Blindness of score-based methods to isolated components and mixing proportions

Li K. Wenliang  Heishiro Kanagawa
Gatsby Computational Neuroscience Unit
University College London
London, UK
kevinwli@outlook.com  heishiro.kanagawa@gmail.com

Abstract

Statistical tasks such as density estimation and approximate Bayesian inference often involve densities with unknown normalising constants. Score-based methods, including score matching, are popular techniques as they are free of normalising constants. Although these methods enjoy theoretical guarantees, a little-known fact is that they exhibit practical failure modes when the unnormalised distribution of interest has isolated components — they cannot discover isolated components or identify the correct mixing proportions between components. We demonstrate these findings using simple distributions and present heuristic attempts to address these issues. We hope to bring the attention of theoreticians and practitioners to these issues when developing new algorithms and applications.

1 Introduction and background

This paper presents a pervasive practical issue of score-based methods. The (Hyvärinen) score function of a differentiable probability density \( p(x) \) is defined by \( s_p(x) := \nabla_x p(x)/p(x) \). Score function does not depend on the normaliser and, therefore, has a broad range of applications in machine learning and Bayesian statistics. Chief among these are the following: (a) training unnormalised density models with score matching (SM) [1], (b) measuring the quality of approximate samplers using Stein discrepancies (SDs) [3, 4, 5, 7, 14], and (c) approximate posterior sampling via Stein variational gradient descent (SVGD) [6]. We show that these theoretically well-motivated score-based algorithms can fail in practice when the unnormalised distribution has isolated components.

In what follows, we exemplify the common failure modes with the following simple setup:

**Example 1 (Gaussian mixtures).** Define the following density functions on \( \mathbb{R} \):

\[
p(x) = \pi_1 p_1(x) + (1 - \pi_1) p_2(x), \quad q(x) = p_1(x), \quad p_1(x) = \mathcal{N}(x; -\mu, \sigma^2), \quad p_2(x) = \mathcal{N}(x; \mu, \sigma^2).
\]

where \( \mu, \sigma > 0 \), and \( \pi_1 \in (0, 1) \) are mixing proportions. In addition, we define \( p'(x) := \pi'_1 p_1(x) + (1 - \pi'_1) p_2(x) \) as the same mixture as \( p(x) \) but with a different mixing proportion \( \pi'_1 \neq \pi_1 \). Instances of these densities are shown in Figure 1. When \( \mu/\sigma^2 \) is large, the components of \( p \) are isolated.

The aforementioned failure modes stem from the following Lemma concerning the distributions in Example 1, the proof of which can be found in Appendix A.1.

**Lemma 1** (Weak dependence of \( s_p \) on \( \pi_1 \)). For the densities \( p \) and \( q \) defined in Example 1, \( s_p(x) \) gets arbitrarily close to \( s_{p_1}(x) = s_q(x) \) regardless of \( \pi_1 \) for \( x \neq 0 \) when \( \mu/\sigma^2 \) gets large.

This property is illustrated in Figure 1 (top) — the score \( s_p(x) \) does not change visibly with \( \pi_1 \). In the following sections, we discuss the consequences of Lemma 1 for the three applications introduced above. To our knowledge, there has not been a synthesised exposition of the common failure modes, except for references [9, 13, 16, 15, 14, 17, 18] which on specific algorithms or other issues. We also propose heuristic remedies to initiate an effort to rectify these issues.

NeurIPS 2021 Workshop Your Model Is Wrong: Robustness and Misspecification in Probabilistic Modeling.
2 Fisher divergence and Stein discrepancy

Consider training an unnormalised density model \( \hat{p}(x) \) with score \( s_p \) on data drawn from \( q(x) \). Hyvärinen [1] proposed SM to train \( \hat{p} \) by minimising the Fisher divergence (FD)

\[
J[q||p] = \frac{1}{2} \int q(x) \| s_q(x) - s_p(x) \|_2^2 \, dx.
\]

The FD is zero if and only if \( p = q \). The densities \( p \) and \( q \), however, can still be “very different” when their FD is close to but not exactly zero, as we show below (see also Appendix A.2).

**Proposition 1** (FD is blind to isolated components). For \( q \) and \( p \) in Example 1, the FD \( J[q||p] \to 0 \) regardless of the mixing proportion \( \pi_1 \) when \( \mu/\sigma^2 \) gets large.

