}\mu(x) \leqslant \tau^2 \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\overline{\pi}_k(x,\,y) \leqslant \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\overline{\pi}_k(x,\,y).$$

Letting $\tau \uparrow 1$, by Fatou's lemma and Proposition 4.2(3),

$$\int_{\mathbb{R}^d} \|\phi^{\overline{\pi}_k}(x) - x\|^2 \,\mathrm{d}\mu(x) \leqslant \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\overline{\pi}_k(x,\,y).$$

Using Propositions 3.2 and 3.3, the first rate follows. Now let $X \sim \mu$ and apply (18) to $\pi = \mathrm{Law}\big(X,\,\phi^{\overline{\pi}_k}(X)\big) \in \Pi(\mu,\,\nu)$: there exists a constant $C > 0$, independent of k, such that

$$\mu\big(\{x \in \mathbb{R}^d,\ \phi^{\overline{\pi}_k}(x) \neq T^\star(x)\}\big) \leqslant C\left(\int_{\mathbb{R}^d} \|\phi^{\overline{\pi}_k}(x) - x\|^2 \,\mathrm{d}\mu(x) - W_2^2(\mu,\nu)\right)^{1/2} = O(k^{-1/2}).$$

Finally, for every j, $\Gamma_{k,\,j} \triangle C_j^\star \subseteq \{x \in \mathbb{R}^d,\ \phi^{\overline{\pi}_k}(x) \neq T^\star(x)\}$, hence $\mu(\Gamma_{k,\,j} \triangle C_j^\star) = O(k^{-1/2})$. □

B.3.1 Non-monotonicity of the transport cost of the terminal map

We build an instance in which the transport cost $\int_{\mathbb{R}^d} \|\phi^{\overline{\pi}_k}(x) - x\|^2 \mathrm{d}\mu(x)$ is non-monotonic in k. We use the below result, which shows that $\phi^{\mu\otimes\nu}$ is the OT map when μ and ν can be factored as the independent product of d one-dimensional measures. This extends Hertrich et al. (2025, Theorem 3(ii)), which tackles non-degenerate Gaussian μ and ν with jointly diagonalizable covariances.

Proposition B.4. *Let $\mu = \alpha_1 \otimes \cdots \otimes \alpha_d$ and $\nu = \beta_1 \otimes \cdots \otimes \beta_d$ with $\alpha_i,\ \beta_i \in \mathcal{P}_2(\mathbb{R})$ for $1 \leqslant i \leqslant d$. Assume that the independent FM flow between μ and ν exists, is unique on $[0,\,1)$, and admits a terminal map $\phi^{\mu\otimes\nu}$ pushing μ to ν. Then*

$$\int_{\mathbb{R}^d} \|\phi^{\mu\otimes\nu}(x) - x\|^2 \,\mathrm{d}\mu(x) = W_2^2(\mu,\,\nu).$$

Proof. Let $(X,\,Y) = (X_1, \ldots, X_d,\, Y_1, \ldots, Y_d) \sim \mu \otimes \nu$ and denote by $(e_i)_{1\leqslant i\leqslant d}$ the canonical basis of $\mathbb{R}^d$. Let $Z_t = (1-t)X + tY$, so that $Z_{t,i} = (1-t)X_i + tY_i$. For fixed $t \in [0,\,1)$ and $\mathcal{L}(Z_t)$-almost every $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$, by Lemma 4.1, the velocity is

$$u_t^{\mu\otimes\nu}(z) = \frac{m_t^{\mu\otimes\nu}(z) - z}{1-t},$$

where, by independence of the pairs $(X_i,\,Y_i)$,

$$\begin{aligned} m_t^{\mu\otimes\nu}(z) &= \mathbb{E}[Y \mid Z_t = z] = \sum_{i=1}^d \mathbb{E}[Y_i \mid Z_t = z]\, e_i = \sum_{i=1}^d \mathbb{E}[Y_i \mid Z_{t,i} = z_i]\, e_i \\ &= \big(m_t^{\alpha_1\otimes\beta_1}(z_1),\, \ldots,\, m_t^{\alpha_d\otimes\beta_d}(z_d)\big). \end{aligned}$$

Thus the d-dimensional ODE decouples into d one-dimensional ODEs. Writing ϕ_t for the corresponding flow maps, uniqueness gives

$$\phi_t^{\mu\otimes\nu}(x_1,\,\ldots,\,x_d) = \big(\phi_t^{\alpha_1\otimes\beta_1}(x_1),\,\ldots,\,\phi_t^{\alpha_d\otimes\beta_d}(x_d)\big)$$

for μ-almost every x and every $0 \leqslant t < 1$. Passing to the terminal limit and using Fubini yields the same identity for the terminal maps, μ-almost everywhere. Moreover, taking coordinate marginals in

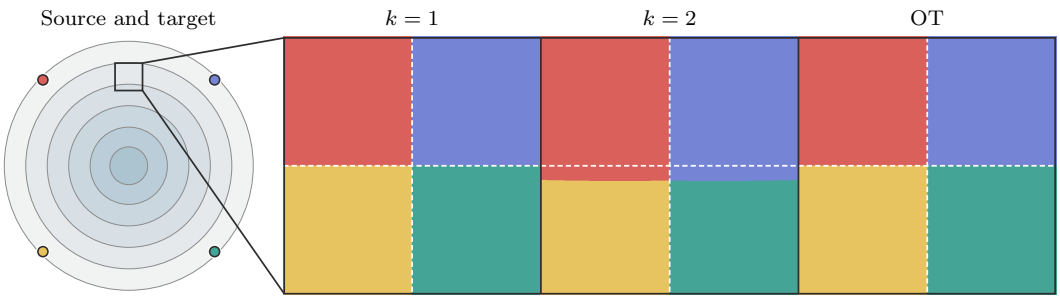


Figure 10: Terminal cells for the counterexample to the monotonicity of the transport cost of $\phi^{\overline{\pi}_k}$.

$\phi^{\mu\otimes\nu}_{\#}\mu = \nu$ shows that $(\phi^{\alpha_i\otimes\beta_i})_{\#}\alpha_i = \beta_i$ for every $1 \leqslant i \leqslant d$. By Pierret et al. (2026b, Note 4.2), in one dimension, the independent FM map is the OT map, hence $\phi^{\alpha_i\otimes\beta_i} = T^{\star}_{\alpha_i\to\beta_i}$ for every $1 \leqslant i \leqslant d$, and

$$\int_{\mathbb{R}^d} \|\phi^{\mu\otimes\nu}(x) - x\|^2 \,\mathrm{d}\mu(x) = \sum_{i=1}^{d} \int_{\mathbb{R}} |\phi^{\alpha_i\otimes\beta_i}(x_i) - x_i|^2 \,\mathrm{d}\alpha_i(x_i) = \sum_{i=1}^{d} W_2^2(\alpha_i,\, \beta_i).$$

Moreover, if $\pi \in \Pi(\mu,\, \nu)$, then its $(x_i,\, y_i)$-marginal belongs to $\Pi(\alpha_i,\, \beta_i)$, and

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\pi(x,\, y) = \sum_{i=1}^{d} \int_{\mathbb{R}^d\times\mathbb{R}^d} |x_i - y_i|^2 \,\mathrm{d}\pi(x,\, y) \geqslant \sum_{i=1}^{d} W_2^2(\alpha_i,\, \beta_i).$$

Taking the infimum over $\pi \in \Pi(\mu,\, \nu)$ gives

$$W_2^2(\mu,\, \nu) \geqslant \sum_{i=1}^{d} W_2^2(\alpha_i,\, \beta_i).$$

Conversely, if $\pi_i^{\star} \in \Pi(\alpha_i,\, \beta_i)$ is an optimal coupling for every $1 \leqslant i \leqslant d$, then $\pi_1^{\star} \otimes \cdots \otimes \pi_d^{\star} \in \Pi(\mu,\, \nu)$ under the canonical identification $(\mathbb{R}\times\mathbb{R})^d \simeq \mathbb{R}^d \times \mathbb{R}^d$, and therefore

$$W_2^2(\mu,\, \nu) \leqslant \sum_{i=1}^{d} W_2^2(\alpha_i,\, \beta_i).$$

It follows that

$$\int_{\mathbb{R}^d} \|\phi^{\mu\otimes\nu}(x) - x\|^2 \,\mathrm{d}\mu(x) = W_2^2(\mu,\, \nu).$$

□

We then present in Figure 10 a numerical counterexample to the monotonicity of $\int_{\mathbb{R}^d} \|\phi^{\overline{\pi}_k}(x) - x\|^2 \,\mathrm{d}\mu(x)$. Let $\mu = \mathcal{N}(0,\, I_2)$ and $\nu = (\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1) \otimes (\frac{17}{20}\delta_{-1} + \frac{3}{20}\delta_1)$. By the preceding proposition, $\phi^{\overline{\pi}_1} = \phi^{\mu\otimes\nu}$ is the OT map, unique by Brenier's theorem. In contrast, we visually remark that $\phi^{\overline{\pi}_2}$ is suboptimal.

B.3.2 Persistent unbounded rays in the four-point example

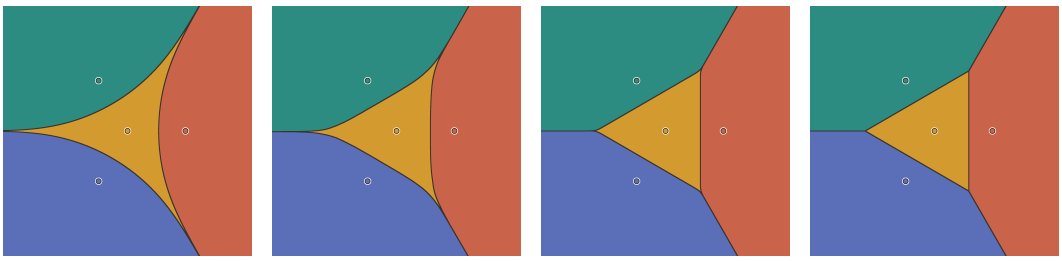
The convergence in Corollary 4.3 is a convergence in μ-measure, not a uniform convergence of the cells. We illustrate this by adapting the four-point counterexample of Pierret et al. (2026a, Section 5.1) to batch OT (see Figure 11). Consider $\nu = \frac{1}{4}\sum_{i=1}^{4} \delta_{v_i} \in \mathcal{P}(\mathbb{R}^2)$ supported on the four atoms

$$v_1 = \left(-\frac{1}{2},\, \frac{\sqrt{3}}{2}\right), \quad v_2 = \left(-\frac{1}{2},\, -\frac{\sqrt{3}}{2}\right), \quad v_3 = (1,\, 0), \quad v_4 = (0,\, 0).$$

For every $k \in \mathbb{N}^*$, the central FM cell $\Gamma_{k,\,4}$ contains the three rays

$$\left\{\left(-c\cos\left(\frac{2q\pi}{3}\right),\, -c\sin\left(\frac{2q\pi}{3}\right)\right),\ c \geqslant 0\right\} \quad \text{for } q \in \{1,\, 2,\, 3\}.$$

In particular, $\Gamma_{k,\,4}$ has an unbounded connected component for every finite k, whereas the corresponding Laguerre cell $C_4^{\star}$ is bounded, as recalled in Pierret et al. (2026a, Section 5.1).



