# On the Wasserstein Gradient Flow Interpretation of Drifting Models

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Recently, Deng et al. (2026) proposed Generative Modeling via Drifting (GMD), a novel framework for generative tasks. This note presents an analysis of GMD through the lens of Wasserstein Gradient Flows (WGF), i.e., the path of steepest descent for a functional in the space of probability measures, equipped with the geometry of optimal transport. Unlike previous WGF-based contributions, GMD can be thought of as directly targeting a fixed point of a specific WGF flow. We demonstrate three main results: first, that one algorithm proposed by Deng et al. can be viewed as targeting the fixed point of the KL Wasserstein gradient flow after replacing the intractable scores by Parzen-smoothed score estimates. Second, that the algorithm actually implemented by Deng et al. corresponds to a different procedure, which bears some resemblance to the fixed point of a WGF on the Sinkhorn divergence, but lacks certain desirable properties of the latter. Third, the same idea can be extended to the limiting point of other WGFs, including the Maximum Mean Discrepancy (MMD), the sliced Wasserstein distance, and GAN critic functions.

## 1 Introduction

Generative Modeling via Drifting (GMD), introduced by Deng et al. (2026), represents a novel and empirically successful approach to generative tasks. The method achieves strong performance without adversarial training, distillation, or multi-step sampling, relying instead on a non-parametric, mean-shift style drift operator to iteratively move generated samples towards the target data distribution during training. Despite these practical strengths, the underlying theoretical mechanics of GMD remain relatively unexplored. In this note, we discuss the relation between drifting models and Wasserstein Gradient Flows (WGF).

We begin in Section 2.1 with a short review of the Generative Modeling via Drifting (GMD) framework proposed by Deng et al., and then cover the basics of WGF (Ambrosio et al., 2008) in Section 2.2. Next, in Section 2.3, we illustrate a re-interpretation of one of the simpler GMD instantiations as a technique to find an approximation to a fixed point of a WGF for the Kullback–Leibler divergence.

The algorithm implemented in practice by Deng et al. (Algorithm 2 in their paper) is more complex, however. In Section 3, we analyse this algorithm (with minor variations), demonstrating a number of results. First, we show that this model bears a resemblance to a drifting model on the WGF on the Sinkhorn divergence, and shares a specific property of this flow; thus, we refer to it as the Sinkhorn Proxy. Second, we prove that the velocity field of the Sinkhorn Proxy is zero if and only if $p$ and $q$ agree when the Gaussian kernel is used, under additional finite-measure assumptions. In addition to these positive results, we also present two negative results for the Gaussian kernel case: first, that the velocity field does not in fact correspond to the Wasserstein gradient of *any* scalar distributional loss (unless a highly implausible condition holds); second, that convergence of the Sinkhorn Proxy does not inherit the good convergence behaviour one would expect of gradient flow on a true optimal transport distance, in the event that source and target distributions have distant modes with non-overlapping support.

We next explore a broader perspective on drifting in Section 4, beginning with a general-purpose formulation of a drifting algorithm. We then consider instantiations for specific measures of divergence. Deng et al. (2026, Appendix C.2) discuss drifting on the maximum mean discrepancy (MMD), which is very close in practice

to their drifting algorithm, and represents the fixed point of the WGF on the MMD (Arbel et al., 2019). We further introduce drifting on the Sliced-Wasserstein flow (Cozzi and Santambrogio, 2025), and on the class of GAN critic functions derived from f-divergences (Nowozin et al., 2016), providing among other results an alternative to the KL-drifting method of Section 2.3.

**Related work.** The interpretation of the drifting algorithm as the stationary point of a WGF was established by Cao et al. (2026); Turan and Ovsjanikov (2026); He et al. (2026). Cao et al. (2026) describes a connection of drifting to the fixed point of a WGF on the KL between Parzen estimates of the densities, although there is a subtle difference with our presentation - we provide more detail in Section 2.3. They further note the interpretation of MMD drifting as the stationary point of the MMD gradient flow. Turan and Ovsjanikov (2026) and Lai et al. (2026) show that for the Gaussian kernel, the velocity field of the initial drifting instantiation is expressed as a score difference, which is zero only when the distributions agree. He et al. (2026) and Turan and Ovsjanikov (2026, Section 6.3, Appendix I.4) each describe a drifting algorithm built on the WGF of the fully converged Sinkhorn divergence. Alternative perspectives on drifting models, complementary to the one presented here, are proposed by Lai et al. (2026) and Li and Zhu (2026). It is notable that the WGF interpretation of the drifting model actually implemented by Deng et al. (2026, Algorithm 2) is not covered in these prior works.

## 2 A Wasserstein Gradient Flow Perspective

### 2.1 Generative Modeling via Drifting models

Let $p$ be the target data distribution on $\mathbb{R}^d$. To approximate $p$, we use a neural network $f_\theta : \mathbb{R}^c \to \mathbb{R}^d$ whose input is some noise $\epsilon \sim \mu$ (e.g., $\epsilon \sim \mathcal{N}(0, I_c)$) and whose output is $x = f_\theta(\epsilon)$. The implicit distribution of $x$ is denoted $q_\theta$. We assume that $f_\theta(\epsilon)$ is differentiable w.r.t. $\theta$ for all $\epsilon \in \mathbb{R}^c$. The drifting-model paradigm evolves $\theta$ during training for $q_\theta$ to match the target $p$. This method relies on a drift field $V_{p,q_\theta} : \mathbb{R}^d \to \mathbb{R}^d$ which we use to displace samples from our current model. We will discuss several instances of drifting fields in subsequent sections. We update parameters $\theta$ by minimizing the regression loss

$$\mathcal{L}(\theta) = \mathbb{E}_{\epsilon\sim\mu}\Big[\|f_\theta(\epsilon) - \mathrm{sg}\big(f_\theta(\epsilon) + \widehat{V}_{p,q_\theta}(f_\theta(\epsilon))\big)\|^2\Big]; \tag{1}$$

i.e. the average mean square error between a sample and its displaced version; see (Deng et al., 2026, Equation (6)). Here $\widehat{V}_{p,q_\theta}$ is an approximation to $V_{p,q_\theta}$.

The stop-gradient $\mathrm{sg}(\cdot)$ operator appearing in (1) is essential: it ensures that optimization moves the generator *toward* the prescribed transport step, rather than changing the drift estimator itself (Deng et al., 2026). Crucially, it also does not require back-propagating through $\widehat{V}_{p,q_\theta}$. Additionally, it is expected that this approach is more robust than minimizing $\mathbb{E}_\epsilon[\|\widehat{V}_{p,q_\theta}(f_\theta(\epsilon))\|^2]$ where the generator can reduce the loss by collapsing mass so attraction and repulsion cancel, spreading mass to make kernel weights flat, or shaping $q_\theta$ so that $\widehat{V}_{p,q_\theta}$ becomes small without matching $p$. In Section 4.1, Proposition 8, we show that the gradient of this loss turns out to be correct under the WGF perspective presented next, and the stop gradient operator is crucial.

### 2.2 Wasserstein Gradient Flows

We will sketch here the basics of WGFs. We refer the reader to (Ambrosio et al., 2008; Santambrogio, 2017) for a comprehensive treatment and to Appendix A in (Arbel et al., 2019) for a gentle introduction[1]

Let $\mathcal{P}_2(\mathbb{R}^d)$ be the space of distributions with finite second-order moments on $\mathbb{R}^d$. Let $\mathcal{F}_p : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be an energy functional minimized at $q = p$; e.g. the Kullback–Leibler divergence $\mathcal{F}_p(q) = \mathrm{KL}(q\|p)$ or the

[1]Throughout the derivation of WGF for KL below, we assume that the distributions under discussion admit positive $C^1$ densities with sufficient decay for the integrations by parts and score fields to be valid. For parametric generators this can be ensured, for example, by taking $c = d$ and assuming $f_\theta$ is a sufficiently regular diffeomorphism, or by adding noise. If $c < d$ or $f_\theta$ is not regular enough, $q_\theta$ may be singular with respect to Lebesgue measure; then the density-level KL derivations should be interpreted formally or after smoothing. The MMD and Sinkhorn statements below are stated at the level of measures whenever possible.

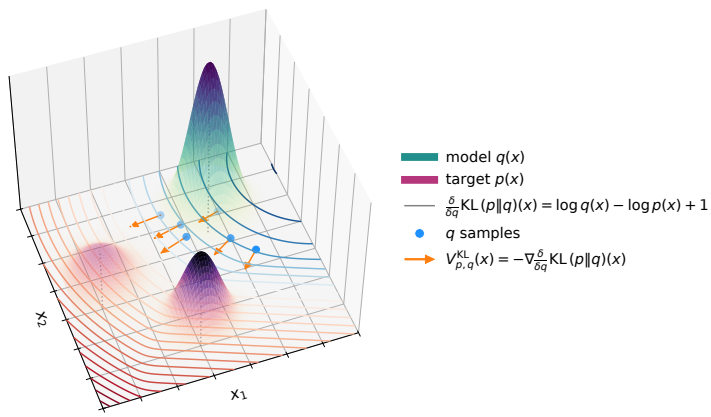


Figure 1: An example of the WGF associated with $\mathcal{F}_p(q) = \mathrm{KL}(q\|p)$, where $p$ is the target distribution and $q$ the model distribution. The contour shows the first variation $\delta\mathrm{KL}(q\|p)/\delta q = \log q - \log p + 1$, where blue is positive and pink is negative. The arrows show the WGF velocity vectors evaluated at samples from $q$, namely $V^{\mathrm{KL}}_{p,q} = \nabla \log p - \nabla \log q$. In practice, the vectors are estimated from samples of unknown $p$ and $q$.

squared Maximum Mean Discrepancy $\mathcal{F}_p(q) = \mathrm{MMD}^2_k(p, q)$ for kernel $k$ (see section 4.2). Under standard regularity assumptions, the Wasserstein-2 ($W_2$) gradient flow of $\mathcal{F}_p$ is characterized by the partial differential equation (PDE) (Ambrosio et al., 2008; Santambrogio, 2017)

$$\partial_t q_t(x) = \nabla \cdot \left( q_t(x) \nabla \frac{\delta \mathcal{F}_p}{\delta q_t}(x) \right), \tag{2}$$

where $\frac{\delta \mathcal{F}_p}{\delta q} : \mathbb{R}^d \to \mathbb{R}$ is the first variation of the functional $\mathcal{F}_p$ with respect to the density $q$ and $\nabla\cdot$ is the divergence operator. If it exists, this first variation is defined as the unique function (up to an additive constant) that satisfies the following linear expansion for any perturbation $\chi$ in the tangent space (i.e. $\int \chi(x)\mathrm{d}x = 0, \int \|x\|^2|\chi(x)|\mathrm{d}x < \infty$), where $h \in \mathbb{R}$

$$\left.\frac{\mathrm{d}}{\mathrm{d}h}\right|_{h=0} \mathcal{F}_p(q + h\chi) = \int \frac{\delta \mathcal{F}_p}{\delta q}(x)\, \chi(x)\mathrm{d}x.$$

The PDE (2) implies that $\mathcal{F}_p(q_t)$ is non-increasing over time as, under regularity conditions, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_p(q_t) &= \int \frac{\delta \mathcal{F}_p}{\delta q_t}(x)\, \partial_t q_t(x)\, \mathrm{d}x \\
&= \int \frac{\delta \mathcal{F}_p}{\delta q_t}(x)\, \nabla \cdot \left( q_t(x) \nabla \frac{\delta \mathcal{F}_p}{\delta q_t}(x) \right) \mathrm{d}x \\
&= -\int q_t(x) \left\| \nabla \frac{\delta \mathcal{F}_p}{\delta q_t}(x) \right\|^2 \mathrm{d}x \leq 0,
\end{aligned}$$

where we have used first the chain rule, then (2) and finally the divergence theorem. The PDE (2) can be rewritten as a continuity equation

$$\partial_t q_t(x) + \nabla \cdot (q_t(x) V_{p,q_t}(x)) = 0, \qquad V_{p,q_t}(x) = -\nabla \frac{\delta \mathcal{F}_p}{\delta q_t}(x). \tag{3}$$

See Figure 1 for an illustration of relevant quantities using the KL divergence as $\mathcal{F}_p$. However, note that this is not a standard continuity equation as the drift not only depends on $x$ but also on the distribution $q_t$ of $x_t$ at time $t$. Formally, this continuity equation corresponds to the following McKean–Vlasov process for individual samples,

$$\mathrm{d}x_t = V_{p,q_t}(x_t)\mathrm{d}t, \qquad x_0 \sim q_0; \tag{4}$$

i.e. this ODE is such that $x_t \sim q_t$. In order to implement such an ODE, we need to discretize it and approximate its drift. For example, using an Euler method leads to

$$x_{t+\eta} = x_t + \eta \widehat{V}_{p,q_t}(x_t), \tag{5}$$

where $\eta > 0$ is a stepsize, and we use $\widehat{V}_{p,q_t}$ estimated from finite samples, which depends on the energy functional being minimized.

### 2.3 Drifting Models as Fixed Point of a Wasserstein Gradient Flow on the KL

We provide two alternative views of Algorithm 1 in Deng et al. (2026) through the score function of distributions. We also clarify the distinction between the pointwise Parzen-smoothed score-difference field and the true WGF of $\mathrm{KL}(q_\tau \| p_\tau)$. The Sinkhorn Proxy method, which better resembles Deng et al.'s Algorithm 2 (main method), is introduced in Section 3. We then show common failure modes of the Sinkhorn Proxy and score objective.

#### 2.3.1 A score-difference drift field

For probability densities $p, q$ on $\mathbb{R}^d$, we first consider the drift field defined by Deng et al. (2026, eq. (8)):

$$V_{p,q}(x) = V_p(x) - V_q(x) \tag{6}$$

with

$$V_p(x) = \frac{\mathbb{E}_{y\sim p}\big[k_\tau(x,y)\, y\big]}{\mathbb{E}_{y\sim p}\big[k_\tau(x,y)\big]} - x, \qquad V_q(x) = \frac{\mathbb{E}_{x'\sim q}\big[k_\tau(x,x')\, x'\big]}{\mathbb{E}_{x'\sim q}\big[k_\tau(x,x')\big]} - x \tag{7}$$

where $k : \mathbb{R}^d \times \mathbb{R}^d \to (0,\infty)$ is a kernel.[2] This defines a *mean-shift* style drifting field (6) built from an "attraction" term (from positives $y \sim p$) and a "repulsion" term (from negatives $x' \sim q$). In particular, for the Gaussian kernel in (Deng et al., 2026, Section 3.3) defined as

$$k_\tau(x,y) \;= (\pi\tau)^{-d/2}\, \exp\Big(-\tau^{-1}\|x-y\|^2\Big), \qquad \tau > 0, \tag{8}$$

we have for $p_\tau(x) \triangleq \mathbb{E}_{y\sim p}[k_\tau(x,y)]$ and $q_\tau(x) \triangleq \mathbb{E}_{x'\sim q}[k_\tau(x,x')]$ that

$$\nabla \log p_\tau(x) = \frac{\nabla p_\tau(x)}{p_\tau(x)} = \frac{2}{\tau}\frac{\mathbb{E}_{y\sim p}\big[(y-x)k_\tau(x,y)\big]}{\mathbb{E}_{y\sim p}\big[k_\tau(x,y)\big]} = \frac{2}{\tau}V_p(x) \tag{9}$$

and similarly $\nabla \log q_\tau(x) = \frac{2}{\tau}V_q(x)$. Note that (9) is an instance of Tweedie's formula. Therefore we have for (8) that

$$V_{p,q}(x) \;=\; \frac{\tau}{2}\Big(\nabla \log p_\tau(x) - \nabla \log q_\tau(x)\Big); \tag{10}$$

that is, up to a constant factor, the drifting field is the difference of the scores of noised versions of $p$ and $q$. This velocity field and its relation to the KL flow was proposed by Weber (2023, Section 3), and later by Lai et al. (2026, Theorem 1) and Turan and Ovsjanikov (2026, Theorem 4.1), where the latter also prove that for the Gaussian kernel, $V_{p,q} = 0$ iff $p = q$ (their Theorem 5.1). Franz et al. (2026) show that when a non-Gaussian kernel is used, the velocity field does not correspond to a Wasserstein gradient, and propose an alternative normalization which restores conservatism.