**Proof sketch.** The FD \( J[q||p] \) is an expectation under \( q \) which has almost all of its mass in \( x < 0 \) as \( \mu/\sigma^2 \) gets large. Lemma 1 implies that \( s_p(x) \to s_{p_1}(x) = s_q(x) \) for \( x < 0 \). Thus, \( J[q||p] \to 0 \).

Another issue arises when two mixtures \( p' \) and \( p \) comprise the same set of components weighted by different mixing proportions. In this case, their FD is almost zero despite the large difference in term of probability mass.

**Proposition 2** (FD is blind to \( \pi \)). For \( p(x) \) and \( p'(x) \) defined by distinct choices of the mixing proportion in Example 1, the FD \( J[p'||p] \to 0 \) as \( \mu/\sigma^2 \) gets large regardless of the mixing proportion.

**Proof sketch.** By Lemma 1, the scores \( s_p \) and \( s_{p'} \) converge to the same limit for \( x \neq 0 \). They differ substantially only for \( x \) close to 0 where \( p' \) puts vanishing mass, so \( J[p'||p] \to 0 \).

In density estimation where \( p \) is the model, Proposition 1 implies that the model can have a mixture component far away from the data \( q \). According to Proposition 2, when isolated components in a data distribution \( p' \) are well-fit individually by the model \( p \), one can obtain another model with small FD by varying the mixing proportions arbitrarily. See Figure 1 (bottom row) for visualisations. We discuss in Section D.1 how previous successes of SM avoided these issues.

Next, we discuss the score-based Stein discrepancies (SDs) [3, 4, 5, 14] which can measure how well samples from \( q \) agree with model \( \hat{p} \). A (Langevin) SD between \( q \) and \( p \) is defined as

\[
SD_F[|q||p] = \sup_{f \in F} \mathbb{E}_{x \sim q} \left[ s_p(x)^T f(x) + \nabla_x^T f(x) \right],
\]

where \( \nabla_x^T \) is the divergence operator, and \( F \) is a class of differentiable vector-valued functions with appropriate boundary conditions (see the foregoing references for precise definitions). Since the SD is upper bounded by the FD (see Appendix A.3), we have the following:

**Proposition 3** (Blindnesses of SD). For \( f \in F \) such that \( \int \| f(x) \|^2_2 q(x) \, dx \leq 1 \) and \( q(x) f(x) \to 0 \) as \( \| x \|_2 \to \infty \), \( SD_F[q||p] \) and \( SD_F[p'||p] \) suffer from the issues of \( J[q||p] \) and \( J[p'||p] \) in Propositions 1 and 2, respectively.

In the case of the kernel SD (KSD), where \( F \) is the unit ball of a reproducing kernel Hilbert space (RKHS) [4, 5], we show in Figure 5 that the best \( f \in F \) witnessing the difference between \( p \) and \( q \) is almost zero around \( x = 0 \). Therefore, diagnostics based on KSD can be misleading. We further discuss this issue in Section D.2 by relating to KSD bounds on integral probability metrics.
Heuristic remedy for KSD: matching a noisy $q$ and a tempered $p$ To better detect isolated components, consider adding noise to $q$ and changing the temperature of $p$: $q_ε(x) := \int q(x')N(x′, \epsilon^2)dx'$ where $\epsilon > 0$, $\tilde{p}^\beta(x) \propto (p(x))^\beta$ where $\beta \in (0, 1]$.

The KSD between $q_ε$ and $\tilde{p}^\beta$ is given in closed-form [4, 5]:

$$KSD^2_{k, ε}(q_ε \parallel p^\beta) = \beta^2 E_{x,x′ \sim q_ε}[s_p(x)^T s_p(x′)k(x,x′)] + \beta E_{x,x′ \sim q_ε}[s_p(x)^T \nabla_x k(x,x′)]$$

$$+ E_{x,x′ \sim q_ε}[\text{tr}[\nabla_x \nabla_x k(x,x′)]]$$

with $\text{tr}$ the matrix trace. Adding noise to $q$ will likely increase this KSD, but we can adjust the temperature in $p$ to compensate for this effect, since both transformations “broaden” the original densities. By noting that (2) is convex and quadratic in $\beta$ for positive-definite $k$, we take its unique infimum over $\beta$ to counter the noise-induced mismatch, giving

$$KSD^2_{k,ε}[q_ε \parallel \tilde{p}^\beta] := \arg\min_{\beta \in (0,1]} KSD^2_{k,ε}(q_ε \parallel \tilde{p}^\beta).$$  (3)