(a) $k = 1$ (b) $k = 25$ (c) $k = 1\,000$ (d) OT plan $\pi^\star$

Figure 11: Terminal map cells in the symmetric four-point setting of Pierret et al. (2026a, Section 5.1).

Proof. We closely follow the proof of Pierret et al. (2026a, Lemma 9). The only change is that the independent softmax weights are replaced by the batch OT posterior weights. By Lemma 4.1,

$$u_t^{\overline{\pi}_k}(x) = \sum_{i=1}^{4} p_i^k(t,\, x)\frac{v_i - x}{1-t} \quad \text{with} \quad p_i^k(t,\, x) = \frac{\varphi\Big(\frac{x-tv_i}{1-t}\Big)\,\overline{a}_i^k\Big(\frac{x-tv_i}{1-t}\Big)}{\sum_{\ell=1}^{4}\varphi\Big(\frac{x-tv_\ell}{1-t}\Big)\,\overline{a}_\ell^k\Big(\frac{x-tv_\ell}{1-t}\Big)}.$$

By symmetry, it suffices to prove the claim for the ray $L_3 = \{(-c,\, 0),\ c \geqslant 0\}$. First, for $x = (b,\, 0)$, symmetry gives $p_1^k(t,\, x) = p_2^k(t,\, x)$, hence $u_t^{\overline{\pi}_k}(b,\, 0) \in \mathbb{R}(1,\, 0)$. Moreover, at the origin one has $p_1^k(t,\, 0) = p_2^k(t,\, 0) = p_3^k(t,\, 0)$, and since $v_1 + v_2 + v_3 = 0$, it follows that $u_t^{\overline{\pi}_k}(0) = 0$. By uniqueness of the pre-terminal flow, $\phi_t^{\overline{\pi}_k}(L_3) \subseteq L_3$. Let $b_0 \leqslant 0$ and write $\phi_t^{\overline{\pi}_k}(b_0,\, 0) = (b_t,\, 0)$: we prove that $b_t \to 0$ as $t \uparrow 1$.

Step 1: for every $\varepsilon > 0$, there exists $t_0 \in (0,\, 1)$ such that $b_t \geqslant -\frac{1}{2} - \varepsilon$ for all $t \geqslant t_0$. Indeed,

$$(1-t)\dot{b}_t = (p_1^k + p_2^k)\left(-\frac{1}{2} - b_t\right) + p_3^k(1 - b_t) - p_4^k b_t \geqslant -\frac{1}{2} - b_t,$$

where the weights are evaluated at $(t, (b_t, 0))$. Hence $\dot{b}_t \geqslant (-\frac{1}{2} - b_t)/(1-t)$, and Gronwall's lemma gives $b_t \geqslant (1-t)b_0 - t/2$, which tends to $-1/2$.

Step 2: we show that for $\varepsilon = 0.4$, there exist $t_1 \in (0,\, 1)$ and $c > 0$ such that, for every $t \geqslant t_1$ and every $b \in -\frac{1}{2} + [-\varepsilon,\, \varepsilon]$, $u_t^{\overline{\pi}_k}(b,\, 0) = \big(w_t^k(b),\, 0\big)$ with $w_t^k(b) \geqslant \frac{c}{1-t}$. Indeed,

$$(1-t)w_t^k(b) \geqslant -b - \frac{1}{2}\big(1 - p_4^k(t,\, (b,\, 0))\big) \geqslant \frac{1}{2} - \varepsilon - \frac{1}{2}\big(1 - p_4^k(t,\, (b,\, 0))\big).$$

It remains to show that $p_4^k(t,\, (b,\, 0)) \to 1$ uniformly on this interval. From the posterior formula,

$$p_4^k(t,(b,0)) = \left(1 + \sum_{i=1}^{3} \frac{\varphi\left(\frac{(b,0)-tv_i}{1-t}\right)\overline{a}_i^k\left(\frac{(b,0)-tv_i}{1-t}\right)}{\varphi\left(\frac{(b,0)-tv_4}{1-t}\right)\overline{a}_4^k\left(\frac{(b,0)-tv_4}{1-t}\right)}\right)^{-1}.$$

By Lemma B.2, $4^{-k} \leqslant \overline{a}_i^k \leqslant 1$, and $\|(b,\, 0) - tv_i\|^2 - \|(b,\, 0) - tv_4\|^2 \geqslant \delta > 0$, uniformly for $b \in -\frac{1}{2} + [-\varepsilon, \varepsilon]$ and t close enough to 1, for $i \in \{1,\, 2,\, 3\}$, hence

$$p_4^k(t,(b,0)) \geqslant \left(1 + 3 \cdot 4^k \exp\left(-\frac{\delta}{2(1-t)^2}\right)\right)^{-1} \xrightarrow[t\to 1]{} 1.$$

Step 3: let $r = \frac{1}{2} - \varepsilon = 0.1$. Since $\operatorname{sep}(\{v_1,\, v_2,\, v_3,\, v_4\}) = 1$, we have $r < \operatorname{sep}(\{v_1,\, v_2,\, v_3,\, v_4\})/8$. Choose t_r so that the trapping argument around v_4 from Step 2 of Appendix B.2.3 applies for the ball $B_r(v_4)$. Set $t_2 = \max(t_0,\, t_1,\, t_r)$. We claim that there exists $t \geqslant t_2$ such that $b_t > -\frac{1}{2} + \varepsilon$. Otherwise, Step 1 gives $b_t \in -\frac{1}{2} + [-\varepsilon,\, \varepsilon]$ for all $t \in [t_2,\, 1)$, and Step 2 gives $\dot{b}_t \geqslant c/(1-t)$. Thus $b_t \geqslant b_{t_2} + c\log((1-t_2)/(1-t))$, which diverges to $+\infty$, a contradiction.

Hence, for some $t \geqslant t_r$, the trajectory belongs to $B_r(v_4)$. By the trapping argument cited above, it converges to v_4. Thus $(b_0,\, 0) \in \Gamma_{k,\,4}$. Since $b_0 \leqslant 0$ was arbitrary, $L_3 \subseteq \Gamma_{k,\,4}$. □

B.4 Posterior concentration and one-step Euler accuracy (Section 4.3)

In the following proposition, we show that in the Gaussian-to-discrete setting, under the independent coupling, there exist positive constants C_1, C_2 (depending only on ν) such that for any $t \in [0, 1)$,

$$0 \leqslant 1 - \mathbb{E}\left[\max_{1\leqslant i\leqslant M} \mathbb{P}(Y = v_i \mid Z_t)\right] \leqslant C_1 \exp\left(-C_2 \frac{t^2}{(1-t)^2}\right).$$

The constant C_2 depends on ν through the separation of its support, which is a key reason why the posterior probabilities concentrate faster when the dimension increases: if the M points v_j are distributed uniformly on $[0, 1]^d$ for instance, then the average minimal separation of two points of ν is of the order $M^{-2/d}$, which increases with d.

Proposition B.5 (Posterior concentration and one-step Euler accuracy)**.** *Let $X \sim \mathcal{N}(0, I_d)$ and $Y \sim \mathrm{Unif}(\mathcal{Y})$ be independent, where $\mathcal{Y} = \{v_1, \dots, v_M\}$ are distinct atoms in $\mathbb{R}^d$. Denote $Z_t = (1-t)X + tY$ and $\mathrm{sep}(\mathcal{Y}) = \min_{1\leqslant i\neq j\leqslant M} \|v_i - v_j\|$. Then, for every $t \in [0, 1)$,*

$$0 \leqslant 1 - \mathbb{E}\left[\max_{1\leqslant j\leqslant M} \mathbb{P}(Y = v_j \mid Z_t)\right] \leqslant (M-1) \exp\left(-\frac{t^2 \,\mathrm{sep}(\mathcal{Y})^2}{8(1-t)^2}\right).$$

If, in addition, $\mathcal{Y} \subseteq [-1, 1]^d$, and if $\rho_t^{\mu\otimes\nu}$ is the distribution of Z_t, then for every $h \in [0, 1-t]$,

$$W_2\left((\mathrm{id} + h u_t^{\mu\otimes\nu})_\sharp \rho_t^{\mu\otimes\nu},\, \rho_{t+h}^{\mu\otimes\nu}\right) \leqslant \frac{2h\sqrt{d(M-1)}}{1-t} \exp\left(-\frac{t^2 \,\mathrm{sep}(\mathcal{Y})^2}{16(1-t)^2}\right).$$

Proof. At $t = 0$, independence gives $\mathbb{P}(Y = v_j \mid Z_0) = 1/M$, so the first inequality follows from $1 - 1/M \leqslant M - 1$. Now let $t \in (0, 1)$. Write $s = \frac{t}{1-t}$ and $W_t = \frac{Z_t}{1-t} = X + sY$. Almost surely, $\mathbb{P}(Y = v_j \mid Z_t) = \mathbb{P}(Y = v_j \mid W_t)$. For $w \in \mathbb{R}^d$, choose a measurable maximizer

$$\hat{\jmath}(w) \in \operatorname*{arg\,max}_{1\leqslant j\leqslant M} p_j(w) \quad \text{where} \quad p_j(w) = \mathbb{P}(Y = v_j \mid W_t = w).$$

Almost surely, $\max_{1\leqslant j\leqslant M} \mathbb{P}(Y = v_j \mid W_t) = \mathbb{P}(Y = v_{\hat{\jmath}(W_t)} \mid W_t)$, and therefore,

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant M} \mathbb{P}(Y = v_j \mid W_t)\right] = \mathbb{P}(Y = v_{\hat{\jmath}(W_t)}).$$

Now, as X and Y are independent,

$$W_t \mid \{Y = v_j\} = X + s v_j \sim \mathcal{N}(s v_j, I_d),$$

hence the density of $W_t \mid \{Y = v_j\}$ is proportional to $\exp(-\frac{1}{2}\|w - s v_j\|^2)$, and

$$\hat{\jmath}(w) \in \operatorname*{arg\,min}_{1\leqslant j\leqslant M} \|w - s v_j\|.$$

Fix $i \in \{1, \dots, M\}$ and condition on $Y = v_i$. On the event that $\hat{\jmath}(W_t) = j$ with $j \neq i$, $\|W_t - s v_j\| \leqslant \|W_t - s v_i\|$, hence

$$\|X + s(v_i - v_j)\| \leqslant \|X\|.$$

Squaring and simplifying, we obtain $\langle X, v_i - v_j\rangle \leqslant -\frac{s}{2}\|v_i - v_j\|^2$. Since X is a standard Gaussian, $\langle X, v_i - v_j\rangle \sim \mathcal{N}(0, \|v_i - v_j\|^2)$, and denoting by Φ the c.d.f. of the standard Gaussian, we obtain

$$\mathbb{P}(\hat{\jmath}(W_t) = j \mid Y = v_i) \leqslant \Phi\left(-\frac{s}{2}\|v_i - v_j\|\right) \leqslant \Phi\left(-\frac{s}{2}\,\mathrm{sep}(\mathcal{Y})\right).$$

Summing over $j \in \{1, \dots, M\}\backslash\{i\}$ gives

$$\mathbb{P}(\hat{\jmath}(W_t) \neq i \mid Y = v_i) \leqslant (M-1)\Phi\left(-\frac{s}{2}\,\mathrm{sep}(\mathcal{Y})\right).$$

Since Y is uniform, averaging over i yields

$$\mathbb{P}(Y \neq v_{\hat{\jmath}(W_t)}) = \frac{1}{M}\sum_{i=1}^{M} \mathbb{P}(\hat{\jmath}(W_t) \neq i \mid Y = v_i) \leqslant (M-1)\Phi\left(-\frac{s}{2}\,\mathrm{sep}(\mathcal{Y})\right).$$

Using $\Phi(-u) \leqslant e^{-u^2/2}$ for $u \geqslant 0$, we conclude that

$$\mathbb{P}(Y = v_{\hat{j}(W_t)}) \geqslant 1 - (M-1)\exp\left(-\frac{s^2}{8}\operatorname{sep}(\mathcal{Y})^2\right).$$

Recalling that $s = \frac{t}{1-t}$ and that conditioning on Z_t or on W_t is equivalent gives the claim. See also Dodson et al. (2026, Lemma B.1) for a related posterior-concentration bound.

