3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–11

Annealed Stein Variational Gradient Descent

Francesco D'Angelo FDANGELO@ETHZ.CH
Vincent Fortuin FORTUIN@INF.ETHZ.CH
ETH Zürich

Abstract

Particle based optimization algorithms have recently been developed as sampling methods that iteratively update a set of particles to approximate a target distribution. In particular Stein variational gradient descent has gained attention in the approximate inference literature for its flexibility and accuracy. We empirically explore the ability of this method to sample from multi-modal distributions and focus on two important issues: (i) the inability of the particles to escape from local modes and (ii) the inefficacy in reproducing the density of the different regions. We propose an annealing schedule to solve these issues and show, through various experiments, how this simple solution leads to significant improvements in mode coverage, without invalidating any theoretical properties of the original algorithm.

1. Introduction

There have been many recent advances on the theoretical properties of sampling algorithms for approximate inference, which changed our interpretation and understanding of them. Particularly worth mentioning is the work of Jordan et al. (1998), who reinterpret Markov Chain Monte Carlo (MCMC) as a gradient flow of the KL divergence over the Wasserstein space of probability measures. This new formulation not only allowed for a deeper understanding of these methods but also inspired the inception of new and more efficient inference strategies. Following this direction, Liu and Wang (2016) recently proposed the Stein Variational Gradient Descent (SVGD) to perform approximate Wasserstein gradient descent. This method belongs to the more general family of particle optimization variational inference (POVI), where a continuous density $p(x)$ is approximated by a set of n particles that evolve over time towards the target. However, a solid understanding of its behavior in the finite particle limit beyond the mean field convergence analysis (Duncan et al., 2019) remains elusive. What is more, there is empirical evidence that SVGD suffers from a degeneracy that compromises the particle diversity under these conditions, making them collapse to a small number of modes (Zhuo et al., 2018). In the following, we discuss how an annealing strategy can significantly mitigate this issue, encourage exploration of the significant modes, and yield better samples from the target density than standard SVGD.

1.1. Related work

Introducing an artificial temperature parameter is a common practice in many approximate inference methods. Indeed, annealing approaches have been shown to be beneficial in both sampling and optimization problems for highly non-convex objectives. In the context of Markov chain Monte Carlo, sampling at different temperatures enhances the mixing times of the chains and thus allows for faster convergence (Marinari and Parisi, 1992; Geyer and

Thompson, 1995). In Bayesian inference, a temperature parameter has been introduced to anneal the likelihood or the posterior of the model (e.g., Wenzel et al., 2020) to escape from poor local minima. A similar effect can be obtained in variational inference and its stochastic counterpart by tempering the KL term (Mandt et al., 2016; Huang et al., 2018; Fu et al., 2019). However, the impact of similar annealing strategies on the SVGD method have not yet been studied. In our work, we show that they can be useful to improve the particle diversity and overcome the mode-collapse problem.

2. Background

Stein variational gradient descent (Liu and Wang, 2016) is a technique to perform approximate inference using a set of particles $q_t(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)}$, with δ_{x_i} being the Dirac measure on particle x_i, to approximate a positive density function $p(x)$ on $\mathcal{X} \in \mathbb{R}^d$. More precisely, SVGD is an efficient numerical technique to discretize the Wasserstein gradient flow of the Kullback-Leibler (KL) divergence functional on a new metric called the Stein geometry (Duncan et al., 2019).

SVGD considers an incremental transformation given by an infinitesimal perturbation of the identity matrix $\mathbf{T}(x) = x + \epsilon\phi(x)$ to move the particles from the initialization to the target. Here, $\phi(x)$ is the direction of the perturbation and ϵ the step size. The former is chosen to maximally decrease the KL divergence between the discrete density of the particles and the final target. As shown in Liu and Wang (2016), closed-form solutions can be obtained when restricting all perturbations ϕ to be from the unit ball of a vector valued reproducing kernel Hilbert space (RKHS) $\mathcal{H}^d = \mathcal{H}_0 \times ... \times \mathcal{H}_0$. Here, $\mathcal{H}_0$ is a scalar-valued RKHS associated with a scalar positive definite kernel $k(x, x'; h)$, and h is the set of kernel hyperparameters. The direction of steepest descent that maximizes the negative gradient of the KL divergence is then given in closed form as:

