Kernel Stein Discrepancy thinning: a theoretical perspective of pathologies and a practical fix with regularization

Clément Bénard, Brian Staber, Sébastien Da Veiga

▶ **To cite this version:**

Clément Bénard, Brian Staber, Sébastien Da Veiga. Kernel Stein Discrepancy thinning: a theoretical perspective of pathologies and a practical fix with regularization. 2023. hal-03962614v1

HAL Id: hal-03962614

https://hal.science/hal-03962614v1

Preprint submitted on 31 Jan 2023 (v1), last revised 25 Oct 2023 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Kernel Stein Discrepancy thinning: a theoretical perspective of pathologies and a practical fix with regularization

Clément Bénard [*1] **Brian Staber** [*1] **Sébastien Da Veiga** [2]

Abstract

Stein thinning is a promising algorithm proposed by (Riabiz et al., 2022) for post-processing outputs of Markov chain Monte Carlo (MCMC). The main principle is to greedily minimize the kernelized Stein discrepancy (KSD), which only requires the gradient of the log-target distribution, and is thus well-suited for Bayesian inference. The main advantages of Stein thinning are the automatic remove of the burn-in period, the correction of the bias introduced by recent MCMC algorithms, and the asymptotic properties of convergence towards the target distribution. Nevertheless, Stein thinning suffers from several empirical pathologies, which may result in poor approximations, as observed in the literature. In this article, we conduct a theoretical analysis of these pathologies, to clearly identify the mechanisms at stake, and suggest improved strategies. Then, we introduce the regularized Stein thinning algorithm to alleviate the identified pathologies. Finally, theoretical guarantees and extensive experiments show the high efficiency of the proposed algorithm.

1. Introduction

Bayesian inference is a powerful approach to solve statistical tasks, and is especially efficient to incorporate prior expert knowledge of the studied system, or to provide uncertainties of the estimated quantities. Bayesian methods have thus demonstrated a high empirical performance for a wide range of applications, in particular in the fields of physics and computational biology, to just name a few. However, the Bayesian framework often leads to the evaluation of expectations with respect to a posterior distribution, which is not tractable (Green et al., 2015), except in the specific case of conjugate prior distribution and likelihood, which hardly occurs in practice. To overcome this issue, Markov chain Monte Carlo (MCMC) is one of the most commonly used computational method to estimate these integrals. Indeed, MCMC algorithms iteratively generate a sample, which follows the targeted posterior distribution, as the Markov chain converges to its stationary state (Robert & Casella, 1999; Brooks et al., 2011). Consequently, the quality of the resulting estimates strongly depends on the convergence of the MCMC and how its output is post-processed.

Standard post-processing procedures of MCMC outputs consist in removing the first iterations, called the burn-in period, and thinning the Markov chain with a constant frequency. Burn-in removal aims at reducing the bias introduced by the random initialization of the Markov chain. The $\hat{R}$ convergence diagnosis of Gelman et al. (1995) is, for instance, a well known method for determining the burn-in period. On the other hand, thinning the Markov chain allows for compressing the MCMC output and may also reduce the correlation between the iteratively selected points. More recently, promising kernel-based procedures were proposed to automatically remove the burn-in period, compress the output, and reduce the asymptotic bias (South et al., 2022). These approaches consist in minimizing a kernel-based discrepancy measure $D(\mathbb{P}, \mathbb{Q}_m)$ between the empirical distribution $\mathbb{Q}_m$ of a subsample of the MCMC output of size m, and the target distribution $\mathbb{P}$. In this respect, minimization of the maximum mean discrepancy (MMD) was investigated by several authors, but these strategies require the full knowledge of the target distribution $\mathbb{P}$, whose density is not tractable in non-conjugate Bayesian inference.

Based on the previous works of Chen et al. (2018) and Chen et al. (2019), Riabiz et al. (2022) propose to minimize the kernelized Stein discrepancy (KSD), to design an efficient kernel-based algorithm to thin MCMC outputs in a non-tractable Bayesian setting. The KSD (Liu et al., 2016) is a score-based discrepancy measure, *i.e.*, it only requires the knowledge of the score function of the target $\mathbb{P}$, which is readily available in our Bayesian framework. Importantly, Gorham & Mackey (2017) showed that under suitable mild conditions, the KSD enjoys good convergence properties. More precisely, the KSD is a valid distance to detect sam-

[*]Equal contribution [1]Safran Tech, Digital Sciences & Technologies, 78114 Magny-Les-Hameaux, France [2]ENSAI, CREST, F-35000 Rennes, France. Correspondence to: Clément Bénard <clement.benard@safrangroup.com>, Brian Staber <brian.staber@safrangroup.com>.

ples drawn form the target distribution, provided that the sample size is large enough. Therefore, KSD thinning is a highly promising tool for post-processing and measuring the quality of MCMC outputs. This article thus focuses on the Stein thinning algorithm proposed by Riabiz et al. (2022), which consists in selecting m points amongst the n iterations of the MCMC output, by greedily minimizing the KSD distance. Thanks to the convergence properties of the KSD, the empirical measure of the selected points weakly converges towards the posterior law $\mathbb{P}$. However, on the practical side, several articles (Wenliang & Kanagawa, 2020; Korba et al., 2021) have noticed empirical limitations of KSD-based sampling algorithms, especially for multimodal target distributions. In fact, these limitations happen to be quite problematic, even in simple experiments, and have been slightly overlooked in the literature so far, in our opinion. Therefore, this article first focuses on the analysis of KSD pathologies in Section 2, taking both an empirical and theoretical point of view. Then, we propose strategies to mitigate the identified problems, and introduce the regularized Stein thinning in Section 3. We show the efficiency of our algorithm through both a theoretical analysis and extensive experiments in Section 4. In the remaining of this initial section, we mathematically formalize the KSD distance and the associated Stein thinning algorithm.

Kernelized Stein discrepancy. Kernelized Stein discrepancy was independently introduced by (Chwialkowski et al., 2016; Liu et al., 2016; Gorham & Mackey, 2017) as a promising tool for measuring dissimilarities between two distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}^d$ with $d \geq 1$, whenever $\mathbb{P}$ admits a continuously differentiable density p, and the normalization constant of p is not tractable. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive semi-definite kernel and let $\mathcal{H}(k)$ be the associated reproducing kernel Hilbert space (RKHS) with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}(k)}$ and norm $\| \cdot \|_{\mathcal{H}(k)}$. Kernelized Stein discrepancy belongs to the family of maximum mean discrepancies (MMD) (Gretton et al., 2006) defined as

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}(k)} \leq 1} |\mathbb{E}[f(\mathbf{X})] - \mathbb{E}[f(\mathbf{Z})]| , \quad (1)$$

where $\mathbf{X} \sim \mathbb{P}$, $\mathbf{Z} \sim \mathbb{Q}$. If the kernel k is characteristic, then the MMD is a distance between probability distributions. In practice, the MMD may not be computable as it involves mathematical expectations with respect to $\mathbb{P}$, whose density is not tractable. To circumvent this issue, Gorham & Mackey (2015) proposed the Stein discrepancy which relies on Stein's method (Stein, 1972). It consists in defining an operator $\mathcal{T}_p$ that maps functions $g : \mathbb{R}^d \to \mathbb{R}^d$ to real-valued functions such that $\mathbb{E}[\mathcal{T}_p g(\mathbf{X})] = 0$, with $\mathbf{X} \sim \mathbb{P}$, for all g in $\mathcal{G}(k) = \{g : \mathbb{R}^d \to \mathbb{R}^d : \sum_{i=1}^d \|g_i\|^2_{\mathcal{H}(k)} \leq 1\}$. The probability measure $\mathbb{P}$ on $\mathbb{R}^d$ is assumed to admit a continuously differentiable Lebesgue density $p \in \mathcal{C}^1(\mathbb{R}^d)$, such that $\mathbb{E}[\|\nabla \log p(\mathbf{X})\|_2^2] < \infty$. The Stein discrepancy is then defined as $\mathrm{SD}(\mathbb{P}, \mathbb{Q}) = \sup_{g \in \mathcal{G}(k)} |\mathbb{E}[(\mathcal{T}_p g)(\mathbf{Z})]|$, where $\mathbf{Z} \sim \mathbb{Q}$. If the Stein operator $\mathcal{T}_p$ is chosen as the Langevin operator $(\mathcal{T}_p g)(\mathbf{x}) = \langle g(\mathbf{x}), \nabla \log p(\mathbf{x}) \rangle + \langle \nabla, g(\mathbf{x}) \rangle$, then Stein's discrepancy has a closed-form expression known as kernelized Stein discrepancy (Chwialkowski et al., 2016; Liu et al., 2016), $\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}[k_p(\mathbf{Z}, \mathbf{Z}')]$, where $\mathbf{Z} \sim \mathbb{Q}, \mathbf{Z}' \sim \mathbb{Q}$, and k_p denotes the Langevin Stein kernel defined from the score function $s_p(\mathbf{x}) = \nabla \log p(\mathbf{x})$, as

$$k_p(\mathbf{x}, \mathbf{x}') = \langle \nabla_{\mathbf{x}}, \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') \rangle + \langle s_p(\mathbf{x}), \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') \rangle + \langle s_p(\mathbf{x}'), \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}') \rangle + \langle s_p(\mathbf{x}), s_p(\mathbf{x}') \rangle k(\mathbf{x}, \mathbf{x}') . \quad (2)$$

The main advantage of the KSD is that it only requires the knowledge of the score function, and does not involve any integration with respect to $\mathbb{P}$. Gorham & Mackey (2017) also established convergence guaranties when the kernel k is chosen as the inverse multi-quadratic (IMQ) kernel function $k(\mathbf{x}, \mathbf{x}') = (c + \|\mathbf{x} - \mathbf{x}'\|^2_\Gamma)^{-\beta}$ with $c > 0$, $\beta \in (0, 1)$, the positive definite matrix Γ is the identity matrix, and the density p is distantly dissipative as defined below. Log-concave distributions outside of a compact set are a typical example of such probability densities.

