Towards Healing the Blindness of Score Matching

Mingtian Zhang
University College London
m.zhang@cs.ucl.ac.uk

Oscar Key
University College London
o.key@cs.ucl.ac.uk

Peter Hayes
University College London
p.hayes@cs.ucl.ac.uk

David Barber
University College London
david.barber@ucl.ac.uk

Brooks Paige
University College London
b.paige@ucl.ac.uk

François-Xavier Briol
University College London
f.briol@ucl.ac.uk

Abstract

Score-based divergences have been widely used in machine learning and statistics applications. Despite their empirical success, a blindness problem has been observed when using these for multi-modal distributions. In this work, we discuss the blindness problem and propose a new family of divergences that can mitigate the blindness problem. We illustrate our proposed divergence in the context of density estimation and report improved performance compared to traditional approaches.

1 Introduction

Score-based divergences such as the Fisher Divergence (FD; also known as score-matching divergence) [10, 9] and Kernel Stein Discrepancy (KSD) [13, 3] are widely used in machine learning and statistics [1, 17]. Their main advantage is that the score function, a derivative of a log-density, can be evaluated without knowledge of the normalization constant of the density and can be applied to problems where other classical divergences (e.g. KL divergence) are intractable. Unfortunately, this advantage can also be a curse in certain scenarios because the score function only provides local information about the slope of a density, but ignores more global information such as the importance of a point relative to another. This has led to a blindness problem in many applications of score-based methods where the densities are multi-modal, including in density estimation [21, 16, 11], MCMC convergence diagnosis [8], Bayesian inference [14, 5]; see [20] for a detailed discussion.

To illustrate this problem, we recall the definition of FD and an example from [20]. Given two distributions with differentiable densities p and q supported on a common domain $\mathcal{X} \subseteq \mathbb{R}^d$, the FD is

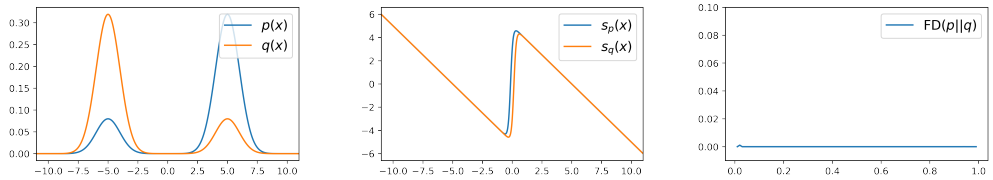
$$\mathrm{FD}(p||q) = \tfrac{1}{2} \int_{\mathcal{X}} p(x)||s_p(x) - s_q(x)||_2^2 dx, \tag{1}$$

where we denote by $s_p(x) \equiv \nabla_x \log p(x)$ and $s_q(x) \equiv \nabla_x \log q(x)$ the score functions of p and q respectively. The classic sufficient conditions [10, 2] for the FD to be a valid statistical divergence (i.e. $\mathrm{FD}(p||q) = 0 \Leftrightarrow p = q$) are: (i) p and q are differentiable with support $\mathcal{X} = \mathbb{R}^d$ and (ii) s_p, s_q are square integrable, i.e. $s_p - s_q \in L^2(p)$, where we denote $f \in L^2(p) \equiv \int_{\mathcal{X}} ||f(x)||_2^2 p(x) dx < \infty$. The blindness problem of the FD can be illustrated through the following example due to [20]. Let p and q be a mixtures with the same components but different mixing weights:

$$p(x) = \alpha_p g_1(x) + (1 - \alpha_p) g_2(x), \quad q(x) = \alpha_q g_1(x) + (1 - \alpha_q) g_2(x), \tag{2}$$

where $\alpha_p \neq \alpha_p$, and g_1, g_2 are Gaussian densities with variance σ^2 and means $-\mu$ and μ respectively. Then $\mathrm{FD}(p||q) \to 0$ when $\mu/\sigma^2 \to \infty$ regardless of the mixture proportions α_p and α_q. To build intuition, we let $\mu = 5$, $\sigma = 1$, $\alpha_p = 0.2$, $\alpha_q = 0.8$ and plot the densities and score functions of p, q

Score-based Method Workshop at Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) Densities of p and q (b) Score functions of p and q (c) $\mathrm{FD}(p||q)$ with different α_q

Figure 1: We plot the densities and score functions of distributions p and q in Figure (a) and (b). Figure (c) shows $\mathrm{FD}(p||q)$ with $\alpha_p = 0.2$ and α_q varies from 0.01 to 0.09 with a grid size 0.01.

in Figure 1a and 1b. We can find the two distributions are very different but their scores are only different around $x = 0$, which has a negligible density value under p. We then fix $\alpha_p = 0.2$ and plot the $\mathrm{FD}(p||q)$ as a function of α in Figure 1c. Here we see the FD is 0 constant function, which shows the FD is 'blind' to the value of the mixture weight. See [14] for a similar example for discrete $\mathcal{X}$.

2 Understanding the Blindness Problem

In the example above, blindness is a numerical problem since the problem occurs despite the fact that the FD is a divergence in that case (i.e. $\mathrm{FD}(p||q) = 0 \Leftrightarrow p = q$ since (i) and (ii) are satisfied). When $\mu/\sigma^2 \to \infty$, although the Gaussian distributions still have the same support, the regions that contain most of the mass of g_1 and g_2 tend to be disjoint, which creates numerical issues. However, the blindness problem is not simply a numerical problem, as illustrated in the following example.

