Workshop at the 7th Symposium on Advances in Approximate Bayesian Inference (non-archival), 2025 1–25

Inclusive KL Minimization: A Wasserstein-Fisher-Rao Gradient Flow Perspective

Jia-Jie Zhu
Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
zhu@wias-berlin.de

Abstract

Otto's Wasserstein gradient flow of the exclusive KL divergence functional provides a powerful and mathematically principled perspective for analyzing machine learning and Bayesian inference algorithms. In contrast, algorithms for the inclusive KL inference, i.e., minimizing $\mathrm{KL}(\pi|\mu)$ w.r.t. μ for some target π, are rarely analyzed using tools from mathematical analysis. This paper shows that a general-purpose approximate inclusive KL inference paradigm can be constructed using the theory of gradient flows derived from PDE analysis. We uncover some precise relationships between the inclusive KL inference and some widely used learning algorithms, including the MMD-minimization and the Wasserstein flow of kernel discrepancies, which are widely used in machine learning applications. For example, a few existing sampling algorithms, such as those based on the Wasserstein flow of kernel discrepancies, can be viewed in a unified manner as inclusive-KL inference with approximate gradient estimators. Finally, we provide the theoretical foundation for the Fisher-Rao type gradient flows for minimizing the inclusive KL divergence.

1. Introduction

Many learning and inference problems can be cast into the framework of minimizing the KL divergence

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\mu|\pi). \tag{1}$$

The functional $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$ is also known as the exclusive KL divergence between μ and π, due to its well-known property commonly referred to as mode-seeking and zero-avoiding. This variational problem forms the foundation of modern Bayesian inference (Zellner, 1988). For example, suppose we have a model $p(\mathrm{Data}|\theta)$ and a prior $p(\theta)$, our goal is to infer the posterior $\pi(\theta) := p(\theta|\mathrm{Data})$. If we further restrict the feasible set A in equation 1 to be the so-called variational family, e.g., the set of all Gaussian distributions, we obtain variational inference (Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2017). Albeit much less popular, there also exists the inference paradigm that minimizes the inclusive KL,

$$\min_{\mu \in A \subset \mathcal{P}} \mathrm{D}_{\mathrm{KL}}(\pi|\mu). \tag{2}$$

For example, algorithms such as expectation propagation (Minka, 2013), (Bishop, 2006, Section 10.7) can be viewed as solving equation 2. Many researchers such as Naesseth et al. (2020); Jerfel et al. (2021); McNamara et al. (2024); Zhang et al. (2022) have argued that the solution of equation 2, if available, offers statistical advantages over equation 1,

© J.-J. Zhu.

e.g., mass-covering with respect to the true posterior, desirable properties for applications requiring conservative uncertainty quantification, avoid light tails that can cause instability. We also refer to the discussion in (Dhaka et al., 2021) about the mode-covering behavior of inclusive KL in moderate-to-high dimensions for variational inference algorithms.

However, many of existing algorithms require adhoc procedures to gain samples from target distributions π or do not have sound mathematical analysis as backbone; see our discussion around the Wasserstein flow equation equation iKL-WGF.

In comparison, there has been significant technical developments for the exclusive KL minimization equation 1 recently. This is mainly due to the injection of rigorous theoretical foundation from analysis of (PDE) gradient flows (Otto, 1996, 2001; Ambrosio et al., 2005; Peletier, 2014; Mielke, 2023) and statistical optimal transport (Chewi et al., 2024; Peyré and Cuturi, 2019; Panaretos and Zemel, 2019). Inference and sampling algorithms based on equation 1 can now be studied under a unified framework and on the rigor level of applied analysis. Can we use such principled theory to study inclusive KL minimization? This paper answers this question affirmatively. Concretely, we list our main *technical contributions* as follows:

1. A major contribution of this paper, is to reveal a fundamental connection between the inclusive KL minimization equation 2 and some widely used paradigms in sampling, inference, and generative models – the MMD minimization problems. While the known connections between those problems are either elementary (e.g., moment-matching) or heuristic, we show that the latter is an approximation (via convolution or mollification) of the former when cast into the rigorous framework of PDE gradient-flow systems.
2. Going beyond Wasserstein geometry, we show several new results regarding the Fisher-Rao (FR) gradient flows, especially the discovery of FR flow of the inclusive KL can be implemented as the MMD-MMD flow. This finding gives both theoretical and practical implications to the learning algorithms.
3. We identify the setting (and the flows) along which MMD globally decays without imposing conditions such as log-Sobolev inequality with a positive constant, and characterized the solution explicitly. We also give an interpretation of the MMD-barycenter problem in the information geometry using the variational characterization.
4. Last but not least, this is the first paper that provides a gradient flow theory foundation for inclusive KL inference. This adds a principled component and future lane of research to the fundamental theory of Bayesian statistics and generative modeling via inclusive KL inference, which has been missing so far.

2. Preliminary: Bayesian inference as Wasserstein gradient flow of KL

An elegant perspective of Bayesian inference is offered by the Wasserstein gradient flow (WGF) framework of Otto (1996), which has attracted much attention from researchers in Bayesian inference; see (Chewi et al., 2024; Trillos and Sanz-Alonso, 2018) for recent surveys. In that framework, one can write a flow equation formally as

$$\dot{\mu} = -\nabla_W F(\mu) = -\mathbb{K}_W(\mu)\frac{\delta F}{\delta \mu}[\mu] = \mathrm{div}(\mu \nabla \frac{\delta F}{\delta \mu}[\mu]). \tag{3}$$

through the Wasserstein Onsager operator $\mathbb{K}_W$, which is defined as the inverse of the Riemannian metric tensor $\mathbb{G}_W$ of the Wasserstein space, i.e., $\mathbb{K}_W(\rho) = \mathbb{G}_W(\rho)^{-1}$. Mathematically, for

the Wasserstein space, $\mathbb{K}_W(\rho) : T^*_\rho\mathcal{M} \to T_\rho\mathcal{M}, \xi \mapsto -\operatorname{div}(\rho\nabla\xi)$, where $T_\rho\mathcal{M}$ is the tangent space of $\mathcal{M}^+$ at ρ and $T^*_\rho\mathcal{M}$ the cotangent space. The terminology Onsager's operator is due to the works of Onsager and Machlup (1953); Onsager (1931). From the mechanics perspective, the dual functions ξ can be interpreted as the generalized thermodynamic forces (Onsager and Machlup, 1953; Mielke et al., 2017). With those ingredients, we can formally define the gradient systems that generate gradient flow equations such as equation 3.

Definition 1 (Gradient system (Otto, 2001; Mielke, 2023)) *We refer to a tuple* $(\mathcal{M}, F, \mathbb{K})$ *as a gradient system. It has the gradient structure identified by:*

1. *a space* $\mathcal{M}$*,*
2. *an energy functional* F*,*
3. *a dissipation geometry given by either: a distance metric defined on* $\mathcal{M}$*, a Riemannian metric tensor* $\mathbb{G}$*, or a symmetric positive-definite Onsager operator* $\mathbb{K} = \mathbb{G}^{-1}$*.*

Note that it is also possible to define dissipation geometry via nonlinear dissipation potential functional; cf. (Mielke et al., 2017).

Regarding Bayesian inference, we choose the energy functional as the exclusive KL divergence as in equation 1, i.e., $F(\mu) = \mathrm{D}_{\mathrm{KL}}(\mu|\pi)$. Through elementary calculation, we obtain from equation 3 the Fokker-Planck equation (FPE)

$$\partial_t\mu = \operatorname{div}\left(\mu\nabla\log\frac{\mu}{\pi}\right) = \Delta\mu - \operatorname{div}\left(\mu\nabla\log\pi\right). \tag{FPE}$$

When we express the target as $\pi(x) = \frac{1}{Z}\exp(-V(x))$ where Z is a normalization constant (partition function), the equation FPE is then $\partial_t\mu = \Delta\mu + \operatorname{div}(\mu\nabla V)$. The fact that the evolution equation does not depend on the partition function Z is often argued to be one of the key advantages of the KL divergence. Viewed as a dynamic system, the KL divergence energy functional dissipates along equation FPE in the steepest descent manner. Based on the formal definition of gradient system equation 1, we say that equation FPE has the *gradient structure* that entails the following key ingredients.

$$\begin{cases} \text{Space :} & \text{prob. space } \mathcal{P} \\ \text{Energy functional :} & F(\cdot) := \mathrm{D}_{\mathrm{KL}}(\cdot|\pi) \\ \text{Dissipation Geometry :} & \text{Wasserstein } \mathbb{K}_W \end{cases} \tag{4}$$

3. Wasserstein gradient flows of inclusive KL and its approximation

Our starting point is the following Wasserstein gradient flow equation of the inclusive KL inference, derived using Otto (2001)'s formal calculation analogous to the exclusive KL case equation 3.