Practically, we are interested in $V_{p,q_\theta}(x) = V_p(x) - V_{q_\theta}(x)$. Here $V_p$ and $V_{q_\theta}$ are approximated by sampling mini-batches $\{y_j\}_{j=1}^{N^+} \overset{\text{i.i.d.}}{\sim} p$ and $\{x_i\}_{i=1}^{N^-} \overset{\text{i.i.d.}}{\sim} q_\theta$. We define normalized weights:

$$\alpha_j^+(x) = \frac{k_\tau(x,y_j)}{\sum_{\ell=1}^{N^+} k_\tau(x,y_\ell)}, \qquad \alpha_i^-(x) = \frac{k_\tau(x,x_i)}{\sum_{\ell=1}^{N^-} k_\tau(x,x_\ell)}.$$

Then, from (7), we have the drift estimate

$$\widehat{V}_{p,q_\theta}(x) = \widehat{V}_p(x) - \widehat{V}_{q_\theta}(x) \;\; \text{with} \;\; \widehat{V}_p(x) = \sum_{j=1}^{N^+} \alpha_j^+(x)(y_j - x), \quad \widehat{V}_{q_\theta}(x) = \sum_{i=1}^{N^-} \alpha_i^-(x)(x_i - x). \tag{11}$$

[2] We emphasize that this is not the algorithm Deng et al. use in practice, which has a different interpretation, and will be discussed in Section 3.

#### 2.3.2 Wasserstein Gradient Flow for Kullback–Leibler (KL) divergence

Consider the WGF associated to the Kullback–Leibler discrepancy $\mathcal{F}_p(q) = \mathrm{KL}(q||p)$. In this case, it can be easily checked that

$$\frac{\delta \mathcal{F}_p}{\delta q}(x) = \log q(x) - \log p(x) + 1 \tag{12}$$

so the ODE (4) becomes

$$\mathrm{d}x_t = (\nabla \log p(x_t) - \nabla \log q_t(x_t))\mathrm{d}t, \quad x_0 \sim q_0, \tag{13}$$

i.e. the drift of (13) is given by the difference of the (intractable) scores of the data distribution $p$ and of $q_t$.[3]

An Euler discretization (i.e. (5)) of (13) with step size $\eta = \tau/2$ is given by

$$x_{t+\frac{\tau}{2}} = x_t + \frac{\tau}{2}(\nabla \log p(x_t) - \nabla \log q_t(x_t)), \qquad x_0 \sim q_0. \tag{15}$$

Now, in particular, this shows that for $x_t \sim q_t = p$, the particle displacement over the time interval of size $\frac{\tau}{2}$ is given by $\frac{\tau}{2}(\nabla \log p(x_t) - \nabla \log q_t(x_t))$, which evaluates exactly to zero. The key point of GMD is that it is not trying to simulate the flow $(x_t)_{t\geq 0}$. It just directly attempts to find its limiting point by setting the following fixed-point equation. For $x_\infty \sim q = p$, the displacement is zero, and we have from (15) that

$$x_\infty = x_\infty + \frac{\tau}{2}(\nabla \log p(x_\infty) - \nabla \log q(x_\infty)), \tag{16}$$

where both $\nabla \log p$ and $\nabla \log q$ are intractable. By making the approximations $\nabla \log p(x) \approx \nabla \log p_\tau(x)$ and $\nabla \log q(x) \approx \nabla \log q_\tau(x)$, we thus get the approximation of fixed-point equation (16)

$$x_\infty = x_\infty + V_{p,q}(x_\infty), \tag{17}$$

where $V_{p,q}(x)$ is defined in (10). Going back to the parametric framework, this suggests identifying $\theta$ as solving the fixed-point equation (17) using

$$f_\theta(\epsilon) = f_\theta(\epsilon) + V_{p,q_\theta}(f_\theta(\epsilon)), \qquad \epsilon \sim \mu;$$

see (Deng et al., 2026, Equation (4)). This is achieved by performing a Monte Carlo approximation $\widehat{V}_{p,q_\theta}$ in (11) of $V_{p,q_\theta}$ and minimizing (1). So, to summarize, drifting models for the drift field defined by (6) and (7) can be re-interpreted as a method for approximating the fixed point of the KL WGF after the intractable scores $\nabla \log p$ and $\nabla \log q$ are replaced by the Parzen-smoothed score estimates $\nabla \log p_\tau$ and $\nabla \log q_\tau$. This should not be confused with the WGF of $\mathrm{KL}(q_\tau \| p_\tau)$, however, whose velocity is given by (18) below.

We now review the related finding of Cao et al. (2026, Corollary 4.10), which states that the drifting model's velocity field is the Wasserstein-2 gradient flow of the KL divergence between smoothed densities, $\mathcal{F}_p^{(\tau)}(q) := \mathrm{KL}(q_\tau||p_\tau)$. Cao et al. (2026, eq. 7) claim that (10) represents this velocity field. The correct expression for the velocity field of this KL divergence is not (10), however, but

$$V_{p,q}^{(\tau)}(y) = -\nabla_y \int k_\tau(x,y) \log \frac{q_\tau(x)}{p_\tau(x)} \mathrm{d}x. \tag{18}$$

The proof is in Appendix A. More precisely, if the kernel $k_\tau(x,y) = \bar{k}_\tau(x-y)$ for some function $\bar{k}$, then we have the following result using a change of variables:

$$V_{p,q}^{(\tau)}(y) = \int k_\tau(x,y) \nabla \log p_\tau(x) \mathrm{d}x - \int k_\tau(x,y) \nabla \log q_\tau(x) \mathrm{d}x \neq \frac{\tau}{2}\left(\nabla \log p_\tau(y) - \nabla \log q_\tau(y)\right).$$

The same issue applies for the additional flows on Parzen window $f$-divergences defined in that work. As a downstream consequence, the dissipation result of Cao et al. (2026, Corollary 4.8 (ii)) is not established, since particles following (10) are not following the velocity field for $\mathcal{F}_p^{(\tau)}(q)$.

[3] This ODE is the probability flow version of the following Langevin diffusion

$$\mathrm{d}x_t = \nabla \log p(x_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t, \quad x_0 \sim q_0 \tag{14}$$

i.e., the processes defined in (13) and (14) have the same marginal distributions.

## 3 A Sinkhorn Proxy for Drifting Models

The score-difference drift of Section 2.3 approximates scores via kernel density estimates. The algorithm used in practice by Deng et al. (2026) follows a different procedure, however (Algorithm 2 in their paper). We develop a closely related drift field—which we term the *Sinkhorn Proxy*—by connecting it to the WGF on the *debiased Sinkhorn divergence* (Ramdas et al., 2017; Feydy et al., 2019):

$$S_\tau(q,p) = \mathrm{OT}_\tau(q,p) - \tfrac{1}{2}\,\mathrm{OT}_\tau(q,q) - \tfrac{1}{2}\,\mathrm{OT}_\tau(p,p),$$

where $\mathrm{OT}_\tau$ denotes entropy-regularized optimal transport as defined by Cuturi (2013), with regularization $\tau > 0$. This section is a summary of a more in-depth exposition in Appendices B-D.

Consider a set of $N^- \in \mathbb{N}_+$ samples $\{x_i\}_{i=1}^{N^-}$ drawn from the model, and a set of $N^+ \in \mathbb{N}_+$ data samples $\{y_j\}_{j=1}^{N^+}$. The WGF associated with $S_\tau$ can be estimated as follows.

**Proposition 1** (Exact Sinkhorn particle velocity)**.** *Let $q = \frac{1}{N^-}\sum_{i=1}^{N^-}\delta_{x_i}$ and $p = \frac{1}{N^+}\sum_{j=1}^{N^+}\delta_{y_j}$ be empirical measures comprising samples from model and the data set, respectively. Let $W^+ \in \mathbb{R}_+^{N^-\times N^+}$ and $W^- \in \mathbb{R}_+^{N^-\times N^-}$ be the row-stochastic conditional optimal transport plans for $\mathrm{OT}_\tau(q,p)$ and $\mathrm{OT}_\tau(q,q)$, respectively, under the transport cost $C(x,y) = \frac{1}{2}\|x-y\|^2$. Then the WGF velocity of the debiased Sinkhorn divergence $S_\tau(q,p)$ at location $x_i$ is*

$$\widehat{V}^{\mathrm{S}}_{p,q}(x_i) = \sum_{j=1}^{N^+} W^+_{ij}\, y_j - \sum_{k=1}^{N^-} W^-_{ik}\, x_k. \tag{19}$$

For reference and to make the paper self-contained, a proof is given in Appendix B.1, using results from (Feydy et al., 2019). The flow velocity depends on $x_i$ through $W^+$ and $W^-$ (and the self-term in the second sum). He et al. (2026) and Turan and Ovsjanikov (2026, Section 6.3, Appendix I.4) propose a drifting algorithm built on the fully converged Sinkhorn solution.

### 3.1 The empirical Sinkhorn proxy

Computing (19) requires running the Sinkhorn algorithm to convergence at each training step. To follow Deng et al.'s Algorithm 2, we instead replace this with a one-shot approximation. Let $z^+_{im} = -\|x_i - y_m\|^2/\tau$ be the scaled affinities between the $i$'th model sample and $m$'th data point, and $z^-_{im} = -\|x_i - x_m\|^2/\tau$ be the scaled affinities within two samples. We define the pseudo-plan

$$A^+_{im} = \sqrt{\frac{\exp(z^+_{im})}{\frac{1}{N^+}\sum_{u=1}^{N^+}\exp(z^+_{iu})}\cdot\frac{\exp(z^+_{im})}{\frac{1}{N^-}\sum_{v=1}^{N^-}\exp(z^+_{vm})}}\,,\quad A^-_{im} = \sqrt{\frac{\exp(z^-_{im})}{\frac{1}{N^-}\sum_{u=1}^{N^-}\exp(z^-_{iu})}\cdot\frac{\exp(z^-_{im})}{\frac{1}{N^-}\sum_{v=1}^{N^-}\exp(z^-_{vm})}}\,, \tag{20}$$

i.e., the geometric mean of the row-wise and column-wise softmax normalizations, with additional $N^+$ and $N^-$ scaling for well-defined population limits. This amounts to composing a *simultaneous Sinkhorn/Ruiz half-iteration* from each side (Appendix C)—when $N^+ = N^-$ and omitting the $N^+N^-$ scaling, repeated iteration of this procedure on $A^+$ converges to the entropy-regularised transport map (Knight et al., 2014, Theorem 3.1). He et al. (2026, Proposition 3.3) show that the simpler drifting method described in section 2.3 corresponds to a gradient flow on the first step of the standard, alternating Sinkhorn-Knopp algorithm.

**Cross-weighting.** Unlike the fully converged solution, both $A^-$ and $A^+$ no longer have unit row or column sums. Deng et al. address the issue via a cross-weighting step, which we describe below. Let $s^+_i = (N^+)^{-1}\sum_j A^+_{ij}$ and $s^-_i = (N^-)^{-1}\sum_k A^-_{ik}$ be the respective normalized row sums, and set

$$\widetilde{W}^+_{ij} = \frac{1}{N^+}A^+_{ij}\cdot s^-_i,\qquad \widetilde{W}^-_{ik} = \frac{1}{N^-}A^-_{ik}\cdot s^+_i.$$

Writing $\tilde{\pi}^+_{ij} = (N^+)^{-1}A^+_{ij}/s^+_i$, $\tilde{\pi}^-_{ik} = (N^-)^{-1}A^-_{ik}/s^-_i$, and $Z_i = s^+_i s^-_i$, the empirical Sinkhorn proxy drift at $x_i$ becomes

$$\widehat{V}^{\mathrm{SP}}_{p,q}(x_i) = Z_i\left(\sum_{j=1}^{N^+}\tilde{\pi}^+_{ij}\, y_j - \sum_{k=1}^{N^-}\tilde{\pi}^-_{ik}\, x_k\right). \tag{21}$$

which shares with the true Sinkhorn flow (19) the property that the rows of $\tilde{\pi}^+_{ij}$ and $\tilde{\pi}^-_{ik}$ sum to 1, and that the $x_i$ cancels across both terms.

This drift for a parametric $q_\theta$ is then used to optimize the $\theta$, similar to the method in section 2.3.2 for the KL drift in section 2.3.1. The complete procedure for computing the drift is summarized in Algorithm 1, where we omit the $N^-$ and $N^+$ terms, which simply result in a common batch-size-dependent scaling factor on the step size $\eta$.[4] As indicated in the algorithm, the Sinkhorn Proxy differs from the practical algorithm of Deng et al. (2026) in the row-normalization scope: we normalize $A^+$ and $A^-$ over $N^+$ and $N^-$ columns respectively, whereas Deng et al. first concatenate positive and negative logits and normalize jointly over $N^+ + N^-$ columns. Our formulation more closely matches the separate marginal structures of $\mathrm{OT}_\tau(q,p)$ and $\mathrm{OT}_\tau(q,q)$ (Appendix B.2).

**Algorithm 1** Computing the Sinkhorn Proxy drift field $\widehat{V}$

**Require:** Model samples $\{x_i\}_{i=1}^{N^-}$, target samples $\{y_j\}_{j=1}^{N^+}$, temperature $\tau$.

$z^+_{ij} = -\|x_i - y_j\|^2/\tau, \quad z^-_{ik} = -\|x_i - x_k\|^2/\tau$ ▷ compute affinities

Iterations are executed for $j = 1,\dots,N^+$, $k = 1,\dots,N^-$ and $m = 1,\dots,N^+ + N^-$ ranges.

| Sinkhorn Proxy (ours) | Deng et al., Algorithm 2 (DA2) |
|---|---|
| **for** $(\diamond,\ell) \in \{(+,j),(-,k)\}$ **do** | $z = [z^+, z^-] \in \mathbb{R}^{N^- \times (N^+ + N^-)}$ ▷ row concat |
| $\alpha^{\mathrm{row},\diamond}_{i\ell} = \frac{\exp(z^\diamond_{i\ell})}{\sum_{u=1}^{N^\diamond}\exp(z^\diamond_{iu})}$ | $\alpha^{\mathrm{row}}_{im} = \frac{\exp(z_{im})}{\sum_{u=1}^{N^+ + N^-}\exp(z_{iu})}$ |
| $\alpha^{\mathrm{col},\diamond}_{i\ell} = \frac{\exp(z^\diamond_{i\ell})}{\sum_{v=1}^{N^-}\exp(z^\diamond_{v\ell})}$ | $\alpha^{\mathrm{col}}_{im} = \frac{\exp(z_{im})}{\sum_{v=1}^{N^-}\exp(z_{vm})}$ |
| $A^\diamond_{i\ell} = \sqrt{\alpha^{\mathrm{row},\diamond}_{i\ell}\cdot\alpha^{\mathrm{col},\diamond}_{i\ell}}$ | $A_{im} = \sqrt{\alpha^{\mathrm{row}}_{im}\cdot\alpha^{\mathrm{col}}_{im}}$; |
| **end for** | $[A^+, A^-] = \mathrm{split_column}(A, [N^+, N^-])$ |

$s^+_i = \sum_{j=1}^{N^+} A^+_{ij}, \quad s^-_i = \sum_{k=1}^{N^-} A^-_{ik}$

$\widetilde{W}^+_{ij} = A^+_{ij}\cdot s^-_i, \quad \widetilde{W}^-_{ik} = A^-_{ik}\cdot s^+_i$ ▷ cross-weighting

**return** $\widehat{V}(x_i) = \sum_{j=1}^{N^+}\widetilde{W}^+_{ij}\,y_j - \sum_{k=1}^{N^-}\widetilde{W}^-_{ik}\,x_k$

### 3.2 Analysis of the Sinkhorn Proxy

We now introduce the population counterpart to the velocity field induced by the one-shot approximation. We show two main results: first, that the velocity is zero if and only if $p = q$ under certain assumptions on $p, q, k$. Second, that the velocity field of the Sinkhorn proxy does not in general correspond to a WGF.

Let $k(x,y) = \exp(-\|x-y\|^2/\tau)$ be the unnormalized Gibbs kernel.[5] In the continuous population limit[6],

$$p_\tau(y) = (k * p)(y), \qquad q_\tau(x) = (k * q)(x).$$

The geometric mean proxy affinities can be written symmetrically as:

$$A^+(x,y) = \frac{k(x,y)}{\sqrt{q_\tau(y)p_\tau(x)}}, \quad A^-(x,x') = \frac{k(x,x')}{\sqrt{q_\tau(x)q_\tau(x')}}.$$

The row integrals of the independent proxies are

$$s^+(x) = \int A^+(x,y)\mathrm{d}p(y), \qquad s^-(x) = \int A^-(x,x')\mathrm{d}q(x').$$

[4] We do not address here the `ignore self` step of Deng et al. (2026, Algorithm 2), where self-interactions are removed, which is important in the implementation.

[5] Note that this is now a positive definite kernel, not a Parzen window, so we do not employ the normalizer in (8). We henceforth write $p_\tau$ and $q_\tau$ to denote convolutions with the unnormalized kernel, with some abuse of notation.

[6] In the original algorithm, the logits for the positive and negative targets are concatenated, which would give a term $\int k(x,x')\mathrm{d}(p+q)(x') = (k*(p+q))(x)$ in place of $p_\tau(x)$. A closer approximation to the Sinkhorn is attained in the equations we use below.