Consider the case where p and q are mixtures whose identical components have disjoint supports. For example, let g_1 and g_2 in Equation 2 have disjoint support sets $\mathcal{X}_1, \mathcal{X}_2 \subseteq \mathbb{R}^d$ respectively with $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$. Then, $g_2(x') = \nabla_x g_2(x') = 0$ for $x' \in \mathcal{X}_1$ and $g_1(x') = \nabla_x g_1(x') = 0$ for $x' \in \mathcal{X}_2$. In this case, the FD is independent of α_q (see Appendix A.1 for a derivation):

$$\mathrm{FD}(p||q) = \tfrac{\alpha_p}{2} \int_{\mathcal{X}_1} g_1(x)||s_{g_1}(x) - s_{g_1}(x)||_2^2 dx + \tfrac{1-\alpha_p}{2} \int_{\mathcal{X}_2} g_2(x)||s_{g_2}(x) - s_{g_2}(x)||_2^2 dx = 0. \quad (3)$$

Therefore, the FD is not a valid divergence here since $\mathrm{FD}(p||q) = 0 \not\Rightarrow p = q$. This example guides us to further study the topology properties of the distributions' support required by the FD. We first extend the Fisher divergence to distributions that have support on the connected space.

Theorem 1 (FD on a connected set)**.** *Assume two distributions (i) have differentiable densities p and q with support on a common*[1] *open connected set $\mathcal{X} \subseteq \mathbb{R}^d$ and (ii) $s_p - s_q \in L^2(p)$. Then, the FD is a valid divergence i.e. $\mathrm{FD}(p||q) = 0 \Leftrightarrow p = q$.*

See Appendix A.2 for a proof. Theorem 1 generalizes the classic FD that is defined on distributions with $\mathcal{X} = \mathbb{R}^d$ [10, 2] ($\mathbb{R}^d$ is a special case of the connected set). Secondly, Theorem 2 shows that *connectedness* of the support is a *necessary* condition to define a valid FD.

Theorem 2 (FD is ill-defined on disconnected sets)**.** *Assume two distributions have common support $\mathcal{X}$ consisting of disjoint sets. Then, the FD is not a valid divergence i.e. $\mathrm{FD}(p||q) = 0 \not\Rightarrow p = q$.*

See Appendix A.3 for a proof. Intuitively, the score function only considers the local derivatives and contains no information of the global normalization constant. If the domain is disconnected, it cannot determine the mass allocation in different domains. This observation can also be extended to the KSD by viewing KSD as a kernelized FD [13, 3], see Appendix A.5 for a detailed discussion.

3 Healing the Blindness Problem with the Mixture Fisher Divergence

In this section, we propose a new variant of the FD which is well-defined in the disconnected scenario. Consider a distribution with density m with support $\mathcal{X}_m = \mathbb{R}^d$ and define the mixtures

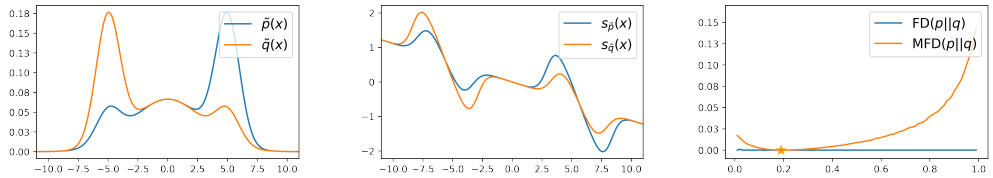
$$\tilde{p}(x) = \beta p(x) + (1-\beta)m(x), \quad \tilde{q}(x) = \beta q(x) + (1-\beta)m(x), \quad (4)$$

where $0 < \beta < 1$. We then define the *Mixture Fisher Divergence* (MFD) as

$$\mathrm{MFD}_{m,\beta}(p||q) \equiv \mathrm{FD}(\tilde{p}||\tilde{q}). \quad (5)$$

Theorem 3 shows the MFD is well-defined when p and q have support on a disconnected space.

[1]The common support condition can be relaxed to $\mathcal{X}_p \subseteq \mathcal{X}_q$, where $\mathcal{X}_p, \mathcal{X}_q$ are the support sets of p and q.



(a) Densities of $\tilde{p}$ and $\tilde{q}$ (b) Score functions of $\tilde{p}$ and $\tilde{q}$ (c) $\mathrm{MFD}(p||q)$ with different α_q

Figure 2: We plot the densities (a) and the score functions (b) of $\tilde{p}$ and $\tilde{q}$. Figure (c) shows $\mathrm{MFD}(p||q)$ with $\alpha_p = 0.2$ and α_q varies from 0.0 to 1.0 with a grid size 0.01. The star mark shows the minima of the MFD is achieved when $\alpha_q = \alpha_p = 0.2$, we also plot the original FD for a comparison.

Theorem 3 (Validity of the MFD)**.** *Consider two distributions with differentiable densities p, q supported on $\mathcal{X}_p, \mathcal{X}_q \subseteq \mathbb{R}^d$ with $s_p, s_q \in L^2(p)$ and a differentiable density m with support $\mathcal{X}_m = \mathbb{R}^d$, $s_m \in L^2(p)$. Then MFD is a valid divergence, i.e.* $\mathrm{MFD}(p||q) = 0 \Leftrightarrow \mathrm{FD}(\tilde{p}||\tilde{q}) \Leftrightarrow p = q$.