For the densities defined in Example 1, we visualise $KSD^2_{k,ε}[q_ε \parallel \tilde{p}^\beta]$ as a function of $\epsilon$ and compare it with $KSD^2_{k,ε}(q_ε \parallel p)$ without temperature matching in Figure 2 (top). For some values of $\epsilon$ $KSD^2_{k,ε}[q_ε \parallel \tilde{p}^\beta]$ is significantly greater than the baseline $KSD^2_{k,ε}(q_ε \parallel p)$, but this is not the case for $KSD^2_{k,ε}[q_ε \parallel \tilde{p}^\beta]$ (middle), suggesting the importance of matching the temperature with noise. Second, $KSD^2_{k,ε}$ reaches a maximum at some $\epsilon$. We define this maximum as the temperature-matched KSD

$$\text{tmKSD}^2_{k,ε}[q_ε \parallel \tilde{p}^\beta] := \max_\epsilon KSD^2_{k,ε}[q_ε \parallel \tilde{p}^\beta].$$

Note that this is not a valid discrepancy (see Appendix B), but it is more sensitive to isolated components than KSD. We show these SDs between a fixed $p'$ and a few $p$ with various $π_1$ in Figure 2 (bottom). Although $\text{tmKSD}^2_{k,ε}[p' \parallel p]$ still cannot reliably identify the correct mixing proportion 0.5, its smaller estimation errors and stronger dependence on $π_1$ compared to KSD are encouraging. We show very similar results for non-Gaussian distributions in Appendix B.

3 Stein variational gradient descent

SVGD [6] approximates an unnormalised distribution $p$ by iteratively updating an empirical distribution $ν_t$ formed by particles. The main idea is to find a direction $φ$ such that the particle update $x ← x + εφ(x)$ lowers the Kullback-Leibler divergence $KL[ν_t \parallel p]$. For $φ$ defined by a function in the RKHS associated with a kernel $k(·, ·)$, the optimal $φ^*$ for a particle $x$ is given by

$$φ^*(x′) = E_{x \sim p}[s_p(x)k(x,x′) + \nabla_x k(x,x′)].$$  (4)

The particles can be initialised as samples drawn from a simple distribution $ν_0$, such as a Gaussian with mean $m_0$ and variance $σ^2_0$. 

Figure 2: Top: densities (left) in Example 1 and their squared KSD (middle) and tmKSD (right) estimated with 500 samples. Bottom: the squared KSD (middle) and tmKSD (right) between $p'$ and a few $p$ with various $π_1$ estimated with 2000 samples. Errorbar shows 1 s.e. estimated by Jackknife.
We stress that the practical failure modes presented here do not diminish, in any way, the theoretical advances of score-based methods; these methods have empowered practitioners to tackle a variety of statistical problems involving intractable distributions with unknown normalisers. Further, the issues discussed here may or may not impact certain downstream applications. For example, the model estimated by SM may still be suitable for local gradient-based methods, such as denoising. In contrast, unconditioned gradient-based sampling over the whole support may suffer from these issues. When using SVGD for Bayesian neural networks, ignoring the trivial posterior components arising from the symmetry of the weights does not affect the predictive distribution. We discuss other score-based methods that do not suffer from these issues in Appendix D.4.

4 Discussion

We have demonstrated that three popular score-based methods fail to detect the isolated components or to identify the correct mixing proportions. The heuristic remedies presented here encourage more principled solutions. We stress that the practical failure modes presented here do not diminish, in any way, the theoretical advances of score-based methods; these methods have empowered practitioners to tackle a variety of statistical problems involving intractable distributions with unknown normalisers. Further, the issues discussed here may or may not impact certain downstream applications. For example, the model estimated by SM may still be suitable for local gradient-based methods, such as denoising. In contrast, unconditioned gradient-based sampling over the whole support may suffer from these issues. When using SVGD for Bayesian neural networks, ignoring the trivial posterior components arising from the symmetry of the weights does not affect the predictive distribution. We discuss other score-based methods that do not suffer from these issues in Appendix D.4.

**Proposition 4** (Blindness of SVGD). For the mixture \( \nu_t = q \) or \( \nu_t = p' \), one has that \( \| \phi^\ast(x') \|^2 \to 0 \) as \( \mu/\sigma^2 \) gets large.