$$\dot{\mu} = \operatorname{div}\left(\mu\nabla\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right). \tag{iKL-WGF}$$

The relation with the exclusive KL gradient flow can be observed by comparing this PDE with equation FPE. The two generalized force functionals agree to the first order near the

equilibrium, i.e., $\log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \approx 1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}$ when $\mathrm{d}\pi/\mathrm{d}\mu \approx 1$. Furthermore, the generator function of the inclusive KL has a larger slope than that of the exclusive KL. This subtle difference will lead to different behaviors of their gradient flows. Rewriting the right-hand side of equation iKL-WGF, we obtain the PDE $\dot{\mu} = -\Delta\pi + \mathrm{div}(\pi\nabla \log \mu)$, which bears similarity to equation FPE but with the position of π and μ exchanged on the right-hand side. Intuitively, the gradient structure of equation iKL-WGF is given by:

$$\begin{cases} \text{Space :} & \mathcal{P} \\ \text{Energy functional :} & F(\cdot) := \mathrm{D}_{\mathrm{KL}}(\pi|\cdot) \\ \text{Dissipation Geometry :} & \text{Wasserstein } \mathbb{K}_W \end{cases} \tag{5}$$

While the gradient structure is clear, a main obstacle to implement the gradient flow is due to that the function $1 - \mathrm{d}\pi/\mathrm{d}\mu$, which may not be accessible or differentiable. To address this, we consider a flow equation with a smooth approximation via the integral operator $\mathcal{T}_{k,\mu}$ defined in equation 10. The resulting kernelized flow equation is given by

$$\dot{\mu} = \mathrm{div}\left(\mu\nabla\mathcal{T}_{k,\mu}\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right). \tag{iKL-WGF-k}$$

Note that, the flow with a kernelized force is not necessarily a gradient flow. However, in this case, equation iKL-WGF-k is indeed a gradient flow that has been applied to machine learning applications. We observe the following.

Theorem 2 (Flow equation equation iKL-WGF-k has a Wasserstein gradient structure) *Suppose that initial condition satisfies $\pi \ll \mu$, i.e., π is absolutely continuous with respect to μ. Then, equation iKL-WGF-k coincides with the Wasserstein gradient flow equation of the MMD equation MMD-WGF,*

$$\dot{\mu} = \mathrm{div}\left(\mu \int \nabla_2 k(x,\cdot)\ \mathrm{d}\left(\mu - \pi\right)(x)\right) \tag{MMD-WGF}$$

where ∇_2 denotes the differentiation with respect to the second variable. Intuitively, equation MMD-WGF and hence equation iKL-WGF-k have the same gradient structure:

$$\begin{cases} \text{Space :} & \mathcal{P} \\ \text{Energy functional :} & F(\cdot) := \frac{1}{2}\,\mathrm{MMD}^2(\cdot,\pi) \\ \text{Dissipation Geometry :} & \text{Wasserstein } \mathbb{K}_W \end{cases} \tag{6}$$

As discussed above, the vanilla equation iKL-WGF cannot be directly used to derive algorithms due to the non-smooth nature of the function $1 - \mathrm{d}\pi/\mathrm{d}\mu$. Now, since equation iKL-WGF-k's flow equation coincides with equation MMD-WGF, our theory suggests that minimizing MMD through equation MMD-WGF is equivalent to simulating a kernelized Wasserstein gradient descent to minimize the inclusive KL. Then, we can make use of numerous implementations that have already been developed for MMD-minimization, see, e.g., (Arbel et al., 2019; Chizat, 2022; Futami et al., 2019; Hagemann et al., 2023; Neumayer et al., 2024; Galashov et al., 2024; Gladin et al., 2024; Chen et al., 2024). Thus,

the dual-force-kernelized gradient flow equation iKL-WGF-k provides an implementable approximation of equation iKL-WGF.

Summarizing the results so far, we offer insights into both score-based (e.g., requiring evaluation of the score function $\nabla \log \pi$) and sample-based (e.g., assuming access to samples from π) (Bayesian) inference and sampling, providing a unifying Wasserstein gradient flow perspective on these methods in Bayesian computation based on inclusive KL inference equation 2. Our insights provide a first-principles interpretation of these methods via gradient flows. We also note that a wider class of gradient flows can be characterized using the kernel Stein discrepancy; see the appendix.

4. Fisher-Rao gradient flows of the inclusive KL functional

While recent machine learning applications primarily focus on the Wasserstein geometry, we emphasize that the gradient flow theory is more general. Prominent examples include the Fisher-Rao and Hellinger geometries (Hellinger, 1909; Kakutani, 1948; Rao, 1945; Bhattacharyya, 1946), which provide a different yet extremely impactful perspective on statistical inference and optimization. They form an important building block for the Wasserstein-Fisher-Rao gradient flow in the next section.

4.1. Fisher-Rao a.k.a. spherical Hellinger gradient flows of inclusive KL functional

In this subsection, we first analyze and uncover a few remarkable properties of the Fisher-Rao gradient flows of the inclusive KL divergence. Then, we establish a precise connection to the MMD gradient flow of the MMD functional. This connection was not previously known, yet machine learning algorithms have already provided empirical implications of such Fisher-Rao gradient flow.

Our starting point is to replace the Wasserstein dissipation geometry in the gradient structure equation 5 with the Fisher-Rao dissipation geometry, defined using the Fisher-Rao Onsager operator, $\mathbb{K}_{\mathsf{FR}}(\rho) : T^*_\rho \mathcal{M} \to T_\rho \mathcal{M}, \xi \mapsto -\rho\left(\xi - \int \xi \, \mathrm{d}\rho\right)$. The resulting gradient structure is

$$\begin{cases} \text{Space :} & \text{prob. measures } \mathcal{P} \\ \text{Energy functional :} & \mathrm{D}_{\mathrm{KL}}(\pi|\cdot) \\ \text{Dissipation Geometry :} & \text{Fisher-Rao } \mathbb{K}_{\mathsf{FR}} \end{cases} \tag{7}$$

Note that the Fisher-Rao space is also referred to as the spherical Hellinger space by Laschos and Mielke (2019) considering the historical development. Interestingly, under the inclusive KL divergence functional, the Hellinger gradient flow over $\mathcal{M}^+$ stays within the probability space $\mathcal{P}$ if initialized therein, i.e., the spherical projection of Laschos and Mielke (2019) is not needed in our case; see the appendix for more details. Previously, flows in the Fisher-Rao space have been studied in ML applications under the name of birth-death dynamics; see (Lu et al., 2019; Rotskoff et al., 2019; Kim and Suzuki, 2024) for applications and further discussions.

We summarize some important properties of the Fisher-Rao gradient systems in the following proposition.

Proposition 3 (FR gradient flow of inclusive KL) *The gradient structure equation 7 generates the flow equation*

$$\dot{\mu} = \pi - \mu. \tag{RevKL-FR-GF}$$

Its closed-form solution is given by

$$\mu_t = e^{-t}\mu_0 + (1 - e^{-t})\pi. \tag{8}$$

This result characterizes an interesting feature of the *inclusive-KL-Fisher-Rao flow: it traverses along a straight line despite the Riemannian structure of the Fisher-Rao geometry.*

A distinctive feature of the inclusive-KL-FR gradient flow is the following.

Theorem 4 (Exponential Decay of inclusive-KL divergence) *There exists a constant $c > 0$ such that the following Polyak-Łojasiewicz functional inequality holds globally.*

$$\left\| 1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu} \right\|_{L^2_\mu}^2 \geq c \cdot \mathrm{D}_{\mathrm{KL}}(\pi|\mu), \quad \forall \mu \in \mathcal{M}^+. \tag{Ł-RKL}$$

Furthermore, the inclusive KL satisfies the exponential decay estimate along the gradient flow

$$\mathrm{D}_{\mathrm{KL}}(\pi|\mu(t)) \leq e^{-t}\mathrm{D}_{\mathrm{KL}}(\pi|\mu_0) \text{ for all } t > 0.$$

We emphasize that Theorem 4 is global and does not require the assumption of a uniform bound on the density ratio $\mathrm{d}\mu_0/\mathrm{d}\pi$ such as in (Lu et al., 2023). This result indicates a remarkable feature of the inclusive KL divergence: its Fisher-Rao gradient flow is capable of creating mass from zero-mass regions of π. In machine learning and statistics, this is a highly desired feature as we often need to locate the support of the target measure π. An intuition of the distinction between the exclusive KL and the inclusive KL (Theorem 4) is indeed given by difference of their entropy generator slopes near the zero-mass region.