We now prove the second inequality for $t \in [0,\,1)$, assuming $\mathcal{Y} \subseteq [-1,\,1]^d$. Let $\tilde{Z}_{t,\,h} = Z_t + hu_t^{\mu\otimes\nu}(Z_t)$. Then $\tilde{Z}_{t,\,h} \sim (\mathrm{id} + hu_t^{\mu\otimes\nu})_\sharp \rho_t^{\mu\otimes\nu}$, and $Z_{t+h} \sim \rho_{t+h}^{\mu\otimes\nu}$. Hence, by the definition of W_2,

$$W_2^2\left((\mathrm{id} + hu_t^{\mu\otimes\nu})_\sharp \rho_t^{\mu\otimes\nu},\, \rho_{t+h}^{\mu\otimes\nu}\right) \leqslant \mathbb{E}[\|\tilde{Z}_{t,\,h} - Z_{t+h}\|^2]. \tag{23}$$

Now $Z_{t+h} = Z_t + h(Y - X)$, and by Lemma 4.1,

$$\tilde{Z}_{t,\,h} - Z_{t+h} = \frac{h}{1-t}\left(m_t^{\mu\otimes\nu}(Z_t) - Y\right).$$

Moreover, denoting $p_\ell(z) = \mathbb{P}(Y = v_\ell \mid Z_t = z)$ and $p_*(z) = \max_\ell p_\ell(z)$, as $\|v_i - v_j\| \leqslant 2\sqrt{d}$,

$$\begin{aligned}\mathbb{E}[\|Y - m_t^{\mu\otimes\nu}(Z_t)\|^2 \mid Z_t = z] &= \sum_{i=1}^M \|v_i - m_t^{\mu\otimes\nu}(z)\|^2 p_i(z) = \frac{1}{2}\sum_{1\leqslant i,j\leqslant M} p_i(z)p_j(z)\|v_i - v_j\|^2 \\ &\leqslant 2d\left(1 - \sum_{j=1}^M p_j(z)^2\right).\end{aligned}$$

Moreover, since $\sum_{j=1}^M p_j(z)^2 \geqslant p_*(z)^2$, we obtain

$$1 - \sum_{j=1}^M p_j(z)^2 \leqslant 1 - p_*(z)^2 = (1 - p_*(z))(1 + p_*(z)) \leqslant 2(1 - p_*(z)).$$

Hence taking expectations,

$$\mathbb{E}[\|Y - m_t^{\mu\otimes\nu}(Z_t)\|^2] \leqslant 4d\left(1 - \mathbb{E}\left[\max_{1\leqslant j\leqslant M}\mathbb{P}(Y = v_j \mid Z_t)\right]\right). \tag{24}$$

Combining (23) and (24), and then applying the first inequality yields the result. □

B.5 Two-atom tractable case (Proposition 4.4)

Classical Euler bounds are poorly adapted to the present regime: they are driven by global Lipschitz estimates on the velocity field, whereas in the binary model the relevant phenomenon is the formation of a sharp transition layer near the origin. We therefore argue directly with the expected batch OT plan in the binary case. From now on, we work with $\mu = \mathcal{N}(0,\,1)$ and $\nu = \mathrm{Unif}\{-1,\,+1\}$.

B.5.1 Auxiliary formulas

For $k \in \mathbb{N}^*$ and $x \in \mathbb{R}$, define $q_k(x) = \mathbb{P}_{(X,\,Y)\sim\overline{\pi}_k}(Y = -1 \mid X = x)$.

Lemma B.6. *For $x \in \mathbb{R}$, let $u = \Phi(x)$, where Φ denotes the c.d.f. of a standard Gaussian. Let $N_- \sim \mathrm{Bin}(k,\,\frac{1}{2})$ and define $B_r = \mathbb{P}(N_- \geqslant r)$ for $r \in \{1,\dots,k\}$. Then $q_k(-x) = 1 - q_k(x)$, and*

$$q_k(x) = \sum_{j=0}^{k-1}\binom{k-1}{j}u^j(1-u)^{k-1-j}B_{j+1}. \tag{25}$$

Proof. The coupling is symmetric under the map $(x,\,y) \mapsto (-x,\,-y)$. Thus if $(X,\,Y) \sim \overline{\pi}_k$, then $(-X,\,-Y) \sim \overline{\pi}_k$, and $q_k(-x) = \mathbb{P}(Y = -1 \mid X = -x) = \mathbb{P}(Y = +1 \mid X = x) = 1 - q_k(x)$.

Since the OT map in dimension 1 is the monotone rearrangement (Peyré and Cuturi, 2019, Section 2.6), a sample from $\overline{\pi}_k$ may be generated as follows. Draw $(X_1,\dots,X_k,\,Y_1,\dots,Y_k) \sim \mu^{\otimes k} \otimes \nu^{\otimes k}$, let

$Y_{(1)} \leqslant \cdots \leqslant Y_{(k)}$ be the sorted target labels, choose an index $J \sim \mathrm{Unif}\{1, \ldots, k\}$ independently, let L be the rank of X_J among $(X_1, \ldots, X_k)$, and output $(X,\, Y) = (X_J,\, Y_{(L)})$.

Condition on the event $\{X = X_J = x\}$. The remaining $k-1$ source samples are still i.i.d. $\mathcal{N}(0,\, 1)$ under this conditioning. Let

$$J_x = \sum_{i \neq J} \mathbb{1}_{\{X_i \leqslant x\}} \sim \mathrm{Bin}(k-1,\, u) \quad \text{and} \quad N_- = \sum_{j=1}^{k} \mathbb{1}_{\{Y_j = -1\}} \sim \mathrm{Bin}(k,\, 1/2).$$

Then the rank of X_J is $L = J_x + 1$, and $Y_{(r)} = -1$ if and only if $r \leqslant N_-$. Therefore, for every $j \in \{0, \ldots, k-1\}$,

$$\mathbb{P}(Y = -1 \mid X = x,\, J_x = j) = \mathbb{P}(Y_{(j+1)} = -1) = \mathbb{P}(N_- \geqslant j+1) = B_{j+1}.$$

Averaging over the law of J_x yields (25). □

For $t \in [0,\, 1)$ and $z \in \mathbb{R}$, set $s = t/(1-t) \in [0,\, \infty)$ and $\xi = z/(1-t) \in \mathbb{R}$.

Lemma B.7. *For every $k \in \mathbb{N}^*$, $t \in [0,\, 1)$ and $z \in \mathbb{R}$,*

$$m_t^{\overline{\pi}_k}(z) = \tanh\!\left(s\xi + \frac{1}{2}\log\frac{q_k(s-\xi)}{q_k(s+\xi)}\right) \quad \textit{and} \quad m_t^{\mu\otimes\nu}(z) = \tanh\!\left(\frac{tz}{(1-t)^2}\right).$$

Proof. Fix $k \in \mathbb{N}^*$, $t \in [0,\, 1)$ and $z \in \mathbb{R}$. Since $Y \in \{-1,\, 1\}$,

$$\begin{aligned} m_t^{\overline{\pi}_k}(z) &= \mathbb{E}[Y \mid Z_t = z] \\ &= \mathbb{P}(Y = +1 \mid Z_t = z) - \mathbb{P}(Y = -1 \mid Z_t = z) \\ &= \tanh\!\left(\frac{1}{2}\log\psi_t^{\overline{\pi}_k}(z)\right) \quad \text{with} \quad \psi_t^{\overline{\pi}_k}(z) = \frac{\mathbb{P}(Y = +1 \mid Z_t = z)}{\mathbb{P}(Y = -1 \mid Z_t = z)}. \end{aligned}$$

For $y \in \{-1,\, 1\}$, applying Bayes' formula gives

$$\mathbb{P}(Y = y \mid Z_t = z) = \frac{f_{Z_t|Y=y}(z)\mathbb{P}(Y = y)}{f_{Z_t}(z)}.$$

The marginal probabilities $\mathbb{P}(Y = y) = 1/2$ and the density $f_{Z_t}(z)$ cancel in the odds ratio. On the event $(Z_t = z,\, Y = y)$, the identity $Z_t = (1-t)X + tY$ forces $x_y = (z - ty)/(1-t)$. Changing variables between Z_t and X in the conditional density gives, up to the common factor $(1-t)^{-1}$,

$$f_{Z_t|Y=y}(z) \propto \varphi(x_y)\mathbb{P}(Y = y \mid X = x_y),$$

where φ is the standard Gaussian density. Consequently,

$$\psi_t^{\overline{\pi}_k}(z) = \frac{\varphi(x_{+1})}{\varphi(x_{-1})}\frac{\mathbb{P}(Y = +1 \mid X = x_{+1})}{\mathbb{P}(Y = -1 \mid X = x_{-1})}. \tag{26}$$

The ratio of Gaussian densities satisfies

$$\log\frac{\varphi(x_{+1})}{\varphi(x_{-1})} = \frac{x_{-1}^2 - x_{+1}^2}{2} = \frac{2tz}{(1-t)^2} = 2s\xi.$$

Moreover, $\mathbb{P}(Y = -1 \mid X = x) = q_k(x)$ and $\mathbb{P}(Y = +1 \mid X = x) = q_k(-x)$ by Lemma B.6. Since $x_{+1} = \xi - s$ and $x_{-1} = \xi + s$, taking logarithms in (26) yields

$$\log\psi_t^{\overline{\pi}_k}(z) = 2s\xi + \log\frac{q_k(s-\xi)}{q_k(s+\xi)}.$$