*Proof sketch.* By the reproducing formula [2, Sec. 4.2], \( \| \phi^\ast(x') \|^2 \leq \sqrt{k(x', x')} \text{SD}_\mathcal{F}[\nu_t\|\mu] \), with \( \mathcal{F} \) the unit ball of the RKHS. The upper bound goes to zero as \( \mu/\sigma^2 \) gets large by Proposition 3.

This means that the particles get stuck at \( q \) or \( p' \), missing a component in \( p \) completely or giving a wrong mixing proportions. We verify this empirically with results shown in Figure 3. The final \( \nu_t \) is highly sensitive to the initial \( \nu_0 \); contrary to intuitions, an overdispersed \( \nu_0 \) alone does not help.

**Heuristic remedy: combine SVGD and Langevin sampling**  
Langevin sampling (LS) targets the same stationary distribution as SVGD while being more exploratory. A heuristic strategy is thus to update the particles according to a combination of LS and SVGD:

\[
x' \leftarrow x' + \epsilon_1 \mathbf{E}_{\nu_t} \left( s_p(x)k(x, x') + \nabla_x k(x, x') \right) + \epsilon_2 s_p(x') + \sqrt{\epsilon_2} \omega, \quad \epsilon_1, \epsilon_2 \geq 0, \quad \epsilon_1 \epsilon_2 \neq 0,
\]

where \( \omega \) is the standard normal. We let \( \epsilon_2 \) decrease gradually while keeping \( \epsilon_1 \) fixed. This is similar to SPOS [17] which reduces both \( \epsilon_1 \) and \( \epsilon_2 \), so we refer to our heuristic by SPOS-dn (decreasing the noisy LS step). We ran LS, SVGD, SPOS and SPOS-dn to sample \( p \) in Example 1 and estimated the final mixing proportions. Figure 4 shows that SPOS and SPOS-dn are better than LS and SVGD, and SPOS-dn converges the fastest. Nonetheless, finding the correct \( \pi \) is still challenging for all algorithms tested. Details and additional results showing robustness to \( \nu_0 \) are in Appendix C.

Another approach proposed by D’Angelo and Fortuin [18] is to run SVGD while annealing \( p \). However, unlike adding noise, annealing does not preserve the masses of isolated components (see Appendix D.3). Thus, this method still produces wrong mixing proportions [18, Fig. 4].
Acknowledgement  We thank Maneesh Sahani and Arthur Gretton for useful discussions. This research is funded by the Gatsby Charitable Foundation.

References

[1] Aapo Hyvärinen. “Estimation of non-normalized statistical models by score matching”. In: Journal of Machine Learning Research 6.Apr (2005), pp. 695–709.
[2] Ingo Steinwart and Andreas Christmann. Support vector machines. Information Science and Statistics. Springer, New York, 2008.
[3] Jackson Gorham and Lester Mackey. “Measuring sample quality with Stein’s method”. In: Advances in Neural Information Processing Systems. 2015.
[4] Kacper Chwialkowski, Heiko Strathmann and Arthur Gretton. “A Kernel Test of Goodness of Fit”. In: ICML. 2016.
[5] Qiang Liu, Jason D. Lee and Michael Jordan. “A Kernelized Stein Discrepancy for Goodness-of-fit Tests”. In: ICML. 2016.
[6] Qiang Liu and Dilin Wang. “Stein variational gradient descent: A general purpose bayesian inference algorithm”. In: Advances in neural information processing systems. 2016.
[7] Jackson Gorham and Lester Mackey. “Measuring sample quality with kernels”. In: International Conference on Machine Learning. PMLR. 2017.
[8] Michael Arbel and Arthur Gretton. “Kernel Conditional Exponential Family”. In: AISTATS. 2018.
[9] Michael Arbel, D. J. Sutherland, Mikolaj Binkowski and Arthur Gretton. “On gradient regularizers for MMD GANs”. In: Advances in Neural Information Processing Systems. 2018.
[10] Murat A. Erdogdu, Lester Mackey and Ohad Shamir. “Global Non-convex Optimization with Discretized Diffusions”. In: Advances in Neural Information Processing Systems. 2018.
[11] Yingzhen Li and Richard E. Turner. “Gradient Estimators for Implicit Models”. In: International Conference on Learning Representations. 2018.
[12] Jiaxin Shi, Shengyang Sun and Jun Zhu. “A Spectral Approach to Gradient Estimation for Implicit Distributions”. In: International Conference on Machine Learning. 2018.
[13] Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen and Bo Zhang. “Message passing Stein variational gradient descent”. In: International Conference on Machine Learning. PMLR. 2018.
[14] Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer and Lester Mackey. “Measuring sample quality with diffusions”. In: The Annals of Applied Probability 29.5 (2019).
[15] Yang Song and Stefano Ermon. “Generative modeling by estimating gradients of the data distribution”. In: Advances in Neural Information Processing Systems. 2019.
[16] Li K. Wenliang, D. J. Sutherland, Heiko Strathmann and Arthur Gretton. “Learning deep kernels for exponential family densities”. In: ICML. 2019.
[17] Jianyi Zhang, Ruiyi Zhang, Lawrence Carin and Changyou Chen. “Stochastic particle-optimization sampling and the non-asymptotic convergence theory”. In: International Conference on Artificial Intelligence and Statistics. PMLR. 2020.
[18] Francesco D’Angelo and Vincent Fortuin. “Annealed stein variational gradient descent”. In: 2021.
A Proof