Definition 1.1 (Distant dissipativity Gorham & Mackey (2017))**.** The density $p \in \mathcal{C}^1(\mathbb{R}^d)$ is distantly dissipative if $\liminf_{r \to \infty} \kappa(r) > 0$, where

$$\kappa(r) = \inf \Big\{ -2 \frac{\langle s_p(\mathbf{x}) - s_p(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle}{\|\mathbf{x} - \mathbf{y}\|_2^2} : \|\mathbf{x} - \mathbf{y}\|_2 = r \Big\}.$$

Stein thinning algorithm. Let $\mathbb{P}$ be a target probability measure that admits density p, and let $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ be a MCMC output. The Stein thinning algorithm (Riabiz et al., 2022) selects $m \leq n$ particles $\mathbf{x}_{\pi_1}, \ldots, \mathbf{x}_{\pi_m}$ by greedily minimizing the kernelized Stein discrepancy. Given $t - 1 < m$ particles $\mathbf{x}_{\pi_1}, \ldots, \mathbf{x}_{\pi_{t-1}}$, the t-th particle is defined as

$$\pi_t \in \operatorname*{argmin}_{i \in \{1, \ldots, n\}} k_p(\mathbf{x}_i, \mathbf{x}_i) + 2 \sum_{j=1}^{t-1} k_p(\mathbf{x}_{\pi_j}, \mathbf{x}_i) ,$$

where the KSD of an empirical distribution has been used to simplify the objective function. The kernel function k is chosen as the IMQ kernel function, defined above, for both its good theoretical properties and empirical efficiency. Several articles (Riabiz et al., 2022; Chen et al., 2018) have led extensive experiments to show the better practical performance of the IMQ kernel over other choices, in particular with $c = 1$, $\beta = 1/2$, and $\mathbf{\Gamma} = \ell^2 \mathbf{I}$, with ℓ given by the median heuristic (*i.e.*, the median of pairwise distances between the particles). Therefore, we consider the IMQ kernel with these settings throughout the article. Also notice that the bandwidth parameter ℓ is quite influential on the algorithm performance, but happens to be very difficult to tune, as highlighted by Chopin & Ducrocq (2021). Indeed, since the normalization constant of the target distribution is unknown, no additional metric is available to assess the precise

performance of the thinning procedure when ℓ varies. Furthermore, the sample quality output by Stein thinning varies in an erratic fashion with respect to ℓ, making the design of heuristic procedures for the choice of ℓ notoriously difficult. Finally, we use the median heuristic for ℓ in the sequel, and refer to Garreau et al. (2017) for an extensive analysis of this approach for kernel methods.

2. Analysis of KSD Pathologies

Although kernelized Stein discrepancy is a highly promising approach to thin MCMC outputs, several empirical studies have highlighted that KSD-based algorithms may suffer from strong pathologies in simple experiments (Wenliang & Kanagawa, 2020; Korba et al., 2021; Riabiz et al., 2022). The most established KSD pathology is that Stein thinning ignores the weights of distant modes of the target distribution, leading to the selection of samples of poor quality by Stein thinning. This problem, called Pathology I throughout the article, is analyzed in Subsection 2.1. Additionally, Korba et al. (2021) also notice that KSD thinning may result in samples concentrated in regions of low probability of p. As opposed to Pathology I, the mechanism leading to this problematic behavior is not well understood in the literature, to our best knowledge. Subsection 2.2 is thus dedicated to the theoretical characterization and illustration of Pathology II. Throughout the article, we illustrate KSD thinning using the running example of a Gaussian mixture, defined in Example 1 below, where initial particles are directly sampled from p to better highlight pathologies. We will come back to the thinning of MCMC outputs in detail in Section 4.

Example 1. Let the density p be a Gaussian mixture model of two components, respectively centered in $(-\mu, \mathbf{0}_{d-1})$ and $(\mu, \mathbf{0}_{d-1})$, of weights w and $1-w$, and of variance $\sigma^2 \mathbf{I_d}$. The initial particles $\{\mathbf{x}_i\}_{i=1}^n$ are drawn from p. The KSD thinning algorithm selects $m < n$ points to approximate p.

2.1. Pathology I: mode proportion blindness

We first focus on Pathology I, which states that Stein thinning is blind to the relative weights of multiple distant modes of a target distribution. Indeed, Wenliang & Kanagawa (2020) show that the score s_p is insensitive to distant mode weights. Consequently, the KSD distance is unable to properly identify samples with different weights than those of the target, in finite sample settings, as long as samples are accurately distributed within each mode. To be more specific, we illustrate this pathology with our Example 1 of a Gaussian mixture. We set $\mu = 3$ and $\sigma = 1$ to enforce the two modes to be well separated, and take an unbalanced proportion $w = 0.2$ for the left mode, and $1 - w = 0.8$ for the right mode. We generate $n = 3000$ observations and select $m = 300$ particles with Stein thinning. Clearly, the red selected sample displayed in Figure 1 has wrong

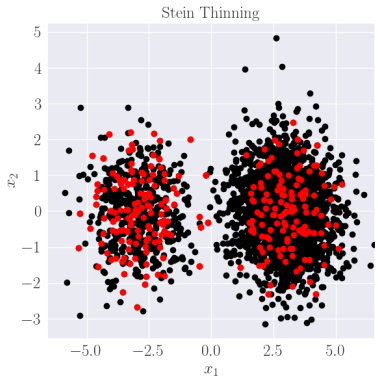

Figure 1. Illustration of Pathology I with the Gaussian mixture of Example 1 ($\mu = 3$, $\sigma = 1$, $w = 0.2$, $n = 3000$, $m = 300$). Initial particles are in black, and the Stein thinning output is red.

proportions, with about half of the particles in each mode, instead of the expected $20 - 80\%$, reflected by the initial black particles sampled from p. More precisely, over 100 repetitions of the Stein thinning algorithm, we obtain an average proportion of 0.53 of particles in the left mode, with a standard deviation of 0.08 across the 100 runs.

Although Wenliang & Kanagawa (2020) clearly show that the KSD distance is insensitive to the mode weights in the specific case of Gaussian mixtures, the mechanism leading to the selection of about half of the particles in each mode for Example 1, remains unexplained in the literature, to our best knowledge. Therefore, we conduct a theoretical analysis in the general case of any mixture distribution with two distant modes, stated in Assumption 2.1 below. For the sake of clarity, we only study the case of a number of modes of two, without loss of generality. Importantly, notice that a finite sample drawn from p with distant modes, takes the form of clusters of particles around each mode, as illustrated in Figure 1. In this case, Stein thinning selects particles among these clusters to approximate p. Then, these particles define an empirical law of a density q with a compact support around each mode. Wenliang & Kanagawa (2020) explain that the score s_p is especially insensitive to the mode weights in these compact areas around modes, leading to output samples with wrong proportions, as in Figure 1 for example. Therefore, Assumption 2.1 below restricts our analysis to densities q with a compact support around each mode.

Assumption 2.1 (Distant bimodal mixture distributions)**.** Let p and q be two mixture distributions in $\mathbb{R}^d$, made of two modes centered in $(-\mu, \mathbf{0}_{d-1})$ and $(\mu, \mathbf{0}_{d-1})$, with $\mu > 0$. The distribution of each mode of $p \in \mathcal{C}^1(\mathbb{R}^d)$ has $\mathbb{R}^d$ as support, whereas each mode distribution of q have a compact support, included in a ball of radius $r > 0$, with $r < \mu$. The left mode of p has weight $w_p \neq 1/2$, and the right mode has weight $1 - w_p$. Similarly, w and $1 - w$ are the mode weights of q. Let $\mathbb{Q}_L$ and $\mathbb{Q}_R$ be the probability measures that

respectively admit the density of the left and right modes of q, and $\mathbb{P}$ and $\mathbb{Q}_w$ be also the probability laws for p and q.

To state our main result about the KSD blindness to mode proportions, we need to formalize the assumption below, which tells that the distributions of the two modes of the mixture q have a close KSD distance with respect to the target p. In particular, this assumption can be easily verified when both p and q have symmetric mode distributions, since the KSD is insensitive to the weights of p.

Assumption 2.2. For distant bimodal mixture distributions q and p satisfying Assumption 2.1, and for $\eta \in (0, 1)$, we have $\left|\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_L)/\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - 1\right| < \eta$.

Theorem 2.3. *Let p and q be two bimodal mixture distributions satisfying Assumption 2.1 and 2.2, for any $\eta \in (0, 1)$. We define $w^\star$ as the optimal mixture weight of q with respect to the KSD distance, i.e., $w^\star = \underset{w\in[0,1]}{\mathrm{argmin}}\, \mathrm{KSD}(\mathbb{P}, \mathbb{Q}_w)$.*

Then, for μ large enough, we have $\left|w^\star - \frac{1}{2}\right| < \frac{\eta}{2(1-\eta)}$.

Theorem 2.3, proved in Appendix B, states that the weight $w^\star$ of the optimal mixture q, which minimizes the KSD distance to the target p, is close to $1/2$ regardless of the true target weight w_p, whenever the distributions of the two modes of the mixture q have a close KSD distance to p, and provided that the two modes are far enough. In particular, this is the case in the experiment of Example 1 and Figure 1, where the two modes are symmetric and well separated. A more specific empirical illustration of Theorem 2.3 can be found in Appendix A.1. Finally, also notice that the experiment of Figure 1, involving the full procedure of Stein thinning, quickly becomes quite unstable as μ increases. Indeed, for higher values of μ, we frequently observe about half of the selected particles in each mode as in Figure 1, but also often almost all particles in the same mode, switching from left to right over repetitions. This behavior of ignored distant mode, mentioned in Riabiz et al. (2022), means that an additional phenomenon than the one highlighted by Theorem 2.3, is at work in the Stein thinning of target densities with distant modes. Since this problem of KSD blindness to mode proportion is a well known drawback of score-based methods, we do not elaborate more about its analysis, to rather focus on the more intriguing Pathology II in the following subsection. Next, we will propose strategies to recover accurate mode proportions and stabilize Stein thinning outputs in Section 3.