This proves the first formula. Then Lemma B.6 gives $q_1 \equiv 1/2$ and the second formula follows. □

Lemma B.8 (Structure of the Euler iterates in the binary case)**.** *Fix $k \in \mathbb{N}^*$ and $n \geqslant 1$. For every $0 \leqslant m \leqslant n$, the map $f_{m,\,n}^{\overline{\pi}_k}$ is odd. In the independent case $\overline{\pi}_1 = \mu \otimes \nu$, the map $f_{m,\,n}^{\mu\otimes\nu}$ is increasing and concave for $x \geqslant 0$. Finally,* $\mathrm{im}(f_n^{\overline{\pi}_k}) \subseteq [-1,\, 1]$. *Therefore, $\mathscr{E}_{n,\,k} = 2\,\mathbb{E}_{X\sim\mu}[(1 - f_n^{\overline{\pi}_k}(X))\mathbb{1}_{\{X>0\}}]$.*

Proof. We first prove oddness. The initial map $f^{\overline{\pi}_k}_{0,\,n} = \mathrm{id}$ is odd. Assume that $f^{\overline{\pi}_k}_{m,\,n}$ is odd for some $0 \leqslant m < n$, and write $r = n - m$ and $t = m/n$. By Lemma 4.1, the Euler recursion (4) becomes

$$f^{\overline{\pi}_k}_{m+1,\,n}(x) = \left(1 - \frac{1}{r}\right) f^{\overline{\pi}_k}_{m,\,n}(x) + \frac{1}{r}\, m_t^{\overline{\pi}_k}\big(f^{\overline{\pi}_k}_{m,\,n}(x)\big)\,.$$

By Lemma B.7, the map $m_t^{\overline{\pi}_k}$ is odd, hence the right-hand side is odd, and induction gives the claim.

We next specialize to $\overline{\pi}_1 = \mu \otimes \nu$. By Lemma B.7, $m_t^{\mu\otimes\nu}(z) = \tanh(\frac{tz}{(1-t)^2})$. For $z \geqslant 0$, this map is increasing and concave. Since $f^{\mu\otimes\nu}_{0,\,n} = \mathrm{id}$ is increasing and concave for $x \geqslant 0$, and since the composition of increasing concave maps is increasing and concave, the recursion shows by induction that $f^{\mu\otimes\nu}_{m,\,n}$ is increasing and concave for $x \geqslant 0$ for every m.

For the terminal bound, take $m = n - 1$, so that $r = 1$. Then $f_n^{\overline{\pi}_k}(x) = m^{\overline{\pi}_k}_{1-1/n}(f^{\overline{\pi}_k}_{n-1,\,n}(x))$. By Lemma B.7, this map takes values in $[-1,\,1]$. Finally, as we are in dimension one, Pierret et al. (2026b, Note 4.2) gives $\phi^{\overline{\pi}_k}(X) = \mathrm{sign}(X)$ almost surely, hence $\mathscr{E}_{n,\,k} = \mathbb{E}\big[|f_n^{\overline{\pi}_k}(X) - \mathrm{sign}(X)|\big]$. For $x > 0$, the terminal bound gives $|f_n^{\overline{\pi}_k}(x) - 1| = 1 - f_n^{\overline{\pi}_k}(x)$. For $x < 0$, oddness and the terminal bound give $|f_n^{\overline{\pi}_k}(x) + 1| = 1 - f_n^{\overline{\pi}_k}(-x)$. Splitting the expectation over the positive and negative half-lines and using the symmetry of the Gaussian density yields the formula. □

B.5.2 One Euler step: the effect of the OT batch size

For one Euler step, the error is entirely statistical: it measures the probability that the expected batch OT plan assigns the wrong sign.

Proof of the asymptotic for $\mathscr{E}_{1,\,k}$ in Proposition 4.4. Since $f_1^{\overline{\pi}_k}(x) = \mathbb{E}[Y \mid X = x]$, we have

$$|f_1^{\overline{\pi}_k}(X) - \mathrm{sign}(X)| = |\mathbb{E}[Y \mid X] - \mathrm{sign}(X)| = \mathbb{E}[|Y - \mathrm{sign}(X)| \mid X].$$

Since both Y and $\mathrm{sign}(X)$ take values in $\{-1,\,+1\}$, the integrand is equal to 0 if $Y = \mathrm{sign}(X)$ and to 2 otherwise. Therefore $\mathscr{E}_{1,\,k} = 2\,\mathbb{P}(Y \neq \mathrm{sign}(X))$.

We compute this probability through the rank representation of $\overline{\pi}_k$. Let $U_1, \dots, U_k, V_1, \dots, V_k$ be i.i.d. uniform random variables on $[0,\,1]$, let $L \sim \mathrm{Unif}\{1, \dots, k\}$ be independent of everything else, and set

$$X = \Phi^{-1}(U_{(L)}) \quad \text{and} \quad Y = \begin{cases} -1, & \text{if } V_{(L)} \leqslant 1/2, \\ +1, & \text{otherwise.} \end{cases}$$

Then $(X,\,Y) \sim \overline{\pi}_k$. Define

$$B = \sum_{i=1}^{k} \mathbb{1}_{\{U_i \leqslant 1/2\}} \quad \text{and} \quad B' = \sum_{i=1}^{k} \mathbb{1}_{\{V_i \leqslant 1/2\}}.$$

Then B and B' are independent variables with law $\mathrm{Bin}(k,\,1/2)$, and

$$\begin{aligned} \mathrm{sign}(X) = -1 &\iff U_{(L)} < 1/2 \iff L \leqslant B, \\ Y = -1 &\iff V_{(L)} \leqslant 1/2 \iff L \leqslant B'. \end{aligned}$$

Therefore, $Y \neq \mathrm{sign}(X)$ if and only if $\min(B,\,B') < L \leqslant \max(B,\,B')$. Conditionally on $(B,\,B')$, the number of such ranks is exactly $|B - B'|$, so

$$\mathbb{P}(Y \neq \mathrm{sign}(X) \mid B,\,B') = \frac{|B - B'|}{k}.$$

Taking expectations gives

$$\mathscr{E}_{1,\,k} = \frac{2}{k}\,\mathbb{E}|B - B'|. \tag{27}$$

Write $B - B' = \sum_{i=1}^{k} D_i$, where the D_i are i.i.d. with

$$\mathbb{P}(D_i = 1) = \mathbb{P}(D_i = -1) = \frac{1}{4} \quad \text{and} \quad \mathbb{P}(D_i = 0) = \frac{1}{2}.$$

Then $\mathbb{E}[D_i] = 0$ and $\mathrm{Var}(D_i) = 1/2$. By the central limit theorem,

$$\frac{B - B'}{\sqrt{k}} \xrightarrow[k\to+\infty]{\mathrm{d}} Z \quad \text{with} \quad Z \sim \mathcal{N}(0,\, 1/2).$$

Since the variables $(B - B')/\sqrt{k}$ have uniformly bounded second moments, their absolute values are uniformly integrable. Therefore

$$\frac{1}{\sqrt{k}}\,\mathbb{E}|B - B'| \xrightarrow[k\to+\infty]{} \mathbb{E}|Z| = \frac{1}{\sqrt{\pi}}.$$

Combining this with (27) yields $\mathscr{E}_{1,\,k} \sim 2/\sqrt{\pi k}$, and taking logarithms gives the result. □

B.5.3 Many Euler steps: the effect of the numerical integration budget

We now fix the coupling to be the independent one, $\overline{\pi}_1 = \mu \otimes \nu$, and study the effect of increasing the number of Euler steps. The Euler map is an odd increasing concave sigmoid approximating the sign map. The error is therefore controlled by the width of the transition layer near the origin, and that width is governed by the derivative at the origin.

From now on, write $f_{m,\,n} = f_{m,\,n}^{\mu\otimes\nu}$ and $f_n = f_{n,\,n}$. By Lemma B.7, for $0 \leqslant m < n$, if $r = n - m$ and $\alpha_{m,\,n} = mn/r^2$, then

$$f_{m+1,\,n}(x) = \left(1 - \frac{1}{r}\right) f_{m,\,n}(x) + \frac{1}{r}\,\tanh(\alpha_{m,\,n} f_{m,\,n}(x))\,. \tag{28}$$

Fix once and for all a constant $A > 2\pi/\sqrt{3}$, and define $x_n = \frac{An^{2/3}}{f'_{n,\,n}(0)}$.

Lemma B.9. *For $0 \leqslant m < n$, writing $r = n - m$,*

$$f'_{m+1,\,n}(0) = c_r\, f'_{m,\,n}(0) \quad \textit{and} \quad c_r = 1 + \frac{n^2 - nr - r^2}{r^3}.$$

Consequently, for $n \geqslant 2$, $\log f'_{n,\,n}(0) = \sum_{r=1}^{n} \log c_r$.

Proof. Differentiate (28) at 0. Since $f_{m,\,n}(0) = 0$ by Lemma B.8 and $\tanh'(0) = 1$,

$$f'_{m+1,\,n}(0) = \left(1 - \frac{1}{r} + \frac{\alpha_{m,\,n}}{r}\right) f'_{m,\,n}(0).$$

Now $\alpha_{m,\,n} = mn/r^2 = n(n-r)/r^2$, which gives the announced coefficient c_r. □

Lemma B.10. $\frac{1}{n^{2/3}} \log f'_{n,\,n}(0) \xrightarrow[n\to\infty]{} \frac{2\pi}{\sqrt{3}}$.

Proof. The $r = 1$ term in Lemma B.9 is $\log(n(n-1)) = O(\log n)$, which is $o(n^{2/3})$. For $r \geqslant 2$, factor $c_r = (1 + u_r)(1 - x_r)$ with $u_r = \frac{n^2}{r^3}$ and $x_r = \frac{n/r^2 + 1/r}{1+u_r}$. Then

$$\log c_r = \log\left(1 + \frac{n^2}{r^3}\right) + \log(1 - x_r).$$

Splitting the sum at $R_n = \lceil n^{2/3} \rceil$ shows that $\sum_{r=2}^{n} |\log(1 - x_r)| = o(n^{2/3})$. Hence

$$\log f'_{n,\,n}(0) = \sum_{r=2}^{n} \log\left(1 + \frac{n^2}{r^3}\right) + o(n^{2/3}).$$

By integral comparison,

$$\sum_{r=2}^{n} \log\left(1 + \frac{n^2}{r^3}\right) = \int_1^n \log\left(1 + \frac{n^2}{u^3}\right)\,\mathrm{d}u + o(n^{2/3}).$$

With the change of variables $u = n^{2/3} s$, the right-hand side becomes

$$n^{2/3} \int_{n^{-2/3}}^{n^{1/3}} \log(1 + s^{-3})\,\mathrm{d}s + o(n^{2/3}).$$

Since $\int_0^\infty \log(1 + s^{-3})\,\mathrm{d}s = 2\pi/\sqrt{3}$, the claim follows. □

Lemma B.11. *For every* $0 \leqslant m \leqslant n-2$, $f'_{m,\,n}(0) \leqslant \frac{8(n-m)}{n^2}\, f'_{n-1,\,n}(0)$.