The cross-weighting step defines the final update weights as

$$W^+(x,y) = s^-(x)A^+(x,y), \qquad W^-(x,x') = s^+(x)A^-(x,x').$$

The population drift vector for the Sinkhorn Proxy is

$$V^{\mathrm{SP}}_{p,q}(x) = \int W^+(x,y)y \,\mathrm{d}p(y) - \int W^-(x,x')x' \,\mathrm{d}q(x'). \tag{22}$$

The velocity field (22) may be written in the following alternative form, as proved in Appendix B.3.

**Proposition 2** (Population form of the Sinkhorn Proxy)**.** *Let* $\mathrm{d}\tilde{p}(y) = q_\tau^{-1/2}(y)\mathrm{d}p(y)$*, and* $\mathrm{d}\tilde{q}(x) = q_\tau^{-1/2}(x)\mathrm{d}q(x)$ *(if* $p$ *and* $q$ *have densities, this is equivalent to the density notation* $\tilde{p}(y) = p(y)/\sqrt{q_\tau(y)}$ *and* $\tilde{q}(y) = q(y)/\sqrt{q_\tau(y)}$*). Then the Sinkhorn Proxy velocity field* (21) *with Gaussian kernel has the population-limit form*

$$V^{\mathrm{SP}}_{p,q}(x) = \frac{\tau}{2}\, Z(x)\Big(\nabla \log(k * \tilde{p})(x) - \nabla \log(k * \tilde{q})(x)\Big), \tag{23}$$

*where* $Z(x) = s^+(x)\, s^-(x)$ *is a spatially varying pre-conditioner.*

This reveals a structural resemblance to the KL score-difference drift of Section 2.3, modulated by $Z(x)$. The velocity field is *consistent*, as shown in Appendix B.4.

**Theorem 3** (Consistency)**.** *Assume that the weighted measures* $\tilde{p}$ *and* $\tilde{q}$ *have finite mass, and that a Gaussian kernel* $k$ *is used. Then* $V = 0$ *for all* $x \in \mathbb{R}^d$ *if and only if* $p = q$*.*

The finite mass of $\tilde{p}(y) = p(y)/\sqrt{q_\tau(y)}$ is not guaranteed. For instance, when $q$ is light-tailed, $p$ is heavy-tailed, both are supported on $\mathbb{R}^d$, and the kernel $k$ performs insufficient blurring on $q$, then the assumption can fail.

As we show in Appendix B.5, however, $V^{\mathrm{SP}}_{p,q}$ does *not* correspond to a valid WGF for the *Gaussian kernel*, similar to the finding of Franz et al. (2026) for the KL-KDE drift with *non-Gaussian kernels* (Section 2.3.1).

**Proposition 4** (Non-gradient structure)**.** *The velocity field* $V^{\mathrm{SP}}_{p,q}$ *in* (23)*, with a Gaussian kernel, corresponds to a valid WGF only when the gradients of the pre-conditioner and the score difference are linearly dependent everywhere:*

$$\partial_i Z(x)\partial_j \Phi(x) = \partial_j Z(x)\partial_i \Phi(x) \quad \textit{for all } i,j,$$

*where* $\Phi(x) = \frac{\tau}{2}(\log(k * \tilde{p})(x) - \log(k * \tilde{q})(x))$*. Equivalently, the wedge product must vanish:* $\nabla Z \wedge \nabla \Phi = 0$*.*

We give here a simple example where Proposition 4 fails.

**A toy example.** Let $d = 2$ and take $p = \delta_b, q = \delta_a, a = (-1,0), b = (1,0)$ and $k(x,y) = \exp\{-\|x-y\|^2/\tau\}$. Then, one can easily check that

$$\Phi(x) = \frac{\tau}{2}\left[\log(k * \widetilde{p})(x) - \log(k * \widetilde{q})(x)\right] = 1 + 2x_1.$$

On the other hand,

$$Z(x) = s^+(x)s^-(x) = \exp(2/\tau)\sqrt{k(x,b)k(x,a)}.$$

So at $x = (0,1)$, simple calculations yield $\nabla\Phi(0,1) = (2,0)$ and $\nabla Z(0,1) = (0, -2Z(0,1)/\tau)$ which are not parallel, or equivalently

$$\partial_1 Z(0,1)\partial_2\Phi(0,1) - \partial_2 Z(0,1)\partial_1\Phi(0,1) = \frac{4Z(0,1)}{\tau} \neq 0.$$

Consequently, the necessary condition in Proposition 4 fails.

### 3.3 Illustration: Sinkhorn Proxy flow vs KL flow and Wasserstein-2 flow

To illustrate a key limitation of both KL-KDE and the Sinkhorn Proxy, we demonstrate that neither is able to shift mass from one mode to another when the modes are far apart relative to kernel width $\tau$. This indicates that the Sinkhorn Proxy does not recover the behaviour one would expect for gradient flow on a mass transport metric.

We consider a synthetic example where the target distribution $p$ and the model distribution $q$ are mixtures of two identical point masses at $+D$ and $-D$, differing only in their mixing proportions $\alpha$ and $\beta$, with $\alpha, \beta \in (0, 1)$ and $\alpha \neq \beta$:

$$p(y) = \alpha\delta_{-D}(y) + (1-\alpha)\delta_{+D}(y),$$
$$q(x) = \beta\delta_{-D}(x) + (1-\beta)\delta_{+D}(x).$$

We assume $\beta < \alpha$, meaning the model has insufficient mass at the left mode $(-D)$ and excess mass at the right mode $(+D)$. We evaluate the continuous vector fields for a model particle currently residing within the overcrowded cluster at $x = +D$. We use the unnormalized Gaussian kernel $k(x, y) = \exp(-\|x-y\|^2/\tau)$. We require the kernel bandwidth $\tau$ to be sufficiently narrow compared to the mode separation $2D$, such that

$$\epsilon = k(D, -D) = \exp\left(-\frac{4D^2}{\tau}\right) \ll 1. \tag{24}$$

For the algorithm to successfully correct the mixing proportions, it must yield a strict *negative* drift at $x = +D$, routing excess mass across the manifold to the starved mode at $-D$. The proofs of all the results in this Section are given in Appendix D. We first describe the failure of KL-KDE.

**Proposition 5.** *For fixed $D > 0$ and fixed $\alpha, \beta \in (0, 1)$ with $\beta < \alpha$, as $\tau \downarrow 0$ (equivalently $\epsilon = \exp(-4D^2/\tau) \downarrow 0$ and (24) holds) the velocity field $V_{\mathrm{KL}}$ of the KL-KDE flow with Gaussian kernel at $x = +D$ is given by*

$$V_{p,q}^{\mathrm{KL}}(D) = -2D\epsilon\left(\frac{\alpha}{1-\alpha} - \frac{\beta}{1-\beta}\right) + o(\epsilon).$$

Unsurprisingly, the velocity vanishes for small $\epsilon$. We now show that the Sinkhorn proxy method has the same failure mode.

**Proposition 6.** *For fixed $D > 0$ and fixed $\alpha, \beta \in (0, 1)$ with $\beta < \alpha$, as $\tau \downarrow 0$ (equivalently $\epsilon = \exp(-4D^2/\tau) \downarrow 0$), the velocity field of the Sinkhorn Proxy flow with Gaussian kernel at $x = D$ is given by*

$$V_{p,q}^{\mathrm{SP}}(D) = -2D\epsilon\underbrace{\left[\sqrt{\frac{1-\alpha}{\beta}}\right]}_{(*)}\left(\frac{\alpha}{1-\alpha} - \frac{\beta}{1-\beta}\right) + o(\epsilon),$$

*where the term (*) represents the difference with the KL-KDE flow.*

We now contrast the Sinkhorn proxy velocity field with the mass transport dynamics of an optimal transport distance, specifically the barycentric velocity at $x = +D$ for the WGF of $\frac{1}{2}W_2^2(q, p)$ (see Appendix D for discussion).

**Proposition 7.** *The barycentric velocity of the WGF on $\frac{1}{2}W_2^2(q, p)$ at $x = D$ is given by*

$$V_{p,q}^{W_2}(D) = -2D\left(\frac{\alpha-\beta}{1-\beta}\right).$$

Velocity increases with D, and there is no scaling by an infinitesimal $\epsilon$.

## 4 Extensions

WGF techniques have been previously proposed to perform generative modeling for various functionals; e.g., Maximum Mean Discrepancy (Arbel et al., 2019), Sobolev witness (Mroueh et al., 2019) and Sliced-Wasserstein distance (Liutkus et al., 2019; Cozzi and Santambrogio, 2025). WGF have also been used

**Algorithm 2** Generic drifted-target method to minimize an energy function $\mathcal{F}_p(q)$

**Require:** data sampler $y \sim p$, noise sampler $\epsilon \sim \mu$, generator $f_\theta$, initial parameter $\theta$

1: **for** $\ell = 0, 1, 2, \dots$ **do**
2: sample $\{y_j\}_{j=1}^{N^+} \overset{\text{i.i.d.}}{\sim} p$ ▷ Real batch
3: sample $\{\epsilon_i\}_{i=1}^{N^-} \overset{\text{i.i.d.}}{\sim} \mu$, set $x_i = f_\theta(\epsilon_i)$ for all $i \in \{1, \dots, N^-\}$ ▷ Model batch
4: compute drifts $\widehat{V}_{p,q_\theta}(x_i) = \widehat{V}_{p,q_\theta}(x_i; \{y_j\}_{j=1}^{N^+}, \{x_k\}_{k=1}^{N^-})$
5: update $\theta$ by minimizing $\sum_{i=1}^{N^-} \|f_\theta(\epsilon_i) - \mathrm{sg}(x_i + \widehat{V}_{p,q_\theta}(x_i))\|^2$
6: **end for**

to solve inverse problems (Crucinio et al., 2024). Contrary to GMD, all these approaches implement time-discretization of the ODE (4). As noted by Cao et al. (2026); Turan and Ovsjanikov (2026), the methodology of Deng et al. (2026) offers an alternative that can be readily implemented for any Wassertein gradient flow. This might prove a useful addition to existing techniques.

### 4.1 Practical algorithm

Given an energy function $\mathcal{F}_p(q)$ such that $\mathcal{F}_p(p) = 0$, the general method to develop a one-stage generator $f_\theta$ to sample from $p$ is as follows:

- identify its first variation $\delta\mathcal{F}_p/\delta q(x)$;
- set $V_{p,q}(x) = -\nabla\delta\mathcal{F}_p/\delta q(x)$;
- approximate $V_{p,q_\theta}(x)$ via $\widehat{V}_{p,q_\theta}(x)$ (e.g. expectations are approximated by mini-batches, approximation of Tweedie's formula etc.). When empirical samples $\{y_j\} \overset{\text{i.i.d.}}{\sim} p$ and $x_k = f_\theta(\epsilon_k)$, $\{\epsilon_k\} \overset{\text{i.i.d.}}{\sim} \mu$ are available, we write $\widehat{V}_{p,q_\theta}(x; \{y_j\}_{j=1}^{N^+}, \{x_k\}_{k=1}^{N^-})$.

Finally instead of minimizing $\mathbb{E}_{\epsilon\sim\mu}[\|\widehat{V}_{p,q_\theta}(f_\theta(\epsilon))\|^2]$ w.r.t. $\theta$, minimize

$$\mathcal{L}(\theta) = \mathbb{E}_{\epsilon\sim\mu}\Big[\|f_\theta(\epsilon) - \mathrm{sg}\big(f_\theta(\epsilon) + \eta\widehat{V}_{p,q_\theta}(f_\theta(\epsilon))\big)\|^2\Big]$$

where $\eta > 0$ is a hyperparameter. The generic algorithm is detailed in Algorithm 2.

The following Proposition, proven in Appendix E, should be read as a surrogate-gradient identity: the target is detached at the current parameter value. The gradient of the drifted target loss evaluated at the current iterate is proportional to the gradient of the energy functional. Therefore, we set the $\eta = 1$ in Algorithm 2.

**Proposition 8** (Surrogate-gradient identity)**.** *Let $q_\theta = (f_\theta)_\#\mu$ be the pushforward of $\mu$ through $f_\theta$, and suppose we have the exact velocity field $V_{p,q_\theta}(x) = -\nabla_x \frac{\delta\mathcal{F}_p}{\delta q_\theta}(x)$ for some functional $\mathcal{F}_p$. Then*

$$\nabla_\theta\mathcal{L}(\theta) = 2\eta\,\nabla_\theta\mathcal{F}_p(q_\theta).$$

### 4.2 Applications

The approximation of $V_{p,q_\theta}$ is very dependent on the energy functional being minimized. We give a few potential applications using different types of flows, in addition to the KL flow in Section 2.3.2.

**Maximum mean discrepancy (MMD) flow** Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a differentiable reproducing kernel (e.g. RBF), and let $\mathcal{H}$ be the corresponding reproducing kernel Hilbert space. The Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) is defined as

$$\begin{aligned}\mathrm{MMD}_k^2(p,q) &= \left(\sup_{g\in\mathcal{H},\|g\|\le 1} \mathbb{E}_{x\sim q}[g(x)] - \mathbb{E}_{y\sim p}[g(y)]\right)^2 \\ &= \mathbb{E}_{x,x'\sim q}[k(x,x')] - 2\mathbb{E}_{x\sim q, y\sim p}[k(x,y)] + \mathbb{E}_{y,y'\sim p}[k(y,y')].\end{aligned}$$

The function $g^*$ attaining the supremum is the MMD witness function. The WGF applied to $\mathcal{F}_p^{\mathrm{MMD}}(q) = \frac{1}{2}\,\mathrm{MMD}_k^2(p,q)$ was first studied by Arbel et al. (2019) who showed that the first variation is

$$\frac{\delta \mathcal{F}_p^{\mathrm{MMD}}}{\delta q}(x) = \mathbb{E}_{x'\sim q}[k(x,x')] - \mathbb{E}_{y\sim p}[k(x,y)] \propto g^*(x),$$

so

$$V_{p,q}^{\mathrm{MMD}}(x) = -\mathbb{E}_{x'\sim q}[\nabla_x k(x,x')] + \mathbb{E}_{y\sim p}[\nabla_x k(x,y)].$$

This can be approximated by the mini-batch estimator

$$\widehat{V}_{p,q}^{\mathrm{MMD}}(x) = \frac{1}{N^+}\sum_{i=1}^{N^+} \nabla_x k(x,y_i) - \frac{1}{N^-}\sum_{j=1}^{N^-} \nabla_x k(x,x_j),$$

which recovers the MMD drifting approach proposed by Deng et al. (2026, Appendix C.2), as demonstrated by Cao et al. (2026, Section 4.5). The meaning of "kernel" is fundamentally different to the KL case: here, it is a reproducing kernel (a positive definite function, which need not be a Parzen window), informing the smoothness of the witness function (first variation). When a characteristic kernel (which includes Gauss and Laplace kernels) is used, the MMD is zero if and only if $p = q$ (Sriperumbudur et al., 2010; 2011).

As a practical matter, WGF on the MMD has much better convergence behaviour when the kernel is chosen adaptively, according to the current location of the $q$ particles relative to those from $p$ (Galashov et al., 2025; Chen et al., 2025). As an alternative to adapting the kernel, a sum of kernels over different bandwidths can be used, as was done by Deng et al. (see Cortes et al., 2009, Section 5, for the multiple kernel learning perspective). A related perspective is the use of the MMD as a distributional loss (critic) in GAN training Li et al. (2015); Dziugaite et al. (2015), and the related diffusion model of Zhou et al. (2025). The present discussion makes explicit the connection with WGFs.

**Sliced-Wasserstein (SW) flow.** Let $\mathbb{S}^{d-1}$ be the unit hypersphere in $\mathbb{R}^d$, and for any direction $\vartheta \in \mathbb{S}^{d-1}$, let $\pi_\vartheta(x) = x\cdot\vartheta$ denote the 1D projection of a point $x$ onto $\vartheta$. The squared Sliced-Wasserstein distance is

$$\mathrm{SW}_2^2(q,p) = \int_{\mathbb{S}^{d-1}} W_2^2\big((\pi_\vartheta)_\# q,\ (\pi_\vartheta)_\# p\big)\,\mathrm{d}\vartheta,$$

where $(\pi_\vartheta)\#q$ denotes the push-forward (i.e., the 1D marginal) of the distribution $q$ along the direction $\vartheta$, and $\mathrm{d}\vartheta$ represents the uniform probability measure on the sphere. We set the energy function to $\mathcal{F}_p^{\mathrm{SW}}(q) \triangleq \frac{1}{2}\,\mathrm{SW}_2^2(q,p)$. Cozzi and Santambrogio (2025) showed that

$$\frac{\delta \mathcal{F}_p^{\mathrm{SW}}}{\delta q}(x) = \int_{\mathbb{S}^{d-1}} \varphi_\vartheta(\vartheta\cdot x)\,\mathrm{d}\vartheta, \quad V_{p,q}^{\mathrm{SW}}(x) = -\nabla_x \frac{\delta \mathcal{F}_p^{\mathrm{SW}}}{\delta q}(x) = \int_{\mathbb{S}^{d-1}} \big(T_\vartheta(\vartheta\cdot x) - \vartheta\cdot x\big)\,\vartheta\,\mathrm{d}\vartheta,$$

where $\varphi_\vartheta$ is the 1D Kantorovich potential and $T_\vartheta : \mathbb{R}\to\mathbb{R}$ is the 1D optimal transport map pushing $(\pi_\vartheta)_\# q$ to $(\pi_\vartheta)_\# p$. This can be approximated using a mini-batch of real data samples and model samples, and $V_{p,q}^{\mathrm{SW}}(x)$ then approximated by uniformly sampled directions $(\vartheta^i)_{i=1}^L$ on $\mathbb{S}^{d-1}$.