$$\phi^*_{q,p}(x') = \underset{\phi}{\operatorname{argmax}} \left\{ -\nabla_\epsilon D_{KL}(q_{[\mathbf{T}]}||p)\big|_{\epsilon\to 0} \right\} \propto \mathbb{E}_{x\sim q}[\mathcal{A}_p k(x, x')] \,, \tag{1}$$

with $\mathcal{A}_p\phi(x) = \phi(x)\nabla_x \log p(x)^\top + \nabla_x \phi(x)$ being the Stein operator. Using this, we can build an iterative procedure that transforms the initial reference distribution q_0 to the target posterior. Practically, we draw a set of particles $\{x_i^0\}_{i=1}^n$ with $x_i^0 \sim q_0$ and subsequently update them using the optimal perturbation in (1):

$$x_i^{t+1} \leftarrow x_i^t + \epsilon_t \hat{\phi}^*(x_i^t) \quad \text{with} \quad \hat{\phi}^*(x) = \frac{1}{n}\sum_{j=1}^{n} [\underbrace{k(x_j^t, x)\nabla_{x_j^t} \log p(x_j^t)}_{\text{driving force}} + \underbrace{\nabla_{x_j^t} k(x_j^T, x)}_{\text{repulsive force}}] \tag{2}$$

Conceptually the update rule attracts the particles to high density regions of the target via the average score function (driving force in (2)) while the repulsive force pushes them away from each other. This avoids a collapse to the MAP estimate and allows a certain degree of diversity among the particles to encourage the exploration of multiple modes and a more faithful reflection of the variance of the target distribution.

2.1. Mode-collapse in SVGD

Despite its theoretical foundations, it has been shown empirically (Zhuo et al., 2018) and theoretically (Zhang et al., 2020) that the particles in SVGD tend to collapse to a few local modes and that this effect is strongly related with the initial distribution of the particles. This issue is also clearly visible in our experiments and seems to already be relevant in low-dimensional problems. Indeed, as shown in Figure 1, it even happens in the case of a one-dimensional mixture of five Gaussians. Here, all particles, independent of the choice of the kernel bandwidth (see Appendix B), end up in the mode closest to the initialization without any possibility of escaping.

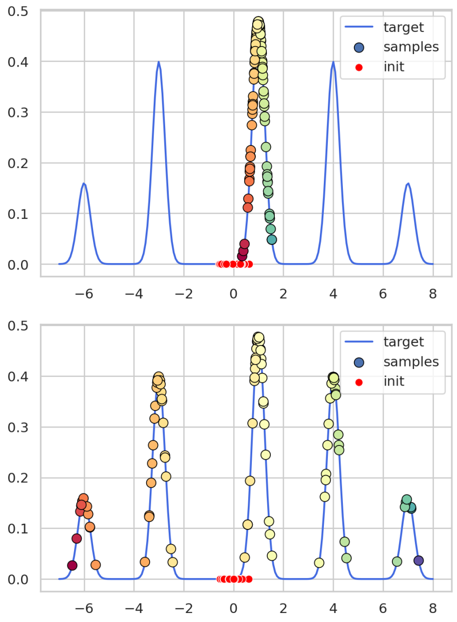

Figure 1: **SVGD mode-collapse.** Comparison of SVGD (top) and our proposed A-SVGD (bottom).