Proof. Let $r = n - m$. By Lemma B.9,

$$f'_{n-1,\,n}(0) = f'_{m,\,n}(0) \prod_{q=2}^{r} c_q.$$

For $q \in \{3, \dots, r\}$, since $n(n-q) \geqslant 0$,

$$c_q = 1 - \frac{1}{q} + \frac{n(n-q)}{q^3} \geqslant 1 - \frac{1}{q} = \frac{q-1}{q}.$$

By telescoping,

$$\prod_{q=3}^{r} c_q \geqslant \prod_{q=3}^{r} \frac{q-1}{q} = \frac{2}{r}.$$

Also $c_2 = (n^2 - 2n + 4)/8 \geqslant n^2/16$. Thus

$$f'_{n-1,\,n}(0) \geqslant f'_{m,\,n}(0)\, \frac{n^2}{8r},$$

which is the desired inequality. □

Lemma B.12. *For all sufficiently large* n, $0 \leqslant 1 - f_{n,\,n}(x_n) \leqslant 2e^{-An^{2/3}}$.

Proof. Fix $0 \leqslant m \leqslant n-2$ and write $r = n - m$. Since $f_{m,\,n}$ is concave for $x \geqslant 0$ and vanishes at 0,

$$0 \leqslant f_{m,\,n}(x_n) \leqslant f'_{m,\,n}(0) x_n.$$

By Lemma B.11, $f'_{m,\,n}(0) \leqslant 8r f'_{n-1,\,n}(0)/n^2$. Using also $\alpha_{m,\,n} \leqslant n^2/r^2$, we obtain

$$\alpha_{m,\,n} f_{m,\,n}(x_n) \leqslant \frac{8}{r}\, f'_{n-1,\,n}(0) x_n.$$

Since $x_n = \frac{An^{2/3}}{f'_{n,\,n}(0)} = \frac{An^{2/3}}{n(n-1)f'_{n-1,\,n}(0)}$, and $r \geqslant 2$,

$$\alpha_{m,\,n} f_{m,\,n}(x_n) \leqslant \frac{4An^{2/3}}{n(n-1)} \leqslant 8An^{-4/3} =: \delta_n.$$

Thus $\delta_n \to 0$, and for all large enough n we have $\delta_n \leqslant 1$. On $[0,\, \delta_n]$,

$$\tanh(u) \geqslant u - \frac{u^3}{3} \geqslant u\left(1 - \frac{\delta_n^2}{3}\right).$$

Applying this estimate in (28) at x_n yields

$$\begin{aligned} f_{m+1,\,n}(x_n) &\geqslant \left(1 - \frac{1}{r}\right) f_{m,\,n}(x_n) + \frac{1}{r}\left(1 - \frac{\delta_n^2}{3}\right) \alpha_{m,\,n} f_{m,\,n}(x_n) \\ &= \left(c_r - \frac{\delta_n^2}{3}\frac{\alpha_{m,\,n}}{r}\right) f_{m,\,n}(x_n). \end{aligned}$$

Since $c_r - \frac{\delta_n^2}{3}\frac{\alpha_{m,\,n}}{r} - \left(1 - \frac{\delta_n^2}{3}\right) c_r = \frac{\delta_n^2}{3}\left(1 - \frac{1}{r}\right) \geqslant 0$, we obtain

$$f_{m+1,\,n}(x_n) \geqslant \left(1 - \frac{\delta_n^2}{3}\right) c_r f_{m,\,n}(x_n).$$

Iterating from $m = 0$ to $m = n-2$ gives

$$f_{n-1,\,n}(x_n) \geqslant \left(1 - \frac{\delta_n^2}{3}\right)^{n-1} \left(\prod_{r=2}^{n} c_r\right) x_n = \left(1 - \frac{\delta_n^2}{3}\right)^{n-1} f'_{n-1,\,n}(0) x_n.$$

Since $n\delta_n^2 \to 0$, the prefactor is at least $1/2$ for all large enough n. Therefore

$$n(n-1) f_{n-1,\,n}(x_n) \geqslant \frac{A}{2} n^{2/3}.$$

Finally, $f_{n,\,n}(x) = \tanh(n(n-1) f_{n-1,\,n}(x))$, and $1 - \tanh(u) \leqslant 2e^{-2u}$ for $u \geqslant 0$. Hence

$$1 - f_{n,\,n}(x_n) \leqslant 2\exp(-2n(n-1) f_{n-1,\,n}(x_n)) \leqslant 2e^{-An^{2/3}}.$$

□

Lemma B.13. *There exist constants $c,\, C > 0$ such that, for all large enough n,*

$$\frac{c}{f'_{n,\,n}(0)} \leqslant \mathscr{E}_{n,\,1} \leqslant \frac{Cn^{2/3}}{f'_{n,\,n}(0)}.$$

In particular, $\log \mathscr{E}_{n,\,1} = -\log f'_{n,\,n}(0) + O(\log n)$.

Proof. We use Lemma B.8, which gives the formula for $\mathscr{E}_{n,\,1}$ and the monotonicity and concavity of $f_{n,\,n}$ for $x \geqslant 0$.

For the lower bound, concavity and $f_{n,\,n}(0) = 0$ give $f_{n,\,n}(x) \leqslant f'_{n,\,n}(0)x$ for $x \geqslant 0$. Set $a_n = 1/(2f'_{n,\,n}(0))$. Then $1 - f_{n,\,n}(x) \geqslant 1/2$ on $[0,\, a_n]$, and

$$\mathscr{E}_{n,\,1} = 2\mathbb{E}_{X\sim\mu}\big[(1 - f_{n,\,n}(X))\mathbb{1}_{\{X>0\}}\big] \geqslant \int_0^{a_n} \varphi(x)\,\mathrm{d}x.$$

By Lemma B.10, $f'_{n,\,n}(0) \to \infty$, so $a_n \to 0$. For large enough n, $a_n \leqslant 1$, hence

$$\mathscr{E}_{n,\,1} \geqslant \frac{\varphi(1)}{2f'_{n,\,n}(0)}.$$

For the upper bound, $f_{n,\,n}$ is increasing and $0 \leqslant 1 - f_{n,\,n}(x) \leqslant 1$ for $x \geqslant 0$. Thus

$$(1 - f_{n,\,n}(X))\mathbb{1}_{\{X>0\}} \leqslant \mathbb{1}_{\{0\leqslant X\leqslant x_n\}} + (1 - f_{n,\,n}(x_n))\mathbb{1}_{\{X\geqslant x_n\}}.$$

Taking expectations and using Lemma B.12,

$$\mathscr{E}_{n,\,1} \leqslant 2\mu([0,\, x_n]) + 4e^{-An^{2/3}} \leqslant 2\varphi(0)\frac{An^{2/3}}{f'_{n,\,n}(0)} + 4e^{-An^{2/3}}.$$

Since $A > 2\pi/\sqrt{3}$, Lemma B.10 implies $e^{-An^{2/3}} \leqslant n^{2/3}/f'_{n,\,n}(0)$ for all large enough n. This proves the upper bound, and the logarithmic equivalence follows. □

Proof of the asymptotic for $\mathscr{E}_{n,\,1}$ in Proposition 4.4. By Lemma B.13, $\log \mathscr{E}_{n,\,1} = -\log f'_{n,\,n}(0) + O(\log n)$. By Lemma B.10, $\log f'_{n,\,n}(0) \sim (2\pi/\sqrt{3})n^{2/3}$. Therefore

$$\log \mathscr{E}_{n,\,1} \sim -\frac{2\pi}{\sqrt{3}}n^{2/3}.$$

□

B.5.4 Numerical verification

Numerically, $\mathscr{E}_{n,\,1}$ decreases stretched-exponentially in n, whereas $\mathscr{E}_{1,\,k}$ decreases only polynomially in k. Moreover, the asymptotic equivalent $2/\sqrt{\pi k}$ is accurate already for $\mathscr{E}_{1,\,k}$, while $\exp(-(2\pi/\sqrt{3})n^{2/3})$ captures the decay rate of $\mathscr{E}_{n,\,1}$ up to a multiplicative constant (Figure 12).

C Experimental details

This appendix gives the implementation details needed to reproduce the experiments.

C.1 Small-scale experiments

These experiments were run on CPU on an Apple M2 MacBook Air with 8 CPU cores and 16 GB RAM, using macOS 26.1, Python 3.9.6, and PyTorch 2.4.1.

C.1.1 Flow map induced by $\phi^{\overline{\pi}_k}$ (Figure 3)

For a given $x \in \mathbb{R}^2$, we approximate $\phi^{\overline{\pi}_k}(x)$ by integrating the flow ODE with 50 uniform second-order Runge–Kutta steps. When $k = 1$, we use the closed-form expression for $\mu \otimes \nu$ in Lemma 4.1. When $k > 1$, we approximate $u_t^{\overline{\pi}_k}$ by Monte Carlo over $R = 2{,}000$ common random batches.

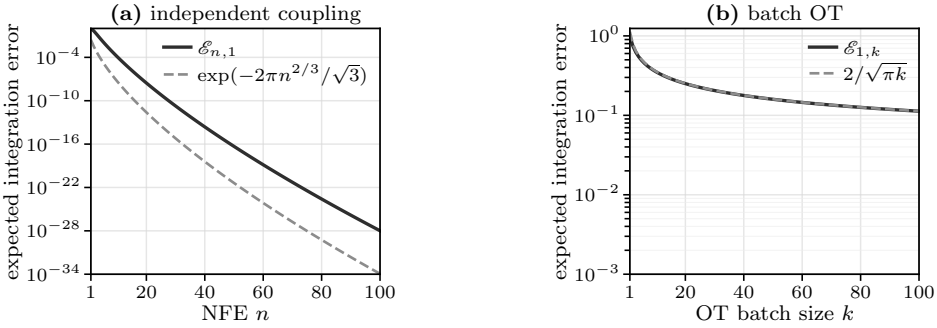


Figure 12: Numerical verification of the two asymptotic regimes in Proposition 4.4.

C.1.2 Posterior concentration vs. k and d (Figure 4)

We consider $M = 100$ target atoms in ν. When varying the OT batch size k, we keep one atom cloud fixed in dimension $d = 20$. When varying the dimension, we draw a new atom cloud uniformly in $[-1,\, 1]^d$ for each value of d. Each curve averages $2{,}000$ explicit Euler trajectories with 100 function evaluations. We approximate $u_t^{\overline{\pi}_k}$ as in Appendix C.1.1. The largest standard error over sampled trajectories is below 0.006 on the plotted 0–1 scale, so standard-error bands are not drawn.

C.1.3 Two-atom case (Figure 5)

In the two-atom case, we do not estimate the expected batch OT velocity by Monte Carlo. We use the closed-form expression for the minibatch OT assignment probability q_k from Lemma B.6, insert it into the closed-form conditional mean in Lemma B.7, and evaluate $\mathscr{E}_{n,\,k}$ using the one-dimensional integral representation in Lemma B.8. Thus the computation is closed form up to a one-dimensional quadrature, which enables the precise curves reported in Figure 5.