There is an interesting *coincidence* of the gradient systems equation 7 and existing machine learning algorithms. Let us consider a seemingly unrelated gradient system where both the energy functional F and the dissipation geometry to be MMD, i.e.,

$$\begin{cases} \text{Space :} & \mathcal{P} \\ \text{Energy functional :} & F(\cdot) := \mathrm{MMD}^2(\cdot, \pi) \\ \text{Dissipation Geometry :} & \mathrm{MMD} \end{cases} \tag{9}$$

Proposition 5 *The MMD-MMD gradient flow equation, generated by the gradient system equation 9, coincides with the inclusive-KL-Fisher-Rao gradient flow equation equation RevKL-FR-GF. Consequently, MMD decays exponentially along the solution μ_t of equation RevKL-FR-GF, i.e.,* $\mathrm{MMD}(\mu_t, \nu) \leq e^{-t} \cdot \mathrm{MMD}(\mu_0, \nu)$.

From a dynamical system perspective, this result also shows that the MMD is a *Lyapunov functional* for the inclusive-KL Fisher-Rao flow. This proposition also shows that the same flow equation equation RevKL-FR-GF has two different gradient structures: MMD and Fisher-Rao.

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savare. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Abdul Fatir Ansari, Ming Liang Ang, and Harold Soh. Refining deep generative models via discriminator gradient flow. (arXiv:2012.00780), June 2021. doi: 10.48550/arXiv.2012.00780.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. December 2019. doi: 10.48550/arxiv.1906.04370.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. (arXiv:2206.08873), June 2022. doi: 10.48550/arxiv.2206.08873.

Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum Stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406, 1946. ISSN 0036-4452.

Miko\laj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018. doi: 10.48550/arxiv.1801.01401.

Christopher Bishop. Pattern recognition and machine learning. *Springer google schola*, 2: 5–43, 2006.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1285773.

José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58:1–53, 2019.

Wilson Ye Chen, Lester Mackey, Jackson Gorham, Francois-Xavier Briol, and Chris Oates. Stein points. In *Proceedings of the 35th International Conference on Machine Learning*, pages 844–853. PMLR, July 2018.

Zonghao Chen, Aratrika Mustafi, Pierre Glaser, Anna Korba, Arthur Gretton, and Bharath K. Sriperumbudur. (de)-regularized maximum mean discrepancy gradient flow. (arXiv:2409.14980), September 2024. doi: 10.48550/arxiv.2409.14980.

Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, and Philippe Rigollet. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. (arXiv:2006.02509), June 2020. doi: 10.48550/arxiv.2006.02509.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport, July 2024.

Lénaïc Chizat. Mean-field Langevin dynamics: Exponential convergence and annealing. (arXiv:2202.01009), August 2022. doi: 10.48550/arxiv.2202.01009.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, February 2018. ISSN 1615-3375, 1615-3383. doi: 10.1007/s10208-016-9331-y.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulation. February 2019. doi: 10.48550/arxiv.1508.05216.

Nicolas Chopin, Francesca R. Crucinio, and Anna Korba. A connection between tempering and entropic mirror descent. (arXiv:2310.11914), March 2024. doi: 10.48550/arXiv.2310.11914.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2606–2615. PMLR, June 2016.

Katy Craig, Karthik Elamvazhuthi, Matt Haberland, and Olga Turanova. A blob method for inhomogeneous diffusion with applications to multi-agent control and sampling. *Mathematics of Computation*, 92(344):2575–2654, November 2023. ISSN 0025-5718, 1088-6842. doi: 10.1090/mcom/3841.

Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

Bo Dai, Niao He, Hanjun Dai, and Le Song. Provable Bayesian inference via particle mirror descent. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 985–994. PMLR, May 2016.

Akash Kumar Dhaka, Alejandro Catalina, Manushi Welandawe, Michael Riis Andersen, Jonathan Huggins, and Aki Vehtari. Challenges and Opportunities in High-dimensional Variational Inference. (arXiv:2103.01085), June 2021. doi: 10.48550/arXiv.2103.01085.

Carles Domingo-Enrich, Samy Jelassi, Arthur Mensch, Grant Rotskoff, and Joan Bruna. A mean-field analysis of two-player zero-sum games. In *Advances in Neural Information Processing Systems*, volume 33, pages 20215–20226. Curran Associates, Inc., 2020.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

Pavel Dvurechensky and Jia-Jie Zhu. Analysis of kernel mirror prox for measure optimization. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2350–2358. PMLR, 2024-05-02/2024-05-04.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015. doi: 10.48550/arxiv.1505.03906.

Jean-Yves Franceschi, Mike Gartrell, Ludovic Dos Santos, Thibaut Issenhuth, Emmanuel de Bézenac, Mickaël Chen, and Alain Rakotomamonjy. Unifying gans and score-based diffusion as generative particle models. *Advances in Neural Information Processing Systems*, 36, 2024.

Futoshi Futami, Zhenghang Cui, Issei Sato, and Masashi Sugiyama. Bayesian posterior approximation via greedy particle optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3606–3613, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33013606.

Alexandre Galashov, Valentin de Bortoli, and Arthur Gretton. Deep MMD gradient flow without adversarial training. (arXiv:2405.06780), May 2024. doi: 10.48550/arxiv.2405.06780.

Thomas O Gallouët and Leonard Monsaingeon. A JKO splitting scheme for kantorovich–fisher–rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

Egor Gladin, Pavel Dvurechensky, Alexander Mielke, and Jia-Jie Zhu. Interaction-force transport gradient flows. *arXiv preprint arXiv:2405.17075*, 2024.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

Paul Hagemann, Johannes Hertrich, Fabian Altekrüger, Robert Beinert, Jannis Chemseddine, and Gabriele Steidl. Posterior sampling based on gradient flows of the mmd with negative distance kernel. *arXiv preprint arXiv:2310.03054*, 2023.

Ye He, Krishnakumar Balasubramanian, Bharath K Sriperumbudur, and Jianfeng Lu. Regularized Stein variational gradient flow. *arXiv preprint arXiv:2211.07861*, 2022.

Matthias Hein and Olivier Bousquet. Kernels, associated structures and generalizations. August 2004.

E. Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, July 1909. ISSN 1435-5345, 0075-4102. doi: 10.1515/crll.1909.136.210.

Alvin Heng, Abdul Fatir Ansari, and Harold Soh. Generative modeling with flow-guided density ratio learning. (arXiv:2303.03714), June 2024. doi: 10.48550/arxiv.2303.03714.

Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed Nash equilibria of generative adversarial networks, October 2018.

Ghassen Jerfel, Serena Wang, Clara Wong-Fannjiang, Katherine A. Heller, Yian Ma, and Michael I. Jordan. Variational refinement for importance sampling using the forward Kullback-Leibler divergence. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 1819–1829. PMLR, December 2021.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, November 1999. ISSN 1573-0565. doi: 10.1023/A:1007665907178.

Shizuo Kakutani. On equivalence of infinite product measures. *Annals of Mathematics*, 49 (1):214–224, 1948.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. *arXiv preprint arXiv:2402.01258*, 2024.

Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. (arXiv:1505.07746), March 2016. doi: 10.48550/arxiv.1505.07746.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel Stein discrepancy descent. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5719–5730. PMLR, July 2021.

Vaios Laschos and Alexander Mielke. Geometric properties of cones with applications on the Hellinger–Kantorovich space, and a new distance on the space of probability measures. 276(11):3529–3576, 2019. doi: 10.1016/j.jfa.2018.12.013.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727. PMLR, 2015.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, March 2018. ISSN 0020-9910, 1432-1297. doi: 10.1007/s00222-017-0759-8.

Qiang Liu, Jason D. Lee, and Michael I. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. (arXiv:1602.03253), July 2016. doi: 10.48550/arXiv.1602.03253.

Yulong Lu, Jianfeng Lu, and J. Nolen. Accelerating Langevin sampling with birth-death. *ArXiv*, May 2019.

Yulong Lu, Dejan Slepčev, and Lihan Wang. Birth–death dynamics for sampling: Global convergence, approximations and their asymptotics. *Nonlinearity*, 36(11):5731, 2023.

Aimee Maurais and Youssef Marzouk. Sampling in unit time with kernel Fisher-Rao flow. (arXiv:2401.03892), February 2024. doi: 10.48550/arxiv.2401.03892.

Declan McNamara, Jackson Loper, and Jeffrey Regier. Sequential monte carlo for inclusive KL minimization in amortized variational inference. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 4312–4320. PMLR, April 2024.

Alexander Mielke. An introduction to the analysis of gradients systems. *arXiv preprint arXiv:2306.05026*, 2023.

Alexander Mielke and Jia-Jie Zhu. Hellinger-kantorovich gradient flows: Global exponential decay of entropy functionals. *arXiv preprint arXiv:2501.17049*, 2025.