Additionally, we noticed in our experiments a connection of how a proper reconstruction of the target mass in the different modes is related to the initialization, as seen in Figure 2. In other words, even if the standard SVGD is able to capture different modes, the particles are not correctly distributed to faithfully reproduce the original mass. Instead, the majority of them tend to end up in the mode closest to their initialization. It is not clear whether this particular issue is a consequence of the mode-seeking limitation of all KL-divergence-based inference methods or of the approximation due to a finite number of particles. Empirically evident instead is that a deterministic update of the samples, like the one characterizing SVGD, in combination with a random initialization, can lead to a catastrophic convergence. In comparison, other methods characterized by a stochastic update like SGLD (Welling and Teh, 2011) can instead always rely on a nonzero probability of escaping from a certain mode, given by the injected noise, and consequently mitigate a bad initialization and encourage exploration. Moreover, it has been shown by Zhang et al. (2020) that injecting noise is also beneficial for the SVGD update to overcome the mode-collapse issues. Finally, it is well established that distance metrics and corresponding kernel-based methods suffer from the curse of dimensionality (Aggarwal et al., 2001; Reddi et al., 2014). That is why we should not hope for this problem to disappear in higher dimensions, but rather expect it to become even worse.

3. Annealed SVGD

Motivated by related work that mitigates similar mode-collapse issues in MCMC methods by tempering the target density to speed up the mixing time and avoid chains to get stuck in a single mode (Neal, 1996), we propose to introduce an annealing schedule in the SVGD

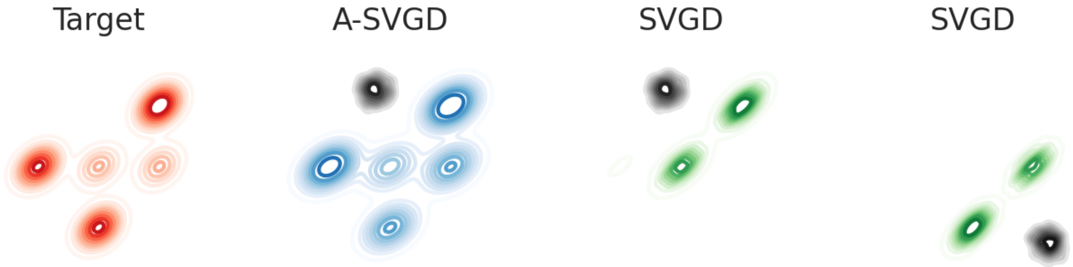

Figure 2: **Mode covering of SVGD.** We compare the final stationary distribution of standard SVGD (green) from two different initialization (black) and A-SVGD (blue) to approximate a mixture of Gaussians (red).

update (A-SVGD). This modification keeps the deterministic nature of the method but is essential to enhance its exploration and mode coverage and ideally would compensate for the limitation of the initialization and the finite number of particles. We introduce an annealing parameter $\gamma(t) \in [0,1]$ depending on the current iteration step and modify the update rule from (2) in the following way:

$$\hat{\phi}^*(x) = \frac{1}{n}\sum_{j=1}^{n}[\underbrace{\gamma(t)k(x_j^t, x)\nabla_{x_j^t}\log p(x_j^t)}_{\text{driving force}} + \underbrace{\nabla_{x_j^t}k(x_j^T, x)}_{\text{repulsive force}}]\,. \tag{3}$$

Intuitively, we can observe two phases in the time evolution of the particles by varying γ in the interval $[0,1]$ with an appropriate schedule: The first phase is exploratory with a predominant repulsive force that pushes the particles away from the initialization and thus allows for a good coverage of the target distribution's support. The second phase is exploitative, where the driving force takes over and shrinks the distribution of the particles to the area around the different modes. From a statistical perspective, our modification corresponds to the introduction of a temperature parameter $T(t) = \frac{1}{\gamma(t)}$ which rescales the target distribution $p(x)^{\frac{1}{T(t)}}$ during the evolution of the approximating density. It is important to notice how the choice of the annealing schedule is fundamental to preserve the convergence properties of SVGD and to keep the final target density unchanged. We ensure this by formulating the annealing schedule in such a way that the final iterations are always performed for the true target density, that is, $\lim_{t\to\infty}\gamma(t) = 1$. From this point of view, our alternative method is formally equivalent to a better parametrization of the initial reference distribution q_0 of the particles that depends on the target distribution. This is due to the fact that even if in the exploratory phase, when the repulsion is dominant, we still have a small component of the driving force that ensures that the particles are not randomly driven far away from the initialization, but are still driven towards high-density regions.