See Appendix A.6 for a proof. For MFD, we no longer require that p, q have *common connected support*, since $\mathcal{X}_m = \mathbb{R}^d$ results in $\tilde{p}, \tilde{q}$ having connected support $\mathbb{R}^d$[2]. The requirements of $m(x)$ are mild and hold for simple choices of distribution e.g. a Gaussian. To avoid the numerical problem mentioned in Section 1, $m(x)$ should be chosen to effectively connect the different component distributions. As an example, for the toy problem described in Figure 1 with components $\mathcal{N}(-5, 1)$ and $\mathcal{N}(-5, 1)$ we can choose $\beta = 0.5$ and $m(x) = \mathcal{N}(0, 9)$ that covers both components. Figure 2 shows the densities and their score functions for $\tilde{p}, \tilde{q}$. We see that the score functions are different on the high-density region of $\tilde{p}$. Figure 2c also shows the minimal value of the $\mathrm{MFD}(p||q)$ is attained when $\alpha_q = \alpha_p$, which indicates that the proposed MFD heals the blindness problem in this example.

4 Density Estimation with Energy-based Models

Given a dataset $\mathcal{X}_{\text{train}} = \{x_1, \cdots, x_N\}$ sampled *i.i.d.* from a data distribution p_d with support $\mathcal{X}_{p_d} \subseteq \mathbb{R}^d$, we would like to learn a model q_θ to approximate p_d. We are interested in a family of models which can only be evaluated up to a normalization constant, e.g. an energy-based model $q_\theta(x) = e^{-f_\theta(x)}/Z(\theta)$, where f_θ is a neural network and $Z(\theta) = \int e^{-f_\theta(x)} dx$. In this case, the standard Maximum Likelihood Estimation (MLE) is not applicable (since $Z(\theta)$ cannot be evaluated during training) and an alternative form of the FD [10] can be applied (see Appendix A.4 for a derivation and additional assumptions)

$$\mathrm{FD}(p_d||q_\theta) = \tfrac{1}{2} \int_{\mathcal{X}_{p_d}} p_d(x) \left(||s_{q_\theta}(x)||_2^2 + 2 \operatorname{Tr}(\nabla_x s_{q_\theta}(x)) \right) dx + const., \tag{6}$$

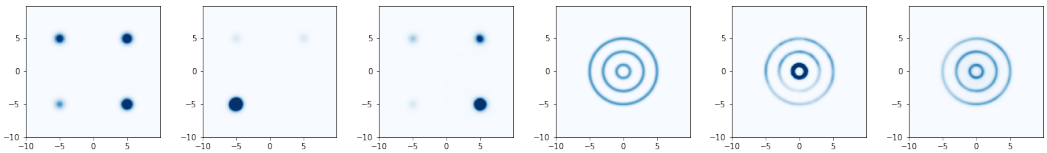
where $\nabla_x s_{q_\theta}(x) = \nabla_x^2 \log q_\theta(x)$ is the Hessian matrix and the constant represents the terms that are independent of θ. The integration over p_d can be approximated by Monte-Carlo using $\mathcal{X}_{train}$. Because both s_{q_θ} and $\nabla_x s_{q_\theta}$ only depend on f_θ, the normalizer $Z_q(\theta)$ is not required during training and we only need to estimate $Z_q(\theta^*)$ once after training. Therefore, density estimation with FD in this setting contains two steps: (1) learn θ^* using Equation 6; (2) estimate $Z(\theta^*)$ to obtain the normalized density $q_\theta(x)$. This scheme can result in blindness in practice [21].

To heal the blindness, we can apply the proposed MFD. However, if we directly minimize MFD in step (1), the score $s_{\tilde{q}_\theta}(x) = \nabla_x \log \left(\beta \exp(-f_\theta(x))/Z_q(\theta) \right) + (1 - \beta) m(x))$ requires estimating $Z_q(\theta)$. This negates the advantage of using score matching because now $Z_q(\theta)$ must be estimated for every gradient step during training (similar to MLE). To avoid this, we propose to instead directly approximate $\tilde{p}_d$ with an energy-based model $\tilde{q}_\theta(x) \equiv e^{-f_\theta(x)}/Z_{\tilde{q}}(\theta)$ and $\tilde{q}_\theta$ can then be trained using

$$\mathrm{FD}(\tilde{p}_d||\tilde{q}_\theta) = \tfrac{1}{2} \int_{\mathbb{R}^d} \tilde{p}_d(x) \left(||s_{\tilde{q}_\theta}(x)||_2^2 + 2 \operatorname{Tr}\left(\nabla_x s_{\tilde{q}_\theta}(x) \right) \right) dx + const., \tag{7}$$

where the integration over $\tilde{p}_d(x)$ can be approximated using the samples from the mixture $\tilde{p}_d(x) = \beta p_d(x) + (1 - \beta) m(x)$. Therefore, the learning of θ is independent of $Z_{\tilde{q}}(\theta)$. Optimally we have

[2] A weaker condition of m can be obtained by requiring the supports of $\tilde{p}, \tilde{q}$, which we denote as $\mathcal{X}_{\tilde{p}}, \mathcal{X}_{\tilde{q}}$, to be connected and $\mathcal{X}_{\tilde{p}} \subseteq \mathcal{X}_{\tilde{q}}$. We here only study the stronger condition that $m(x)$ has support $\mathbb{R}^d$ for simplicity.



(a) True (b) FD (c) MFD (d) True (e) FD (f) MFD

Figure 3: Density estimation comparisons with FD and MFD for the energy-based model. The $\mathrm{KL}(p_d||p_\theta)$ evaluations are 3.52/0.22 (b/e) for FD and 0.17/0.01 (c/f) for MFD, lower is better.