Alexander Mielke, Robert I. A. Patterson, Mark A. Peletier, and D. R. Michiel Renger. Non-equilibrium thermodynamical principles for chemical reactions with mass-action kinetics. *SIAM Journal on Applied Mathematics*, 77(4):1562–1585, January 2017. ISSN 0036-1399, 1095-712X. doi: 10.1137/16M1102240.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. (arXiv:1301.2294), January 2013. doi: 10.48550/arxiv.1301.2294.

Youssef Mroueh and Mattia Rigotti. Unbalanced Sobolev descent, September 2020.

Christian Naesseth, Fredrik Lindsten, and David Blei. Markovian score climbing: variational inference with KL(p||q). In *Advances in Neural Information Processing Systems*, volume 33, pages 15499–15510. Curran Associates, Inc., 2020.

Sebastian Neumayer, Viktor Stein, and Gabriele Steidl. Wasserstein gradient flows for moreau envelopes of f-divergences in reproducing kernel hilbert spaces. *arXiv preprint arXiv:2402.04613*, 2024.

Nikolas Nüsken. Stein transport for Bayesian inference, September 2024.

L. Onsager and S. Machlup. Fluctuations and irreversible processes. *Physical Review*, 91(6): 1505–1512, September 1953. doi: 10.1103/PhysRev.91.1505.

Lars Onsager. Reciprocal relations in irreversible processes. i. *Physical Review*, 37(4):405–426, February 1931. doi: 10.1103/PhysRev.37.405.

Felix Otto. Double degenerate diffusion equations as steepest descent. 1996.

Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001. doi: 10.1081/PDE-100002243.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, March 2019. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-030718-104938.

Mark A. Peletier. Variational modelling: Energies, gradient flows, and large deviations. February 2014. doi: 10.48550/arxiv.1402.1990.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073.

C Radhakrishna Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945.

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. In *International Conference on Machine Learning*, pages 5508–5517. PMLR, 2019.

Vladimir Spokoiny. Nonparametric estimation: Parametric view. 2016.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35(3):363–417, June 2012. ISSN 1432-0940. doi: 10.1007/s00365-012-9153-3.

James Stewart. Positive definite functions and generalizations, an historical survey. *The Rocky Mountain Journal of Mathematics*, 6(3):409–434, 1976. ISSN 0035-7596.

Nicolas Garcia Trillos and Daniel Sanz-Alonso. The Bayesian update: Variational formulations and gradient flows. (arXiv:1705.07382), November 2018. doi: 10.48550/arxiv.1705.07382.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 2009. ISBN 978-0-387-79051-0 978-0-387-79052-7. doi: 10.1007/b13794.

Francisco Vargas, Shreyas Padhy, Denis Blessing, and Nikolas Nusken. Transport meets variational inference: Controlled monte carlo diffusions. In *The Twelfth International Conference on Learning Representations*. May 2024.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Guillaume Wang and Lénaïc Chizat. An exponentially converging particle method for the mixed nash equilibrium of continuous games, November 2022.

Guillaume Wang and Lénaïc Chizat. Local convergence of gradient methods for min-max games: Partial curvature generically suffices, November 2023.

Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, December 2004. ISBN 978-1-139-45665-4.

Yuling Yan, Kaizheng Wang, and Philippe Rigollet. Learning gaussian mixtures using the Wasserstein-Fisher-Rao gradient flow. (arXiv:2301.01766), January 2023. doi: 10.48550/arXiv.2301.01766.

Mingxuan Yi and Song Liu. Bridging the gap between variational inference and Wasserstein gradient flows. 2023. doi: 10.48550/ARXIV.2310.20090.

Mingxuan Yi, Zhanxing Zhu, and Song Liu. Monoflow: Rethinking divergence gans via the perspective of Wasserstein gradient flows. In *Proceedings of the 40th International Conference on Machine Learning*, pages 39984–40000. PMLR, July 2023.

Arnold Zellner. Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

Liyi Zhang, David M. Blei, and Christian A. Naesseth. Transport score climbing: Variational inference using forward KL and adaptive neural transport. (arXiv:2202.01841), September 2022. doi: 10.48550/arXiv.2202.01841.

Jia-Jie Zhu and Alexander Mielke. Kernel approximation of Fisher-Rao gradient flows. *arXiv preprint arXiv:2410.20622*, 2024.

Appendix A. Integral operator and maximum-mean discrepancy

Given a positive measure ρ on $\mathbb{R}^d$ and a positive-definite kernel k, the integral operator $\mathcal{T}_{k,\rho} : L^2_\rho \to \mathcal{H}$ is defined by

$$\mathcal{T}_{k,\rho} g(x) := \int k\left(x, x'\right) g\left(x'\right) d\rho\left(x'\right) \quad \text{for} \quad g \in L^2_\rho, \tag{10}$$

where $\mathcal{H}$ is the reproducing kernel Hilbert space associated with the kernel k. With a slight abuse of terminology, the following compositional operator $\mathcal{K}_\rho := \mathrm{Id} \circ \mathcal{T}_{k,\rho}$ is also referred to as the integral operator, albeit defined for $L^2(\rho) \to L^2(\rho)$. $\mathcal{K}_\rho$ is compact, (semi-)positive, self-adjoint, and nuclear; cf. (Steinwart and Christmann, 2008; Hein and Bousquet, 2004; Steinwart and Scovel, 2012). The adjoint of $\mathcal{T}_{k,\rho}$ is the embedding operator $\mathrm{Id} : \mathcal{H} \to L^2(\rho)$, i.e., $\langle \mathrm{Id}\, f, g \rangle_{L^2(\rho)} = \langle f, \mathcal{T}_{k,\rho} g \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $g \in L^2(\rho)$. When using a kernel such as the Gaussian kernel, the image $\mathcal{T}_{k,\rho} g$ can be regarded as a smooth approximation of g, which is sometimes referred to as approximation by convolution or mollification (Wendland, 2004). An assumption we will generally make throughout the paper is that the kernel k is bounded, symmetric, and satisfies the *integrally strict positive-definite* (ISPD) condition (Sriperumbudur et al., 2010; Steinwart and Christmann, 2008; Stewart, 1976): $\int k(x, x')\, \mathrm{d}\rho(x)\, \mathrm{d}\rho(x') > 0$ for any non-zero signed measure ρ. The purpose of this condition is to ensure that the integral operator $\mathcal{T}_{k,\rho}$ is strictly positive-definite and also for technical reasons in terms of PDE analysis. In this paper, the kernel maximum mean discrepancy (MMD) between two positive measures μ and ν is defined as

$$\mathrm{MMD}^2(\mu, \nu) := \int \int k(x, x')\, \mathrm{d}(\mu - \nu)(x)\, \mathrm{d}(\mu - \nu)(x')$$

The MMD and integral operator of a kernel are the central tools for understanding the approximate gradient flows of the inclusive KL for practical applications.

Appendix B. Unbalanced transport: Wasserstein-Fisher-Rao gradient flows

The Wasserstein geometry endows us with the mechanism to transport mass. On the other hand, the Fisher-Rao geometry lets us create and destroy mass. One major development in optimal transport theory is the combination of both via unbalanced transport, invented independently by Chizat et al. (2018, 2019); Liero et al. (2018); Kondratyev et al. (2016). The resulting metric between two non-negative measures is known as the Wasserstein-Fisher-Rao (WFR) distance, also known as the Hellinger-Kantorovich distance, defined via the entropic transport problem (Liero et al., 2018)

$$\mathsf{WFR}^2(\mu_1, \mu_2) = \min_{\Pi \in \Gamma(\mu_1, \mu_2)} \left\{ \alpha \int c \, \mathrm{d}\Pi + \beta \mathrm{D}_{\mathrm{KL}}(\pi_1 | \mu_1) + \beta \mathrm{D}_{\mathrm{KL}}(\pi_2 | \mu_2) \right\}$$

where α and β are two scaling parameters. $\Gamma(\mu_1, \mu_2)$ is the set of all positive measures with marginals μ_1 and μ_2. c is the transport cost in the standard Wasserstein distance and D_φ is the φ-divergence (defined in equation 13). In this paper, we define the WFR gradient structure via the Onsager operator: the WFR Riemannian metric tensor is an inf-convolution of the Wasserstein tensor and the Fisher-Rao tensor $\mathbb{G}_{\mathsf{WFR}}(\mu) = \mathbb{G}_W(\mu) \square \mathbb{G}_{\mathsf{FR}}(\mu)$ (Chizat et al., 2019; Liero et al., 2018; Gallouët and Monsaingeon, 2017). By the Legendre transform, its inverse, the Onsager operator, is given by the sum $\mathbb{K}_{\mathsf{WFR}}(\mu) = \mathbb{K}_W(\mu) + \mathbb{K}_{\mathsf{FR}}(\mu)$. For conciseness, we only focus on the case of WFR distance restricted to the space of probability measures by default. Therefore, the WFR distance should technically be referred to as the spherical Hellinger-Kantorovich distance. Let us now consider the following gradient structure in the WFR space

$$\begin{cases} \text{Space :} & \mathcal{P} \\ \text{Energy functional :} & \text{inclusive KL: } \mathrm{D}_{\mathrm{KL}}(\pi | \cdot) \\ \text{Dissipation Geometry :} & \text{(spherical) WFR } \mathbb{K}_{\mathsf{WFR}} \end{cases} \tag{11}$$

Using equation iKL-WFR-GF and the results in the previous two sections, the WFR gradient flow equation generated by equation 11 is given by the reaction-diffusion-type PDE

$$\dot{\mu} = \underbrace{\alpha \operatorname{div}\left(\mu \nabla \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)}_{\text{Wasserstein: transport}} - \underbrace{\beta \mu \cdot \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)}_{\text{Fisher-Rao: birth-death}} . \tag{iKL-WFR-GF}$$

The derivation is standard; cf. the aforementioned references. Exploiting the unique properties established in Theorem 4, we can conclude the following.