**Dual Forms of $f$-Divergences.** The $f$-divergences are the standard divergence measures used in generative adversarial networks (GANs). As shown by Nowozin et al. (2016), critics based on $f$-divergences can be formulated using the Fenchel dual of their corresponding divergences. This dual formulation introduces an optimal witness function $g^*$ whose gradient is the Wasserstein gradient of the $f$-divergence. Building upon this, Glaser et al. (2021) proposed the KL Approximate Lower-bound Estimator (KALE) as an alternative to the KL gradient flow discussed in Section 2.3. KALE operates by regularizing the dual KL divergence over a designated function class $\mathcal{H}$. For a regularization parameter $\lambda > 0$, the estimator of Glaser et al. (2021, Proposition 2) is:

$$\mathcal{F}_p^{\mathrm{KALE}}(q) = (1+\lambda)\max_{g\in\mathcal{H}}\left\{\mathbb{E}_{y\sim q}[g(y)] - \mathbb{E}_{x\sim p}[e^{g(x)}] + 1 - \frac{\lambda}{2}\|g\|_{\mathcal{H}}^2\right\}. \tag{25}$$

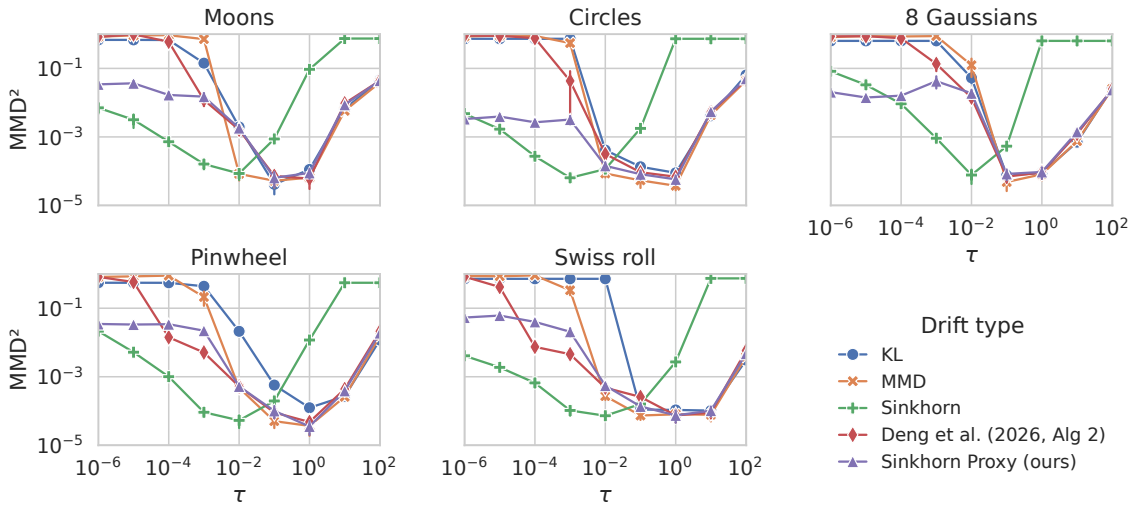


Figure 2: MMD between true and generated samples trained by different drift types.

This is computed via a quadratic program when $\mathcal{H}$ is a reproducing kernel Hilbert space (Glaser et al., 2021), and by SGD on the weights $\phi$ of $g_\phi$ when $\mathcal{H}$ is a neural network (Nowozin et al., 2016; Galashov et al., 2025). As for the MMD, $\nabla \frac{\delta \mathcal{F}_p}{\delta q}(x) = (1+\lambda)\nabla g^*$, where $g^* \in \mathcal{H}$ achieves the supremum in (25). A neural network implementation of the KL gradient flow, which parameterizes the critic function with a deep architecture, is given by (Galashov et al., 2025, Section 6 and Appendix D). A potential advantage of the Fenchel dual instantiation of the Wasserstein Gradient on the KL, over the flow described in Section 2.3, is that the former does not require kernel density estimates in its definition. As an alternative to the KL, a gradient flow on the dual approximation to the $\chi^2$ divergence is given by Chen et al. (2025), which has closed-form updates.

### 4.3 Illustration

As a proof of concept, we ran experiments to train generators using 6 flow objectives above on 5 synthetic 2D datasets. We use $\tau$ to denote hyperparameters in different flow, and vary its value over a wide range to test the sensitivity and effectiveness of the different approaches. For the KL and MMD flows, $\tau$ is the kernel bandwidth in (8). For Sinkhorn-based flows, including Deng et al. (2026)'s Algorithm 2 (DA2) and the approximate Sinkhorn algorithm we derived, $\tau$ denotes the entropic regularization weight. The plain Sinkhorn flows used 100 Sinkhorn-Knopp iterations to evaluate the potentials. All Sinkhorn-based algorithms operate in log-probability space for numerical stability. For the SW flow, we varied the number of random slices $L$ from 1–100 but did not find any effect of $L$ on sample quality. The generator network architecture is a feedforward ResNet, with parameters optimized by the Adam algorithm with learning rate $10^{-4}$, trained for $10^6$ steps. Each data-drift type combination is repeated 20 times with different random seeds for the dataset and parameter initializations.

We used the $\text{MMD}^2$ under a median-heuristic Gaussian kernel as the metric for evaluating sample quality at the end of training. The key results are shown in Figure 2. Sliced-Wasserstein is omitted because its hyperparameter $L$ does not have any effect on sample quality. All flows can attain similarly best performances at some sweet spot of their relevant hyperparameter. The Sinkhorn drift can tolerate a small $\tau$ and only gets worse for large $\tau$. Interestingly, our Sinkhorn Proxy has better tolerance to smaller values of $\tau$ compared to DA2. Qualitatively, Figures 3-7 in Appendix F show the true and the generated sample histograms. For KL, a too small noise level leads to mode-collapse common for score-based methods (Wenliang and Kanagawa, 2020), and a too large noise level causes the learned distribution to ignore local noise structure (Figures 3, 5 and 4). The mode-collapse is not as pronounced for the Sinkhorn Proxy or DA2, which might be due to the "hub-supression" effect discussed in Appendix C.

## References


Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser.

Arbel, M., Korba, A., Salim, A., and Gretton, A. (2019). Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems*.

Cao, J., Wei, Z., and Liu, Y. (2026). Gradient flow drifting: Generative modeling via Wasserstein gradient flows of KDE-approximated divergences. *arXiv preprint arXiv:2603.10592*.

Chen, Z., Mustafi, A., Glaser, P., Korba, A., Gretton, A., and Sriperumbudur, B. K. (2025). (De)-regularized maximum mean discrepancy gradient flow. *Journal of Machine Learning Research*, 26(235):1–77.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2009). L2 regularization for learning kernels. In *Uncertainty in Artificial Intelligence*.

Cozzi, G. and Santambrogio, F. (2025). Long-time asymptotics of the sliced-Wasserstein flow. *SIAM Journal on Imaging Sciences*, 18(1):1–19.

Crucinio, F. R., De Bortoli, V., Doucet, A., and Johansen, A. M. (2024). Solving Fredholm integral equations of the first kind via Wasserstein gradient flows. *Stochastic Processes and Their Applications*, 173.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*.

Deng, M., Li, H., Li, T., Du, Y., and He, K. (2026). Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence*.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. (2019). Interpolating between optimal transport and MMD using Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*.

Franz, L., Hoffmann, S., and Martius, G. (2026). Drifting fields are not conservative. *arXiv preprint arXiv:2604.06333*.

Galashov, A., De Bortoli, V., and Gretton, A. (2025). Deep MMD gradient flow without adversarial training. In *International Conference on Learning Representations*.

Glaser, P., Arbel, M., and Gretton, A. (2021). KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. In *Advances in Neural Information Processing Systems*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. J. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13.

He, P., Khangaonkar, O., Pirsiavash, H., Bai, Y., and Kolouri, S. (2026). Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*.

Knight, P. A., Ruiz, D., and Uçar, B. (2014). A symmetry preserving algorithm for matrix scaling. *SIAM Journal on Matrix Analysis and Applications*, 35(3):931–955. hal-00569250.

Lai, C.-H., Nguyen, B., Murata, N., Takida, Y., Uesaka, T., Mitsufuji, Y., Ermon, S., and Tao, M. (2026). A unified view of drifting and score-based models. *arXiv preprint arXiv:2603.07514*.

Li, Y., Swersky, K., and Zemel, R. (2015). Generative moment matching networks. In *International Conference on Machine Learning*.

Li, Z. and Zhu, B. (2026). A long-short flow-map perspective for drifting models. *arXiv preprint arXiv:2602.20463*.

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. (2019). Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*.

Mroueh, Y., Sercu, T., and Raj, A. (2019). Sobolev descent. In *International Conference on Artificial Intelligence and Statistics*.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-GAN: training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*.

Ramdas, A., Trillos, N., and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2).

Santambrogio, F. (2017). {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561.

Turan, E. and Ovsjanikov, M. (2026). Generative drifting is secretly score matching: a spectral and variational perspective. *arXiv preprint arXiv:2603.09936*.

Weber, R. M. (2023). The score-difference flow for implicit generative modeling. *Transactions on Machine Learning Research*.

Wenliang, L. K. and Kanagawa, H. (2020). Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*.

Zhou, L., Ermon, S., and Song, J. (2025). Inductive moment matching. In *International Conference on Machine Learning*.

## A Wasserstein Gradient flow of KL on Parzen Window density estimates

We derive the first variation and the WGF velocity for the KDE-smoothed forward KL divergence $\mathcal{F}_p^{(\tau)}(q) := \mathrm{KL}(q_\tau \| p_\tau)$, where $q_\tau = k_\tau * q$ and $p_\tau = k_\tau * p$ are Parzen window density estimates with a symmetric kernel $k_\tau$.

The functional $\mathcal{F}_p^{(\tau)}$ is a composition: $q$ first maps to $q_\tau$ via the linear convolution $q \mapsto k_\tau * q$, and then the standard KL divergence is evaluated. By (12), the first variation of $\mathrm{KL}(\cdot\|p_\tau)$ with respect to its first argument is

$$\frac{\delta\,\mathrm{KL}(\cdot\|p_\tau)}{\delta q_\tau}(x) = \log q_\tau(x) - \log p_\tau(x) + 1.$$

To obtain the first variation of $\mathcal{F}_p^{(\tau)}$ with respect to $q$, we apply the chain rule through the convolution. For any perturbation $\chi$ with $\int \chi(y)\,\mathrm{d}y = 0$, linearity gives $k_\tau * (q + h\chi) = k_\tau * q + h\,(k_\tau * \chi) = q_\tau + h\,(k_\tau * \chi)$, so

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}h}\bigg|_{h=0} \mathcal{F}_p^{(\tau)}(q + h\chi) &= \frac{\mathrm{d}}{\mathrm{d}h}\bigg|_{h=0} \mathrm{KL}\big(q_\tau + h\,(k_\tau * \chi) \,\big\|\, p_\tau\big) \\
&= \int \frac{\delta\,\mathrm{KL}(\cdot\|p_\tau)}{\delta q_\tau}(x)\;(k_\tau * \chi)(x)\,\mathrm{d}x \\
&= \int \Big(\log q_\tau(x) - \log p_\tau(x) + 1\Big) \int k_\tau(x,y)\,\chi(y)\,\mathrm{d}y\,\mathrm{d}x,
\end{aligned}$$

where the second line is the definition of first variation applied to $\mathrm{KL}(\cdot\|p_\tau)$ with perturbation $k_\tau * \chi$. Exchanging the order of integration (Fubini) and using the symmetry $k_\tau(x,y) = k_\tau(y,x)$, this equals

$$\int \underbrace{\left[\int k_\tau(y,x)\Big(\log q_\tau(x) - \log p_\tau(x) + 1\Big)\,\mathrm{d}x\right]}_{=\ \dfrac{\delta \mathcal{F}_p^{(\tau)}}{\delta q}(y)} \chi(y)\,\mathrm{d}y.$$

Because $\int k_\tau(y,x)\,\mathrm{d}x$ is constant in $y$ for translation-invariant kernels, the $+1$ term contributes only a constant and vanishes under $\nabla_y$. The WGF velocity of the KDE-smoothed KL divergence is therefore

$$V_{p,q}^{(\tau)}(y) = -\nabla_y \frac{\delta \mathcal{F}_p^{(\tau)}}{\delta q}(y) = -\nabla_y \int k_\tau(y,x) \log \frac{q_\tau(x)}{p_\tau(x)}\,\mathrm{d}x.$$

## B Details of the Sinkhorn Proxy

This appendix contains the full derivation and analysis of the Sinkhorn Proxy drift field summarized in Section 3. We will switch from the generic notation $V_{p,q}$ to method-specific notation, such as $V_{p,q}^{\mathrm{KL}}$, $V_{p,q}^{\mathrm{SP}}$, in the following analysis to avoid ambiguity.

### B.1 Drifting Models as Fixed Point of a Wasserstein Gradient Flow on an Approximate Sinkhorn Divergence

As noted in their paper, the drifting model as implemented does not use (Deng et al., 2026, eq. (8)) in practice, but the approach described in (Deng et al., 2026, Algorithm 2, Appendix A). We show that this represents a fixed point of a probability flow on an approximation to the Sinkhorn Optimal Transport distance, where the velocity field is zero everywhere if and only if $p$ and $q$ coincide under the stated assumptions. Despite this resemblance to a Sinkhorn distance, however, the resulting velocity field is not generally conservative and therefore cannot, for generic pairs $(p,q)$ in dimension at least two, be the Wasserstein gradient of a scalar distributional loss. Its transport dynamics are in certain important respects closer to a gradient flow on the KL.

We begin with a brief review of the Sinkhorn distance and its associated WGF on a set of particles. Let $q, p \in \mathcal{P}_2(\mathbb{R}^d)$ be probability measures with bounded second moment and let $C(x,y) = \frac{1}{2}\|x-y\|^2$. The *debiased Sinkhorn divergence* (Ramdas et al., 2017) is defined as:

$$S_\tau(q,p) = \mathrm{OT}_\tau(q,p) - \frac{1}{2}\mathrm{OT}_\tau(q,q) - \frac{1}{2}\mathrm{OT}_\tau(p,p),$$

where $\mathrm{OT}_\tau(q,p)$ is the entropy-regularized OT distance (Cuturi, 2013), and $\tau > 0$ is the entropic regularization parameter.