$\tilde{q}_{\theta^*}(x) = \tilde{p}_d(x) = \beta p_d(x) + (1-\beta)m(x)$. To obtain a model of the underlying true density $q^* = p_d$, we need to remove the mixture component from $\tilde{q}_{\theta^*}$, which can be done through a 'correction step':

$$q^*(x) = \frac{1}{\beta}\left(\tilde{q}_{\theta^*}(x) - (1-\beta)m(x)\right) = \frac{1}{\beta}\left(\frac{e^{-f_{\theta^*}(x)}}{Z_{\tilde{q}}(\theta^*)} - (1-\beta)m(x)\right). \tag{8}$$

This procedure for obtaining q^* is equivalent to $q^*(x) = \arg\min_q \mathrm{MFD}(p_d(x)||q(x))$ and when $\mathrm{MFD}(p_d(x)||q(x)) = 0$, we have $q^*(x) = p_d(x)$. Therefore, density estimation with MFD in this setting contains three steps: (1) learn θ^* by minimizing Equation 7; (2) estimate $Z_{\tilde{q}}(\theta^*)$; and (3) apply the correction step (Equation 8) to obtain q^*. Compared to FD, the additional correction step has negligible computation cost.

Choice of m and β: As we discussed in Section 3, a good m should have support $\mathbb{R}^d$ and be able to bridge disconnected component distributions. For a given set of data samples $\{x_1, \cdots, x_N\} \sim p_d$, we can simply choose $m(x) = \mathcal{N}(\bar{\mu}, \bar{\Sigma})$, where $\bar{\mu}$ and $\bar{\Sigma}$ are the empirical mean and covariance of the available training data: $\bar{\mu} = \frac{1}{N}\sum_{n=1}^N x_n$, $\bar{\Sigma} = \frac{1}{N}\sum_{n=1}^N x_n x_n^T$, which corresponds to an empirical moment matching approximation of p_d and can thus cover different components. The β is treated as a hyper-parameter in our method. Intuitively, a large beta means that the proportion of data points from p_d is small, and the model is learning m. On the other hand, a small value means we may still have the numerical version of the blindness issue. In this experiment, we use $\beta = 0.8$ can find it can empirically heal blindness. We leave the theoretical study of choosing the β into future work.

Demonstration: We apply the proposed method to train a deep energy-based model and examine the performance against two target densities with multiple isolated components: 1) a weighted mixture of four Gaussians $p_d(x) = 0.1g_1(x) + 0.2g_2(x) + 0.3g_3(x) + 0.4g_4(x)$, where g_1, g_2, g_3, g_4 are 2D Gaussians with identity covariance matrix and mean $[-5,-5], [-5,5], [5,5], [5,-5]$ respectively; and 2) a mixture of 3 concentric circles as proposed in [21]. We use Simpson's rule for the 2D numerical integration to estimate the normalization constant for both methods. The model specifications and training details can be found in Appendix B. In Figure 3 we plot the ground truth and the estimated density with classic FD and the proposed MFD methods. We also provide the corresponding KL evaluation (see Appendix B) between the ground truth density p_d and the estimated model p_θ. We find the proposed MFD method can significantly improve performance and heal the blindness problem.

5 Related Work

In addition to the mixture construction, conducting a Gaussian convolution on both p_d and q_θ can also bridge the disjoint components and defines a valid divergence [22]. However, the score function is generally intractable for a deep energy-based model q, see Appendix D for a detailed discussion.

Paper [16] proposes to add Gaussian noise with variance σ^2 *only* to p_d and anneal $\sigma^2 \to 0$ during training. This helps alleviate the blindness problem in the early stage of training but when $\sigma^2 \approx 0$, the blindness phenomenon will be observed again, see Appendix C for an example.

Paper [7] proposes to transform p and q with a common differentiable invertible function before defining the FD, which is also shown to be equivalent to [2]. However, since the invertible transformation is a homeomorphism and will not change the topology of its domain [4, 23], the invertible transformation will not fix the blindness caused by the disconnected support sets in principle.

The blindness problem also exists in other score-based applications. As discussed in Section 3, directly applying the MFD requires knowing the normalizer, which potentially sheds light on choosing score-based methods. We leave the case-by-case study of how to heal the blindness to future work.

References

[1] A. Anastasiou, A. Barp, F.-X. Briol, B. Ebner, R. E. Gaunt, F. Ghaderinezhad, J. Gorham, A. Gretton, C. Ley, Q. Liu, et al. Stein's method meets statistics: A review of some recent developments. *arXiv preprint arXiv:2105.03481*, 2021.

[2] A. Barp, F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey. Minimum stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32, 2019.

[3] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International conference on machine learning*, pages 2606–2615. PMLR, 2016.

[4] R. Cornish, A. Caterini, G. Deligiannidis, and A. Doucet. Relaxing bijectivity constraints with continuously indexed normalising flows. In *International conference on machine learning*, pages 2133–2143. PMLR, 2020.

[5] F. D'Angelo and V. Fortuin. Annealed stein variational gradient descent. *arXiv preprint arXiv:2101.09815*, 2021.

[6] G. B. Folland. *Advanced calculus*. Pearson, 2001.

[7] W. Gong and Y. Li. Interpreting diffusion score matching using normalizing flow. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, June 2021.

[8] J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019.

[9] A. Hyvärinen. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

[10] A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[11] A. Jolicoeur-Martineau, R. Piché-Taillefer, I. Mitliagkas, and R. T. des Combes. Adversarial score matching and improved sampling for image generation. In *International Conference on Learning Representations*, 2020.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

[14] T. Matsubara, J. Knoblauch, F.-X. Briol, and C. J. Oates. Generalised Bayesian inference for discrete intractable likelihood. *arXiv:2206.08420*, 2022.

[15] P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[16] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

[17] Y. Song and D. P. Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

[18] T. Tao. Analysis ii, texts and readings in mathematics, 2015.