Corollary 6 *The inclusive KL divergence functional decays exponentially towards zero along the solution of the PDE equation iKL-WFR-GF.*

While this result renders the WFR gradient flow equation an attractive candidate for algorithm design, we again cannot simulate equation iKL-WFR-GF due to the function $1 - \mathrm{d}\pi / \mathrm{d}\mu$. To address this, we now follow equation iKL-WGF-k to kernelize the generalized

force in the transport velocity

$$\dot{\mu} = \alpha \cdot \operatorname{div}\left(\mu \int \nabla_2 k(x, \cdot)\ \mathrm{d}\left(\mu - \pi\right)(x)\right) - \beta \cdot (\mu - \pi)\,. \tag{IFT-GF}$$

Due to Proposition 5, we immediately find that:

Corollary 7 *equation IFT-GF is the gradient flow equation of the squared MMD functional, i.e., with the gradient structure*

$$\begin{cases} \textit{Space :} & \mathcal{P} \\ \textit{Energy functional :} & \frac{1}{2}\,\mathrm{MMD}^2(\cdot, \pi) \\ \textit{Dissipation Geometry :} & \textit{Interaction-force transport (IFT) (Gladin et al., 2024)} \end{cases} \tag{12}$$

This has recently been studied under the name of interaction-force transport (IFT) gradient flow by Gladin et al. (2024). It has been shown to practically accelerate and improve the performance of the MMD minimization task with proven guarantees. Here, we have shown that *the IFT gradient flow in Gladin et al. (2024) is an approximation to the Wasserstein-Fisher-Rao gradient flow of the inclusive-KL functional.* Gladin et al. (2024) have shown that MMD decays exponentially along the solution μ_t to the PDE equation IFT-GF, i.e., $\mathrm{MMD}(\mu_t, \nu) \leq e^{-\beta t} \cdot \mathrm{MMD}(\mu_0, \nu)$. An important aspect is that this convergence does not rely on the so-called log-concavity of the target distribution π; cf. (Chewi et al., 2024).

Appendix C. Additional derivations and proofs

By default, we work on the base space $\mathbb{R}^d$. The measures that appear in this paper are by default assumed to be absolutely continuous with respect to the Lebesgue measure. In formal derivation, we use measures and their density interchangeably, i.e., $\int f \cdot \mu$ means the integral w.r.t. the measure μ. We use the notation $\mathcal{P}, \mathcal{M}^+$ to denote the space of probability and non-negative measures on the closed, bounded, convex subset of $\mathbb{R}^d$. For many of our results, this domain can be generalized to $\mathbb{R}^d$ and the measures can be generalized to atomic measures; see (Ambrosio et al., 2005) for more details. The first variation of a functional F at $\mu \in \mathcal{M}^+$ is defined as a function $\frac{\delta F}{\delta \mu}[\mu]$ such that $\frac{\mathrm{d}}{\mathrm{d}\epsilon} F(\mu + \epsilon \cdot v)|_{\epsilon=0} = \int \frac{\delta F}{\delta \mu}[\mu](x)\ \mathrm{d}v(x)$ for any valid perturbation in measure v such that $\mu + \epsilon \cdot v \in \mathcal{M}^+$ when working with gradient flows over $\mathcal{M}^+$ and $\mu + \epsilon \cdot v \in \mathcal{P}$ over $\mathcal{P}$. The mathematical proofs are formal. To avoid confusion, we refer to the forward KL divergence $\mathrm{D}_{\mathrm{KL}}(\pi|\mu)$ as the inclusive KL; the reverse KL divergence $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$ as the exclusive KL. The technicalities of PDEs such as solution uniqueness, existence, and regularity are beyond the scope of this paper. Without further specification, the duality pairing $\langle f, g \rangle$ is the L^2 inner product $\displaystyle\int f(x) g(x)\ \mathrm{d}x$.

The φ-divergence, also known as the f-divergence (Csiszár, 1967), is a class of statistical divergences that measure the difference between a pair of measures, defined as

$$D_\varphi(\mu|\pi) = \int \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(x)\right)\ \mathrm{d}\pi(x),\ \text{if } \mu \ll \pi \tag{13}$$

and $+\infty$ otherwise; $\mathrm{d}\mu/\,\mathrm{d}\pi$ is the Radon-Nikodym derivative. The entropy generator function φ is a convex function satisfying $\varphi(1) = \varphi'(1) = 0,\ \varphi''(1) = 1$. Different choices of φ lead to various well-known divergences, such as exclusive KL: $\varphi_{\mathrm{KL}}(s) := s\log s - s + 1$, inclusive KL: $\varphi_{\mathrm{revKL}}(s) = s - 1 - \log s$, Hellinger: $\varphi_{\mathrm{H}}(s) = (\sqrt{s}-1)^2$, χ^2: $\varphi_{\chi^2}(s) = \frac{1}{2}(s-1)^2$. Note that our definition for measures that are not necessarily probability measures. As a central topic of this paper, we will focus on the KL divergence evaluated in both the direction of $\mathrm{D}_{\mathrm{KL}}(\mu|\pi)$ and $\mathrm{D}_{\mathrm{KL}}(\pi|\mu)$.

Proof [Proof of Theorem 2] The verification is a straightforward calculation. From the right-hand side of equation iKL-WGF-k, we have

$$\begin{aligned}\operatorname{div}\left(\mu\nabla\mathcal{T}_{k,\mu}\left(1-\frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right) &= \operatorname{div}\left(\mu\nabla\left(\int k(x,x')\mu(x')\left(1-\frac{\mathrm{d}\pi}{\mathrm{d}\mu}(x')\right)\,\mathrm{d}x'\right)\right)\\ &= \operatorname{div}\left(\mu\nabla\left(\int k(x,x')\left(\mu(x')-\pi(x')\right)\,\mathrm{d}x'\right)\right),\end{aligned}$$

which coincides with the right-hand side of equation MMD-WGF. ■

Proof [Proof of Proposition 3] The calculation of the flow equation is straightforward via Otto's formalism.

$$\dot{\mu} = -\mathbb{K}_{\mathsf{FR}}(\mu)\left(1-\frac{\mathrm{d}\pi}{\mathrm{d}\mu}-Z\right) \tag{14}$$

where Z is the normalization constant. Then,

$$\dot{\mu} = -\mu\left(1-\frac{\mathrm{d}\pi}{\mathrm{d}\mu}-Z\right) = -(\mu-\pi) \tag{15}$$

where Z disappears due to that the gradient flow is already mass-preserving. Therefore, the flow equation is indeed equation RevKL-FR-GF. The ODE solution is obvious. ■

Proof [Proof of Theorem 4] This is a corollary of the more general result by Mielke and Zhu (2025). There, they proved that the PL inequality holds for a large class of relative entropy functionals including the squared Hellinger distance, the inclusive KL divergence, and the reverse χ^2 divergence. Therefore, equation Ł-RKL holds globally.