For notational simplicity we define $\pi_2 := 1 - \pi_1$.

A.1 Proof of Lemma 1

The score function is

$$\nabla_x \log p(x) = \frac{\pi_1 \nabla_x p_1(x) + \pi_2 \nabla_x p_2(x)}{\pi_1 p_1(x) + \pi_2 p_2(x)}$$

$$= -\pi_1 \exp \left[ -\frac{(x+\mu)^2}{2\sigma^2} + \pi_2 \exp \left[-\frac{(x-\mu)^2}{2\sigma^2}\right]\right]$$

$$= -\frac{\pi_1 (x+\mu) + \pi_2 \exp \left[\frac{(x+\mu)^2}{2\sigma^2} - \frac{(x-\mu)^2}{2\sigma^2}\right]}{\pi_1 + \pi_2 \exp \left[\frac{(x+\mu)^2}{2\sigma^2} - \frac{(x-\mu)^2}{2\sigma^2}\right]}$$

$$= -\frac{\pi_1 (x+\mu) + \pi_2 \exp \left[\frac{2 \mu}{\sigma^2} x - \frac{(x-\mu)^2}{2\sigma^2}\right]}{\pi_1 + \pi_2 \exp \left[\frac{2 \mu}{\sigma^2} x - \frac{(x-\mu)^2}{2\sigma^2}\right]}$$

$$= -\frac{x}{\sigma^2} + \frac{\pi_1 - \pi_2 \exp \left[\frac{2 \mu}{\sigma^2} x - \frac{(x-\mu)^2}{2\sigma^2}\right]}{\pi_1 + \pi_2 \exp \left[\frac{2 \mu}{\sigma^2} x - \frac{(x-\mu)^2}{2\sigma^2}\right]} \mu$$

In the limit of $\mu, \sigma^2$ such that $\mu/\sigma^2 \to \infty$, we can see that

$$\left| \nabla_x \log p(x) - \left( -\frac{x + \mu}{\sigma^2} \right) \right| \to 0 \text{ for } x < 0;$$

$$\left| \nabla_x \log p(x) - \left( -\frac{x - \mu}{\sigma^2} \right) \right| \to 0 \text{ for } x > 0.$$

The two limits can be identified with $\nabla_x \log p_1(x) = -(x + \mu)/\sigma^2$ for $x < 0$ and $\nabla_x \log p_2(x) = -(x - \mu)/\sigma^2$ for $x > 0$.

Remark 1. A similar result holds for an arbitrary number of components without the Gaussian assumption. Consider a mixture of $K$ components with conditional likelihoods $p(x|z = k)$ for $k \in \{1, \ldots, K\}$ that may differ across components. We say that the components are isolated if $p(z|x)$ is concentrated on a single component for all data points. In this case, we can write $s_p(x)$ as

$$s_p(x) = \frac{\nabla_x \int p(x|z)p(z)dz}{p(x)} = \int \nabla_x \log p(x|z)p(x|z)p(z)dz = \sum_{k=1}^{K} p(z = k|x) \nabla_x \log p(x|z = k).$$

For a given $x$, if the posterior is concentrated on $z = k$, then it is clear that $s_p$ is approximately equal to $\nabla_x \log p(x|z = k)$, the score of the $k$th component.