[19] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

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

[21] L. Wenliang, D. J. Sutherland, H. Strathmann, and A. Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, pages 6737–6746. PMLR, 2019.

[22] M. Zhang, P. Hayes, T. Bird, R. Habib, and D. Barber. Spread divergence. In *International Conference on Machine Learning*, pages 11106–11116. PMLR, 2020.

[23] M. Zhang, Y. Sun, S. McDonagh, and C. Zhang. Flow based models for manifold data. *arXiv preprint arXiv:2109.14216*, 2021.

A Derivations and Proofs

A.1 Derivation of Equation 3

Let two differentiable densities g_1 and g_2 have disjoint supports $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$ and

$$p(x) = \alpha_p g_1(x) + (1-\alpha_p) g_2(x), \quad q(x) = \alpha_q g_1(x) + (1-\alpha_q) g_2(x). \tag{9}$$

The FD between p and q can be written as

$$\mathrm{FD}(p||q) = \tfrac{\alpha_p}{2} \int_{\mathcal{X}_1} g_1(x) ||s_p(x) - s_q(x)||_2^2 dx + \tfrac{1-\alpha_p}{2} \int_{\mathcal{X}_2} g_2(x) ||s_p(x) - s_q(x)||_2^2 dx. \tag{10}$$

Since g_1 and g_2 has disjoint support, so g_2 will be a zero function on the support of g_1, so $g_2(x') = \nabla_x g_2(x') = 0$ for $x' \in \mathcal{X}_1$. We then have

$$s_p(x') = \frac{\alpha_p \nabla g_1(x') + \cancel{(1-\alpha_p) \nabla g_2(x')}}{\alpha_p g_1(x') + \cancel{(1-\alpha) g_2(x')}} = \frac{\alpha_p \nabla_x g_1(x')}{\alpha_p g_1(x')} = s_{g_1}(x'), \tag{11}$$

and

$$s_q(x') = \frac{\alpha_q \nabla g_1(x') + \cancel{(1-\alpha_q) \nabla g_2(x')}}{\alpha_q g_1(x') + \cancel{(1-\alpha) g_2(x')}} = \frac{\alpha_q \nabla_x g_1(x')}{\alpha_q g_1(x')} = s_{g_1}(x'), \tag{12}$$

Similarly, for $x' \in \mathcal{X}_2$ we have $s_p(x') = s_q(x') = s_{g_2}(x')$. Therefore, the FD is equivalent to

$$\mathrm{FD}(p||q) = \tfrac{\alpha_p}{2} \int_{\mathcal{X}_1} g_1(x) ||s_{g_1}(x) - s_{g_1}(x)||_2^2 dx + \tfrac{1-\alpha_p}{2} \int_{\mathcal{X}_2} g_2(x) ||s_{g_2}(x) - s_{g_2}(x)||_2^2 dx = 0, \tag{13}$$

which is independent of α_q.

A.2 Proof of Theorem 1

The following two lemmas can be found in Folland [6, Corollary 2.41 and Theorem 2.42]. For completeness, we also provide simplified proofs.

Lemma 4. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is differentiable on an open convex set $\mathcal{X} \subseteq \mathbb{R}^d$ and $\nabla_x f(x) = 0$ for all $x \in \mathcal{X}$, then f is a constant on $\mathcal{X}$.*

Proof. For any two points $x_1, x_2 \in \mathcal{X}$, we denote the the line segment that connects a, b as L_{x_1,x_2}. Since $\mathcal{X}$ is a convex set, then $L_{x_1,x_2} \subseteq \mathcal{X}$. By the Mean Value Theorem (see Folland [6, Theorem 2.39]), there exists a point $x_3 \in L_{x_1,x_2}$ such that $f(x_2) - f(x_1) = \nabla_x f(x_3)(x_2 - x_1)$. Since $x_3 \in S$, so $\nabla_x f(x_3) = 0$ thus $f(x_2) = f(x_1)$. Therefore, f has to be a constant function. □

Lemma 5. *Suppose $f : \mathcal{X} \to \mathbb{R}$ is differentiable on a connected open set $\mathcal{X} \subseteq \mathbb{R}^d$ and $\nabla_x f(x) = 0$ for all $x \in \mathcal{X}$, then f is a constant on $\mathcal{X}$.*

Proof. For any point $a \in \mathcal{X}$, we define $\mathcal{X}_1 = \{x \in \mathcal{X} : f(x) = f(a)\}$ and $\mathcal{X}_2 = \{x \in \mathcal{X} : f(x) \neq f(a)\}$, so $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ by construction. For every $x \in \mathcal{X}_1$, there is a ball $B \in S$ centred at x. Since B is convex, we have $B \in \mathcal{X}_1$ by Lemma 4. Therefore, every point $x \in \mathcal{X}_1$ is an interior point of $\mathcal{X}_1$, so $\mathcal{X}_1$ is an open set. The image of $\mathcal{X}_2$ under f: $\mathbb{R} \setminus \{f(a)\}$ is an open set, so $\mathcal{X}_2$ is a open set since f is a continuous function (see Folland [6, Theorem 1.33]). We thus have both $\mathcal{X}_1$ and $\mathcal{X}_2$ are open sets and $\mathcal{X}_1$ is non-empty (it contains a). Since any connected space cannot be written as an union of two disjoint non-empty sets (see Tao [18, Definition 2.4.1]), so $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ indicates $\mathcal{X}_2 = \emptyset$. Therefore, f is a constant function. □

We can then prove the Theorem 1. For two a.c. distributions that are supported on a connected space $\mathcal{X} \subseteq \mathbb{R}^d$ with differentiable density p and q. Then $\mathrm{FD}(p||q) = 0 \Leftrightarrow \nabla_x \log p(x) = \nabla_x \log q(x)$ for $x \in S$. We define function $f(x) = \log p(x) - \log q(x)$, so $f(x)$ differentiable on $\mathcal{X}$ and $\nabla_x f(x) = 0$. By Lemma 5, we have f as a constant function (we denote as c) so we have $p = q \exp(c)$. Since p and q are densities, we have $\int q(x) \exp(c) dx = 1 \Leftrightarrow c = 0$. Therefore, $\mathrm{FD}(p||q) = 0 \Leftrightarrow p = q$.