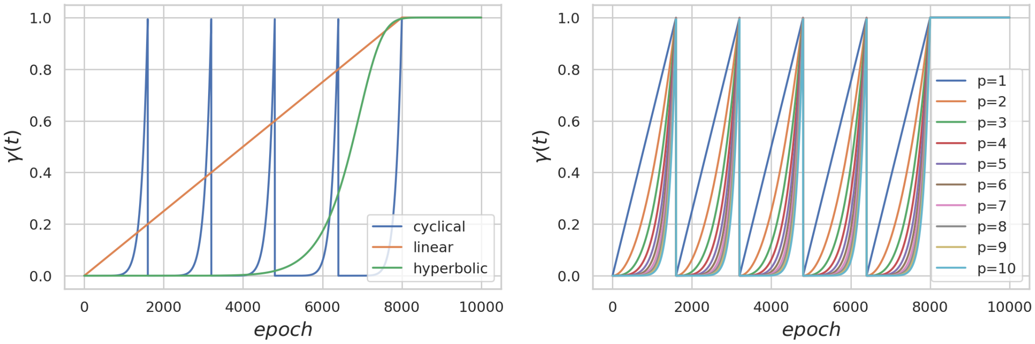

Figure 3: **Annealing schedules.** An illustration of the proposed annealing schedules.

3.1. The annealing schedule

As mentioned before, the choice of the annealing schedule is fundamental for the convergence to multiple modes while yielding samples from the proper target distribution. For this purpose, we introduced and tested different annealing schedules as shown in Figure 3. The simplest idea is a linear annealing on the interval $[0, 1]$; however, our experiments showed that this choice is not optimal. Indeed, linear tempering of the density leads to slow particle dynamics that are beneficial to neither the exploration nor the convergence. For this reason we chose to make the transition between the two inference phases steeper, so that the majority of the evolution happens in one of the two phases and not in between. To do so, we construct the annealing schedule using the hyperbolic tangent: $\gamma(t) = \tanh\left[(1.3\frac{t}{T})^p\right]$, with t being the current time step and T the total number of steps. Despite its good exploration, this method might encourage modes very far from the initialization due to the flattening of the target density in the initial exploratory phase. To achieve a tradeoff and have a more "target-guided" exploration, we follow the cyclical annealing schedule idea proposed in Loshchilov and Hutter (2016) and Huang et al. (2017), which has already been used for MCMC sampling in Zhang et al. (2019). We adapted this technique to have a sequence of C cycles of exploratory and converging phases, obtaining the following expression:

$$\gamma(t) = \left(\frac{mod(t, T/C)}{T/C}\right)^p, \tag{4}$$

with T being the total number of time steps, C the number of cycles and p an exponent determining the speed of the transition between the two phases (as shown in Figure 3).

4. Experiments

We demonstrate the advantages of our method in several synthetic experiments. In all the experiments we used SVGD with a standard RBF kernel.

Univariate Gaussian mixture. We first assessed the ability of A-SVGD to sample from a multi-modal univariate distribution given by a mixture of five Gaussians. The step size was fixed to $\epsilon = 0.1$ and we used the hyperbolic annealing schedule. We see in Figure 1 that