Consequently, calculating the time-derivative of the inclusive KL divergence, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\mathrm{KL}}(\pi|\mu) = \langle 1-\frac{\mathrm{d}\pi}{\mathrm{d}\mu},\dot{\mu}\rangle = -\langle 1-\frac{\mathrm{d}\pi}{\mathrm{d}\mu},\mu\cdot(1-\frac{\mathrm{d}\pi}{\mathrm{d}\mu})\rangle \overset{equation\ -RKL}{\leq} -c\cdot\mathrm{D}_{\mathrm{KL}}(\pi|\mu). \tag{16}$$

By Grönwall's Lemma, we obtain the desired estimate. ■

Proof [Proof of Proposition 5] First, the equivalence between the flow equations is by direct identification – the flow equations coincide. This is a consequence of Theorem 3.4 of (Gladin et al., 2024). Then, using this equivalence, the MMD-decay statement follows from Theorem 3.5 of (Gladin et al., 2024) and their equation (12). ■

Proof [Proof of Proposition **??**] We calculate the optimality condition of the following optimization problem equation **??**.

$$1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu} + \frac{1}{\eta}\left(1 - \frac{\mathrm{d}\mu^l}{\mathrm{d}\mu}\right) = 0. \tag{17}$$

By the ISPD condition of the kernel k, the integral operator $\mathcal{T}_{k,\mu}$ is strictly positive-definite. Therefore, let $\mathcal{T}_{k,\mu}$ act on the both sides of the equation above, we have

$$\mathcal{T}_{k,\mu}\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right) + \frac{1}{\eta}\mathcal{T}_{k,\mu}\left(1 - \frac{\mathrm{d}\mu^l}{\mathrm{d}\mu}\right) = 0, \tag{18}$$

which coincides with the optimality condition of the variational problem equation **??** given Radon-Nikodym derivatives exist. ■

Corollary 8 [1] *The inclusive KL divergence functional decays exponentially towards zero along the solution of the PDE equation iKL-WFR-GF.*

Proof [Proof of Corollary 6] The proof is by exploiting the inf-convolution structure of the WFR flow. By taking the time-derivative of the inclusive KL divergence, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\mathrm{KL}}(\pi|\mu) = \langle 1 - \ \mathrm{d}\pi/\ \mathrm{d}\mu, \dot{\mu}\rangle = -\alpha\|\nabla\left(1 - \ \mathrm{d}\pi/\ \mathrm{d}\mu\right)\|^2_{L^2(\mu)} - \beta\|1 - \ \mathrm{d}\pi/\ \mathrm{d}\mu\|^2_{L^2(\mu)} \\ \leq -\beta\|1 - \ \mathrm{d}\pi/\ \mathrm{d}\mu\|^2_{L^2(\mu)}.\end{aligned}$$

By the functional inequality equation Ł-RKL in Theorem 4, we obtain the decay result for the inclusive KL functional. ■

Chewi et al. (2020)'s kernelized WGF of χ^2-divergence Previously, Chewi et al. (2020) proposed a kernelized Wasserstein gradient flow of the χ^2-divergence. They considered the following kernelized gradient flow equation:

$$\dot{\mu} = \mathrm{div}\left(\mu\mathcal{K}_\mu\nabla\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right), \tag{19}$$

where $\mathcal{K}_\mu$ should be taken as the integral operator defined by $\mathcal{K}_\mu f = \mathrm{Id}\circ\mathcal{T}_{k,\mu}$. However, in their implemented algorithm, they switched the order of the operators ∇ and $\mathcal{K}_\mu$, either as a heuristic or practical means. That is, what they actually implemented (in (Chewi et al., 2020, Section 4)) is

$$\dot{\mu} = \mathrm{div}\left(\mu\nabla\mathcal{K}_\pi\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right). \tag{20}$$

From this paper's perspective, this is kernelizing the genralized thermodynamic force, rather than the velocity function $\nabla\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right)$.

1. We note that there was an error in the original statement of Corollary 6. We now correct the statement and provide a proof.

Using this paper's technique, we can now derive a force-kernelized WGF of the χ^2-divergence from the first principle. Consider the gradient flow equation

$$\dot{\mu} = \operatorname{div}\left(\mu \nabla \mathcal{T}_{k,\pi}\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right)\right). \tag{21}$$

Note that the integral operator $\mathcal{T}_{k,\pi}$ is associated with the target measure π, rather than the measure μ as in equation iKL-WGF-k. Nonetheless, a simple observation is that equation 21 formally coincides with equation iKL-WGF-k. Therefore, we conclude that a *principled force-kernelized flow of the χ^2-divergence WGF equation 21 is equivalent to the WGF of the MMD studied by (Arbel et al., 2019; Korba et al., 2021)*, which is straightforward to implement and in contrast to using the ad-hoc scheme of (Chewi et al., 2020).

Local nonparametric regression formulation equation iKL-WGF-k can be viewed as the following local regression estimator of the WGF of inclusive KL equation iKL-WGF.

$$f = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \int \mu(x') k(x' - x) \left| \theta - \frac{\delta F}{\delta \mu}\left[\mu_t\right](x') \right|^2 \mathrm{d}x' \right\}$$

where the energy functional the inclusive KL divergence $F = \mathrm{D}_{\mathrm{KL}}(\pi|\cdot)$. Using standard local regression results (Tsybakov, 2009; Spokoiny, 2016; Zhu and Mielke, 2024), we obtain the closed-form estimator

$$f(x) = \int \mu(x') \frac{k(x' - x)}{\int \mu(x') k(x' - x)\ \mathrm{d}x'} \frac{\delta F}{\delta \mu}\left[\mu_t\right](x')\ \mathrm{d}x', \tag{22}$$

and a fintie-sample Nadaraya-Watson estimator $\hat{f}(x) = \sum_{i=1}^{N} \frac{k(x_i - x)}{\sum_{i=1}^{N} k(x_i - x)} \cdot \frac{\delta F}{\delta \mu}\left[\mu_t\right](x_i)$.

In particular, in the inclusive KL setting, we obtain

$$f(x') = \int \frac{k(x' - x)}{\int \mu(x') k(x' - x)\ \mathrm{d}x'} \Big(\mu(x') - \pi(x')\Big)\ \mathrm{d}x'.$$

Given two samples $\{y_i\}_{i=1}^N \sim \mu$ and $\{z_i\}_{i=1}^M \sim \pi$, a finite-sample estimator is the difference between two kernel density estimators

$$\hat{f}(x) = \sum_{i=1}^{N} \frac{k(y_i - x)}{\sum_{i=1}^{N} k(y_i - x)} - \sum_{i=1}^{M} \frac{k(z_i - x)}{\sum_{i=1}^{M} k(z_i - x)}.$$

The resulting gradient flow equation is given by

$$\dot{\mu} = \operatorname{div}\left(\mu \cdot \nabla \int \frac{1}{Z} k(x' - x)\ \mathrm{d}\Big(\mu(x') - \pi(x')\Big)\right) \text{ for } Z = \int \mu(x') k(x' - x)\ \mathrm{d}x'. \tag{23}$$

Comparing with equation MMD-WGF and equation iKL-WGF-k, we can see that the local regression estimator only differs by a constant scaling factor Z. Therefore, the force-kernelized gradient flow equation equation iKL-WGF-k, and hence equation MMD-WGF, can be interpreted as a flow matching implementation of the Wasserstein gradient flow of the inclusive KL divergence equation iKL-WGF.

Remark 9 (Approximation limit) *Suppose the kernel k in equation iKL-WGF-k is a Gaussian kernel with bandwidth σ. One might conjecture that, as the bandwidth σ approaches zero and the integral operator tends to an identity map, equation MMD-WGF recovers equation iKL-WGF. This potential connection could provide a new link between the two gradient flows. However, proving rigorous Γ-convergence in this setting is mathematically non-trivial and left for future research. We refer interested readers to relevant works on Γ-convergence, e.g., (Craig et al., 2023; Carrillo et al., 2019; Lu et al., 2023; Zhu and Mielke, 2024).*

Appendix D. Further background on Wasserstein gradient flows

We provide further background on Wasserstein gradient flows, especially on the pseudo-Riemannian structure of the Wasserstein space.

The Onsager operator, as well as the Riemannian metric tensor $\mathbb{G}_W = \mathbb{K}_W^{-1}$, induces a duality pairing between the tangent and cotangent spaces. We use the unweighted space for simplicity. Note that the calculation can also be made in the weighted space $L^2(\rho)$.

$$\text{duality pairing: } {}_{\text{dual}}\langle \xi, \mathbb{K}_W(\rho)\ \zeta\rangle_{\text{primal}} = \langle \xi, \mathbb{K}_W(\rho)\ \zeta\rangle_{L^2} = \int \xi \cdot \mathbb{K}_W(\rho)\ \zeta. \tag{24}$$

The Stein geometry can also be characterized in this way. Duncan et al. (2019) proposed the following Onsager operator that is a modification of the Otto's Wasserstein formalism,

$$\mathbb{K}_{\text{Stein}}(\rho) : T_\rho^*\mathcal{M} \to T_\rho\mathcal{M}, \xi \mapsto -\operatorname{div}(\rho \cdot \operatorname{Id} \circ \mathcal{T}_{k,\rho}\nabla\xi). \tag{25}$$

The resulting Stein gradient flow equation is given by

$$\partial_t \mu = -\mathbb{K}_{\text{Stein}}(\mu) \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} = \operatorname{div}\left(\mu \mathcal{K}_\mu \nabla\left(V + \log \mu\right)\right).$$

We now look at the Wasserstein gradient flow of the inclusive KL divergence. The gradient flow equation can be given by the Otto's formal calculation,

$$\nabla_W \mathrm{D}_{\mathrm{KL}}(\pi\|\mu) = \ \mathbb{K}_W \partial \mathrm{D}_{\mathrm{KL}}(\pi\|\mu) = -\operatorname{div}\left(\mu\nabla\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right)$$

where $\mathbb{K}_W$ is the Wasserstein Onsager operator, i.e., the inverse of the Riemannian metric tensor $\mathbb{G}_W$ of the Wasserstein manifold.