A.3 Proof of Theorem 2

Since we can always represents a distribution with disjoint support set as a mixture distribution with components supported on several connected subsets, we can then prove the theorem by Proposition 1.

Proposition 1 (FD is ill-defined on disconnected sets)**.** *Let a set of a.c. distributions have differentiable densities $\{g_1, \cdots, g_K\}$ with mutual disjoint (disconnected) support sets $\{\mathcal{X}_1, \cdots, \mathcal{X}_K\}$: $\mathcal{X}_i \bigcap \mathcal{X}_j = \emptyset$ for any $i \neq j$ and each support $\mathcal{X}_i$ is connected. Let two densities $p = \sum_k \alpha_p^k g_k$ and $q = \sum_k \alpha_q^k g_k$ with positive coefficients $\sum_k \alpha_p^k = 1$ and $\sum_{k=1} \alpha_q^k = 1$. Then $\mathrm{FD}(p||q) = 0 \Leftrightarrow \alpha_p^k = \alpha_q^k e^{c_k}$, where $\{c_1, \cdots, c_K\}$ is a set of constants with constraints $\sum_k e^{c_k} = 1$.*

We can decompose $\mathrm{FD}(p||q) = \frac{1}{2}\sum_{k=1}^K \alpha_p^k \int_{\mathcal{X}_k} g_k(x)||s_p(x) - s_q(x)||_2^2 dx$. Since α_p^k and g_k are positive, $\mathrm{FD}(p||q) = 0 \Rightarrow \int_{\mathcal{X}_k} g_k(x)||s_p(x) - s_q(x)||_2^2 dx = 0$ for any k, so $\nabla_x \log p(x) = \nabla_x \log q(x)$ for $x \in \bigcup_{k=1}^K \mathcal{X}_k$. Since $\mathcal{X}_k$ is connected, by Lemma 5, we have for $x \in \mathcal{X}_k$, $\log p(x) - \log q(x) = c_k \Leftrightarrow p(x) = q(x)e^{c_k} \Leftrightarrow \alpha_p^k g_k(x) = \alpha_q^k g_k(x) e^{c_k} \Leftrightarrow \alpha_p^k = \alpha_q^k e^{c_k}$, where $\{c_1, \cdots, c_K\}$ is a set of constants. Since $\sum_k \alpha_p^k = \sum_k \alpha_q^k e^{c_k} = 1$ and $\sum_k \alpha_q^k = 1$, we then have the constrain $\sum_k e^{c_k} = 1$.

A.4 Derivation of Score Matching

Let p_d and q_θ are differentiable densities with a common support $\mathcal{X} \subseteq \mathbb{R}^d$ and assume q_θ is twice differentiable, we can rewrite the FD as [10]

$$\begin{aligned}
\mathrm{FD}(p_d(x)||q_\theta(x)) &= \tfrac{1}{2}\textstyle\int_{\mathcal{X}} p_d(x)||s_{p_d}(x) - s_{q_\theta(x)}||_2^2 dx && (14)\\
&= \tfrac{1}{2}\textstyle\int_{\mathcal{X}} p_d(x)\left(s_{p_d}^2(x) + s_{q_\theta}^2(x) - 2 s_{p_d}(x) s_{q_\theta}(x)\right) dx && (15)\\
&= \tfrac{1}{2}\textstyle\int_{\mathcal{X}} p_d(x)\left(s_{q_\theta}^2(x) - 2 s_{p_d}(x) s_{q_\theta}(x)\right) dx + const., && (16)
\end{aligned}$$

where the constant terms are independent of the model parameters θ. Using the log-trick, we have

$$\textstyle\int_{\mathcal{X}} p_d(x) s_{p_d}(x) s_{p_\theta}(x) dx = \int_{\mathcal{X}} \nabla_x p_d(x) s_{q_\theta}(x) dx. \quad (17)$$

For simplicity, we assume $\mathcal{X} = \mathbb{R}$ and $p_d(x)s_{p_\theta}(x)$ vanishes at $-\infty$ and ∞, using integration by parts, we have

$$\textstyle\int_{\mathcal{X}} \nabla_x p_d(x) s_{q_\theta}(x) dx = \underbrace{p_d(x) s_{q_\theta}(x)\Big|_{-\infty}^{+\infty}}_{=0} - \int_{\mathcal{X}} p_d(x) \nabla_x s_{q_\theta}(x) dx. \quad (18)$$

In general, this holds for $\mathcal{X} = \mathbb{R}^d$ and $\lim_{||x||\to\infty} p_d(x) s_{q_\theta}(x) = 0$ or $\mathcal{X} \subseteq \mathbb{R}^d$ is a compact subset of $\mathbb{R}^d$ and $f(x)p(x) = 0$ for $x \in \partial\mathcal{X}$ where $\partial\mathcal{X}$ is the piecewise smooth boundary of $\mathcal{X}$ (by the divergence theorem [6, Theorem 5.34]), also see [13] for a similar discussion. Therefore, we have

$$\mathrm{FD}(p_d(x)||q_\theta(x)) = \tfrac{1}{2}\textstyle\int_{\mathcal{X}} p_d(x)\left(s_{q_\theta}^2(x) + 2\nabla_x s_{q_\theta}(x)\right) dx.. \quad (19)$$