A.2 Proof of Proposition 1

We split the integral in the definition of the Fisher divergence into the positive and negative parts. For the positive part, by the definition of $q$, we have

$$\int_0^\infty q(x)(s_q(x) - s_p(x))^2 dx = \int_0^\infty q(x)(s_p(x) - s_{p_1}(x))^2 dx.$$

From the proof of Lemma 1, one can check that

$$\int q(x)(s_p(x) - s_{p_1}(x))^2 dx = 4 \int q(x) \left( \frac{\pi_2 \mu/\sigma^2}{\pi_1 \exp \left[2 \mu/\sigma^2\right] + \pi_2} \right)^2 dx.$$
Since $\mu/\sigma^2 > 0$, for each point $x \in (0, \infty)$, the integrand converges to 0 as $\mu/\sigma^2 \to \infty$ and is bounded as
\[
\left( \frac{\pi_2 \mu/\sigma^2}{\pi_1 \exp \left[ 2x \pi_2/\sigma^2 \right] + \pi_2} \right)^2 \leq \frac{1}{2x} \frac{W(\pi_2/\pi_1 \cdot e^{-1}) + 1}{W(\pi_2/\pi_1 \cdot e^{-1}) + 2x + 1},
\]
with $W$ the Lambert W function. The upper bound is integrable with respect to the distribution $q$. Thus, by the dominated convergence theorem, we have
\[
\int_0^\infty q(x) (s_p(x) - s_{q}(x))^2 \, dx \to 0 \text{ as } \frac{\mu}{\sigma^2} \to \infty.
\]
The same conclusion can be similarly shown for the negative part, and therefore we have $J[q||p] \to 0$ regardless of the mixing proportion $\pi_1$ when $\mu/\sigma^2 \to \infty$.

### A.3 Proof of Proposition 3

Under the stated class $\mathcal{F}$, observe that
\[
S_F[q||p] = \sup_{f \in \mathcal{F}} |E_q[s_p(x)^T f(x) + \nabla_x f(x)]|
= \sup_{f \in \mathcal{F}} |E_q[(s_p(x) - s_q(x))^T f(x)]|
\leq \sup_{\{f : \|f(x)\|_q(x)dx \leq 1\}} |E_q[(s_p(x) - s_q(x))^T f(x)]|
\leq \sqrt{\int \|s_p(x) - s_q(x)\|^2 q(x)dx} \sup_{\{f : \|f(x)\|_q(x)dx \leq 1\}} \sqrt{\int \|f(x)\|^2 q(x)dx}
\leq \sqrt{2J[q||p]}.
\]
The second line is by integration by parts, the third line is from the integral assumption on $\mathcal{F}$, and the fourth line follows from the Cauchy-Schwartz inequality. As $J[q||p]$ tends to zero as $\mu/\sigma^2 \to \infty$, so does the lower bound $S_F[q||p]$.

To provide more intuition, we computed the optimal $f$ (witness function) for densities in Example 1 when $\mathcal{F}$ is given by the RKHS associated with a kernel $k$ (KSD)
\[
E_q[s_p(x)^T k(x, \cdot) + \nabla_x^T k(x, \cdot)] = E_q[(s_p(x) - s_q(x))^T k(x, \cdot)].
\]
We repeated this with both the Gaussian and IMQ kernels [7] with various bandwidths (0.5, 1.0, 2.0, 5.0 and 10.0). The results are shown in Figure 5 in Figure 5 for $q$ and $p$ and Figure 6 for $p$ and $p'$ with in Figure 6. For all kernels and bandwidths, the witness functions are almost zero.

![Figure 5](image)

Figure 5: Top row, densities $p$ (blue) and $q$ (orange). Middle and bottom rows show witness functions for KSD $[q||p]$, given by Gaussian and IMQ kernels, respectively. Darker colour indicates wider bandwidths of the kernels.
A.4 Proof of Proposition 4

For $x \in \mathbb{R}^d$ and a reproducing kernel $k$ with associated RKHS $\mathcal{H}$, define

$$\xi_i(x, \cdot) := \mathbb{E}_{x \sim \nu_t} [s_{p,i}(x)k(x, \cdot) + \nabla_x k(x, \cdot)], \quad i \in \{1, \ldots, d\}$$

By the reproducing property and Cauchy-Schwarz, we have

$$\|\phi^*(x')\|^2_2 = \sum_{i=1}^d \xi_i^2(x, x') = \sum_{i=1}^d (k(x', \cdot), \xi_i(x, \cdot))^2_{\mathcal{H}}$$

$$\leq \|k(x', \cdot)\|^2_{\mathcal{H}} \sum_{i=1}^d \|\xi_i(x, \cdot)\|^2_{\mathcal{H}} = k(x', x')KSD^2(\nu_t || p),$$

where we followed the definition of $\mathcal{H}^d$ and KSD in [4].