**Proposition 1** (Exact Sinkhorn particle velocity)**.** *Let $q = \frac{1}{N^-}\sum_{i=1}^{N^-}\delta_{x_i}$ and $p = \frac{1}{N^+}\sum_{j=1}^{N^+}\delta_{y_j}$ be empirical measures comprising samples from model and the data set, respectively. Let $W^+ \in \mathbb{R}_+^{N^-\times N^+}$ and $W^- \in \mathbb{R}_+^{N^-\times N^-}$ be the row-stochastic conditional optimal transport plans for $\mathrm{OT}_\tau(q,p)$ and $\mathrm{OT}_\tau(q,q)$, respectively, under the transport cost $C(x,y) = \frac{1}{2}\|x-y\|^2$. Then the WGF velocity of the debiased Sinkhorn divergence $S_\tau(q,p)$ at location $x_i$ is*

$$\widehat{V}_{p,q}^{\mathrm{S}}(x_i) = \sum_{j=1}^{N^+} W_{ij}^+\, y_j - \sum_{k=1}^{N^-} W_{ik}^-\, x_k. \tag{19}$$

*Proof.* By the envelope theorem (Feydy et al., 2019, Proposition 1), the first variation of $S_\tau$ with respect to its first argument is the difference of optimal dual potentials,

$$\frac{\delta S_\tau(q,p)}{\delta q}(x) = g_{q,p}(x) - g_{q,q}(x),$$

where $g_{q,p}$ and $g_{q,q}$ are the optimal potentials on the $q$-side of $\mathrm{OT}_\tau(q,p)$ and $\mathrm{OT}_\tau(q,q)$, respectively. For the self term $\mathrm{OT}_\tau(q,q)$, the variation of $q$ affects both marginals; the derivative of $\frac{1}{2}\mathrm{OT}_\tau(q,q)$ is $\frac{1}{2}(f_{q,q}+g_{q,q})$. Since the two marginals and the cost are identical, the optimal potentials may be chosen with $f_{q,q}=g_{q,q}$ up to an additive constant, so the self-term contribution is $g_{q,q}$. The WGF velocity is therefore

$$\widehat{V}^{\mathrm{S}}_{p,q}(x) = -\nabla_x \frac{\delta S_\tau(q,p)}{\delta q}(x) = -\nabla_x g_{q,p}(x) + \nabla_x g_{q,q}(x).$$

It remains to compute these gradients. For the $g_{q,p}$ term, the joint optimal transport plan $\pi^+$ and the dual potentials $g_{q,p}$, $h_{q,p}$ are related by (Feydy et al., 2019, Equation 9)

$$\pi^+_{ij} = \frac{1}{N^+N^-}\exp\left(\frac{g_{q,p}(x_i)+h_{q,p}(y_j)-\frac{1}{2}\|x_i-y_j\|^2}{\tau}\right).$$

The plan $\pi^+$ satisfies $\sum_j \pi^+_{ij} = \frac{1}{N^-}$ and $\sum_i \pi^+_{ij} = \frac{1}{N^+}$. The potentials satisfy the coupled fixed-point equations:

$$g_{q,p}(x_i) = -\tau \log \sum_{j=1}^{N^+} \frac{1}{N^+}\exp\left(\frac{h_{q,p}(y_j)-\frac{1}{2}\|x_i-y_j\|^2}{\tau}\right),$$
$$h_{q,p}(y_j) = -\tau \log \sum_{i=1}^{N^-} \frac{1}{N^-}\exp\left(\frac{g_{q,p}(x_i)-\frac{1}{2}\|x_i-y_j\|^2}{\tau}\right).$$

Taking $\nabla_{x_i}$ of the first equation, the $\frac{1}{N^+}$ factors appear in both numerator and denominator and cancel:

$$\nabla_{x_i} g_{q,p}(x_i) \;=\; -\tau \frac{\sum_{j=1}^{N^+}\frac{1}{N^+}\exp\left(\frac{h_{q,p}(y_j)-\frac{1}{2}\|x_i-y_j\|^2}{\tau}\right)\frac{-(x_i-y_j)}{\tau}}{\sum_{j'=1}^{N^+}\frac{1}{N^+}\exp\left(\frac{h_{q,p}(y_{j'})-\frac{1}{2}\|x_i-y_{j'}\|^2}{\tau}\right)} \;=\; \sum_{j=1}^{N^+} W^+_{ij}(x_i-y_j),$$

where $W^+_{ij} \triangleq \pi^+_{ij}/\sum_{j'}\pi^+_{ij'} = N^-\pi^+_{ij}$ is the row-stochastic *conditional* plan (with $\sum_j W^+_{ij}=1$), and the last equality holds because the softmax ratio equals $W^+_{ij}$. An identical calculation for the self-term gives $\nabla_{x_i} g_{q,q}(x_i) = \sum_k W^-_{ik}(x_i-x_k)$. Substituting back:

$$\widehat{V}^{\mathrm{S}}_{p,q}(x_i) = -\sum_{j=1}^{N^+} W^+_{ij}(x_i-y_j) + \sum_{k=1}^{N^-} W^-_{ik}(x_i-x_k) = \sum_{j=1}^{N^+} W^+_{ij}\,y_j - \sum_{k=1}^{N^-} W^-_{ik}\,x_k, \tag{26}$$

since $W^+$ and $W^-$ both have unit row sums ($\sum_j W^+_{ij} = \sum_k W^-_{ik} = 1$). □

### B.2 The One-Shot Sinkhorn Proxy

Equation (19) requires computing the true conditional optimal transport plans (i.e., running Sinkhorn to convergence) at every timestep. A drifting model using the fully converged Sinkhorn method was proposed by He et al. (2026) and Turan and Ovsjanikov (2026, Section 6.3, Appendix I.4). To recover (Deng et al., 2026, Algorithm 2, Appendix A) to a close approximation, we replace the exact solution with a one-shot approximation.

Let $z^+_{im} = -C^+_{im}/\tau$ and $z^-_{im} = -C^-_{im}/\tau$ be the scaled affinities, where $i$ is an index of a $q$ sample, and $m$ may be either a $p$-sample index or a $q$-sample index, depending on whether we refer to $C^+_{ij}$ or $C^-_{ij}$, respectively. We will focus on the case of the squared Euclidean distance,[7] $C^+_{im} = \|x_i-y_m\|^2$. We approximate the

[7] Note that the exact Sinkhorn flow derived in Proposition 1 uses the standard optimal transport ground cost $C(x,y) = \frac{1}{2}\|x-y\|^2$. The empirical algorithm's use of the unscaled distance $\|x-y\|^2$ introduces a constant factor of 2 into the spatial gradient, which in practice is absorbed by the step-size hyperparameter $\eta$ during optimization.

Sinkhorn solution via the geometric mean of the row-wise and column-wise softmax normalizations (with additional $N^+$ and $N^-$ scaling to ensure the population limits),[8]

$$A^+_{im} = \sqrt{\frac{\exp(z^+_{im})}{(N^+)^{-1}\sum_{u=1}^{N^+}\exp(z^+_{iu})} \cdot \frac{\exp(z^+_{im})}{(N^-)^{-1}\sum_{v=1}^{N^-}\exp(z^+_{vm})}},$$
$$A^-_{im} = \sqrt{\frac{\exp(z^-_{im})}{(N^-)^{-1}\sum_{u=1}^{N^-}\exp(z^-_{iu})} \cdot \frac{\exp(z^-_{im})}{(N^-)^{-1}\sum_{v=1}^{N^-}\exp(z^-_{vm})}},$$

where we again use $A^+_{im}$ when $m$ is a $p$-sample index, and $A^-_{im}$ for a $q$-sample index $m$. This serves as a pseudo-plan, used instead of the fully converged Sinkhorn solution. The relation of this approximation to the full Sinkhorn algorithm is described in Appendix C. Setting aside the scaling by $N^+$ and $N^-$, we emphasize that unlike the full Sinkhorn solution, the rows and columns of the unscaled $A$ are no longer guaranteed to sum to 1, and the use of the envelope theorem (Feydy et al., 2019) in obtaining the spatial gradient is no longer valid.

Because $A_{im}$ does not perfectly enforce the normalization conditions, we may not directly use $A^+$ and $A^-$ to compute the velocity field. This is addressed by a cross-weighting trick. Define

$$s^+_i = \frac{1}{N^+}\sum_j A^+_{ij} \quad \text{and} \quad s^-_i = \frac{1}{N^-}\sum_k A^-_{ik}$$

Writing

$$\begin{aligned}\widetilde{W}^+_{ij} &= \frac{1}{N^+} A^+_{ij} \cdot s^-_i \\ &= \frac{1}{N^+}\left(\frac{A^+_{ij}}{s^+_i}\right) \times \left(s^+_i \cdot s^-_i\right) \\ &= \underbrace{\left(\frac{A^+_{ij}}{\sum_j A^+_{ij}}\right)}_{\text{conditional plan } (\tilde{\pi}^+_{ij})} \times \underbrace{\left(s^+_i \cdot s^-_i\right)}_{\text{scaling}(Z_i)}\end{aligned}$$

and defining $Z_i = s^+_i s^-_i$, then

$$\begin{aligned}\widehat{V}^{\text{SP}}_{p,q}(x_i) &= -\sum_{j=1}^{N^+}\widetilde{W}^+_{ij}(x_i - y_j) + \sum_{k=1}^{N^-}\widetilde{W}^-_{ik}(x_i - x_k) \\ &= -s^-_i\left[\frac{1}{N^+}\sum_{j=1}^{N^+}A^+_{ij}(x_i - y_j)\right] + s^+_i\left[\frac{1}{N^-}\sum_{k=1}^{N^-}A^-_{ik}(x_i - x_k)\right] \\ &= Z_i\left(\sum_j \tilde{\pi}^+_{ij} y_j - \sum_k \tilde{\pi}^-_{ik} x_k\right)\end{aligned}$$

which resembles (26), scaled by a particle-specific $Z_i$. The drift vector $\widetilde{V}_i$ corresponds to the probability flow of the approximate Sinkhorn solution. We emphasize that due to the cross-weighting, the cancellation of $x_i$ in (26) still holds.

[8] In the original paper, the expression was

$$A^+_{im} = \sqrt{\left(\frac{\exp(z_{im})}{\sum_p^{N_+ + N_-}\exp(z_{ip})}\right)\left(\frac{\exp(z_{im})}{\sum_q^{N_-}\exp(z_{qm})}\right)}.$$

We replace the sum over $N_+ + N_-$ samples with a sum over only one set of samples, for reasons that will become clear.

### B.3 Alternative form, Sinkhorn proxy velocity field

An alternative form for the population velocity field is given in the following proposition.

**Proposition 2** (Population form of the Sinkhorn Proxy)**.** *Let* $\mathrm{d}\tilde{p}(y) = q_\tau^{-1/2}(y)\mathrm{d}p(y)$*, and* $\mathrm{d}\tilde{q}(x) = q_\tau^{-1/2}(x)\mathrm{d}q(x)$ *(if* $p$ *and* $q$ *have densities, this is equivalent to the density notation* $\tilde{p}(y) = p(y)/\sqrt{q_\tau(y)}$ *and* $\tilde{q}(y) = q(y)/\sqrt{q_\tau(y)}$*). Then the Sinkhorn Proxy velocity field (21) with Gaussian kernel has the population-limit form*

$$V^{\mathrm{SP}}_{p,q}(x) = \frac{\tau}{2}\, Z(x)\Big(\nabla \log(k * \tilde{p})(x) - \nabla \log(k * \tilde{q})(x)\Big), \tag{23}$$

*where* $Z(x) = s^+(x)\, s^-(x)$ *is a spatially varying pre-conditioner.*

*Proof.* Recall the original expression of the drift vector from (22),

$$V(x) = s^-(x) \int A^+(x,y) y \, dp(y) - s^+(x) \int A^-(x,x') x' \, dq(x'), \tag{27}$$

where

$$A^+(x,y) = \frac{k(x,y)}{\sqrt{p_\tau(x) q_\tau(y)}}, \qquad A^-(x,x') = \frac{k(x,x')}{\sqrt{q_\tau(x) q_\tau(x')}}.$$

and

$$s^+(x) = \int A^+(x,y)\, dp(y), \qquad s^-(x) = \int A^-(x,x')\, dq(x').$$

Using the definitions $d\tilde{p}(y) = q_\tau^{-1/2}(y)dp(y)$ and $d\tilde{q}(x) = q_\tau^{-1/2}(x)dq(x)$, we can express the row sums as kernel convolutions over these modified measures:

$$s^+(x) = \frac{1}{\sqrt{p_\tau(x)}} \int \frac{k(x,y)}{\sqrt{q_\tau(y)}}\, dp(y) = \frac{1}{\sqrt{p_\tau(x)}} (k * \tilde{p})(x), \tag{28}$$

$$s^-(x) = \frac{1}{\sqrt{q_\tau(x)}} \int \frac{k(x,x')}{\sqrt{q_\tau(x')}}\, dq(x') = \frac{1}{\sqrt{q_\tau(x)}} (k * \tilde{q})(x). \tag{29}$$

Next, we replace the proxy affinities $A^+$ and $A^-$ inside the expected destination integrals in (27), yielding:

$$V(x) = \frac{s^-(x)}{\sqrt{p_\tau(x)}} \int y k(x,y) d\tilde{p}(y) - \frac{s^+(x)}{\sqrt{q_\tau(x)}} \int x' k(x,x') d\tilde{q}(x'). \tag{30}$$

To evaluate these spatial integrals, we use Tweedie's formula for the unnormalized Gaussian kernel $k(x,z) = \exp(-\|x-z\|^2/\tau)$, which for a measure $\mu(z)$, yields

$$\frac{\int z k(x,z) d\mu(z)}{\int k(x,z) d\mu(z)} = x + \frac{\tau}{2} \nabla_x \log(k * \mu)(x).$$

Multiplying both sides by the denominator $\int k(x,z) d\mu(z) = (k * \mu)(x)$ yields:

$$\int z k(x,z) d\mu(z) = (k * \mu)(x) \left[ x + \frac{\tau}{2} \nabla_x \log(k * \mu)(x) \right].$$

We apply this result to (30) for both modified measures $\tilde{p}$ and $\tilde{q}$. From our row sum derivations in (28) and (29), we have $(k * \tilde{p})(x) = \sqrt{p_\tau(x)} s^+(x)$ and $(k * \tilde{q})(x) = \sqrt{q_\tau(x)} s^-(x)$. Substituting these relations into the right-hand side of Tweedie's formula gives:

$$\int y k(x,y) d\tilde{p}(y) = \sqrt{p_\tau(x)} s^+(x) \left[ x + \frac{\tau}{2} \nabla_x \log(k * \tilde{p})(x) \right],$$

$$\int x' k(x,x') d\tilde{q}(x') = \sqrt{q_\tau(x)} s^-(x) \left[ x + \frac{\tau}{2} \nabla_x \log(k * \tilde{q})(x) \right].$$

Finally, we substitute these evaluated integrals back into the drift vector expression (30):

$$\begin{aligned} V(x) &= \frac{s^-(x)}{\sqrt{p_\tau(x)}}\left(\sqrt{p_\tau(x)}s^+(x)\left[x+\frac{\tau}{2}\nabla_x \log(k*\tilde{p})(x)\right]\right) \\ &\quad - \frac{s^+(x)}{\sqrt{q_\tau(x)}}\left(\sqrt{q_\tau(x)}s^-(x)\left[x+\frac{\tau}{2}\nabla_x \log(k*\tilde{q})(x)\right]\right) \\ &= s^-(x)s^+(x)\left[x+\frac{\tau}{2}\nabla_x \log(k*\tilde{p})(x)\right] - s^+(x)s^-(x)\left[x+\frac{\tau}{2}\nabla_x \log(k*\tilde{q})(x)\right]. \end{aligned}$$

Defining $Z(x) = s^+(x)s^-(x)$ and expanding the brackets, the coordinates $x$ cancel:

$$V(x) = \frac{\tau}{2}Z(x)\big(\nabla_x \log(k*\tilde{p})(x) - \nabla_x \log(k*\tilde{q})(x)\big).$$

□

### B.4 Consistency of the Sinkhorn Proxy

**Theorem 3** (Consistency)**.** *Assume that the weighted measures $\tilde{p}$ and $\tilde{q}$ have finite mass, and that a Gaussian kernel $k$ is used. Then $V = 0$ for all $x \in \mathbb{R}^d$ if and only if $p = q$.*

*Proof.* We analyze the condition under which the cross-weighted vector field $V^{\mathrm{SP}}_{p,q}(x)$ vanishes everywhere. ( $\Longleftarrow$ ) Assume $p = q$. It follows that the modified measures are identical:

$$d\tilde{p}(y) = q_\tau^{-1/2}(y)dp(y) = q_\tau^{-1/2}(x)dq(x) = d\tilde{q}(y) \quad \forall y.$$

Consequently, their convolutions with the kernel $k$ are identical, yielding $\nabla_x \log(k*\tilde{p})(x) = \nabla_x \log(k*\tilde{q})(x)$ for all $x$, and the velocity field is identically zero everywhere.