C.2 Larger-scale experiments

These experiments were run on several machines running AlmaLinux 9.7, each equipped with an Intel Core i7-14700K CPU, 62 GiB RAM, and a single NVIDIA RTX 4000 Ada Generation GPU with 20 GB VRAM, using Python 3.9.25 and PyTorch 2.5.0 (Paszke et al., 2019) with CUDA 12.4 and cuDNN 9.1. Their wall-clock time is reported in Table 1.

Table 1: Total wall-clock time for the experiments, including exploratory runs.

Task	**Dataset**	**Wall-clock time**
Training	CIFAR-10	4,884.3 h
	SVHN	2,151.0 h
Inference & FID	CIFAR-10	957.8 h
	SVHN	248.9 h
Estimation of $\mathscr{E}_{n,\,k}$	Synthetic	$\approx$ 540 h
	Total	$\approx$ 8,780 h

C.2.1 Rates on expected batch OT cost and plan (Figure 2)

To avoid using stochastic semidiscrete solvers to compute $W_2^2(\mu,\, \nu)$, we follow the idea of Hundrieser et al. (2024, Section 4): we first discretize μ by a large empirical measure, and then choose the target weights so that the nearest-neighbor projection is optimal for this empirical reference problem. Specifically, fix $N_{\mathrm{ref}} = 2 \times 10^6$, draw $X_1, \ldots, X_{N_{\mathrm{ref}}} \sim \mu$ independently, and let

$i^\star(x) \in \arg\min_{1 \leqslant i \leqslant M} \|x - v_i\|^2$. We set

$$\mu_{\mathrm{ref}} = \frac{1}{N_{\mathrm{ref}}} \sum_{\ell=1}^{N_{\mathrm{ref}}} \delta_{X_\ell}, \qquad \widehat{a}_i = \frac{1}{N_{\mathrm{ref}}} \sum_{\ell=1}^{N_{\mathrm{ref}}} \mathbb{1}_{\{i^\star(X_\ell)=i\}}, \qquad \nu = \sum_{i=1}^{M} \widehat{a}_i \delta_{v_i}.$$

By construction, the map $T^\star : X \mapsto v_{i^\star(X)}$ pushes μ_{ref} forward to ν. Since each $v_{i^\star(X_\ell)}$ is a nearest neighbor of X_ℓ, this map is optimal, and if N_{ref} is large enough,

$$W_2^2(\mu,\, \nu) \approx W_2^2(\mu_{\mathrm{ref}},\, \nu) = \frac{1}{N_{\mathrm{ref}}} \sum_{\ell=1}^{N_{\mathrm{ref}}} \|X_\ell - v_{i^\star(X_\ell)}\|^2.$$

For each Monte Carlo repetition $r \in \{1, \dots, 2000\}$ and each $k \in \{2, \lfloor 2^{1.25} \rfloor, \lfloor 2^{1.5} \rfloor, \dots, 2^{15}\}$, we draw batches $(\mathbf{x}_k^{(r)},\, \mathbf{y}_k^{(r)}) \sim \mu^{\otimes k} \otimes \nu^{\otimes k}$. We then compute the OT plan $\widehat{\pi}_k^{(r)}$ between $\widehat{\mu}_k^{(r)}$ and $\widehat{\nu}_k^{(r)}$, and store $W_2^2(\widehat{\mu}_k^{(r)},\, \widehat{\nu}_k^{(r)})$ and $\mathbb{E}_{(X,\,Y)\sim\widehat{\pi}_k^{(r)}}[\|Y - T^\star(X)\|^2]$.

C.2.2 Multi-atom synthetic case (Figure 6)

Atoms are sampled once uniformly in $[-1,\, 1]^{20}$ and then kept fixed. For each pair $(n,\, k)$, we consider the target velocity field $u_t^{\overline{\pi}_k}$, starting from 10^4 independent samples $X \sim \mathcal{N}(0,\, I_{20})$. We compare a coarse explicit Euler integration with n function evaluations to a second-order Runge–Kutta reference integration with 200 function evaluations. We report the empirical mean error

$$\widehat{\mathscr{E}}_{n,\,k} = \frac{1}{10^4} \sum_{r=1}^{10^4} \|f_n^{\overline{\pi}_k}(X_r) - \widetilde{f}_{200}^{\overline{\pi}_k}(X_r)\|,$$

where $\widetilde{f}_{200}^{\overline{\pi}_k}$ denotes the Runge–Kutta reference map. We approximate $u_t^{\overline{\pi}_k}$ as in Appendix C.1.1. The error bars in Figure 6 are standard errors over the sampled initial conditions.

C.2.3 Expected batch OT cost on image datasets (Figure 7)

For each $k \in \{1,\, 2,\, 4,\, 8, \dots, 8192\}$, we estimate the expected batch OT cost by averaging over $B_k = \lceil \frac{2^{20}}{k} \rceil$ independent OT problems of size k. Thus, for every k, the total number of sampled pairs is kept close to 2^{20}. We report the empirical mean of the empirical OT costs over these B_k batches, where each batch matches k i.i.d. Gaussian samples in $\mathbb{R}^{3072}$ to k samples drawn uniformly from the training set. In the left panel of Figure 7 standard errors over these B_k independent OT batches are below plotting resolution.

C.2.4 Image generation experiments (Figure 7)

The flow model is the U-Net from the TorchCFM library (Tong et al., 2024) and the FID is computed using `clean-fid` (Parmar et al., 2022). See Table 3 for more details.

We measure the computational tradeoff between increasing OT batch size k during training and using more function evaluations at inference. On CIFAR-10, we measure the wall-clock time of one training step as a function of the OT batch size k, and the wall-clock time of one U-Net evaluation during sampling. For $k = 1$, we use the independent coupling, obtained by pairing a Gaussian noise batch with a CIFAR-10 data batch at random, without solving OT. For $k > 1$, we construct pairs by solving exact OT between empirical batches of size k, using POT (Flamary et al., 2021). Each timed training step includes pair construction, linear interpolation, U-Net forward pass, loss computation, backward pass, gradient clipping, Adam update, scheduler step, and EMA update.

Up to $k = 1024$, training time for CIFAR-10 is nearly unchanged and increases by about 3% relative to the independent coupling. At $k = 8192$, training becomes 55% more expensive. One U-Net function evaluation takes 0.861 ± 0.0002 ms per image. Therefore, on our hardware, for a $400{,}000$-step run, if N function evaluations are saved at inference, the extra training cost is amortized after roughly $\frac{45.5}{N}$ million generated images.

Final-checkpoint FID values for both image datasets are reported in Table 4. Larger OT batch sizes are associated with lower training loss and lower FID during training (Figure 13). We display a rolling mean with window $w = 100$ for clarity, which does not change the ordering of the curves.

C.3 Software licenses

The experiments were run using `POT` (Flamary et al. (2021), MIT license), `torchcfm` 1.0.7 (Tong et al. (2024), MIT), `torchdiffeq` 0.2.5 (Chen et al. (2018), MIT), `torchdyn` 1.0.6 (Poli et al. (2021), Apache-2.0), `torchvision` 0.20.0 (TorchVision maintainers and contributors (2024), BSD-3-Clause), `numpy` 2.0.2 (Harris et al. (2020), BSD-style), `scipy` 1.13.1 (Virtanen et al. (2020), BSD-style), and `clean-fid` 0.1.35 (Parmar et al. (2022), MIT).

D Broader impact

This work is primarily theoretical: it studies the coupling induced by repeated minibatch optimal transport and its consequences for flow matching. It may help make generative modeling methods easier to analyze and optimize, but any resulting gains are likely to be realized unevenly in a field shaped by unequal access to compute and deployment infrastructure. Improvements in generative modeling can also be used in systems for spam and deceptive content, although this paper does not release a high-capability generative model.

Table 2: CIFAR-10 wall-clock training cost as a function of OT batch size. Costs assume $400{,}000$ training steps on one GPU. Values are mean $\pm$ standard error.

k	Step time (ms)	Training cost (GPU-h)	Extra cost vs. $k = 1$ (GPU-h)
1	180.79 ± 0.03	20.09 ± 0.00	0.00
2	186.33 ± 0.04	20.70 ± 0.00	0.62 ± 0.01
4	184.07 ± 0.01	20.45 ± 0.00	0.36 ± 0.00
8	182.83 ± 0.15	20.31 ± 0.02	0.23 ± 0.02
16	182.37 ± 0.14	20.26 ± 0.02	0.18 ± 0.02
32	182.24 ± 0.03	20.25 ± 0.00	0.16 ± 0.00
64	181.94 ± 0.03	20.22 ± 0.00	0.13 ± 0.00
128	181.81 ± 0.01	20.20 ± 0.00	0.11 ± 0.00
256	182.09 ± 0.02	20.23 ± 0.00	0.14 ± 0.00
512	183.06 ± 0.02	20.34 ± 0.00	0.25 ± 0.00
1024	185.80 ± 0.02	20.64 ± 0.00	0.56 ± 0.00
2048	195.89 ± 0.18	21.77 ± 0.02	1.68 ± 0.02
4096	221.50 ± 0.25	24.61 ± 0.03	4.52 ± 0.03
8192	278.81 ± 0.17	30.98 ± 0.02	10.89 ± 0.02

Table 3: Flow model training and FID evaluation protocol.