2.2. Pathology II: spurious minimum

The core of this section is dedicated to the theoretical characterization of Pathology II. We first need to introduce additional notations to formalize our analysis. We thus define $\mathcal{M}_{s_0}$, the region of the input space where the score norm is lower than the threshold $s_0 \geq 0$, formally

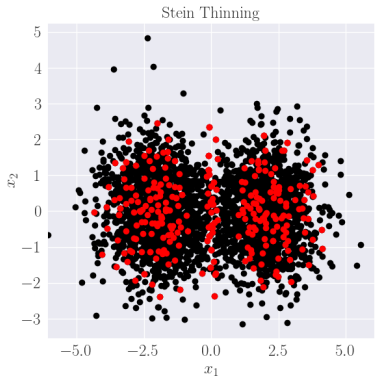

Figure 2. Illustration of Pathology II in the case of the Gaussian mixture from Example 1 ($\mu = 2$, $\sigma = 1$, $w = 0.5$, $n = 3000$, $m = 300$): many particles are selected around the line $x^{(1)} = 0$.

$\mathcal{M}_{s_0} = \{\mathbf{x} \in \mathbb{R}^d : \|s_p(\mathbf{x})\|_2 \leq s_0\}$. We also introduce an independent and identically distributed (iid) sample $\mathbf{X}_1, \ldots, \mathbf{X}_m$ of $\mathbb{P}$, with $\mathbb{P}_m$ the associated empirical measure for a positive integer m, and $X^{(j)}$ the j-th component of $\mathbf{X}$. Now, we can state our main result about Pathology II.

Theorem 2.4 (KSD spurious minimum)**.** *Let $\{\mathbf{x}_i\}_{i=1}^m \subset \mathcal{M}_{s_0} = \{\mathbf{x} \in \mathbb{R}^d : \|s_p(\mathbf{x})\|_2 \leq s_0\}$ be a fixed set of points of empirical measure $\mathbb{Q}_m = \frac{1}{m}\sum_{i=1}^m \delta(\mathbf{x}_i)$, with $s_0 \geq 0$ and $m \geq 2$. We have*

$$\mathrm{KSD}^2\big(\mathbb{P}, \mathbb{Q}_m\big) < \mathbb{E}[\mathrm{KSD}^2\big(\mathbb{P}, \mathbb{P}_m\big)],$$

if the score threshold s_0 and the sample size m are small enough to satisfy $m < 1 + \frac{\mathbb{E}[\|s_p(\mathbf{X})\|_2^2] - s_0^2}{d/\ell^2 + s_0/\ell + s_0^2}$.

Theorem 2.4 shows that samples concentrated in regions of the input space where the norm of the score is low, have smaller KSD than samples drawn from the true target distribution, for small sample size. Additionally, the score norm is low around stationary points of the density p, including local minimum and saddle points, as shown in Corollary 2.5 below. However, samples concentrated in local minimum of p are bad approximations of the target distribution by definition. Therefore, these results highlight that pathological samples may be generated by Stein thinning, which minimizes the empirical KSD, and thus explains Pathology II observed by Korba et al. (2021), and shown in Figure 2.

Corollary 2.5 (Low KSD samples at density minimum)**.** *Let p be a density with at least one local minimum or one saddle point. For $m \geq 2$, if $\{\mathbf{x}_i\}_{i=1}^m \subset \mathbb{R}^d$ is a set of points, all located at local minimum or saddle points of p, then we have $\mathrm{KSD}^2\big(\mathbb{P}, \mathbb{Q}_m\big) < \mathbb{E}[\mathrm{KSD}^2\big(\mathbb{P}, \mathbb{P}_m\big)]$ if $m < 1 + \frac{\ell^2}{d}\mathbb{E}[\|s_p(\mathbf{X})\|_2^2]$.*

The proofs of Theorem 2.4 and Corollary 2.5, reported in Appendix C, are built on the idea that the KSD of the empirical law of $\{\mathbf{x}_i\}_{i=1}^n$, has a bias of the form

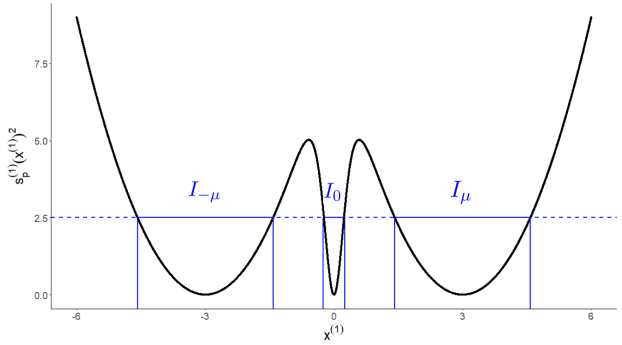

Figure 3. Squared first component of the score $s_p(\mathbf{x})$ along $x^{(1)}$ for a Gaussian mixture, as defined in Example 1 ($\mu = 3$, $\sigma = 1$).

$\sum_{i=1}^m \|s_p(\mathbf{x}_i)\|_2^2/m^2$. Consequently, when m is small, the bias has a strong influence on KSD estimates, which favor samples concentrated in regions of low score norm, as stationary points of p. This mechanism is illustrated in Figure 2 and Corollary 2.6 for Gaussian mixtures. In this case, Stein thinning aligns a large number of particles around the line defined by $x^{(1)} = 0$, an area of low probability of the targeted mixture distribution, because of the variations of the score function, displayed in Figure 3.

Corollary 2.6 (KSD spurious minimum for Gaussian mixtures)**.** *Let the density p be a Gaussian mixture model of two components with equal weights, respectively centered in $(-\mu, \mathbf{0}_{d-1})$ and $(\mu, \mathbf{0}_{d-1})$, of variance $\sigma^2\mathbf{I_d}$, and let $\nu = \mu/\sigma$. If $\nu > 1$ and $0 \leq s_0 < \left[\nu\sqrt{\nu^2-1} - \ln(\nu + \sqrt{\nu^2-1})\right]/\mu$, then for any $\{\mathbf{x}_i\}_{i=1}^m \subset \mathcal{M}_{s_0}$ of empirical measure $\mathbb{Q}_m$, we have*

(i) $\mathrm{KSD}^2\big(\mathbb{P}, \mathbb{Q}_m\big) < \mathbb{E}[\mathrm{KSD}^2\big(\mathbb{P}, \mathbb{P}_m\big)]$ *if m and s_0 satisfy* $m < 1 + \frac{\mathbb{E}[\|s_p(\mathbf{X})\|_2^2] - s_0^2}{d/\ell^2 + s_0/\ell + s_0^2}$,

(ii) it exists three disjoint intervals $I_{-\mu}, I_0, I_\mu \subset \mathbb{R}$, respectively centered around $-\mu$, 0, and μ, such that $x_1^{(1)}, \ldots, x_m^{(1)} \in I_{-\mu} \cup I_0 \cup I_\mu$.

Remark 2.7. Corollary 2.6 states that for a given target distribution of Gaussian mixture, Pathology II appears if the sample size m is small enough. From another perspective, for any sample size m, it exists a Gaussian mixture with μ/σ large enough, such that Pathology II occurs, when the bandwidth ℓ is chosen with the heuristic of the median distance of initial particles. Therefore, Pathology II can appear for arbitrarily large samples m, depending on the target distribution properties.

3. Regularized Stein Thinning

Stein thinning suffers from two main pathologies, analyzed in Section 2. In a word, Pathology I comes from the insensitivity of the score to the relative weights of distant modes, whereas Pathology II originates from the variations of the score norm, which do not differentiate local minimum from local maximum of the target distribution. We propose to regularize the KSD distance to fix these two problems, using terms that are highly sensitive to the type of stationary point and the relative weights of modes. The proposed algorithm is first introduced in Subsection 3.1, then theoretical properties are discussed in Subsection 3.2, and finally the good empirical performance will be shown in Section 4.

3.1. Algorithm

Entropic regularization. In order to compensate the blindness of the KSD to mode proportions in multimodal distributions, we introduce the following entropic regularized KSD, denoted by KSD_λ, and defined as

$$\mathrm{KSD}_\lambda^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}[k_p(\mathbf{Z}, \mathbf{Z}')] - \lambda\mathbb{E}[\log(p(\mathbf{Z}))],$$

where $\mathbf{Z}$ and $\mathbf{Z}'$ have probability law $\mathbb{Q}$, and $\mathbb{P}$ admits the density p. In our Bayesian setting, $\mathbb{E}[\log(p(\mathbf{Z}))]$ is known up to an additive constant since the normalization factor of p is not tractable. However, it is possible to use $\mathrm{KSD}_\lambda^2(\mathbb{P}, \mathbb{Q})$ as the objective function of the Stein thinning algorithm, as the greedy selection of particles to optimize this quantity does not rely on the unknown additive constant. The main idea of this entropic regularization is that $-\log(p(\mathbf{x}))$ takes higher values in modes of smaller probability, and therefore provides the relative mode weight information, which is missing in the KSD distance.