( $\Longrightarrow$ ) Assume $V^{\mathrm{SP}}_{p,q}(x) = 0$ for all $x$. The pre-conditioner $Z(x)$ is strictly positive everywhere (as the Gaussian $k$ is supported on $\mathbb{R}^d$), hence we must have:

$$\nabla_x \log(k*\tilde{p})(x) = \nabla_x \log(k*\tilde{q})(x) \quad \forall x.$$

Integrating this spatial gradient implies that the kernel-smoothed measures are proportional up to a global constant $\lambda > 0$:

$$(k*\tilde{p})(x) = \lambda(k*\tilde{q})(x) \quad \forall x.$$

When $k$ is characteristic (Sriperumbudur et al., 2010) and translation invariant on $\mathbb{R}^d$, then the mean embedding map $\mu \mapsto k*\mu$ is injective for finite signed measures (Sriperumbudur et al., 2011, Proposition 2 and Section 3.2). This property holds for the Gaussian kernel, since its Fourier transform is everywhere non-zero. Assuming that the modified measures $\tilde{p}$ and $\tilde{q}$ have finite mass, then injectivity applies. Therefore, if the smoothed measures are proportional, the underlying modified measures must themselves be proportional:

$$d\tilde{p}(y) = \lambda d\tilde{q}(y) \quad \forall y.$$

Substituting the definitions of $\tilde{p}$ and $\tilde{q}$ yields:

$$q_\tau^{-1/2}(y)dp(y) = \lambda q_\tau^{-1/2}(x)dq(x)$$

Because the kernel-smoothed model density is strictly positive ($q_\tau(y) > 0$), then

$$dp(y) = \lambda dq(y) \quad \forall y.$$

Since $p$ and $q$ are both valid probability density functions representing the target data distribution and the model distribution, respectively, they must integrate to 1,

$$\int p(y)\,dy = \lambda \int q(y)\,dy \implies \lambda = 1.$$

□

### B.5 The Sinkhorn Proxy Velocity does not generally correspond to a Wasserstein Gradient Flow

**Proposition 4** (Non-gradient structure)**.** *The velocity field $V^{\mathrm{SP}}_{p,q}$ in (23), with a Gaussian kernel, corresponds to a valid WGF only when the gradients of the pre-conditioner and the score difference are linearly dependent everywhere:*

$$\partial_i Z(x)\partial_j \Phi(x) = \partial_j Z(x)\partial_i \Phi(x) \quad \textit{for all } i, j,$$

*where $\Phi(x) = \frac{\tau}{2}(\log(k * \tilde{p})(x) - \log(k * \tilde{q})(x))$. Equivalently, the wedge product must vanish: $\nabla Z \wedge \nabla \Phi = 0$.*

*Proof.* To be a valid Wasserstein gradient flow, the associated velocity vector field $V^{\mathrm{SP}}_{p,q}(x)$ driving the particles must take the form of a pure gradient field (see (3)). Thus, there must exist a scalar function (the first variation $\frac{\delta \mathcal{F}_p}{\delta q}(x)$) whose gradient produces this field.

We start with the form in proposition 2. Define the term inside the parenthesis as

$$\Phi(x) = \frac{\tau}{2}\Big(\log(k * \tilde{p})(x) - \log(k * \tilde{q})(x)\Big).$$

We can then rewrite the Sinkhorn Proxy drift concisely as:

$$V^{\mathrm{SP}}_{p,q}(x) = Z(x)\nabla_x \Phi(x).$$

By Poincaré's Lemma, a continuously differentiable vector field on a simply connected domain is a pure gradient field if and only if its Jacobian matrix is symmetric. Let $V_i$ denote the $i$-th component of $V^{\mathrm{SP}}_{p,q}$. The symmetry condition $\partial_j V_i = \partial_i V_j$ implies:

$$\begin{aligned} \partial_j\big(Z(x)\partial_i \Phi(x)\big) &= \partial_i\big(Z(x)\partial_j \Phi(x)\big) \\ \partial_j Z(x)\partial_i \Phi(x) + Z(x)\partial_j\partial_i \Phi(x) &= \partial_i Z(x)\partial_j \Phi(x) + Z(x)\partial_i\partial_j \Phi(x). \end{aligned}$$

By Schwarz's theorem, the mixed partial derivatives of $\Phi$ are equal ($\partial_j\partial_i \Phi = \partial_i\partial_j \Phi$), so

$$\partial_j Z(x)\partial_i \Phi(x) = \partial_i Z(x)\partial_j \Phi(x).$$

In other words, the gradient of the scalar pre-conditioner $\nabla Z(x)$ must be perfectly parallel to the gradient of the score difference $\nabla \Phi(x)$ at all points in space. Given their distinct definitions, this geometric condition will not hold in general. □

## C Geometric mean as approximation to Sinkhorn solution

We demonstrate the relation between $A^+$ and the optimal transport plan obtained via Sinkhorn iteration.

In entropy-regularized optimal transport (Cuturi, 2013), the interaction between points is governed by a strictly positive Gibbs kernel matrix $K$. The algorithm defines the independent pairwise logits as $Z^+_{ij} = -C^+_{ij}/\tau$. Exponentiating these logits yields the Gibbs kernel:

$$K_{ij} = \exp(Z^+_{ij}) = \exp\left(-\frac{C^+_{ij}}{\tau}\right)$$

The objective of the Sinkhorn algorithm is to find the strictly positive vectors $u$ and $v$ such that the resulting transport plan $P = \mathrm{diag}(u)K\mathrm{diag}(v)$ satisfies target marginal constraints. To recover the Sinkhorn proxy algorithm, these marginals are assumed to be *uniform*,

$$(P\mathbf{1})^\top = \mathbf{1}^\top P = \mathbf{1}.$$

The resulting optimal transport plan $P^*$ takes the form

$$P^*_{ij} = u_i K_{ij} v_j$$

The standard iterative Sinkhorn algorithm finds $u$ and $v$ by repeatedly alternating between dividing by row sums and column sums until the marginals match the target distributions.

Assume first that Sinkhorn initializes the column scaling vector as a vector of ones: $v^{(0)} = \mathbf{1}$. The first update for the source scaling vector $u$ is computed by dividing the target marginals by the row sums of the current scaled kernel,

$$u_i^{(1)} = \frac{1}{(Kv^{(0)})_i} = \frac{1}{(K\mathbf{1})_i} = \frac{1}{r_i}$$

Applying this scaling alone yields $u_i^{(1)} K_{ij} = K_{ij}/r_i$, which corresponds exactly to the `softmax_row` operation. Consider now the case that we begin the Sinkhorn iterations with $(u^{(0)} = \mathbf{1})$. We would then make the update

$$v_j^{(1)} = \frac{1}{(K^T u^{(0)})_j} = \frac{1}{(K^T\mathbf{1})_j} = \frac{1}{c_j}.$$

Turning now to the approach of Deng et al. (2026, Algorithm 2), we examine the relation between the matrix $A^+$ and the Sinkhorn structure. First, the row-wise softmax of the logits $Z^+$ is equivalent to the Gibbs kernel divided by its row sums. Let $r_i = \sum_m K_{im}$ be the $i$-th row sum:

$$\mathrm{softmax}_{\mathrm{row}}(Z^+)_{ij} = \frac{\exp(Z_{ij}^+)}{\sum_m \exp(Z_{im}^+)} = \frac{K_{ij}}{r_i}$$

Similarly, the column-wise softmax is the Gibbs kernel divided by its column sums. Let $c_j = \sum_n K_{nj}$ be the $j$-th column sum:

$$\mathrm{softmax}_{\mathrm{col}}(Z^+)_{ij} = \frac{\exp(Z_{ij}^+)}{\sum_n \exp(Z_{nj}^+)} = \frac{K_{ij}}{c_j}$$

The update of Deng et al. corresponds to a geometric mean of these normalization operations,[9]

$$A_{ij}^+ = \frac{K_{ij}}{\sqrt{r_i \cdot c_j}}$$

rather than iterating between them. When $N^+ = N^-$, then repeated iteration of this procedure converges to the entropy-regularised transport map (Knight et al., 2014, Theorem 3.1), as with the better known Sinkhorn approach that alternates row and column normalization.

While the above correction does not guarantee the marginal constraints of a fully converged Sinkhorn approach (which would require infinite sequential iterations), it does capture the property of "hub suppression:" If a target $y_j$ is a hub (a region of high density surrounded by many source points), its column sum $c_j$ will be large. Consequently, the column scaling factor $1/\sqrt{c_j}$ becomes small, which down-scales the affinity of particles to that hub.

## D Failure mode on a synthetic example

We first describe the velocity field of the KL-KDE in the regime where $D, \alpha, \beta$ are fixed and $\tau \downarrow 0$.

**Proposition 5.** *For fixed $D > 0$ and fixed $\alpha, \beta \in (0,1)$ with $\beta < \alpha$, as $\tau \downarrow 0$ (equivalently $\epsilon = \exp(-4D^2/\tau) \downarrow 0$ and (24) holds) the velocity field $V_{\mathrm{KL}}$ of the KL-KDE flow with Gaussian kernel at $x = +D$ is given by*

$$V_{p,q}^{\mathrm{KL}}(D) = -2D\epsilon\left(\frac{\alpha}{1-\alpha} - \frac{\beta}{1-\beta}\right) + o(\epsilon).$$

*Proof.* The continuous KL-KDE drift evaluates the difference between the scores of the kernel-smoothed distributions:

$$V_{p,q}^{\mathrm{KL}}(x) = \frac{\tau}{2}\big(\nabla_x \log p_\tau(x) - \nabla_x \log q_\tau(x)\big).$$

[9] In (20) there is additional scaling by $N^+$ and $N^-$, which we omit here - see discussion in main text.

We evaluate the score of the target distribution $p_\tau$ at $x = +D$. The smoothed density is

$$p_\tau(D) = (1-\alpha)k(D,D) + \alpha k(D,-D) = (1-\alpha) + \alpha\epsilon.$$

The gradient of the unnormalized Gaussian kernel $k(x,z) = \exp(-\|x-z\|^2/\tau)$ is $\nabla_x k(x,z) = -\frac{2(x-z)}{\tau}k(x,z)$. Evaluating this at $x = +D$ gives $\nabla_x k(D,D) = 0$ and $\nabla_x k(D,-D) = -\frac{4D}{\tau}\epsilon$. The gradient of the smoothed target distribution is

$$\nabla_x p_\tau(D) = (1-\alpha)\nabla_x k(D,D) + \alpha\nabla_x k(D,-D) = \alpha\left(-\frac{4D}{\tau}\epsilon\right).$$

The score function is the ratio of these two quantities:

$$\nabla_x \log p_\tau(D) = \frac{\nabla_x p_\tau(D)}{p_\tau(D)} = \frac{-\frac{4D}{\tau}\alpha\epsilon}{(1-\alpha)+\alpha\epsilon}.$$

Assuming the mixing proportion $\alpha \neq 1$),

$$\frac{1}{(1-\alpha)+\alpha\epsilon} = \frac{1}{1-\alpha}\left(\frac{1}{1+\frac{\alpha}{1-\alpha}\epsilon}\right).$$

Expanding $(1+z)^{-1} = 1 - z + \mathcal{O}(z^2)$ for small $z$,

$$\begin{aligned}\nabla_x \log p_\tau(D) &= \left(-\frac{4D}{\tau}\alpha\epsilon\right)\frac{1}{1-\alpha}\left(1 - \frac{\alpha}{1-\alpha}\epsilon + \mathcal{O}(\epsilon^2)\right)\\ &= -\frac{4D}{\tau}\epsilon\left(\frac{\alpha}{1-\alpha}\right) + \mathcal{O}\left(\frac{\epsilon^2}{\tau}\right).\end{aligned}$$

To verify that the remainder $\mathcal{O}(\frac{\epsilon^2}{\tau})$ is indeed $o(\epsilon)$, we must show that $\lim_{\tau\to 0}\frac{\epsilon^2/\tau}{\epsilon} = 0$. Using the fact that $\epsilon = \exp(-4D^2/\tau)$, we can express $\tau$ as $\tau = -\frac{4D^2}{\ln\epsilon}$. Then:

$$\lim_{\epsilon\to 0}\frac{\epsilon^2/\tau}{\epsilon} = \lim_{\epsilon\to 0}\frac{\epsilon}{\tau} = \lim_{\epsilon\to 0}\frac{-\epsilon\ln\epsilon}{4D^2} = 0. \tag{31}$$

Since the limit is zero, the remainder decays faster than $\epsilon$. Therefore, $\mathcal{O}(\frac{\epsilon^2}{\tau}) \subset o(\epsilon)$, and thus,

$$\begin{aligned}\frac{\tau}{2}\nabla_x \log p_\tau(D) &= \frac{\tau}{2}\left[-\frac{4D}{\tau}\epsilon\left(\frac{\alpha}{1-\alpha}\right) + \mathcal{O}\left(\frac{\epsilon^2}{\tau}\right)\right]\\ &= -2D\epsilon\left(\frac{\alpha}{1-\alpha}\right) + o(\epsilon).\end{aligned}$$

Applying the same reasoning to the model distribution $q_\tau$, with mixing proportion $\beta$, yields:

$$\frac{\tau}{2}\nabla_x \log q_\tau(D) = -2D\epsilon\left(\frac{\beta}{1-\beta}\right) + o(\epsilon).$$

Subtracting the two scaled scores gives the net KL-KDE drift:

$$\begin{aligned}V_{p,q}^{\mathrm{KL}}(D) &= \frac{\tau}{2}\nabla_x \log p_\tau(D) - \frac{\tau}{2}\nabla_x \log q_\tau(D)\\ &= \left[-2D\epsilon\left(\frac{\alpha}{1-\alpha}\right) + o(\epsilon)\right] - \left[-2D\epsilon\left(\frac{\beta}{1-\beta}\right) + o(\epsilon)\right]\\ &= -2D\epsilon\left(\frac{\alpha}{1-\alpha} - \frac{\beta}{1-\beta}\right) + o(\epsilon).\end{aligned}$$

□

The drift points in the correct direction (negative, since we assume $\beta < \alpha$), but its magnitude is scaled by $\epsilon$. Because $\tau$ is narrow relative to the mode separation $2D$, $\epsilon$ may drop below machine precision ($\epsilon \approx 0$). We now show that the Sinkhorn proxy method has the same failure mode.

**Proposition 6.** *For fixed $D > 0$ and fixed $\alpha, \beta \in (0,1)$ with $\beta < \alpha$, as $\tau \downarrow 0$ (equivalently $\epsilon = \exp(-4D^2/\tau) \downarrow 0$), the velocity field of the Sinkhorn Proxy flow with Gaussian kernel at $x = D$ is given by*

$$V^{\mathrm{SP}}_{p,q}(D) = -2D\epsilon \underbrace{\left[\sqrt{\frac{1-\alpha}{\beta}}\right]}_{(*)} \left(\frac{\alpha}{1-\alpha} - \frac{\beta}{1-\beta}\right) + o(\epsilon),$$

*where the term (*) represents the difference with the KL-KDE flow.*

*Proof.* We begin with the population-limit velocity field established in proposition 2,

$$V^{\mathrm{SP}}_{p,q}(x) = \frac{\tau}{2} Z(x)\big(\nabla_x \log(k * \tilde{p})(x) - \nabla_x \log(k * \tilde{q})(x)\big),$$

where $d\tilde{p}(y) = q_\tau^{-1/2}(y)dp(y)$ and $d\tilde{q}(x) = q_\tau^{-1/2}(x)dq(x)$. We evaluate the algorithmic drift $V^{\mathrm{SP}}_{p,q}(x)$ at the right-hand overcrowded mode $x = +D$.