A standard characterization of the Wasserstein gradient flow is the following energy dissipation equality in the inclusive KL setting

$$\mathrm{D}_{\mathrm{KL}}(\pi|\mu_t) - \mathrm{D}_{\mathrm{KL}}(\pi|\mu_s) = -\int_s^t \left\|\nabla\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu_r}\right)\right\|_{L^2(\mu_r)}^2 \mathrm{d}r. \tag{26}$$

The dissipation of the inclusive KL divergence energy, a.k.a. the production of the relative entropy, equals the integral of the Sobolev norm of the differential of the inclusive KL along

the curve μ_r. For completeness, we provide a standard characterization via the following differential energy dissipation equality

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D}_{\mathrm{KL}}(\pi|\mu_t) = \langle 1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}, \dot{\mu}_t\rangle = \langle 1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}, \mathbb{K}_W \partial \mathrm{D}_{\mathrm{KL}}(\pi|\mu)\rangle$$
$$= \langle 1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}, \mathrm{div}(\mu\nabla\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right))\rangle = -\left\|\nabla\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu}\right)\right\|^2_{L^2(\mu)}. \quad (27)$$

Integrating both sides, the integral form of EDE is then given by equation 26.

Appendix E. Further background on Fisher-Rao and Hellinger gradient flows

We provide further background on Fisher-Rao and Hellinger gradient flows, especially on the technicalities of the Hellinger flows over positive measures $\mathcal{M}^+$.

We first consider the Hellinger flow of the exclusive (reverse) KL divergence over the positive measures $\mathcal{M}^+$. Its gradient flow equation, the reaction equation, is given by

$$\dot{\mu} = -\mu \cdot \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}. \quad (28)$$

The gradient structure is given by

$$\begin{cases} \text{Space :} & \text{positive measures } \mathcal{M}^+ \\ \text{Energy functional :} & \text{exclusive KL: } \mathrm{D}_{\mathrm{KL}}(\cdot|\pi) \\ \text{Dissipation Geometry :} & \text{Hellinger} \end{cases} \quad (29)$$

One can further restrict the gradient flow to the probability measures by modifying the dynamics in equation 28 with a projection onto the probability measures, i.e.,

$$\dot{\mu} = -\mu \cdot \left(\log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} - \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}\, \mathrm{d}\mu\right). \quad (30)$$

The resulting ODE is the gradient flow equation over the Fisher-Rao manifold of the probability measures, also known as the spherical Hellinger manifold (Laschos and Mielke, 2019). That is, it has the following gradient structure:

$$\begin{cases} \text{Space :} & \text{probability measures } \mathcal{P} \\ \text{Energy functional :} & \text{exclusive KL: } \mathrm{D}_{\mathrm{KL}}(\cdot|\pi) \\ \text{Dissipation Geometry :} & \text{spherical Hellinger a.k.a. Fisher-Rao} \end{cases} \quad (31)$$

For the inclusive (forward) KL divergence, as discussed in the main text, the Hellinger flow over the positive measures $\mathcal{M}^+$ coincides with the Fisher-Rao flow over the probability measures $\mathcal{P}$, given the same initial condition. Specifically, the Hellinger gradient structure over the positive measures $\mathcal{M}^+$ is given by

$$\begin{cases} \text{Space :} & \text{positive measures } \mathcal{M}^+ \\ \text{Energy functional :} & F(\cdot) := \mathrm{D}_{\mathrm{KL}}(\pi|\cdot) \\ \text{Dissipation Geometry :} & \text{Hellinger} \end{cases} \quad (32)$$

This flow actually contains the flow equation 7 if initialized as probability measures.

For the Hellinger flows, an interesting and known analysis result is that the following Polyak-Łojasiewicz functional inequality cannot hold globally for exclusive KL divergence functional. in the Hellinger geometry,

$$\left\| \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi} \right\|_{L^2_\mu}^2 \geq c \cdot \mathrm{D}_{\mathrm{KL}}(\mu(t) \| \pi). \tag{Ł-KL}$$

Inequality equation Ł-KL differs from the typical log-Sobolev inequality in that no Sobolev norm is involved. An elementary proof was provided by Mielke and Zhu (2025). Consequently, we obtain the following lemma regarding the property of the Hellinger flows of the exclusive KL by Grönwall's Lemma. This is in sharp contrast to the case of the inclusive KL as discussed in the main text.

Lemma 10 (No global Łojasiewicz condition in Fisher-Rao flows of KL) *There exists no $c > 0$ such that equation Ł-KL holds along the Hellinger gradient flow of the exclusive KL divergence functional.*

Remark 11 *Strictly speaking, we make the following distinction:*

- *Fisher-Rao (FR): a Bregman divergence between parameters of the (exponential-family) distributions generated by the suitable generator functions.*
- *Hellinger (He): a special φ-divergence/distance defined in our paper.*
- *spherical Hellinger (SHe): a distance induced by restricting the Hellinger geodesics to the probability measures; also called Bhattacharya distance by Rao (1945) after its first introduction by Bhattacharyya (1946). We can recover the equivalence between SHe and FR if we consider the trivial parameterization of the probability measure by itself (infinite-dimensional).*

For more details, see the discussion in (Mielke and Zhu, 2025).

Appendix F. Kernel Stein discrepancy descent as inclusive KL inference

The original implementation of equation MMD-WGF by Arbel et al. (2019) suffers from a few drawbacks such as mode collapse or slow convergence. Furthermore, their algorithm requires samples from the target distribution π, which may be impractical in some applications. For example, in Bayesian inference, one typically only has access to the posterior via the score function $\nabla \log \pi$. Instead of the MMD, authors such as Korba et al. (2021); Chen et al. (2018); Barp et al. (2019) advocated for minimizing the kernel Stein discrepancy (KSD) (Gorham and Mackey, 2017; Liu et al., 2016; Chwialkowski et al., 2016) for inference. From the optimization perspective, we replace the MMD objective with the KSD objective $\frac{1}{2}\mathrm{KSD}^2(\mu|\pi)$ The KSD can be viewed as a special case of the MMD associated with the Stein kernel (Gorham and Mackey, 2017; Liu et al., 2016; Chwialkowski et al., 2016). The Wasserstein gradient flow equation of the KSD can be straightforwardly calculated as noted by Korba et al. (2021).

$$\dot{\mu} = \mathrm{div}(\mu \cdot \int \nabla_2 s_\pi(x, \cdot) \, \mathrm{d}\mu(x)), \tag{KSD-WGF}$$

where s_π is the Stein kernel; see the appendix. Unlike equation MMD-WGF, to implement a discrete-time algorithm that simulates equation KSD-WGF, we only need to evaluate the score function $\nabla \log \pi$ without needing the samples from π. As KSD can be viewed as a special case of MMD with the Stein kernel, using our characterization of the MMD-WGF in Theorem 2, we obtain the following insight.

Corollary 12 (Formal equivalence between KSD-WGF and inclusive KL inference) *The WGF equation of KSD equation KSD-WGF is equivalent to equation iKL-WGF-k, which is the kernelized WGF of the inclusive KL divergence energy functional when the kernel is the Stein kernel s_π.*

Notably, Korba et al. (2021) empirically demonstrated that the KSD-based flow significantly outperforms the MMD-based flow in practice. The finding in this paper unifies the KSD-based flows and the MMD-based flows. This forms a unified framework for inclusive KL minimization via kernelized Wasserstein gradient flows.

Appendix G. Examples and algorithmic implications

In this section, we demonstrate our gradient flow theory in stylized examples from machine learning and statistical inference.

G.1. Inference and sampling algorithms via force-kernelized Wasserstein flows

The goal of this section is to show a general-purpose inference algorithm for the inclusive KL inference, analog to the SVGD for the forward KL inference, can be constructed using our gradient flow theory. Suppose our goal is to approximate a target distribution π via the inclusive KL minimization equation 2. We consider two settings: **(1)** we have access to samples from the target $y^i \sim \pi$, e.g., in generative modeling; **(2)** we have access to the score function $\nabla \log \pi$, e.g., in inference and sampling. Our scheme is based on discretizing the force-kernelized Wasserstein gradient flow equation equation iKL-WGF-k, obtaining the discret-time update scheme

$$X_{t+1} = X_t - \tau \nabla \int \nabla_2 k(x', x) \frac{\delta F}{\delta \mu}[\mu_t](x') \ \mathrm{d}\mu_t(x'). \tag{33}$$

An interacting particle system can be simulated by considering particle approximation to the measure, $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i},\ x_i \in \mathbb{R}^d$.