A.5 Kernelized Stein Discrepancy Extensions

For two a.c. distributions p and q, the Kernelized Stein Discrepancy [13, 3] can be defined as (see [13, Definition 3.2])

$$\mathrm{KSD}(p||q) = \mathbb{E}_{x,x'\sim p}\left[(s_p(x) - s_q(x))k(x,x')(s_p(x') - s_q(x'))\right], \quad (20)$$

where k is an integrally strictly positive kernel (see [13, Definition 3.1]) and x, x' are *i.i.d.* samples from $p(x)$. The $\mathrm{KSD}(p||q) = 0$ if and only if $s_p = s_q$ (see [13, 3]). Therefore, when p and q are supported on a connected open set, by Lemma 5, we have $\mathrm{KSD}(p||q) = 0 \Leftrightarrow s_p = s_q \Leftrightarrow p = q$. When p and q are supported on a disconnected space, we have $\mathrm{KSD}(p||q) = 0 \not\Rightarrow p = q$. This is because the KSD can be upper bounded by a (positively) scaled FD [13, Theorem 5.1]:

$$|\mathrm{KSD}(p||q)| \leq \sqrt{\mathbb{E}_{x,x'\sim p}[k(x,x')^2]} \times \mathrm{FD}(p||q), \quad (21)$$

we then have $\mathrm{FD}(p||q) = 0 \Rightarrow \mathrm{KSD}(p||q) = 0$. When p and q are supported on a disconnected space, we have $\mathrm{FD}(p||q) = 0 \not\Rightarrow p = q$ (Theorem 2), so $\mathrm{KSD}(p||q) = 0 \not\Rightarrow p = q$.

A.6 Proof of Theorem 3

Since the support of m as $\mathcal{X}_m = \mathbb{R}^d$ then $\tilde{p}$ and $\tilde{q}$ have the same support $\mathcal{X} = \mathbb{R}^d$. For the score functions, we also have

$$\int_{\mathcal{X}} ||s_{\tilde{p}}(x)||_2^2 \tilde{p}(x)dx = \int_{\mathcal{X}} \|\nabla_x \log(\beta p(x) + (1-\beta)m(x))\|_2^2 \, \tilde{p}(x)dx \tag{22}$$

$$= \int_{\mathcal{X}} \left\| \frac{\beta \nabla_x p(x) + (1-\beta)\nabla_x m(x)}{\beta p(x) + (1-\beta)m(x)} \right\|_2^2 \tilde{p}(x)dx \tag{23}$$

$$\leq \int_{\mathcal{X}} \left\| \frac{\beta \nabla_x p(x)}{\beta p(x) + (1-\beta)m(x)} \right\|_2^2 \tilde{p}(x)dx + \int_{\mathcal{X}} \left\| \frac{(1-\beta)\nabla_x m(x)}{\beta p(x) + (1-\beta)m(x)} \right\|_2^2 \tilde{p}(x)dx \tag{24}$$

$$\leq \int_{\mathcal{X}} ||s_p||_2^2 \tilde{p}(x)dx + \int_{\mathcal{X}} ||s_m||_2^2 \tilde{p}(x)dx \leq \int_{\mathcal{X}} ||s_p||_2^2 p(x)dx + \int_{\mathcal{X}} ||s_m||_2^2 p(x)dx < \infty, \tag{25}$$

so $s_{\tilde{p}} \in L^2(\tilde{p})$ and similarly $s_{\tilde{q}} \in L^2(\tilde{p})$. Therefore, the FD between $\tilde{p}$ and $\tilde{q}$ is a valid divergence i.e. $\mathrm{FD}(\tilde{p}||\tilde{q}) = 0 \Leftrightarrow \tilde{p} = \tilde{q} \Leftrightarrow \beta p(x) + (1-\beta)m(x) = \beta q(x) + (1-\beta)m(x) \Leftrightarrow p(x) = q(x)$, thus $\mathrm{MFD}(p||q) = 0 \Leftrightarrow \mathrm{FD}(\tilde{p}||\tilde{q}) = 0 \Leftrightarrow p = q$.

B Experiment Details

For both experiments, we sample 100k data from p_d as our training datasets. The energy network $f_\theta(x)$ is a 3-layer feedforward network with 200 hidden units and swish activation functions [15]. We train the model for 30k iterations with the Adam optimizer [12] and batch-size 300. For the numerical integration we use Simpson's rule provided in the package [19]. We use a Monte-Carlo approximation to estimate the KL divergence evaluations $\widehat{\mathrm{KL}}(p_d(x)||q_\theta(x)) = \frac{1}{K}\sum_{k=1}^{K} \log p_d(x_k) - \log p_\theta(x_k)$, where we use $K = 10000$.

C Data Noise Annealing Doesn't Help

In this section, we empirically show that only adding noise to the data and annealing the noise to 0 during training won't fix the blindness problem in practice. We use a deep energy-based model with a 3-layer feedforward neural network with 30 hidden units and tanh activation function to learn the toy mixture of two Gaussian distributions described in Section 1. We train the model with Adam optimizer with a learning rate $3e^{-4}$ for 10k iterations and batch size 300. We add convolutional Gaussian noise to the data samples with a standard deviation of 3.0 and anneal to 0 by multiplying by 0.9999 at each iteration. The noise at the end of training has a standard deviation less than 0.001. In Figure 4 we plot the learned density during training. We find that when the noise is big the model can identify the correct mixture co-efficient, but when the noise is close to 0, the model fails to capture the correct mixing proportions. We also plot the density estimation results with vanilla FD and the proposed MFD in Figure 5a and 5b and we find that the density estimation with MFD achieves the best performance.