First, we evaluate the kernel-smoothed model density $q_\tau$ at both modes, expressing the result as $1 + \mathcal{O}(\epsilon)$ perturbations:

$$\begin{aligned} q_\tau(D) &= \beta k(D,-D) + (1-\beta)k(D,D) = (1-\beta) + \beta\epsilon = (1-\beta)(1+\mathcal{O}(\epsilon)) \\ q_\tau(-D) &= \beta k(-D,-D) + (1-\beta)k(-D,D) = \beta + (1-\beta)\epsilon = \beta(1+\mathcal{O}(\epsilon)) \end{aligned}$$

Using the Taylor expansion $(1+\mathcal{O}(\epsilon))^{-1/2} = 1 + \mathcal{O}(\epsilon)$, we substitute these into the modified distributions $\tilde{p}$ and $\tilde{q}$:

$$\begin{aligned} d\tilde{p}(y) &= \frac{\alpha}{\sqrt{q_\tau(-D)}}\delta_{-D}(y) + \frac{1-\alpha}{\sqrt{q_\tau(D)}}\delta_{+D}(y) = \left(\frac{\alpha}{\sqrt{\beta}} + \mathcal{O}(\epsilon)\right)\delta_{-D}(y) + \left(\frac{1-\alpha}{\sqrt{1-\beta}} + \mathcal{O}(\epsilon)\right)\delta_{+D}(y) \\ d\tilde{q}(x) &= \frac{\beta}{\sqrt{q_\tau(-D)}}\delta_{-D}(x) + \frac{1-\beta}{\sqrt{q_\tau(D)}}\delta_{+D}(x) = \left(\sqrt{\beta} + \mathcal{O}(\epsilon)\right)\delta_{-D}(x) + \left(\sqrt{1-\beta} + \mathcal{O}(\epsilon)\right)\delta_{+D}(x) \end{aligned}$$

Next, we evaluate the kernel-smoothed modified measures at $x = +D$. Since $k(D,-D) = \epsilon$ and $k(D,D) = 1$:

$$\begin{aligned} (k * \tilde{p})(D) &= \left(\frac{\alpha}{\sqrt{\beta}} + \mathcal{O}(\epsilon)\right)\epsilon + \left(\frac{1-\alpha}{\sqrt{1-\beta}} + \mathcal{O}(\epsilon)\right)(1) = \frac{1-\alpha}{\sqrt{1-\beta}} + \mathcal{O}(\epsilon) \\ (k * \tilde{q})(D) &= \left(\sqrt{\beta} + \mathcal{O}(\epsilon)\right)\epsilon + \left(\sqrt{1-\beta} + \mathcal{O}(\epsilon)\right)(1) = \sqrt{1-\beta} + \mathcal{O}(\epsilon) \end{aligned} \tag{32}$$

To compute the score functions, we evaluate the gradients at $x = +D$. For the unnormalized Gaussian kernel $k(x,z) = \exp(-\|x-z\|^2/\tau)$, the gradients are $\nabla_x k(x,-D)|_{x=D} = -\frac{4D}{\tau}\epsilon$ and $\nabla_x k(x,D)|_{x=D} = 0$. This yields:

$$\begin{aligned} \nabla_x(k * \tilde{p})(D) &= \left(\frac{\alpha}{\sqrt{\beta}} + \mathcal{O}(\epsilon)\right)\left(-\frac{4D}{\tau}\epsilon\right) + \left(\frac{1-\alpha}{\sqrt{1-\beta}} + \mathcal{O}(\epsilon)\right)(0) = -\frac{4D}{\tau}\epsilon\frac{\alpha}{\sqrt{\beta}} + \mathcal{O}\left(\frac{\epsilon^2}{\tau}\right) \\ \nabla_x(k * \tilde{q})(D) &= \left(\sqrt{\beta} + \mathcal{O}(\epsilon)\right)\left(-\frac{4D}{\tau}\epsilon\right) + \left(\sqrt{1-\beta} + \mathcal{O}(\epsilon)\right)(0) = -\frac{4D}{\tau}\epsilon\sqrt{\beta} + \mathcal{O}\left(\frac{\epsilon^2}{\tau}\right) \end{aligned}$$

The score functions at $x = +D$ are then the ratio of the gradients to the densities. Using the quotient Taylor expansion $\frac{A\epsilon+\mathcal{O}(\epsilon^2/\tau)}{B+\mathcal{O}(\epsilon)} = \frac{A}{B}\epsilon + \mathcal{O}\left(\frac{\epsilon^2}{\tau}\right)$ and that $\mathcal{O}(\epsilon^2/\tau) = o(\epsilon)$ from (31),

$$\nabla_x \log(k * \tilde{p})(D) = \frac{-\frac{4D}{\tau}\epsilon\frac{\alpha}{\sqrt{\beta}} + \mathcal{O}\left(\frac{\epsilon^2}{\tau}\right)}{\frac{1-\alpha}{\sqrt{1-\beta}} + \mathcal{O}(\epsilon)} = -\frac{4D}{\tau}\epsilon\left(\frac{\alpha}{1-\alpha}\right)\sqrt{\frac{1-\beta}{\beta}} + o(\epsilon)$$

$$\nabla_x \log(k * \tilde{q})(D) = \frac{-\frac{4D}{\tau}\epsilon\sqrt{\beta} + \mathcal{O}\left(\frac{\epsilon^2}{\tau}\right)}{\sqrt{1-\beta} + \mathcal{O}(\epsilon)} = -\frac{4D}{\tau}\epsilon\sqrt{\frac{\beta}{1-\beta}} + o(\epsilon)$$

Thus

$$\frac{\tau}{2}\left(\nabla_x \log(k * \tilde{p})(D) - \nabla_x \log(k * \tilde{q})(D)\right) = -2D\epsilon\sqrt{\frac{1-\beta}{\beta}}\left[\frac{\alpha}{1-\alpha} - \frac{\beta}{1-\beta}\right] + \frac{\tau}{2}o(\epsilon)$$

Finally, we consider the scaling term

$$Z(D) = s^+(D)s^-(D) = \frac{1}{\sqrt{p_\tau(D)}}(k * \tilde{p})(D) \cdot \frac{1}{\sqrt{q_\tau(D)}}(k * \tilde{q})(D)$$

We have

$$\frac{1}{\sqrt{p_\tau(D)}} = \frac{1}{\sqrt{1-\alpha}} + \mathcal{O}(\epsilon), \qquad \frac{1}{\sqrt{q_\tau(D)}} = \frac{1}{\sqrt{1-\beta}} + \mathcal{O}(\epsilon),$$

and thus, from previous results (32) on $(k * \tilde{p})(D)$ and $(k * \tilde{q})(D)$, that

$$s^+(D) = \frac{\sqrt{1-\alpha}}{\sqrt{1-\beta}} + \mathcal{O}(\epsilon), \qquad s^-(D) = 1 + \mathcal{O}(\epsilon).$$

It follows that

$$Z(D) = \frac{\sqrt{1-\alpha}}{\sqrt{1-\beta}} + \mathcal{O}(\epsilon).$$

Substituting these into the drift equation gives:

$$V^{\mathrm{SP}}_{p,q}(D) = -2D\epsilon\left[\sqrt{\frac{1-\alpha}{\beta}}\right]\left(\frac{\alpha}{1-\alpha} - \frac{\beta}{1-\beta}\right) + o(\epsilon).$$

□

This confirms that in the present setting, despite its superficial resemblance to a Sinkhorn flow, the Sinkhorn Proxy algorithm does not share the mass transport dynamics of an Earth Mover's Distance, but behaves more like a flow on KL. We suspect that Deng et al.'s Algorithm 2 also has this problem.

**Proposition 7.** *The barycentric velocity of the WGF on $\frac{1}{2}W_2^2(q,p)$ at $x = D$ is given by*

$$V^{W_2}_{p,q}(D) = -2D\left(\frac{\alpha-\beta}{1-\beta}\right).$$

*Proof.* We compute the barycentric velocity at $x = +D$ for the Wasserstein gradient flow of $\frac{1}{2}W_2^2(q,p)$. Because the measures are atomic and the optimal plan can split the mass at $+D$, this should be understood as the barycentric projection of the optimal plan, or as the limit obtained by replacing the atoms by small absolutely continuous clusters. As defined in the main text, the target data distribution $p$ and the model distribution $q$ are mixtures of two isolated point masses:

$$p(y) = \alpha\delta_{-D}(y) + (1-\alpha)\delta_{+D}(y),$$
$$q(x) = \beta\delta_{-D}(x) + (1-\beta)\delta_{+D}(x).$$

We assume $\beta < \alpha$. This means the model $q$ is "starved" of mass at the left mode $(-D)$ and "overcrowded" at the right mode $(+D)$.

To find the Wasserstein-2 velocity, we must first find the optimal transport plan $\pi(x, y)$ that minimizes the quadratic cost $C(x, y) = \frac{1}{2}\|x - y\|^2$ to move mass from $q$ to $p$. The joint distribution $\pi(x, y)$ must satisfy the following marginal constraints:

$$\sum_y \pi(x, y) = q(x), \qquad \sum_x \pi(x, y) = p(y).$$

To minimize the cost, mass should remain in place wherever possible (cost = 0).

- All mass $\beta$ at $x = -D$ in $q$ will stay at $y = -D$ in $p$. Thus, $\pi(-D, -D) = \beta$.
- Because the target $p$ requires a total mass of $\alpha$ at $-D$, the remaining deficit of $(\alpha - \beta)$ must be transported from $x = +D$. Thus, $\pi(+D, -D) = \alpha - \beta$.
- The remaining mass at $x = +D$ stays at $y = +D$. Thus, $\pi(+D, +D) = (1 - \beta) - (\alpha - \beta) = 1 - \alpha$.
- Finally, to obtain that the probability sum at one, we have $\pi(-D, +D) = 0$.

For probability mass at $x = +D$, the Wasserstein gradient flow velocity is

$$V_{p,q}^{W_2}(x) = -\int \nabla_x C(x, y)\, d\pi(y|x),$$

with the expectation over the optimal *conditional* transport plan $d\pi(y|x)$ (Ambrosio et al., 2008).

We compute the conditional probability $\pi(y \mid x) = \frac{\pi(x,y)}{q(x)}$ for the overcrowded mode $x = +D$:

The total available mass at $x = +D$ is $q(+D) = 1 - \beta$.

- Fraction of mass moving to $-D$: $\pi(-D \mid +D) = \frac{\alpha-\beta}{1-\beta}$
- Fraction of mass staying at $+D$: $\pi(+D \mid +D) = \frac{1-\alpha}{1-\beta}$

The velocity vector $V_{p,q}^{W_2}$ for the mass at $x$ is the difference between its expected target destination and its current position:

$$V_{p,q}^{W_2}(x) = \mathbb{E}_{Y \sim \pi(\cdot|x)}[Y] - x$$

Evaluating this at $x = +D$:

$$V_{p,q}^{W_2}(+D) = [\pi(-D \mid +D)(-D) + \pi(+D \mid +D)(+D)] - D.$$

Substituting the conditional probabilities,

$$\begin{aligned} V_{p,q}^{W_2}(+D) &= \left[\left(\frac{\alpha - \beta}{1 - \beta}\right)(-D) + \left(\frac{1 - \alpha}{1 - \beta}\right)(+D)\right] - D \\ &= D\left[\frac{-(\alpha - \beta) + (1 - \alpha) - (1 - \beta)}{1 - \beta}\right] \\ &= -2D\left(\frac{\alpha - \beta}{1 - \beta}\right) \end{aligned}$$

which concludes the proof.

□

## E Proof of Proposition 8

**Proposition 8** (Surrogate-gradient identity)**.** *Let $q_\theta = (f_\theta)_\# \mu$ be the pushforward of $\mu$ through $f_\theta$, and suppose we have the exact velocity field $V_{p,q_\theta}(x) = -\nabla_x \frac{\delta \mathcal{F}_p}{\delta q_\theta}(x)$ for some functional $\mathcal{F}_p$. Then*

$$\nabla_\theta \mathcal{L}(\theta) = 2\eta \, \nabla_\theta \mathcal{F}_p(q_\theta).$$

*Proof.* Write $x = f_\theta(\epsilon)$ and let $t = \mathrm{sg}(x + \eta V_{p,q_\theta}(x))$ denote the stop-gradient target, which is treated as a constant with respect to $\theta$. Then $\mathcal{L}(\theta) = \mathbb{E}_{\epsilon\sim\mu}[\|f_\theta(\epsilon) - t\|^2]$, and applying the chain rule yields:

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \mathbb{E}_{\epsilon\sim\mu}\Big[2\big(f_\theta(\epsilon) - t\big)^\top \nabla_\theta f_\theta(\epsilon)\Big] \\
&= \mathbb{E}_{\epsilon\sim\mu}\Big[2\big(f_\theta(\epsilon) - f_\theta(\epsilon) - \eta V_{p,q_\theta}(f_\theta(\epsilon))\big)^\top \nabla_\theta f_\theta(\epsilon)\Big] \\
&= -2\eta \, \mathbb{E}_{\epsilon\sim\mu}\Big[V_{p,q_\theta}\big(f_\theta(\epsilon)\big)^\top \nabla_\theta f_\theta(\epsilon)\Big].
\end{aligned}$$

A perturbation $\delta\theta$ changes the functional through the marginal distribution $q_\theta$. By the definition of the first variation:

$$\nabla_\theta \mathcal{F}_p(q_\theta) = \int \frac{\delta \mathcal{F}_p}{\delta q_\theta}(x) \, \nabla_\theta q_\theta(x) \, \mathrm{d}x.$$

Since $q_\theta = (f_\theta)_\# \mu$, the parameter perturbation $\delta\theta$ induces a displacement field on the particles in data space. Because multiple noise instances $\epsilon$ could map to the same spatial point $x$ (if $f_\theta$ is non-invertible), we define the effective Eulerian velocity field at spatial location $x$ via the conditional expectation:

$$v_\theta(x) = \mathbb{E}_{\epsilon\sim\mu}\big[\nabla_\theta f_\theta(\epsilon) \mid f_\theta(\epsilon) = x\big].$$

The evolution of the density $q_\theta$ with respect to $\theta$ is mathematically governed by the continuity equation $\nabla_\theta q_\theta(x) = -\nabla_x \cdot (q_\theta(x) v_\theta(x))$. Substituting this into the first variation and applying the divergence theorem (assuming boundary terms vanish) gives:

$$\begin{aligned}
\nabla_\theta \mathcal{F}_p(q_\theta) &= -\int \frac{\delta \mathcal{F}_p}{\delta q_\theta}(x) \, \nabla_x \cdot \big(q_\theta(x) \, v_\theta(x)\big) \, \mathrm{d}x \\
&= \int \nabla_x \frac{\delta \mathcal{F}_p}{\delta q_\theta}(x)^\top v_\theta(x) \; q_\theta(x) \, \mathrm{d}x.
\end{aligned}$$

Using the law of total expectation (the tower property), we can rewrite this integral over the spatial domain $\mathbb{R}^d$ as an expectation directly over the latent noise distribution $\mu$:

$$\begin{aligned}
\nabla_\theta \mathcal{F}_p(q_\theta) &= \mathbb{E}_{x\sim q_\theta}\Big[\nabla_x \frac{\delta \mathcal{F}_p}{\delta q_\theta}(x)^\top \mathbb{E}_{\epsilon\sim\mu}\big[\nabla_\theta f_\theta(\epsilon) \mid f_\theta(\epsilon) = x\big]\Big] \\
&= \mathbb{E}_{\epsilon\sim\mu}\Big[\nabla_x \frac{\delta \mathcal{F}_p}{\delta q_\theta}\big(f_\theta(\epsilon)\big)^\top \nabla_\theta f_\theta(\epsilon)\Big]. \qquad (33)
\end{aligned}$$

Finally, substituting the definition of the Wasserstein gradient $V_{p,q_\theta}(x) = -\nabla_x \frac{\delta \mathcal{F}_p}{\delta q_\theta}(x)$ into (33) directly recovers the relation $\nabla_\theta \mathcal{L}(\theta) = 2\eta \, \nabla_\theta \mathcal{F}_p(q_\theta)$. □

## F Additional results

In Figures 3-7, we present experimental results for various toy examples.

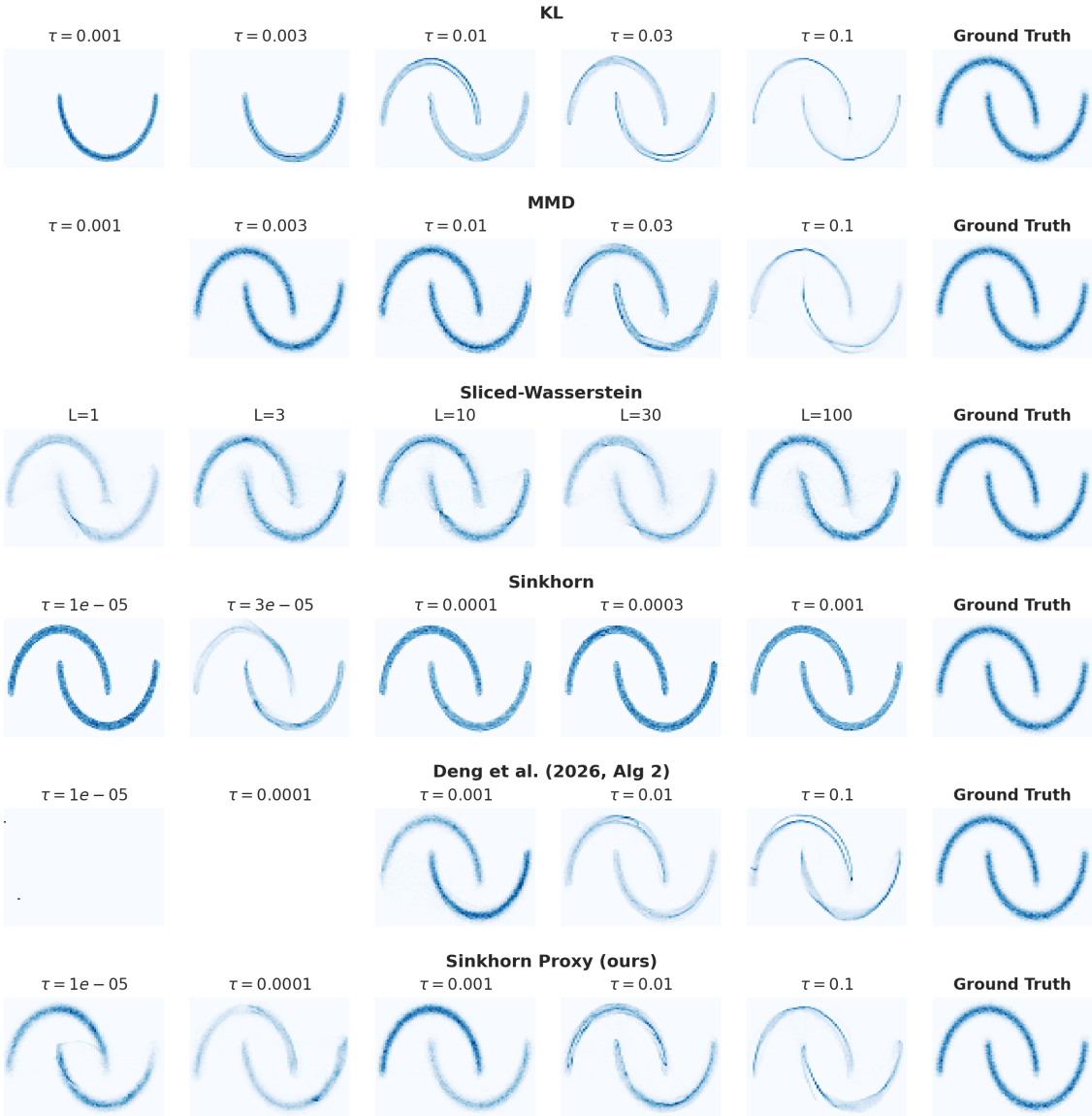


Figure 3: True and generated samples for different types of drift and hyperparameters. Empty panel means the samples have diverged.