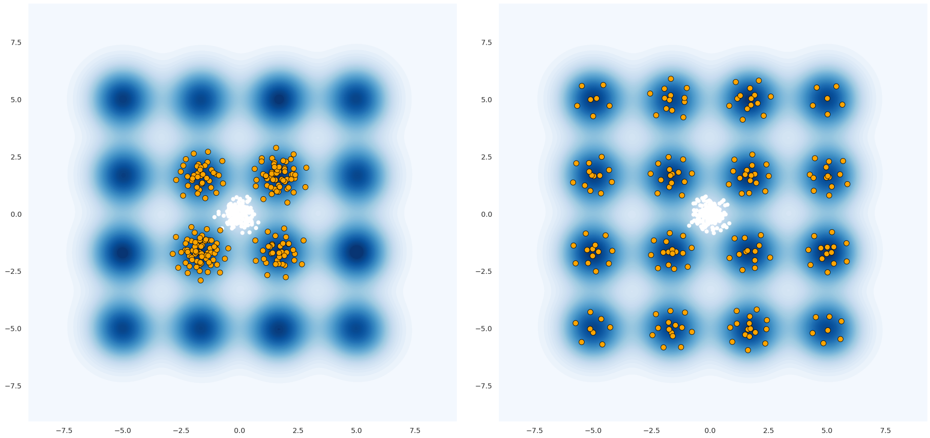

Figure 4: **SVGD on multi-modal 2D data.** We show the final samples of SVGD with annealing (right) and without (left) starting from the same initial distribution (white dots).

our proposed method successfully covers all the modes, while the standard SVGD collapses to just a single mode.

Bivariate regular Gaussian mixture. Secondly, we tested our method on a 2D mixture of 16 Gaussians with means equally distributed on a 4×4 grid and standard deviation $\sigma = 0.5$. In this experiment we used the cyclical annealing schedule from (4). As reported in Figure 4, we observe that the standard SVGD gets trapped in four of the modes, neighboring the initialization. In contrast, our method is able to find and characterize all modes, independently of the initial position.

Bivariate irregular Gaussian mixture. In our last experiment we studied the ability of SVGD to reproduce the weights of the mixture components of 2D Gaussians. The cyclical annealing schedule is used for the A-SVGD and for the standard SVGD we show two different initializations to illustrate their impact. We see in Figure 4 that our proposed method not only covers all the modes, but also approximately recovers their mixture weights, while the standard SVGD again collapses to the modes that are closest to its initialization.

5. Conclusion

In this work, we discussed the mode-collapse issue of SVGD for approximate inference and proposed an annealing strategy to overcome these limitations. We illustrated the impact of the initialization on a deterministic sampling algorithm like SVGD, highlighting two major drawbacks, namely (i) a tendency of the particles to fall into the neighboring modes without any possibility of escape and (ii) the difficulty of the particles in reproducing the effective local density of any given mode. We found that the introduction of a temperature parameter and an annealing schedule can help alleviate these undesirable behaviors, leading to better samples that can effectively capture multi-modal densities.

References

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001.

A Duncan, Nikolas Nuesken, and Lukasz Szpruch. On the geometry of stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.

Charles J Geyer and Elizabeth A Thompson. Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90 (431):909–920, 1995.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Chin-Wei Huang, Shawn Tan, Alexandre Lacoste, and Aaron C Courville. Improving explorability in variational inference with annealed variational objectives. In *Advances in Neural Information Processing Systems*, pages 9701–9711, 2018.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in neural information processing systems*, pages 2378–2386, 2016.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Stephan Mandt, James McInerney, Farhan Abrol, Rajesh Ranganath, and David Blei. Variational tempering. In *Artificial Intelligence and Statistics*, pages 704–712, 2016.

Enzo Marinari and Giorgio Parisi. Simulated tempering: a new monte carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.

Radford M Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366, 1996.

Sashank J Reddi, Aaditya Ramdas, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. *arXiv preprint arXiv:1406.2083*, 2014.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

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*, pages 1877–1887, 2020.

Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902.03932*, 2019.

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*, pages 6018–6027. PMLR, 2018.

Appendix A. Additional results on multi-modal data

In extension to the multi-modal synthetic data presented in section 4 we present here the extreme case for which the initialization is exactly in one of the mode of the target distribution. As shown in Figure A.1, the SVGD with annealing is remarkably able to escape from the initialization, covering all the modes characterizing the target density. On the other hand, the standard SVGD is not able to model anything besides the mode in which the particles have been initialized.