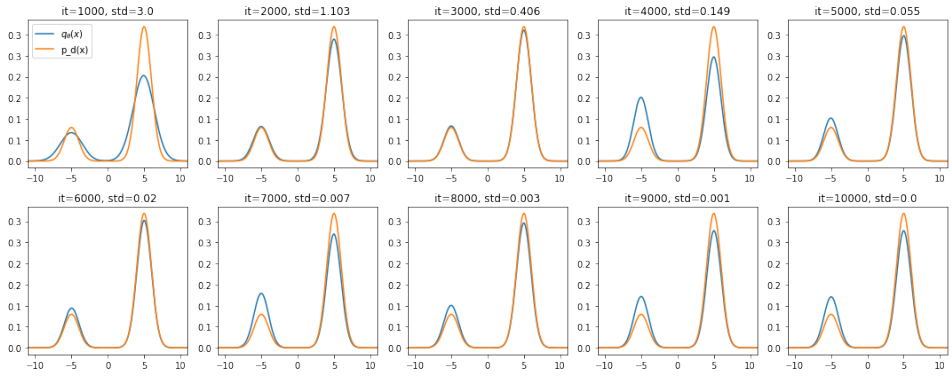
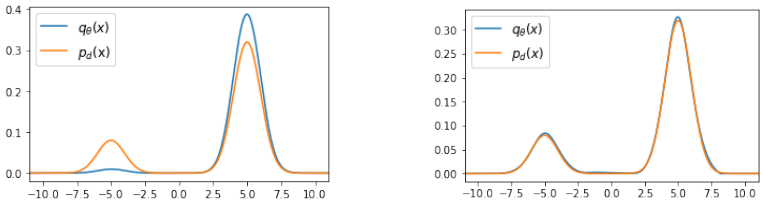


Figure 4: FD with training data noise annealing



(a) FD (b) MFD

Figure 5: Density Estimation with FD and MFD.

D Spread Fisher Divergence

For two distributions with densities $p_d(x)$ and $q_\theta(x)$ with supports $\mathcal{X}_{p_d}, \mathcal{X}_{q_\theta} \subseteq \mathbb{R}^d$, we can choose $k(\tilde{x}|x) = \mathrm{N}(x, \sigma^2)$ and let

$$\tilde{p}_d(\tilde{x}) = \int_{\mathcal{X}_{p_d}} k(\tilde{x}|x) p_d(x) dx \quad \tilde{q}_\theta(\tilde{x}) = \int_{\mathcal{X}_{q_\theta}} k(\tilde{x}|x) q_\theta(x) dx \tag{26}$$

We follow the spread f-divergence [22] and define the *Spread Fisher Divergence* ($\widetilde{\mathrm{FD}}$) as

$$\widetilde{\mathrm{FD}}_k(p_d||q_\theta) \equiv \mathrm{FD}(\tilde{p}_d||\tilde{q}_\theta), \tag{27}$$

The convolution transform makes $\tilde{p}_d$ and $\tilde{q}_\theta$ have support $X_{\tilde{p}_d} = X_{\tilde{q}_\theta} = \mathbb{R}^d$ (which is a connected space) and $\widetilde{\mathrm{FD}}_k(p_d||q_\theta) \equiv \mathrm{FD}(\tilde{p}_d||\tilde{q}_\theta)$ is a valid discrepancy, i.e. $\widetilde{\mathrm{FD}}_k(p_d||q_\theta) = 0 \Leftrightarrow \tilde{p}_d = \tilde{q}_\theta \Leftrightarrow p_d = q_\theta$. The spread Fisher divergence is also well-defined for the singular distributions (distributions that are not a.c. w.r.t Lebesgue measure), see [22] for a detailed discussion.

Similar to the FD, we can rewrite the $\widetilde{\mathrm{FD}}$ as

$$\widetilde{\mathrm{FD}}_k(p_d||p_\theta) = \tfrac{1}{2} \int_{\mathbb{R}^d} \tilde{p}_d(\tilde{x}) \left\| s_{\tilde{p}_d}(\tilde{x}) - s_{\tilde{q}_\theta}(\tilde{x}) \right\|_2^2 d\tilde{x} \tag{28}$$

$$= \tfrac{1}{2} \int_{\mathbb{R}^d} \tilde{p}_d(\tilde{x}) \left(s_{\tilde{q}_\theta}^2(\tilde{x}) + 2\nabla_{\tilde{x}} s_{\tilde{q}_\theta}(\tilde{x}) \right) d\tilde{x} + const., \tag{29}$$

where the constant terms are independent of the model parameters. For an energy-based model $q_\theta(x) = e^{-f_\theta(x)}/Z(\theta)$, the spread model $\tilde{q}_\theta(\tilde{x}) = \frac{1}{Z(\theta)\sqrt{2\pi\sigma^2}} \int e^{-f_\theta(x) - \frac{1}{2\sigma^2}(\tilde{x}-x)^2} dx$ has an intractable score. Additionally, unlike the mixture construction, if we directly assume $\tilde{q}_\theta(\tilde{x}) = e^{-f_\theta(\tilde{x})}/Z(\theta)$, the underlying 'correct' model $q_\theta(x)$ can not be recovered from $\tilde{q}_\theta(\tilde{x})$ even if we know $Z(\theta)$. Therefore, the $\widetilde{\mathrm{FD}}$ is not directly applicable in this case.