	Setting	**Value**
TRAINING	Architecture	U-Net from `torchcfm` (Tong et al., 2024)
	Input size	$3 \times 32 \times 32$
	Base channels	128
	Residual blocks per level	2
	Channel multipliers	$[1, 2, 2, 2]$
	Attention heads	4
	Head channels	64
	Attention resolution	16
	Dropout	0.1
	OT batch size	$k \in \{1,\ 2,\ 4,\ 8, \ldots, 8192\}$
	Optimization minibatch size	64
	Optimizer	Adam, learning rate 2×10^{-4}
	Learning-rate schedule	Linear warmup for 5,000 steps, then constant
	Gradient clipping	Global gradient norm clipped to 1.0
	Exponential moving average	Decay 0.9999
	Other fixed hyperparameters	$\sigma = 0$, total training steps $= 400{,}000$
FID EVAL.	Implementation	`clean-fid` (Parmar et al., 2022)
	Sampler	Forward Euler
	Discretization	NFE for CIFAR-10: $\{10, 20, 50, 100\}$; for SVHN: $\{5, 10, 50, 100\}$
	Number of generated samples	50,000
	FID batch size	1024
	Image conversion	Clamp to $[-1, 1]$ and convert to $[0, 255]$

Table 4: FID at the final checkpoint. Values are mean ± std. dev. over six training seeds.

	k	NFE 10	NFE 20	NFE 50	NFE 100
CIFAR-10	1	14.69 ± 0.27	9.05 ± 0.14	6.28 ± 0.10	5.34 ± 0.10
	2	14.17 ± 0.24	9.06 ± 0.12	6.40 ± 0.09	5.46 ± 0.08
	4	13.89 ± 0.24	9.14 ± 0.19	6.50 ± 0.15	5.56 ± 0.13
	8	13.57 ± 0.11	9.14 ± 0.08	6.52 ± 0.07	5.59 ± 0.07
	16	13.39 ± 0.12	9.14 ± 0.13	6.57 ± 0.11	5.63 ± 0.10
	32	13.42 ± 0.22	9.24 ± 0.14	6.64 ± 0.07	5.68 ± 0.05
	64	13.14 ± 0.14	9.11 ± 0.11	6.53 ± 0.10	5.58 ± 0.09
	128	13.18 ± 0.16	9.20 ± 0.14	6.59 ± 0.10	5.63 ± 0.08
	256	12.92 ± 0.16	9.09 ± 0.17	6.53 ± 0.16	5.57 ± 0.14
	512	12.76 ± 0.23	9.03 ± 0.19	6.51 ± 0.12	5.54 ± 0.09
	1024	12.67 ± 0.20	9.04 ± 0.21	6.50 ± 0.17	5.50 ± 0.14
	2048	12.41 ± 0.15	8.95 ± 0.12	6.43 ± 0.10	5.46 ± 0.07
	4096	12.31 ± 0.16	8.86 ± 0.13	6.39 ± 0.12	5.42 ± 0.11
	8192	12.06 ± 0.12	8.66 ± 0.11	6.24 ± 0.09	5.28 ± 0.08
	k	NFE 5	NFE 10	NFE 50	NFE 100
SVHN	1	60.99 ± 0.89	31.34 ± 0.54	5.67 ± 0.14	3.42 ± 0.12
	2	59.89 ± 0.44	31.49 ± 0.31	5.76 ± 0.05	3.43 ± 0.03
	4	57.18 ± 0.73	30.92 ± 0.57	5.69 ± 0.21	3.40 ± 0.12
	8	55.59 ± 0.59	30.34 ± 0.33	5.65 ± 0.12	3.40 ± 0.09
	16	54.08 ± 0.44	29.99 ± 0.33	5.73 ± 0.13	3.47 ± 0.06
	32	53.32 ± 0.89	29.91 ± 0.78	5.77 ± 0.28	3.48 ± 0.17
	64	52.15 ± 0.74	29.61 ± 0.50	5.80 ± 0.17	3.50 ± 0.08
	128	51.83 ± 0.92	29.98 ± 0.78	6.02 ± 0.21	3.68 ± 0.12
	256	50.93 ± 0.59	29.47 ± 0.43	5.99 ± 0.18	3.69 ± 0.13
	512	49.83 ± 0.77	29.16 ± 0.66	6.15 ± 0.24	3.82 ± 0.16
	1024	49.82 ± 0.40	29.64 ± 0.35	6.52 ± 0.14	4.09 ± 0.10
	2048	49.54 ± 0.64	29.84 ± 0.69	6.80 ± 0.23	4.39 ± 0.14
	4096	49.81 ± 0.46	30.05 ± 0.55	7.13 ± 0.23	4.68 ± 0.18
	8192	50.10 ± 0.66	30.19 ± 0.60	7.06 ± 0.33	4.59 ± 0.21

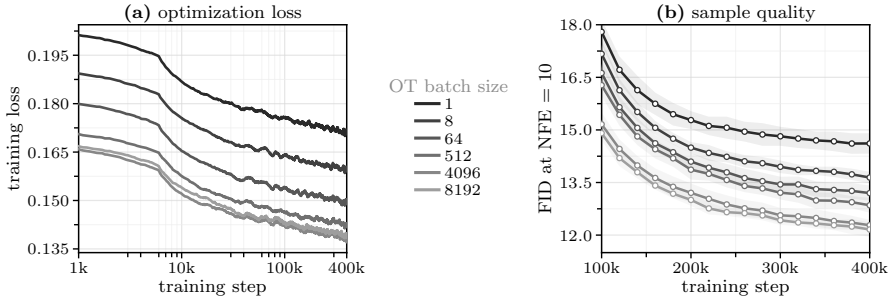


Figure 13: Training curves on CIFAR-10 for a representative subset of OT batch sizes; training loss and FID at fixed $\mathrm{NFE} = 10$ versus optimization step. Larger values of k achieve lower loss and lower FID during training. Loss curves show rolling means over available runs; FID bands show ± 1 empirical standard deviation over seeds.

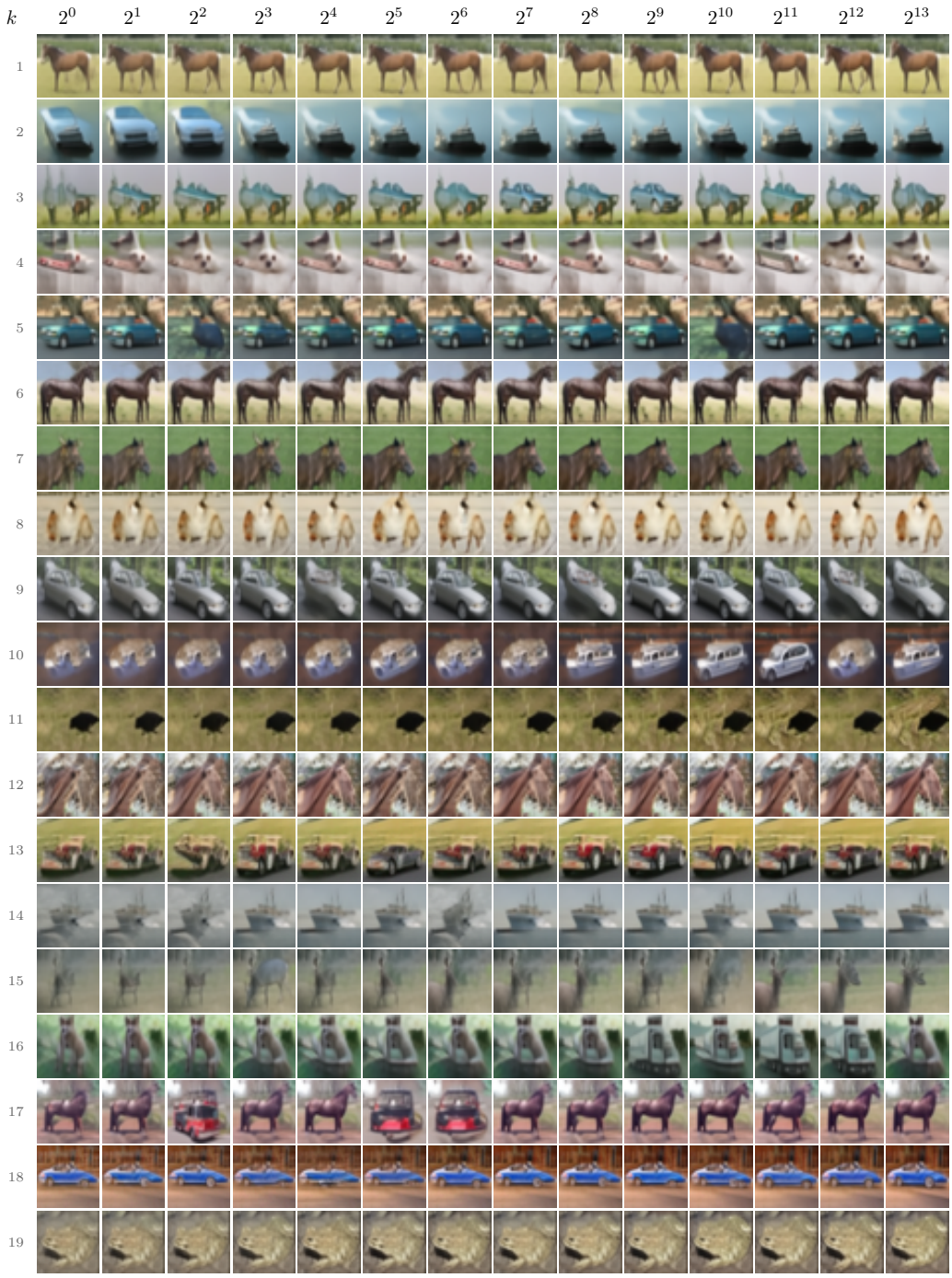


Figure 14: Generated CIFAR-10 samples for $k \in \{1, 2, 4, \ldots, 8192\}$ at NFE $= 10$.

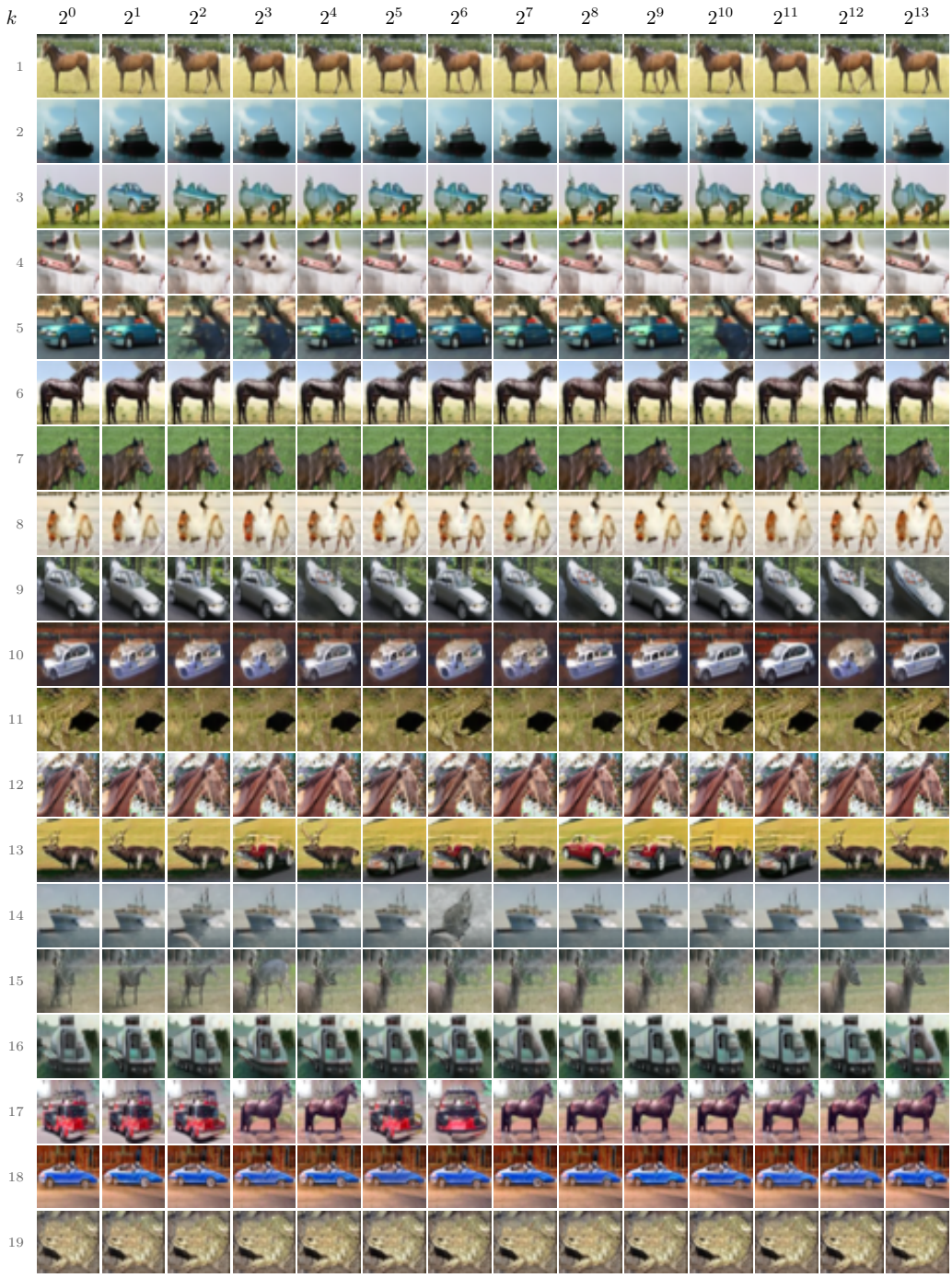


Figure 15: Generated CIFAR-10 samples for $k \in \{1, 2, 4, \ldots, 8192\}$ at NFE $= 100$.