Laplacian correction. Chen et al. (2018) and Riabiz et al. (2022) have noticed that the term $k_p(\mathbf{x}_i, \mathbf{x}_i)$, which naturally appears in the empirical kernelized Stein discrepancy with the Langevin operator, can be interpreted as a regularization term. For example, Stein thinning does not select particles in the burn-in period of an MCMC output thanks to this regularization. However, this term $k_p(\mathbf{x}_i, \mathbf{x}_i)$ is also responsible for Pathology II, of samples concentrated in stationary points of p, as shown in Theorem 2.4. Therefore, we add a second regularization term to compensate the weaknesses of $k_p(\mathbf{x}_i, \mathbf{x}_i)$, by penalizing particles located at local minimum and saddle points of the density p. Such points are located in areas of convexity of the target distribution, which can thus be detected with the positive values of the Laplacian of the density. Therefore, using the truncated Laplacian operator $\Delta^+ f(\mathbf{x}) \overset{def}{=} \sum_{j=1}^d \left(\frac{\partial^2 f(\mathbf{x})}{\partial x^{(j)2}}\right)^+$ for a function $f \in \mathcal{C}^2(\mathbb{R}^d)$, we propose the L-KSD estimate with a Laplacian correction for densities $p \in \mathcal{C}^2(\mathbb{R}^d)$, defined by

$$\text{L-KSD}^2(\mathbb{P}, \mathbb{Q}_m) = \frac{1}{m^2}\sum_{i\neq j}^m k_p(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{m^2}\sum_{i=1}^m \left[k_p(\mathbf{x}_i, \mathbf{x}_i) + \Delta^+ \log(p(\mathbf{x}_i))\right].$$

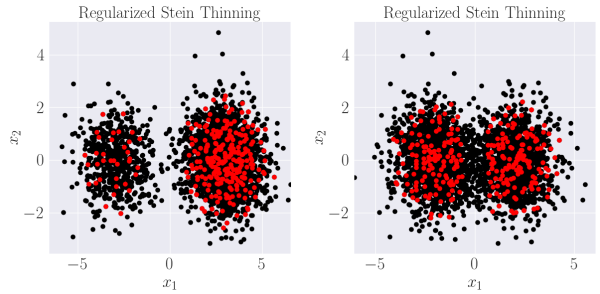

Figure 4. Pathologies fixed by the regularized Stein thinning.

Regularized Stein thinning. Overall, we obtain the following estimate for the entropic regularized KSD with Laplacian correction

$$\text{L-KSD}^2_\lambda(\mathbb{P}, \mathbb{Q}_m) = \text{L-KSD}^2(\mathbb{P}, \mathbb{Q}_m) - \frac{\lambda}{m}\sum_{i=1}^{m} \log(p(\mathbf{x}_i)).$$

Then, at each iteration $t \in \{1, \dots, m\}$, the regularized Stein thinning greedily minimizes

$$\pi_t \in \underset{i\in\{1,\dots,n\}}{\text{argmin}}\; k_p(\mathbf{x}_i,\mathbf{x}_i) + 2\sum_{j=1}^{t-1} k_p(\mathbf{x}_{\pi_j}, \mathbf{x}_i) + \Delta^+ \log(p(\mathbf{x}_i)) - \lambda t \log(p(\mathbf{x}_i))\,.$$

Finally, Figure 4 illustrates the performance of regularized Stein thinning to fix the two pathologies analyzed in Section 2, in the case of Example 1 with Gaussian mixtures. Indeed, the left panel of Figure 4 shows that the majority of particles are selected in the right mode, as expected from the target distribution with $w = 0.2$. More precisely, an average proportion of 0.11 of the particles are located in the left mode over 100 repetitions of the procedure (0.89 in the right mode), with a standard deviation of 0.03. For the value choice of λ, we refer to the next subsection and the experimental Section 4. On the right panel of Figure 4, we observe that no particle is now selected on the line $x^{(1)} = 0$, as expected from the target Gaussian mixture distribution.

Remark 3.1. The Laplacian correction of k_p introduces second-order derivatives of p in the Stein discrepancy, and therefore enables to differentiate local minimum and saddle points of the density p from its local maximum. A natural approach to introduce second-order derivatives of p in KSD estimates, is to define the Stein discrepancy using second-order operators. A Laplacian Stein operator (Oates et al., 2017) is derived in Appendix G, but experiments show that this strategy is not efficient to fix Pathologies I & II.

3.2. Theoretical properties

This subsection is dedicated to the theoretical analysis of regularized Stein thinning. First, we show that the proposed algorithm now enjoys good theoretical properties regarding Pathologies I and II, and thus mitigates the identified problems of the original Stein thinning. Secondly, we extend the convergence analysis of Riabiz et al. (2022) for the post-treatment of MCMC output, to show the weak convergence of the empirical law output by regularized Stein thinning towards the target probability measure.

Entropic regularization. In the previous section, Theorem 2.3 highlights how Pathology I of mode proportion blindness originates from the score insensitivity to mode weights. On the other hand, the entropic regularization is directly built on the target density, and therefore strongly depends on the mode weights. In the same setting of Assumption 2.1 for Theorem 2.3 with distant bimodal mixture distributions, the following theorem shows that the entropic regularized KSD is minimized for the appropriate target weight, with the suitable regularization strength λ.

Theorem 3.2. *Let p and q be two bimodal mixture distributions satisfying Assumption 2.1. We define $w^\star_\lambda$ as the optimal mixture weight of q with respect to the entropic regularized KSD distance, i.e., $w^\star_\lambda = \underset{w\in[0,1]}{\text{argmin}}\, \text{KSD}_\lambda(\mathbb{P}, \mathbb{Q}_w)$.*

If $\mathbb{E}[\log(p(\mathbf{Z}_L))] \neq \mathbb{E}[\log(p(\mathbf{Z}_R))]$ where $\mathbf{Z}_L \sim \mathbb{Q}_L$ and $\mathbf{Z}_L \sim \mathbb{Q}_R$, it exists $\lambda \in \mathbb{R}$ such that $w^\star_\lambda = w_p$.

Notice that Theorem 3.2, proved in Appendix D, is valid if $\mathbb{E}[\log(p(\mathbf{Z}_L))] \neq \mathbb{E}[\log(p(\mathbf{Z}_R))]$, otherwise the impact of the entropic regularization on $w^\star_\lambda$ vanishes. However, as $w_p \neq 1/2$ is required in Assumption 2.1 for Pathology I to occur, p is asymmetric, and $\mathbb{E}[\log(p(\mathbf{Z}_L))] = \mathbb{E}[\log(p(\mathbf{Z}_R))]$ is only possible in very specific cases. Theorem 3.2 clearly shows that the regularized entropic KSD is sensitive to the weights of distant modes. Efficient strategies to choose the regularization strength will be first discussed in the asymptotic analysis below, and then in the experiments of the next section.

Laplacian correction. First, we stress that the L-KSD is a strongly consistent estimate of the KSD distance, where the proof follows from the law of large numbers. Therefore, the Laplacian correction introduced in the L-KSD estimate does not undermine the good asymptotic properties of the KSD distance. Secondly, the following theorem shows that samples concentrated in local minimum or saddle points of the target distribution and of low density values, are well identified by the L-KSD as samples of worse quality than those truly sampled from the target. Consequently, the Laplacian correction fixes Pathology II, previously formalized in Theorem 2.4.

Theorem 3.3. *For $m \geq 2$, let $\{\mathbf{x}_i\}_{i=1}^m \subset \mathbb{R}^d$ be a set of points concentrated at $\mathbf{x}_0$, a local minimum or saddle point of p, where the density is lower than the following positive*

threshold,

$$p(\mathbf{x}_0) < \frac{\Delta^+ p(\mathbf{x}_0)}{\mathbb{E}[\|s_p(\mathbf{X})\|_2^2] + \mathbb{E}[\Delta^+ \log p(\mathbf{X})]}.$$

Then, with $\mathbb{Q}_m$ the empirical measure of $\{\mathbf{x}_i\}_{i=1}^m$, we have

$$\text{L-KSD}^2\big(\mathbb{P}, \mathbb{Q}_m\big) > \mathbb{E}[\text{L-KSD}^2\big(\mathbb{P}, \mathbb{P}_m\big)].$$

Convergence of regularized Stein thinning. While regularized Stein thinning fixes finite sample size pathologies, the asymptotic properties of Stein thinning are also preserved. Indeed, if the initial set of particles is drawn from a different distribution than the target using a Markov chain Monte Carlo, Theorem 3.5 states that the empirical measure of the sample obtained with regularized Stein Thinning, converges towards the target measure $\mathbb{P}$, and thus extends the results of Riabiz et al. (2022). Notice that the weak convergence of a sequence of probability measure is denoted by $\Rightarrow$, and that distantly dissipative distributions are defined in Definition 1.1. The required assumption below, essentially states mild integrability conditions, and enforces that the MCMC output is not too far from a sample drawn from p.