B Temperature-matched Kernel Stein Discrepancy

B.1 Relative magnitude of tmKSD

In the example given in Figure 2, we obtained a zero tmKSD between a Gaussian $q$ and itself. This is because adding zero-mean Gaussian noise to and changing the temperature of a Gaussian $q$ both yield another Gaussian, so it is always possible to find a value of $\beta$ such that $q_\beta = q$, giving a zero tmKSD as desired. If two Gaussian distributions $q$ and $p$ differ only in their variance, then KSD can easily detect the difference, while tmKSD cannot. Thus, tmKSD is better used to find specifically for isolated components after the usual KSD test.

More generally, when distributions are not restricted to Gaussians, then adding noise to and changing the temperature of the same distribution may result in nonzero KSDs. Thus, tmKSD is not a proper metric, and the absolute magnitude may not be indicative of goodness-of-fit. However, we see empirically that $KSD^2_{\nu_t, q} || q$ is lower than $KSD^2_{\nu_t, q} || p$ when $q \neq p$, which suggests that tmKSD may be used as a relative measure.

B.2 Experiments on other mixture distributions

To further validate the proposed tmKSD, we ran additional experiments on Laplace and Student-t distributions, which have heavier tails and more dispersed samples. The same experimental procedures as Figure 2 are used here. The results for Laplace distributions are shown in Figure 7, and those for Student-t (d.o.f 5.0) in Figure 8. They are largely consistent with the results on the Gaussian distributions. Note that, for these distributions, the squared tmKSD between $q$ and $p$ is never smaller than that between $q$ and itself.
Figure 7: Same as Figure 2 but for Laplace distribution and its mixtures.

Figure 8: Same as Figure 2 but for student-T distribution and its mixtures.

C Combining SVGD with LS

Empirically, we found that SVGD gave good solutions even with $\epsilon_1$ fixed at 1.0, so $\epsilon_1$ does not need to be decreased. Intuitively, LS contributes by mixing the initial particles to explore for isolated components, and SVGD “fine-tunes” the final particles thanks to the coupling between different particles.

We report the procedure and hyperparameters used for these experiments. For SVGD, we used $\epsilon = 1.0$. For LS, we reduced the step size linearly from 1.0 to 0.0 at steps of 0.01. For the original SPOS, we reduced both $\epsilon_1$ and $\epsilon_2$ using the same linear schedule. For our SPOS-dn, we applied this linear schedule only to $\epsilon_2$ while keeping $\epsilon_1$ fixed at 1.0. All kernels used are squared-exponential with unit bandwidth. For each mixing proportion, we repeated each algorithm 20 times with different random seeds. To evaluate the effective mixing proportion of the final particles, we calculated the fraction of samples below 0.0 as $\hat{\pi}_1$ and report the error

$$\Delta \pi_1 := |\hat{\pi}_1 - \pi_1|$$

where $\pi_1$ is the true mixing proportion in $p$.

The results in Figure 4 were obtained when the initial distribution is $\nu_0 = \mathcal{N}(0, 1)$. We ran more simulations with $\nu_0$ sampled from $\mathcal{N}(-10.0, 1)$ or $\mathcal{N}(10.0, 1)$ and report all results in Figure 9. In all settings, the proposed SPOS-dn gave the best estimated mixing proportions.
Figure 9: The trajectory of $\Delta \pi_1$ in (6) for SVGD, LS, SPOS and the proposed SPOS-dn to approximate the density $p$ in Example 1. Three rows are results from initial particles distributed as Gaussians with the indicated means and unit variance. Lines are mean ± se estimated from 20 independent runs.

D Further discussions

D.1 Previous successes with score matching

Previous successes on energy-based models required additional constraints preprocessing. To build a full probabilistic density model that supports all downstream statistical applications (e.g. density estimation, empirical Bayes, parameter interpretation, etc.), the issue of Proposition 1 can be partially alleviated by controlling the tail behaviour of $p$ [8, 16], although the notion of tails in a mixture distribution with isolated components may be harder to define. The issue of Proposition 2 can be partially addressed by fitting to each component after clustering the data [16]. Song and Ermon [15] initiated a novel approach to training a sequence of score functions for sample generation, which gave very impressive results. However, this approach is not yet a generic learning algorithm for any given energy-based model for full downstream applications. More explicitly, the method of training a sequence of score functions is at odds with training a single energy-based model with a fixed architecture.