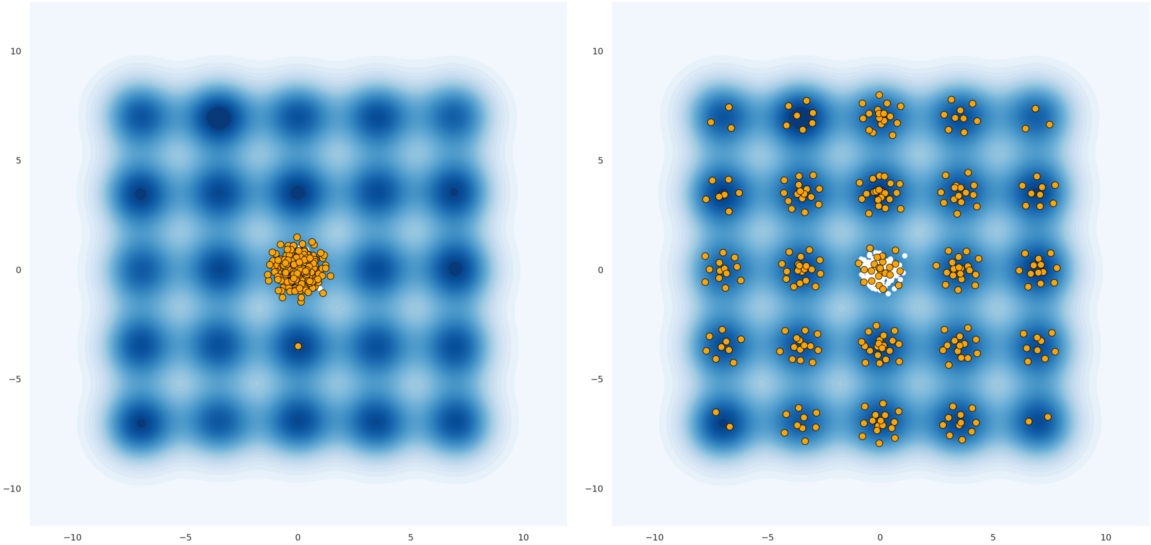

Figure A.1: **SVGD on multi-modal data with central initialization.** We show the final samples of SVGD with annealing (right) and without (left) starting from the same initial distribution (white dots). In this particular case the particles are initialized in the central mode. This particular initialization perfectly shows how the standard SVGD is not able to efficiently cover the entire target density, but instead remains trapped in the initialization.

Appendix B. Different RBF kernel bandwidths

We compared the effect of different bandwidths for the RBF kernel to show if this parameter affects the issues illustrated in 2.1. We used the synthetic multi-modal data from Figure 2 to test the following bandwidth values $h \in \{0.001, 0.01, 0.1, 1, 10, 100, \text{median}\}$ where median is the median heuristic. As illustrated in Figure B.1, none of the bandwidths shows significant improvement and, in contrast to our A-SVGD, the inference is still limited to the neighboring modes.