Assumption 3.4. Let $\mathbb{Q}$ be a probability distribution on $\mathbb{R}^d$, such that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$. Let $\{\mathbf{Z}_i\}_{i\in\mathbb{N}} \subset \mathbb{R}^d$ be a $\mathbb{Q}$-invariant, time-homogeneous Markov chain, generated using a V-uniformly ergodic transition kernel, such that $V(\mathbf{x}) \geq \frac{d\mathbb{P}}{d\mathbb{Q}}\sqrt{d/\ell^2 + \|s_p(\mathbf{x})\|_2^2}$. Suppose that, for some $\gamma > 0$,

$$\sup_{i\in\mathbb{N}} \mathbb{E}\big[e^{\gamma \max(1, \frac{d\mathbb{P}}{d\mathbb{Q}}(\mathbf{Z}_i)^2)(d/\ell^2 + \|s_p(\mathbf{Z}_i)\|_2^2)}\big] < \infty,$$

$$\sup_{i\in\mathbb{N}} \mathbb{E}[e^{\gamma|\log(p(\mathbf{Z}_i))|}] < \infty, \quad \sup_{i\in\mathbb{N}} \mathbb{E}[e^{\gamma\Delta^+ \log p(\mathbf{Z}_i)}] < \infty,$$

$$\sup_{i\in\mathbb{N}} \mathbb{E}\Big[\frac{d\mathbb{P}}{d\mathbb{Q}}(\mathbf{Z}_i)\sqrt{d/\ell^2 + \|s_p(\mathbf{Z}_i)\|_2^2 V(\mathbf{Z}_i)}\Big] < \infty.$$

Theorem 3.5. *Let $\mathbb{P}$ be a distantly dissipative probability measure, that admits the density $p \in \mathcal{C}^2(\mathbb{R}^d)$, k_p be the Stein kernel associated with the IMQ kernel, $\{\mathbf{Z}_i\}_{i\in\mathbb{N}} \subset \mathbb{R}^d$ be a Markov chain satisfying Assumption A3.4, π be the index sequence of length m_n generated by regularized Stein thinning, and $\mathbb{Q}_{m_n}$ be the associated empirical measure of $\{\mathbf{Z}_{\pi_i}\}_{i=1,\dots,m_n}$. If $\log(n)^\beta < m_n < n$, with any $\beta > 1$, and $\lambda_{m_n} = o(\log(m_n)/m_n)$, then we have almost surely*

$$\mathbb{Q}_{m_n} \underset{n\to\infty}{\Longrightarrow} \mathbb{P}.$$

Theorem 3.5, proved in Appendix F, provides us with interesting insights about the entropic regularization strength λ. Indeed, for our ultimate application of MCMC post-processing, the Stein thinning sample distribution converges towards the target, provided that the regularization parameter satisfies $\lambda = o(\log(m)/m)$. This strongly suggests that λ should be chosen with a rate at least as fast as $O(1/m)$. In practice, it is not possible to tune this regularization parameter λ for computational reasons, and because no metric is available to assess the Stein thinning quality for various values of λ on real cases in our Bayesian setting. However, we will see in experiments of the following section that the effect of the entropic regularization vanishes with faster decay rate of λ with respect to m. Therefore, we set $\lambda = 1/m$ in the regularized Stein thinning, to ensure the algorithm convergence and good empirical performance.

4. Empirical Assessment

This section shows how regularized Stein thinning outperforms the original algorithm through three batches of experiments, namely: a Gaussian mixture with four modes, a mixture of banana-shaped distributions with t-tails, and Bayesian logistic regression on real datasets. Two Metropolis-Hastings samplers are considered: the Metropolis-Adjusted Langevin Algorithm (MALA) and the No-U-Turn sampler (NUTS). Notice that additional details and experiments are provided in Appendix A, and that the code is available in the Supplementary Material.

4.1. Gaussian and banana-shaped mixtures

As a first experiment, we consider a d-dimensional Gaussian mixture of four modes of equal weight, illutrated in Figure 5, where the two modes at the top have a variance four times higher than the two bottom modes. We sample the target Gaussian mixture with both MALA and NUTS using three different step sizes ε and 10^4 iterations. The generated samples are post-processed with the Stein thinning and regularized Stein thinning algorithms, and their performances are compared with the MMD between the post-processed samples and samples drawn from the known target Gaussian mixture. This experiment is run for various thinning sizes m and dimensions d, with 20 repetitions to quantify uncertainties. The results obtained with the MALA sampler are shown in Figure 6, and it is found that the regularized Stein thinning yields lower MMD distances than the vanilla Stein thinning. An example of post-processed MALA output is depicted in Figure 5, together with a heatmap of the Laplacian correction. On the left panel of Figure 5, we see that Pathologies I & II are especially strong in this experiment, with a large number of particles lying between modes in regions of low probability, and highly unbalanced mode weights. On the right panels, we observe that the entropic regularization and the Laplacian correction fix both pathologies. Similar results were obtained with NUTS and are reported in Appendix A for brevity. Besides, we take advantage of this first experiment to explore other regularization rates than our default $\lambda = 1/m$. Figure 11 in Appendix A shows that a rate of $\lambda = 1/\log(m)$, which violates the

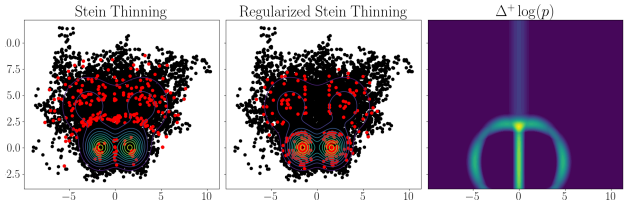

Figure 5. (Gaussian mixture) First two panels: solutions obtained with Stein thinning and regularized Stein thinning with contour lines of the target distribution. Last panel: heatmap of the Laplace correction $\Delta^+ \log(p)$.

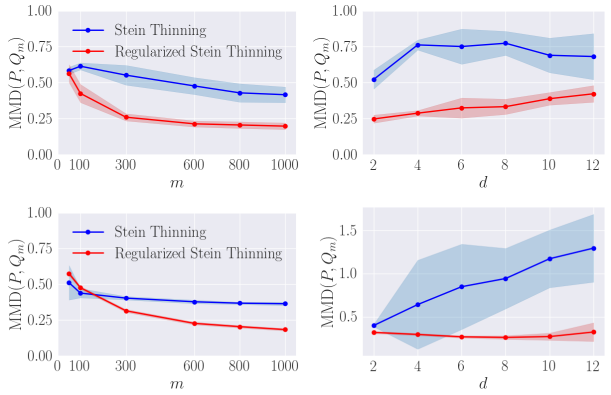

Figure 6. (MALA) Graphs of the MMD with respect to the thinning size m (with $d = 2$) and with respect to d (with $m = 300$). Top row: Gaussian mixture. Bottom row: banana-shaped mixture.

convergence assumptions of Theorem 3.5, has significantly worse performance than the original Stein thinning. On the other hand, with a faster rate than $1/m$ such as $1/m^2$, the effect of the entropic regularization vanishes, and results are similar to the original Stein thinning. Therefore, this justifies the default value of $\lambda = 1/m$. Finally, we conduct the same experiment for a mixture of d-dimensional banana-shaped distributions with t-tails (Haario et al., 1999; Pompe et al., 2020), with details in Appendix A. Again, Figure 6 shows the better performance of regularized Stein thinning.

4.2. Bayesian logistic regression

We now compare the two Stein thinning algorithms in the Bayesian logistic regression setting for binary classification, since such problem usually involves multimodal posterior—see, *e.g.*, Gershman et al. (2012); Liu & Wang (2016); Fong et al. (2019); Korba et al. (2021). Given a dataset $\mathcal{D}_N = \{(\mathbf{X}_i, Y_i)\}_{i=1}^N$ made of N pairs of features $\mathbf{X}_i \in \mathbb{R}^d$ and labels $Y_i \in \{0, 1\}$, the probability that Y_i is of class 1 is given by $p(Y_i = 1|\mathbf{X}_i, \boldsymbol{\beta}, \beta_0) = 1/(1 + \exp(-\beta_0 - \boldsymbol{\beta}^T \mathbf{X}_i))$, for some parameters $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{d+1}$. The prior distributions of the weight vector $\boldsymbol{\theta}$ is assumed to be Gaussian, $p(\beta^{(j)}|\gamma^{(j)}) = \mathcal{N}(\beta^{(j)}|, 0, 1/\gamma^{(j)})$, and a Gamma prior with parameters (a, b) is chosen for the precision $\gamma^{(j)}$. Upon marginalizing, it is found that $\beta^{(j)}$ is distributed as the non-standardized t-distribution Student-t$(2a, 0, b/a)$ (Bishop & Nasrabadi, 2006). Following (Fong et al., 2019), the hyper-parameters a and b are chosen as $a = b = 1$. The posterior distribution of the weights $\boldsymbol{\theta}$ is sampled with both MALA and NUTS using 48 independent chains, of respectively 10^4 and 10^5 iterations, and four step sizes ε are considered along with three thinning sizes m. Each MCMC sample is post-processed with the two Stein thinning algorithms. For a new input $\mathbf{x}^\star$, the resulting thinned samples are used to approximate the posterior predictive distribution defined by $p(Y = 1|\mathbf{x}^\star, \mathcal{D}_N) = \int p(Y = 1|\mathbf{x}^\star, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}_N) d\boldsymbol{\theta}$. The performances of Stein thinning algorithms are assessed using the standard AUC metric for classification problems, estimated with 10-fold cross-validation and 10 repetitions for uncertainties. Table 1 gathers the results for five public datasets from the UCI repository (Dua & Graff, 2017), and described in Appendix A, where the best AUC obtained for each algorithm over the four MCMC step sizes are reported. Clearly, regularized Stein thinning significantly improves the performance of Bayesian logistic regression.

Table 1. AUCs obtained with NUTS sampler with Stein Thinning (ST) and Regularized Stein Thinning (RST), estimated with 10-fold cross-validation (10 repetitions for uncertainties).

	$m = 50$		$m = 300$	
Dataset	ST	RST	ST	RST
Breast W.	0.88 (0.02)	**0.96** (0.00)	0.93 (0.01)	**0.96** (0.00)
Diabetes	**0.52** (0.01)	0.50 (0.02)	0.53 (0.02)	**0.57** (0.02)
Haberman	0.51 (0.04)	**0.53** (0.02)	0.53 (0.03)	**0.58** (0.02)
Liver	0.53 (0.04)	**0.69** (0.01)	0.61 (0.04)	**0.70** (0.01)
Sonar	0.80 (0.02)	**0.81** (0.01)	0.81 (0.01)	0.81 (0.01)

5. Conclusion

Stein thinning has raised a high interest in recent years, as a powerful tool to post-process MCMC outputs, by the greedy minimization of the kernelized Stein discrepancy. Unfortunately, empirical studies have shown that KSD-based algorithms suffer from strong pathologies. We have conducted an in-depth theoretical analysis to identify the mechanisms at stake. From this understanding, we propose an improved Stein thinning algorithm relying on entropic regularization and Laplacian correction. This approach exhibits relevant theoretical properties regarding pathologies, as well as highly improved empirical performance. Also notice that the scope of this article was restricted to moderate dimensions of MCMC outputs. The analysis of higher dimensional problems seems a challenging research direction with a wide range of critical applications.