D.2 Isolated modes and Stein discrepancy bounds on integral probability metrics

Diffusion-based Stein discrepancies are known to upper-bound integral probability metrics (IPMs) such as the $L^1$-Wasserstein distance [3, 14] or the Dudley metric [7]. The key assumption in those results is that the diffusion has a fast Wasserstein decay rate as detailed in Section 2.2 of [14]; dissipativity conditions are sufficient for this requirement [14, Section 3]. A Gaussian mixture with a fixed shared variance satisfies the distant dissipativity condition. The $L^1$-Wasserstein rate of a diffusion targeting the distribution, however, can be slow, as shown in Proposition 3.4 of [10]; the rate has a factor exponential in the maximum distance between modes. Therefore, constants (known as Stein factors) appearing in the upper-bounds in the aforementioned papers can be large, and thus a small Stein discrepancy value might not imply the closeness in the IPM. This observation reflects the blindness of Stein discrepancies Proposition 3 – two Gaussian mixtures with largely different mixture weights should have different means and hence a large value of the $L^1$-Wasserstein distance.

D.3 Annealing does not preserve probability mass

D’Angelo and Fortuin [18] recently introduced this idea to SVGD and produced better samples. However, the mixing proportions are still incorrectly estimated. This is because annealing cannot preserve the mixing proportions between different temperatures.
Consider the mixture of two Gaussian with density shown in Figure 10 (left). The probability masses round the two components change substantially as temperature varies, and they stay isolated. We manually pick a density threshold (middle, dashed black) below which we consider as low-density. When $\beta < 0.5$, the density at $x = 0.0$ is above the threshold for the support shown, and there are no isolated components; SVGD will correctly sample the distribution. When $\beta > 0.5$, the components become isolated as the density at $x = 0$ falls below the threshold in the green region (right), but the masses of the two components in the red and blue regions are still converging slowly to masses in the original mixture when $\beta = 1$. Thus, when the particles are well-mixed at $\beta = 0.5$, the proportions of particles allocated to $x < 0$ and $x > 0$ do not agree with the correct mixing proportions. The wrong mixing proportion are carried over into lower temperature up to $\beta = 1$ (right, dotted lines). Note that this happens regardless of the annealing schedule.

D.4 Entropy gradient estimation does not suffer from the blindness

The score function appears in estimating the gradient of entropy of implicit distributions, e.g. [11, 12]. Consider an implicit distribution $p_\phi(x)$ defined by the mapping $f_\phi : z \mapsto x$, $z \sim \zeta$ where $\zeta$ is some simple distribution and $f_\phi$ is a flexible function parametrised by $\phi$. The gradient of the entropy satisfies

$$\nabla_\phi H[p_\phi(x)] = \mathbb{E}_\zeta(z)[\nabla_x \log p_\phi(x) \nabla_\phi f_\phi(z)].$$

There are two reasons why this application does not suffer from the blindness discussed here. First, samples from $p(x)$ can be easily drawn from the implicit distribution, unlike when $p$ is an energy-based model. Second, the expectation above is an expectation of a $p_\phi(x)$-dependent function under $p_\phi(x)$ itself (through $\zeta$), which cannot be blind to itself. This is unlike SM or SD where the expectation involves two different density functions $p(x)$ and $q(x)$. 

Figure 10: Left: densities of a Gaussian mixture at different temperatures $\beta$. Middle: log normalised densities of the mixtures. The dashed line delineates a threshold such that densities below this threshold is considered low, and very few samples exist. For illustrative purpose, this threshold is taken as the Gaussian density evaluated at 4 standard deviations from the mean. During annealing, the density at $x = 0$ falls below the threshold as $\beta$ exceeds 0.5. Right, the mass in the middle green region becomes negligible when $\beta > 0.5$ even when the masses in the red and blue regions are changing substantially. This means that during annealing there are hardly any samples that would appear in or transition through the green region. The relative sample proportion as $\beta$ increase beyond 0.5 is almost the same as when $\beta = 0.5$, following the dotted lines, while the true proportions continue to change, following the solid lines.