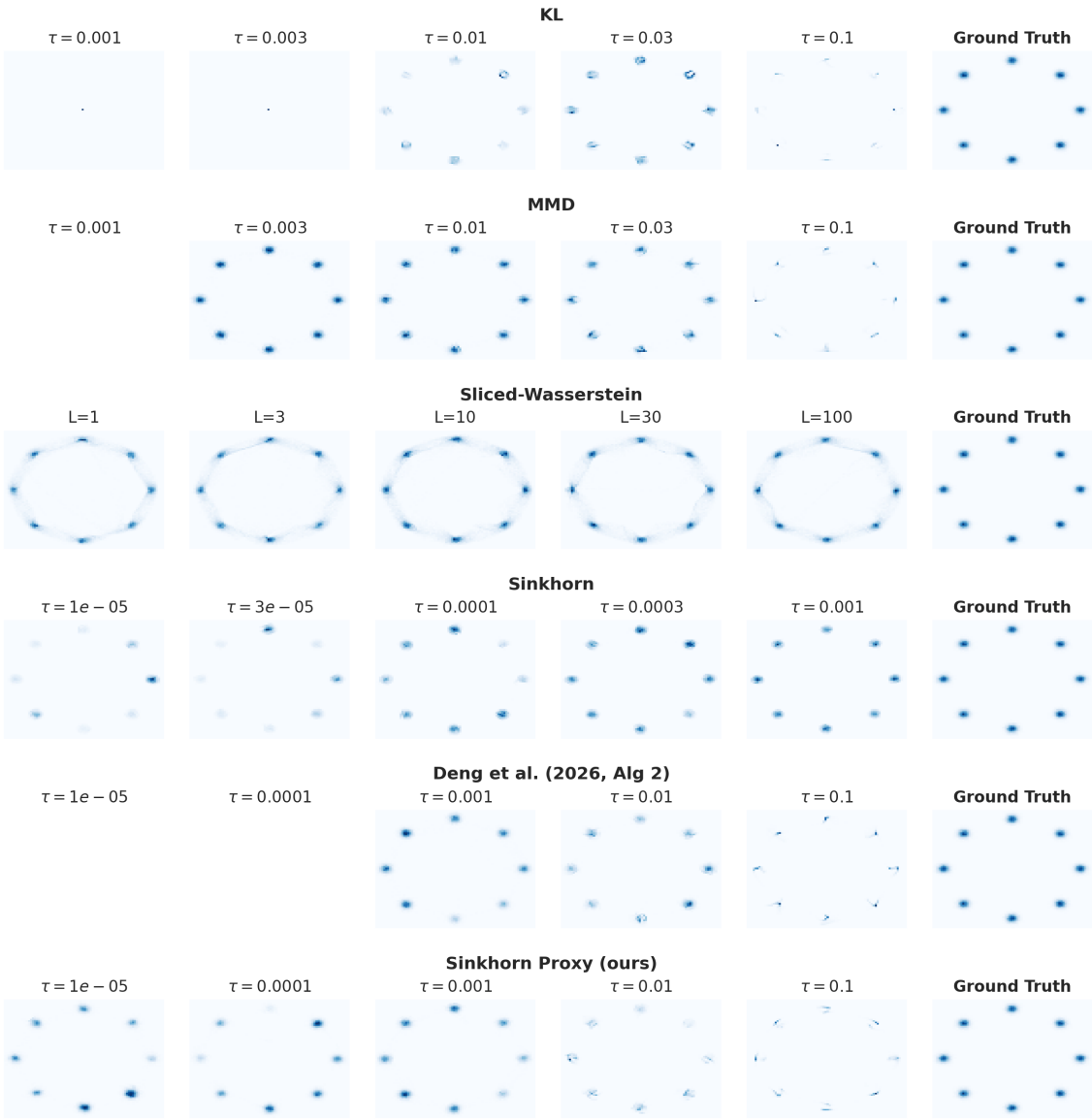


Figure 4: Results for the 8 Gaussian dataset. Empty panel means the samples have diverged.

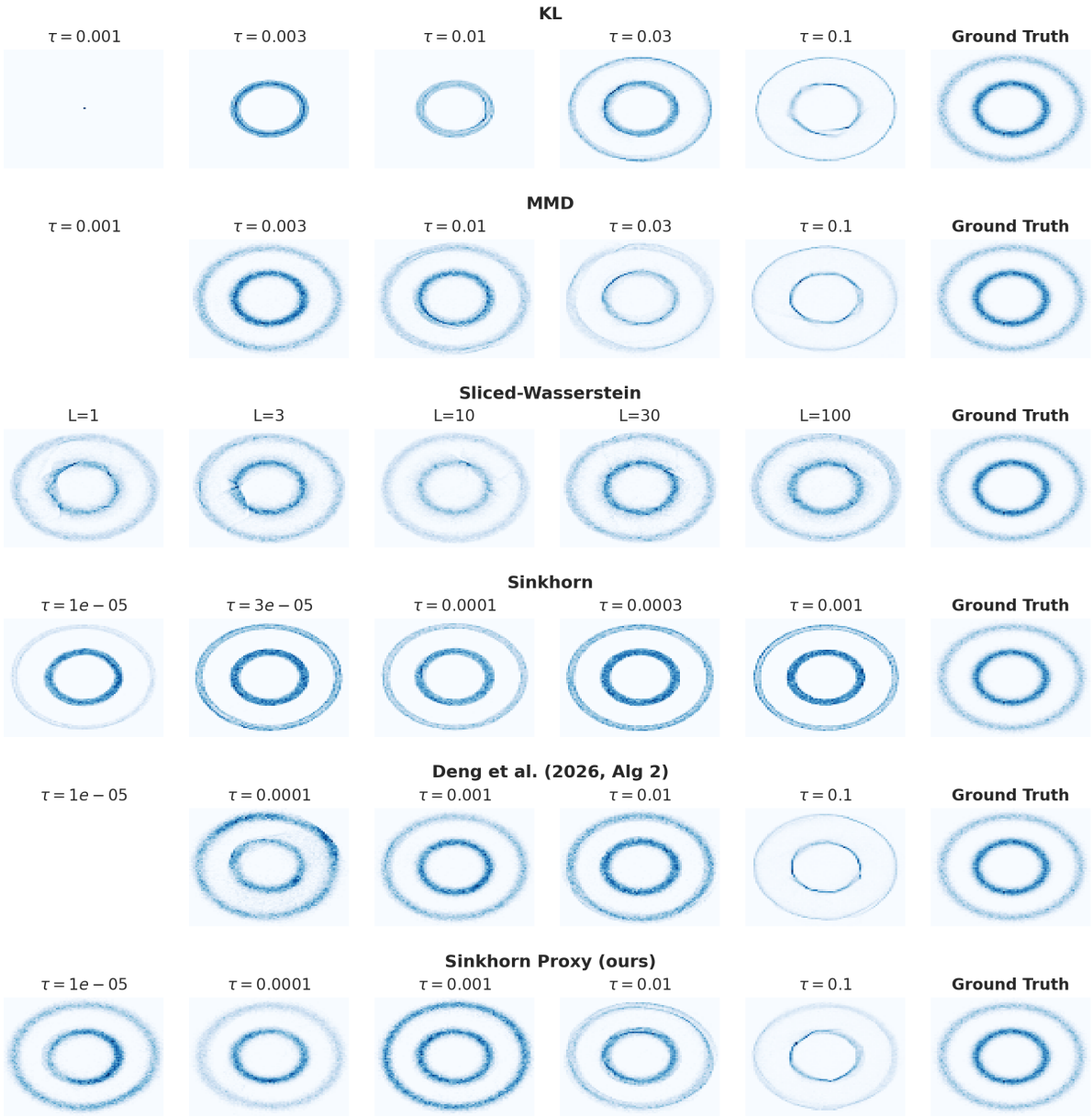


Figure 5: Results for the Circles dataset. Empty panel means the samples have diverged.

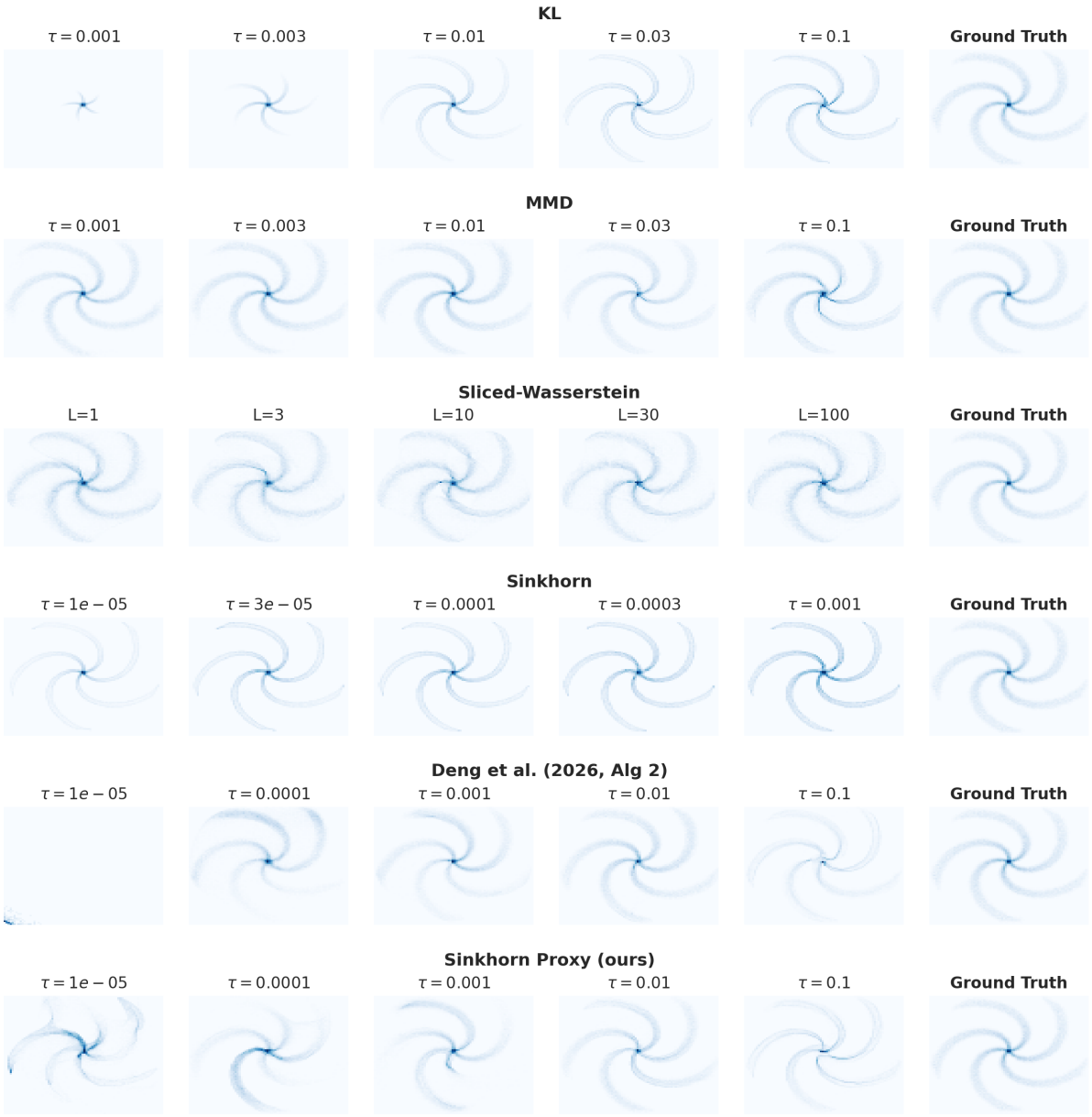


Figure 6: Results for the Pinwheel dataset. Empty panel means the samples have diverged.

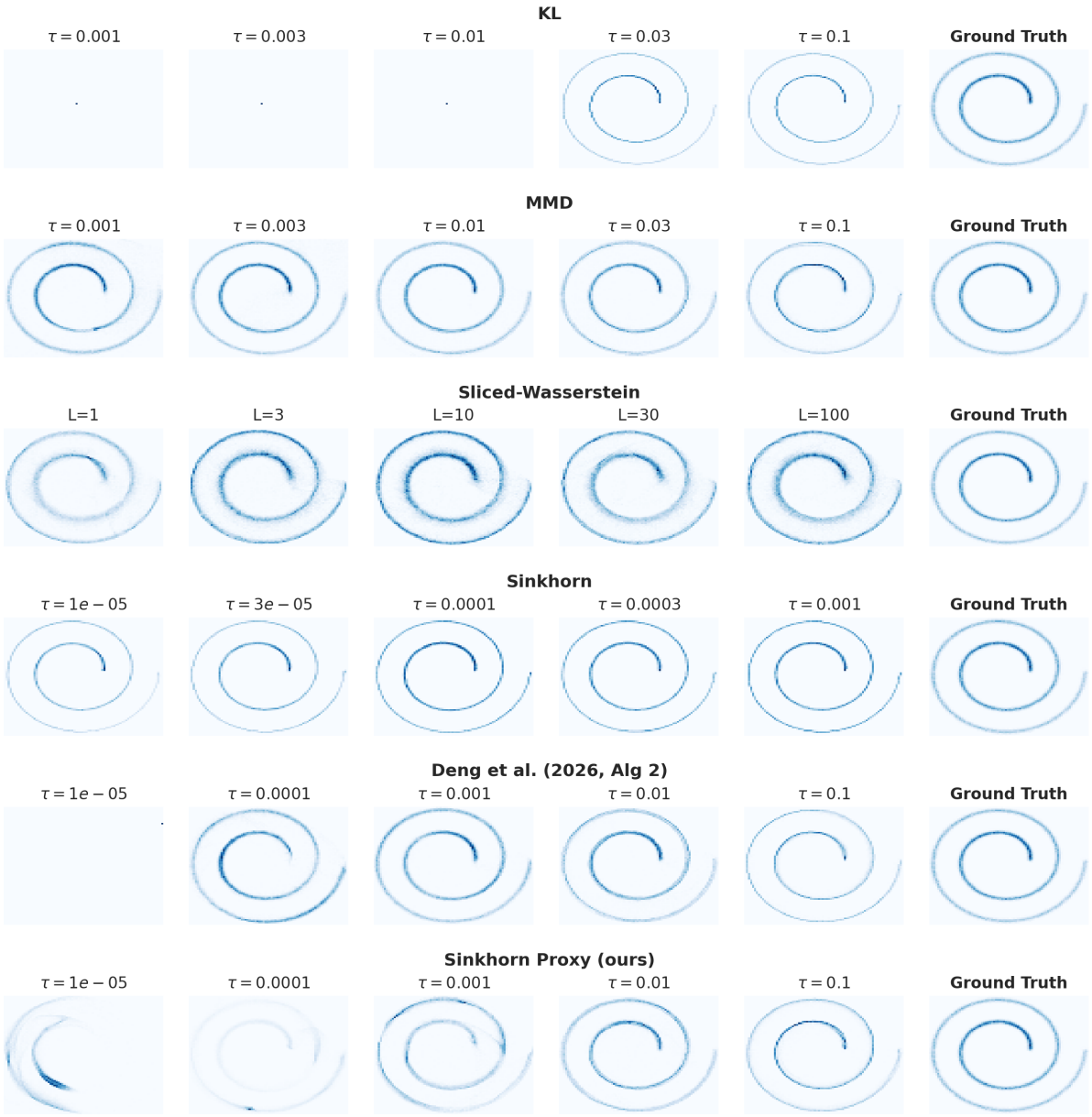


Figure 7: Results for the Swiss roll dataset. Empty panel means the samples have diverged.