Setting (1): sample-based setting with flows of MMD In general, for Wasserstein gradient flow of the energy functional F, one may implement a practical algorithm that discretizes the PDE equation 3. As discussed in the beginning of Section 3, in the vanilla Wasserstein gradient flow of the inclusive KL divergence equation iKL-WGF, the velocity field $\nabla\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\mu_t}(X_t)\right)$ cannot be implemented out of the box. Based on Theorem 2, we now resort to equation 33 which is algorithmicly equivalent to Arbel et al. (2019)'s algorithm which they termed MMD-flow. This amounts to simulating (in discrete time) an interacting particle system:

$$X_{t+1}^i = X_t^i - \tau \left(\frac{1}{N}\sum_{j=1}^N \nabla_2 k(X_t^j, X_t^i) - \frac{1}{M}\sum_{j=1}^M \nabla_2 k(Y_t^j, X_t^i) \right), \tag{34}$$

where X_t^i are samples from the distribution μ_t; cf. (Arbel et al., 2019) for the experimental results.

Setting (2): score-based setting with flows of KSD In variational inference, we typically have access to the target π in the form of the score function $\nabla \log \pi$ without samples. Discretizing the PDE equation KSD-WGF, we have

$$X_{t+1} = X_t - \tau \int \nabla_2 s_\pi(x, \cdot) \, \mathrm{d}\mu(X_t). \tag{35}$$

A sample-based implementation of the above algorithm is then given by

$$X_{t+1}^i = X_t^i - \tau \left(\frac{1}{N} \sum_{j=1}^N \nabla_2 s_\pi(X_t^j, X_t^i) \right). \tag{36}$$

In summary, to solve the inclusive KL minimization equation 2, one can apply the general-purpose algorithm via the discrete time scheme equation 33. This is implementable in practice in both sample-based and score-based settings; cf. empirical results in (Korba et al., 2021; Arbel et al., 2019). Hence, we have established a previously missed link between inclusive KL inference and the gradient flows of MMD-type functionals.

G.2. Generative modeling

Recently, there is a surge of interest in formulating GANs in the fashion of Wasserstein gradient flows Promising empirical results have been reported by Ansari et al. (2021); Yi et al. (2023); Yi and Liu (2023); Franceschi et al. (2024); Heng et al. (2024). In addition, there have also been a series of paper that present theoretical analysis of GAN training dynamics as *interacting gradient flows* by, e.g., Hsieh et al. (2018); Domingo-Enrich et al. (2020); Wang and Chizat (2022, 2023); Dvurechensky and Zhu (2024-05-02/2024-05-04). Using this paper's insight, we now uncover some previously unknown connections between generative models and the Wasserstein gradient flow of the inclusive KL functional.

Our starting point is the standard divergence-based generative modeling training, which solves the optimization problem

$$\min_\theta \mathrm{D}_{\mathrm{KL}}(\pi_{\mathrm{data}} | g_{\theta\#} P_Z), \tag{37}$$

where P_Z is the latent variable distribution, e.g., standard Gaussian, and g_θ is the generator network. Note that it is also possible to work with numerous other geneartive formulations such as models without explicit generator parameterization such as discriminator flows as discussed in Franceschi et al. (2024). Following our force-kernelized WGF framework as in equation iKL-WGF-k, consider a force-kernelized projected gradient flow for the inclusive KL minimization equation 37.

$$\dot{\theta} = -\Pi_\Theta \left(-\operatorname{div} \left(\mu_\theta \nabla \mathcal{T}_{k,\mu_\theta} \left(1 - \frac{d\pi}{d\mu_\theta} \right) \right) \right), \quad \mu_\theta = g_{\theta\#} P_Z \tag{38}$$

where $\mathbb{K}_W$ again denotes the Wasserstein Onsager operator. Π_Θ is the projection (of the Riemannian gradient) onto the parameter space Θ. Form equation 38, we immediately observe that the flow can be written in the form of a flow-matching model

$$\dot{\theta} = -\Pi_\Theta \left(-\operatorname{div}\left(\mu_\theta \nabla f^*(x)\right)\right), \quad f^*(x) = \int k(x', x) \left(\mu_\theta(x') - d\pi(x')\right) \, \mathrm{d}x'. \tag{39}$$

Discretizing the above PDE, we have

$$\theta^{l+1} \leftarrow \theta^l - \eta^l \Pi_\Theta \left(-\operatorname{div}\left(\mu_\theta \nabla f^*(x)\right)\right) \tag{40}$$

which corresponds to the training dynamics of MMD-GANs (Li et al., 2015; Dziugaite et al., 2015; Li et al., 2017; Bińkowski et al., 2018) using the optimal test function f^*. The insight of our paper is that, through the lens of equation 38, we can view the MMD-GAN training dynamics as performing inclusive KL inference using an approixmate Wasserstein gradient flow.

G.3. Discrete-time mirror descent

Recently, there have a few studies using mirror descent of the exclusive KL divergence such as (Chopin et al., 2024; Aubin-Frankowski et al., 2022). We now provide the details of inclusive KL minimization via mirror descent. Consider an explicit Euler scheme for the gradient flow

$$\min_{\rho \in \mathcal{P}} \langle \partial \mathrm{D}_{\mathrm{KL}}(\pi|\rho), \rho \rangle + \frac{1}{\tau} \mathrm{D}_{\mathrm{KL}}(\rho|\rho^l). \tag{41}$$

where $\langle , \rangle$ is the L^2 inner product. Using the optimality condition of this optimization problem, we can derive the following mirror descent update:

$$\rho^{l+1}(x) \leftarrow \frac{1}{Z^l} \rho^l(x) \cdot \exp\left(-\tau \left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\rho^l}\right)\right) \text{ for all } x \in \mathbb{R}^d, \tag{42}$$

where Z^l is the normalization constant. We now again apply the kernel approximation of this paper, obtaining the update rule

$$\rho^{l+1}(x) \leftarrow \frac{1}{Z^l} \rho^l(x) \cdot \exp\left(-\tau \cdot \mathcal{T}_{k,\rho^l}\left(1 - \frac{\mathrm{d}\pi}{\mathrm{d}\rho^l}\right)\right) = \frac{1}{Z^l} \rho^l(x) \cdot \exp\left(-\tau \cdot \int k(x, y) \left(\rho^l(y) - \pi(y)\right) \, \mathrm{d}y\right).$$

Similarly, using Stein's method, we can also perform update only via the score function of the target $\nabla \log \pi$.

$$\rho^{l+1} \leftarrow \frac{1}{Z^l} \rho^l \cdot \exp\left(-\tau \cdot \int s_\pi(x, y) \rho^l(y) \, \mathrm{d}y\right). \tag{43}$$

Appendix H. Other related works

Outside machine learning, there exists many works in interacting particle systems that uses similar approximation methods as kernelization, such as the blob method (Carrillo et al., 2019; Craig et al., 2023). In the optimization literature, there are a few related works using

particle-based gradient descent methods, such as (Dai et al., 2016) and (Chizat, 2022), albeit they are not concerned with the inclusive KL divergence. For Bayesian inference, Trillos and Sanz-Alonso (2018) provide a variational perspective for Bayesian update, framing it in terms of gradient flows. Therein, they primarily consider the KL, χ^2, and Dirichlet energy functional. Maurais and Marzouk (2024)'s algorithm seeks a velocity field for the ODE to match the behavior of the Fisher-Rao flow. It is also worth noting that they assume the setting of importance sampling where one has access to the density ratio. Vargas et al. (2024) proposed a framework that governs many existing variational Bayesian methods. While not directly related to our work, one part of their reversal framework is indeed a inclusive KL inference problem. Chewi et al. (2020) propose a perspective that views SVGD as a kernelized Wasserstein gradient flow. In the appendix, we expand on the precise relation between this paper and the algorithm they actually implemented, which switched the order or gradient and kernelization operation. Then, we show that the χ^2 flow can be cast into our framework without that heuristic implementation. In addition to the Wasserstein gradient flow, there exist several works that are based on unbalanced transport and its variants, such as (Lu et al., 2019; Mroueh and Rigotti, 2020; Lu et al., 2023; Yan et al., 2023; Gladin et al., 2024). Instead of approximation via integral operator, a ridge-regression type of gradient flow approximation can also be considered; cf. (He et al., 2022; Zhu and Mielke, 2024; Nüsken, 2024).

Acknowledgements

The author acknowledges the funding support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the priority programme "Theoretical foundations of deep learning" (project number: 543963649). I thank Wittawat Jitkrittum and Thomas Möllenhoff for the helpful discussions.