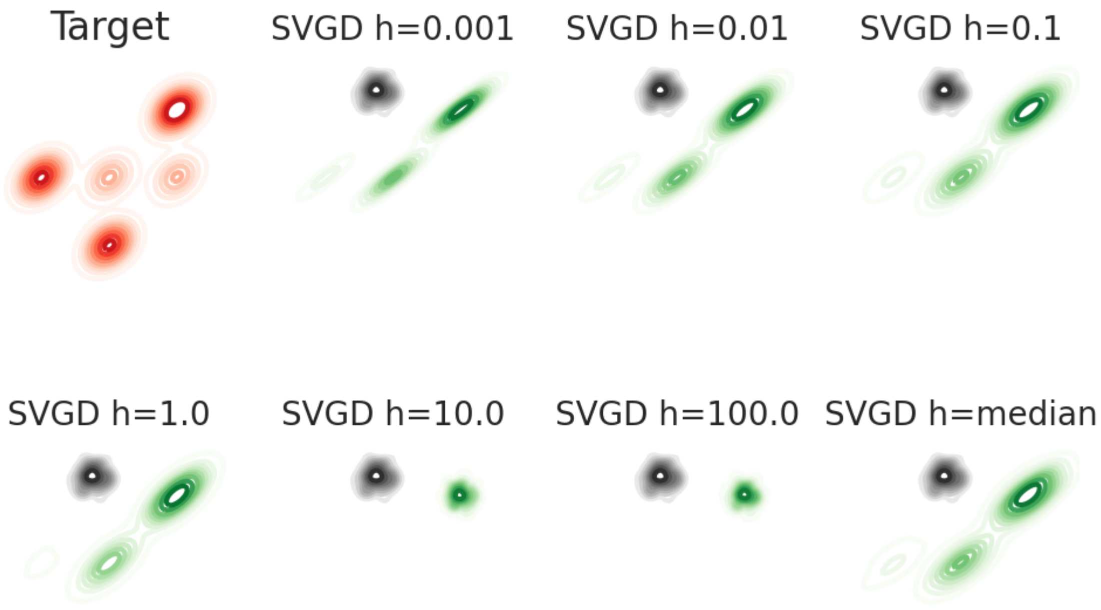

Figure B.1: **Mode covering of SVGD for different bandwidth.** We compare the final stationary distribution of standard SVGD (green) using different bandwidth (h).

Appendix C. Different annealing schedules

We compared the different annealing schedule proposed in section 3.1 on the bivariate irregular Gaussian mixture as reported in Figure C.1. We also computed the maximum mean discrepancy (MMD) (Gretton et al., 2012) during the evolution of the particles as reported in Figure C.2.

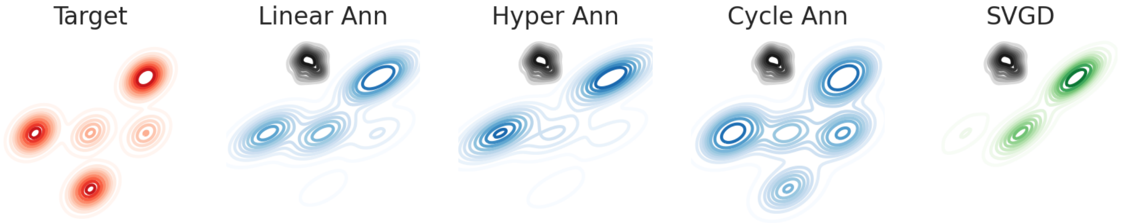

Figure C.1: **Mode covering of different annealing schedules.**We compare the final stationary distribution of standard SVGD (green) and A-SVGD (blue) using the three different annealing schedules to approximate a mixture of Gaussians (red)

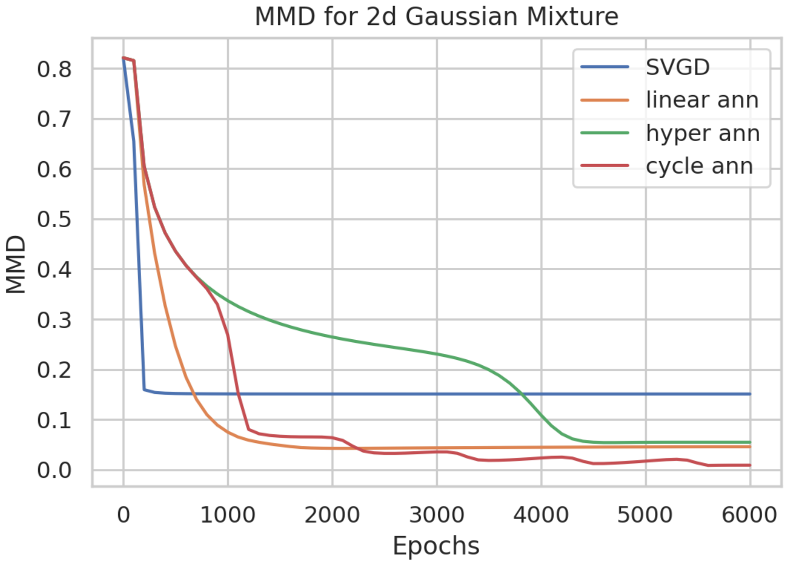

Figure C.2: **MMD for different annealing schedules.** We compute MMD during the evolution of the particles for the three different proposed annealing schedules