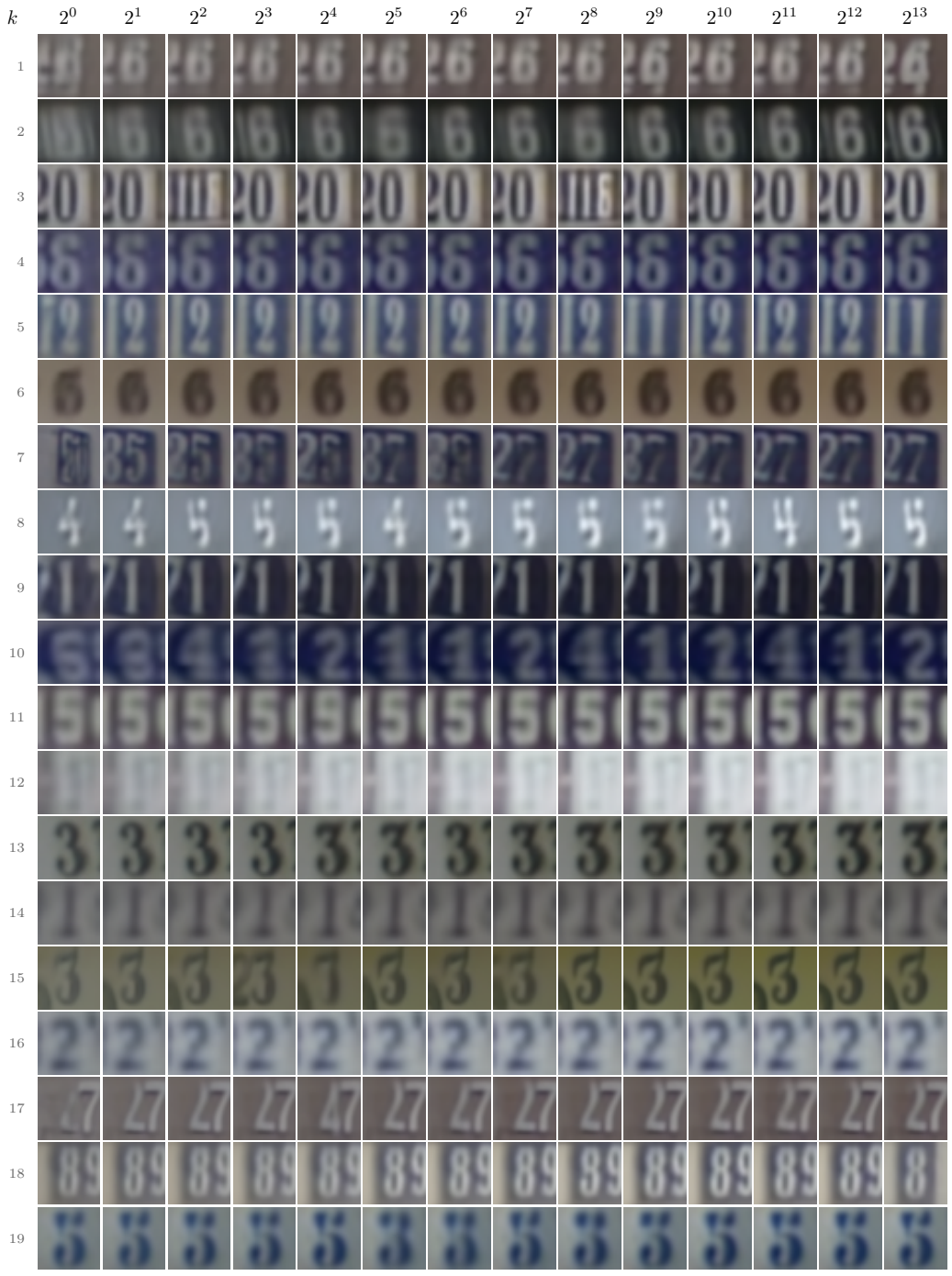


Figure 16: Generated SVHN samples for $k \in \{1, 2, 4, \ldots, 8192\}$ at NFE $= 5$.

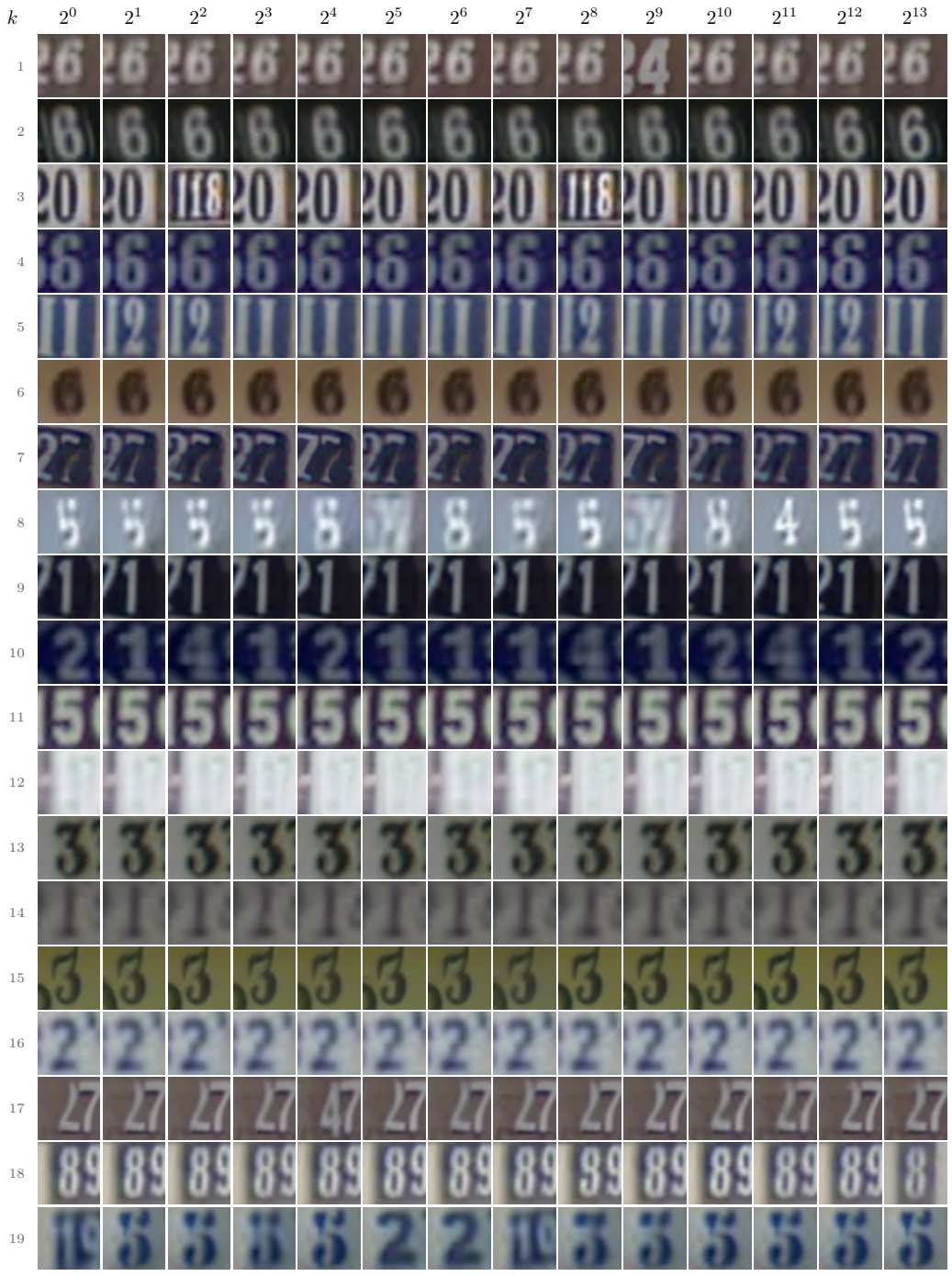


Figure 17: Generated SVHN samples for $k \in \{1, 2, 4, \ldots, 8192\}$ at NFE $= 100$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The claims match the contributions and are supported by Sections 3 and 4 and the experiments in Figures 2, 6 and 7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The "Limitations and future directions" paragraph in the conclusion discusses the semidiscrete scope of the quantitative results, the restriction of the theory to exact velocities, and the lack of general bounds for $\mathscr{E}_{n,k}$.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: Assumptions are stated with the theoretical results, and complete proofs are provided in Appendices A and B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: Experimental protocols, hyperparameters, evaluation procedures, software versions, and compute resources are described in Appendix C. The introduction links to the code repository, and the accompanying README gives reproduction instructions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The introduction links to the code repository: the accompanying code includes dependencies, reproduction scripts, and saved results. The experiments use public CIFAR-10 and SVHN datasets, which the training scripts download.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Training and evaluation details are specified in Appendix C and the accompanying code, including preprocessing, architecture, optimizer, OT batch sizes, NFE values, and FID protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: Error bars and bands are defined in the figure captions and in Appendix C. We report standard errors for Monte Carlo estimates and standard deviations over training seeds for FID.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Compute resources, hardware, software versions, wall-clock times, and exploratory compute are reported in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work conforms to the NeurIPS Code of Ethics: it uses public benchmark datasets and does not involve human subjects, crowdsourcing, private data, surveillance data, or security-sensitive systems.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes].

Justification: Broader impacts are discussed in Appendix D, including possible benefits for generative modeling and risks related to misuse and unequal access to compute.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper does not release high-risk pretrained models, scraped datasets, private data, or security-sensitive systems.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: Existing datasets and software assets are cited, and versions and licenses for the main software dependencies are listed in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The accompanying code includes a README with installation and reproduction instructions, a dependency specification, and an MIT license.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.