References

Bishop, C. M. and Nasrabadi, N. *Pattern recognition and machine learning*. Springer, 2006.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. *Handbook of markov chain monte carlo*. CRC press, 2011.

Chen, W., Mackey, L., Gorham, J., Briol, F., and Oates, C. Stein points. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2018.

Chen, W., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. Stein point markov chain monte carlo. In *International Conference on Machine Learning*, pp. 1011–1021. PMLR, 2019.

Chopin, N. and Ducrocq, G. Fast compression of MCMC output. *Entropy*, 23:1017, 2021.

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

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Fong, E., Lyddon, S., and Holmes, C. Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *International Conference on Machine Learning*, pp. 1952–1962. PMLR, 2019.

Garreau, D., Jitkrittum, W., and Kanagawa, M. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Gershman, S., Hoffman, M., and Blei, D. Nonparametric variational inference. *arXiv preprint arXiv:1206.4665*, 2012.

Gorham, J. and Mackey, L. Measuring sample quality with stein's method. *Advances in Neural Information Processing Systems*, 28, 2015.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pp. 1292–1301. PMLR, 2017.

Green, P., Łatuszyński, K., Pereyra, M., and Robert, C. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25:835–862, 2015.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Haario, H., Saksman, E., and Tamminen, J. Adaptive proposal distribution for random walk metropolis algorithm. *Computational statistics*, 14:375–395, 1999.

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. Kernel stein discrepancy descent. In *International Conference on Machine Learning*, pp. 5719–5730. PMLR, 2021.

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

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

Oates, C., Barp, A., Girolami, M., et al. Posterior integration on a riemannian manifold. *arXiv preprint arXiv:1712.01793*, 2017.

Pompe, E., Holmes, C., and Łatuszyński, K. A framework for adaptive mcmc targeting multimodal distributions. *The Annals of Statistics*, 48:2930–2952, 2020.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S., Mackey, L., and Oates, C. Optimal thinning of MCMC output. *Journal of the Royal Statistical Society: Series B*, in press, 2022.

Robert, C. and Casella, G. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pp. 2263–2291, 2013.

South, L., Riabiz, M., Teymur, O., and Oates, C. Postprocessing of mcmc. *Annual Review of Statistics and Its Application*, 9:529–555, 2022.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pp. 583–603. University of California Press, 1972.

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

Supplementary Material for Regularized Stein thinning

A. Additional Experiments

A.1. Illustration of Theorem 2.3

To better illustrate Theorem 2.3, we run an additional experiment, where p is still defined as in Figure 1 from Example 1, with unbalanced mode weights of 0.2 and 0.8. The density q is distributed as p, but each mode is truncated outside a circle of two standard deviation radius, and q has weight w. Next, for various values of $w \in [0.1, 0.9]$, we draw two samples of size $n = 3000$ from p and q, and compute $\mathrm{KSD}(\mathbb{P}, \mathbb{Q}_w)$ (with 30 repetitions for each w value). The result is displayed in Figure 7, and shows that the optimal weight is close to $1/2$, as predicted by Theorem 2.3, since $|\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_L)/\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - 1|$ is estimated as 0.01 in this case.

A.2. Gaussian and Banana-shaped Mixtures

This appendix gathers additional results and details for the Gaussian and banana-shaped mixtures, as well as the MMD distance used to evaluate thinning performance, and the regularization parameter λ.

Maximum Mean Discrepancy. When the target distribution is known, the efficiency of the Stein thinning algorithms are assessed by computing the MMD distance between a sample drawn from the target distribution and the thinned samples. The kernel function in Equation (1) is chosen as the distance-induced kernel studied by Sejdinovic et al. (2013). More specifically, we use the following closed-form expression of the MMD (Gretton et al., 2006) with $\mathbf{Z} \sim \mathbb{P}$ and $\mathbf{Z}' \sim \mathbb{Q}$,

$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}[k(\mathbf{Z}, \mathbf{Z})] + \mathbb{E}[k(\mathbf{Z}', \mathbf{Z}')] - 2\mathbb{E}[k(\mathbf{Z}, \mathbf{Z}')]\,, \tag{3}$$

where $k(\mathbf{X}, \mathbf{X}') = \|\mathbf{X}\|_2 + \|\mathbf{X}'\|_2 - \|\mathbf{X} - \mathbf{X}'\|_2$. In this setting, the MMD reduces to the well known energy distance, as shown by Sejdinovic et al. (2013).

Gaussian mixture. The first batch of experiments in Section 4 considers a d-dimensional Gaussian mixture of four modes of equal weight, with $d \geq 2$, illustrated in Figure 8. The center of modes are chosen as $(-1.6, 0)$, $(1.6, 0)$, $(-3, 4)$, and $(3, 4)$, and null values for the higher dimension coordinates. The first two modes have an identity covariance matrix, while the remaining two modes have a diagonal covariance matrix with variance equal to 4. The results for regularized Stein thinning and the original Stein thinning are provided in Figure 9 for MALA sampler, and in Figure 10 for NUTS sampler. In both figures, the three tested step size ε are displayed, with a small impact on the resulting performance.

Regularization parameter λ. Figure 11 displays the MMD obtained with regularization parameters λ, set as $\lambda = 1/m^2$ and $\lambda = 1/\log(m)$. These results should be compared with the ones shown in Figures 9 and 10, which were obtained with a regularization parameter $\lambda = 1/m$. These additional experiments show the importance of choosing the regularization parameter such that $\lambda = o(\log(m)/m)$, as suggested by Theorem 3.5, and that faster rates than $\lambda = 1/m$ tend to remove the effect of the entropic regularization.

Banana-shaped mixture with t-tails. The second batch of experiments in Section 4 considers a banana-shaped mixture with t-tails, defined as follows. Let $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ be the transformation defined by $\varphi_i(x) = x_i$ if $i \neq 2$, and $\varphi_2(x) = x_2 + bx_1^2 - 100b$. Let $\mathbf{Z}$ be a random variable that follows the multivariate t-Student distribution with degrees of freedom 7. Then, the random variable $\mathbf{X} = \varphi(\mathbf{Z}) + \boldsymbol{\mu}$ follows a t-banana-shaped distribution centered at $\boldsymbol{\mu}$. We consider a mixture of two t-banana-shaped distributions with weights $w_1 = 0.25$ and $w_2 = 0.75$, respectively, which is illustrated in Figure 12 for $d = 2$. The results for regularized Stein thinning and the original Stein thinning are provided in Figure 13 for MALA sampler, and in Figure 14 for NUTS sampler. In both figures, the three tested step size ε are displayed, with a quite small impact on the resulting performance.

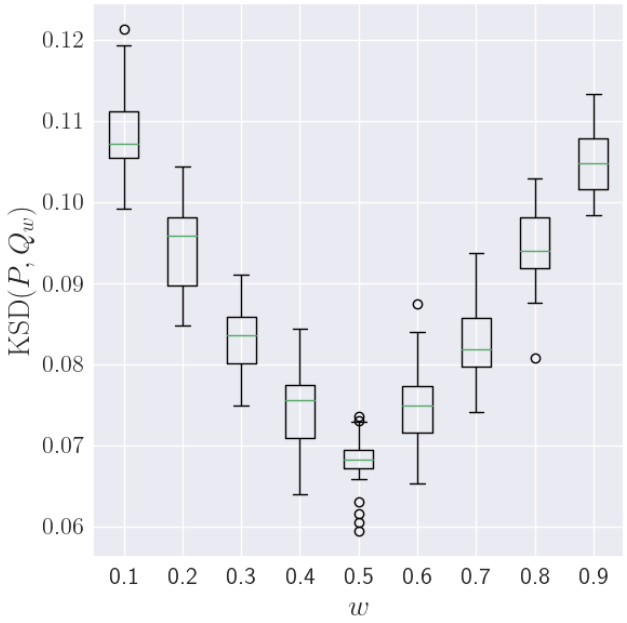

Figure 7. KSD$(\mathbb{P}, \mathbb{Q}_w)$ for p as defined in Example 1 with $\mu = 3$, $\sigma = 1$, $w_p = 0.2$, and q a truncation of p and with weight w. The KSD is estimated with $n = 3000$ and 30 repetitions for each w value.

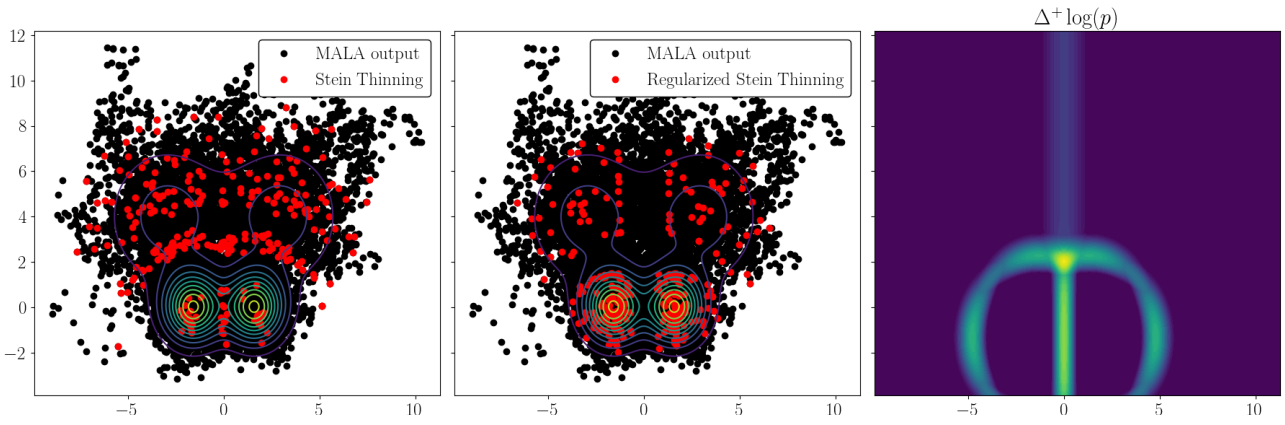

Figure 8. (Gaussian mixture) First two panels: solutions obtained with Stein thinning and regularized Stein thinning with contour lines of the target distribution. Last panel: heatmap of the Laplace correction $\Delta^+ \log(p)$.

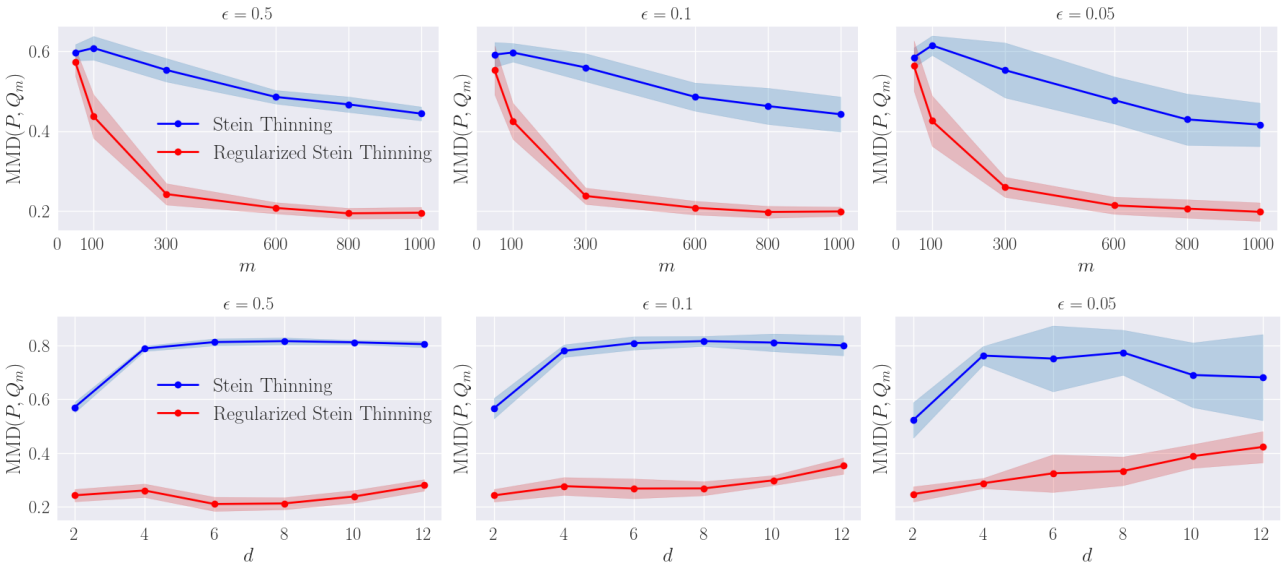

Figure 9. (Gaussian mixture, MALA) Graphs of the MMD distance with respect to the thinning size m (with $d = 2$) and with respect to d (with $m = 300$) for various step sizes ε.

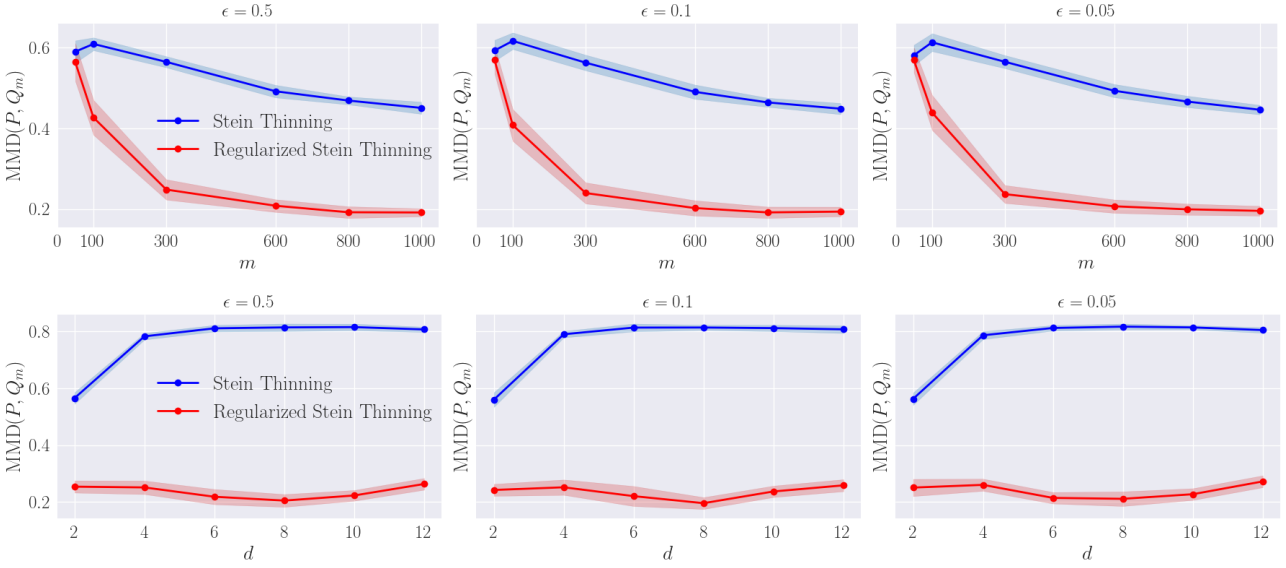

Figure 10. (Gaussian mixture, NUTS) Graphs of the MMD distance with respect to the thinning size m (with $d = 2$) and with respect to d (with $m = 300$) for various step sizes ε.

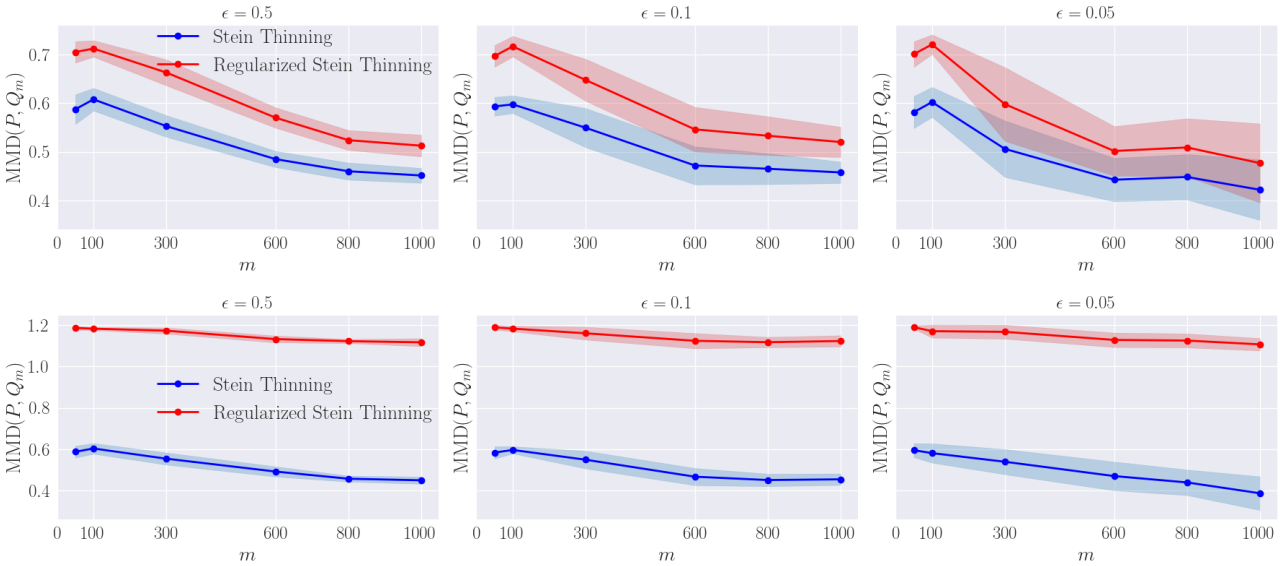

Figure 11. (Gaussian mixture, MALA) Graphs of the MMD distance with respect to the thinning size m (with $d = 2$) and with respect to d (with $m = 300$) for various step sizes ε. For the first row, we set $\lambda = 1/m^2$, and we observe that the effect of entropic regularization almost vanishes. For the second row, we set $\lambda = 1/\log(m)$, violating the convergence assumption, and resulting in bad thinned samples.

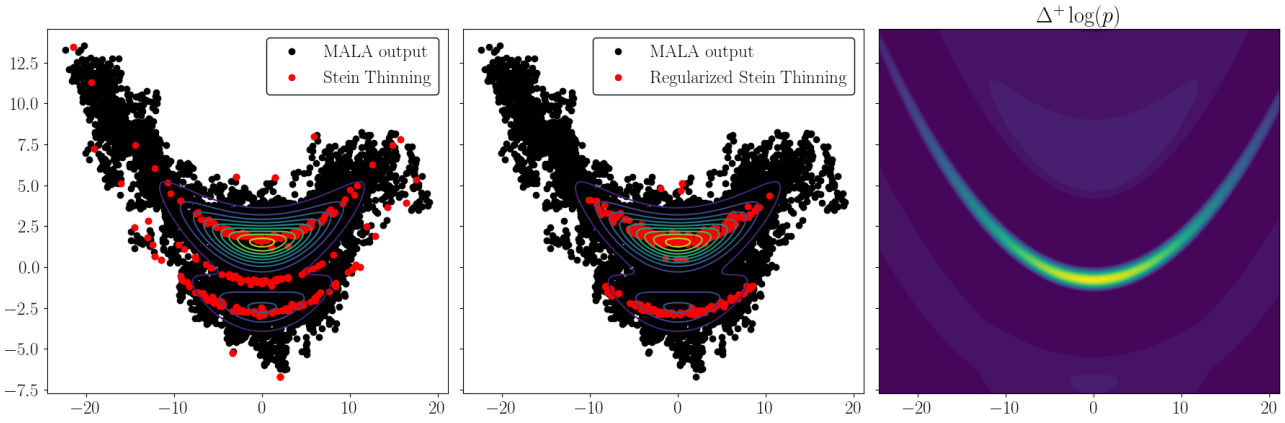

Figure 12. (t banana-shaped mixture) First two panels: solutions obtained with Stein thinning and regularized Stein thinning with contour lines of the target distribution. Last panel: heatmap of the Laplace correction $\Delta^+ \log(p)$.

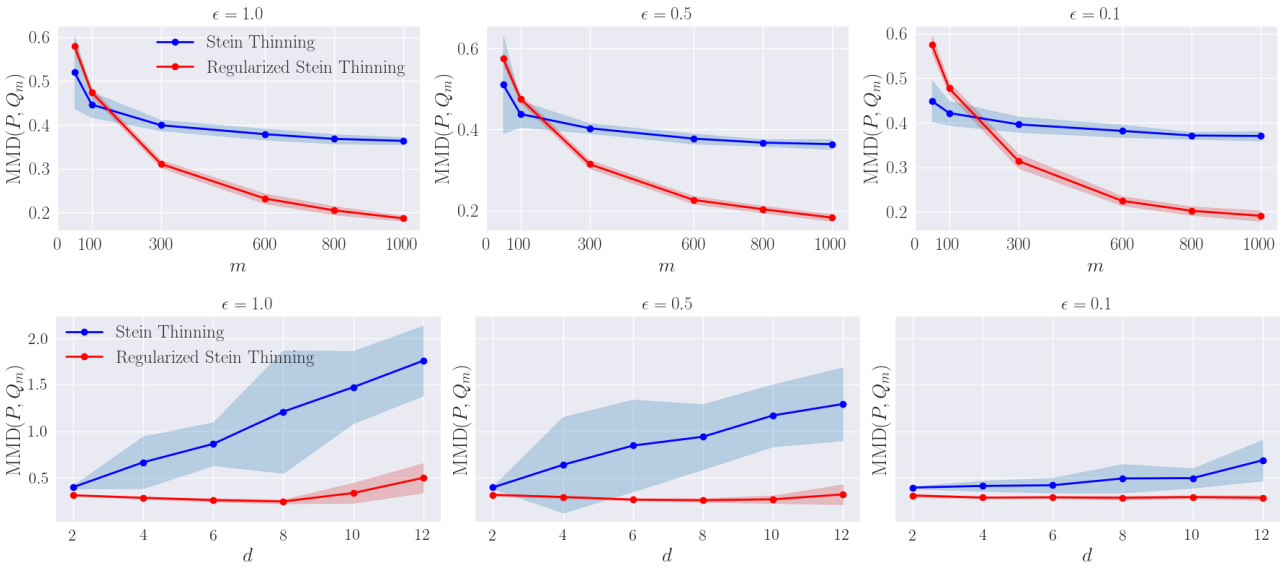

Figure 13. (Mixture of t-banana-shaped distributions, MALA) Graphs of the MMD distance with respect to the thinning size m (with $d = 2$) and with respect to d (with $m = 300$) for various step sizes ε.

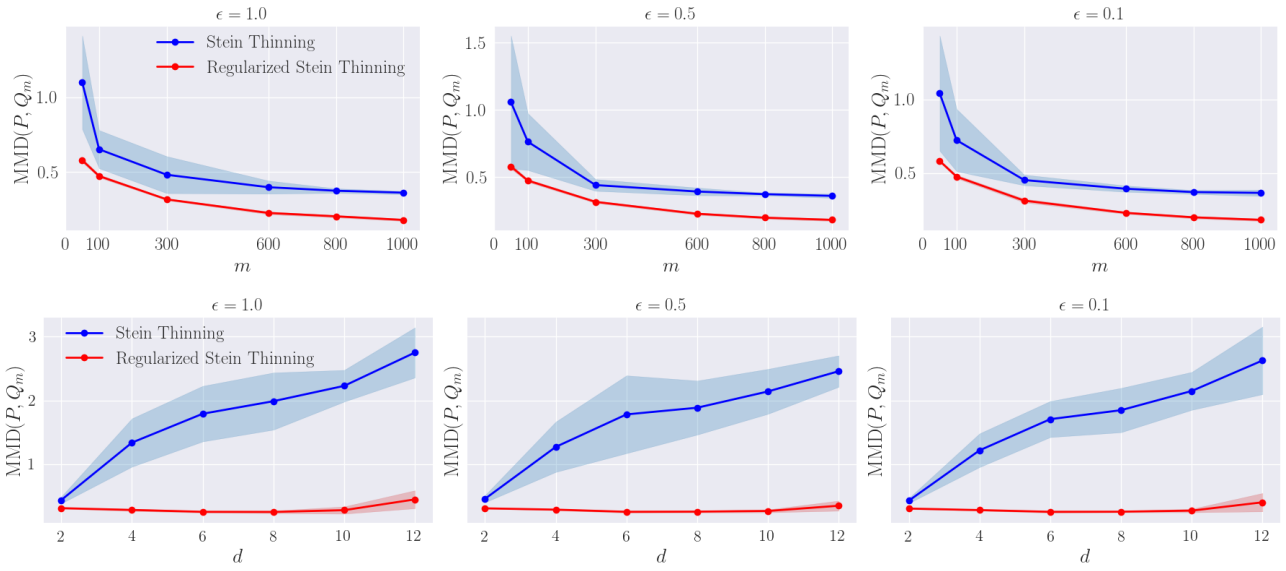

Figure 14. (Mixture of t-banana-shaped distributions, NUTS) Graphs of the MMD distance with respect to the thinning size m (with $d = 2$) and with respect to d (with $m = 300$) for various step sizes ε.

Table 2. Description of UCI datasets

Dataset	Sample size	Dimension
Breast Wisconsin	569	30
Diabetes	768	8
Haberman	306	3
Liver Disorders	345	6
Sonar	208	60

Table 3. AUCs obtained by Stein Thinning (ST) and Regularized Stein Thinning (RST). A 10-fold cross-validation is performed and the experiments are repeated 10 times to provide uncertainties.

NUTS Sampler						
	$m = 50$		$m = 100$		$m = 300$	
Dataset	ST	RST	ST	RST	ST	RST
Breast W.	0.88 (0.020)	**0.96** (0.004)	0.91 (0.023)	**0.96** (0.003)	0.93 (0.008)	**0.96** (0.004)
Diabetes	**0.52** (0.009)	0.50 (0.019)	0.52 (0.021)	**0.55** (0.018)	0.53 (0.015)	**0.57** (0.019)
Haberman	0.51 (0.038)	**0.53** (0.023)	0.54 (0.033)	**0.58** (0.035)	0.53 (0.034)	**0.58** (0.017)
Liver	0.53 (0.044)	**0.69** (0.014)	0.56 (0.038)	**0.70** (0.013)	0.61 (0.039)	**0.70** (0.011)
Sonar	0.80 (0.021)	**0.81** (0.007)	0.81 (0.009)	**0.82** (0.011)	0.81 (0.011)	0.81 (0.009)
MALA Sampler						
	$m = 50$		$m = 100$		$m = 300$	
Dataset	ST	RST	ST	RST	ST	RST
Breast W.	0.68 (0.044)	**0.93** (0.010)	0.72 (0.048)	**0.93** (0.007)	0.72 (0.037)	**0.88** (0.026)
Diabetes	**0.51** (0.012)	0.48 (0.010)	**0.53** (0.028)	0.51 (0.014)	0.53 (0.016)	**0.56** (0.016)
Haberman	0.52 (0.034)	**0.60** (0.027)	0.53 (0.024)	**0.58** (0.017)	0.55 (0.024)	**0.61** (0.013)
Liver	0.54 (0.033)	**0.70** (0.008)	0.55 (0.034)	**0.69** (0.005)	0.57 (0.024)	**0.62** (0.032)
Sonar	0.80 (0.019)	0.80 (0.019)	0.80 (0.010)	0.80 (0.010)	**0.81** (0.013)	0.80 (0.010)

A.3. Bayesian Logistic Regression

This appendix gathers additional results for Bayesian logistic regression. In particular, Table 2 provides a description of the tested datasets. Table 3 gives the resulting AUC, for $m = 50, 100, 300$, using NUTS or MALA sampler. We recall that only the best AUC over the four tested MCMC step size ε is reported.

B. Proof of Theorem 2.3

Lemma B.1. *Let k_p be the Stein kernel associated with the multi-quadratic kernel for $c = 1$, $\beta = 1/2$, and $\ell > 0$. For $\delta > 0$, if $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, such that $\|\mathbf{x} - \mathbf{y}\|_2 > \delta$, $s_p(\mathbf{x}) < s_0$ and $s_p(\mathbf{y}) < s_0$, we have*

$$|k_p(\mathbf{x}, \mathbf{y})| < \frac{s_0^2 \ell}{\delta} + \frac{2 s_0 \ell}{\delta^2} + \frac{(d+3)\ell}{\delta^3}.$$

Proof of Theorem 2.3. We consider the mixture distributions p and q satisfying Assumption 2.1, for $\mu > 0$ and $r > 0$, and assume that Assumption 2.2 is satisfied for $\eta \in (0, 1)$. More precisely, we denote by q_L the distribution of the left mode of the mixture q, and similarly, q_R is the distribution of the right mode of q. The probability measures $\mathbb{Q}_L$ and $\mathbb{Q}_R$ respectively admits the densities q_L and q_R.

By definition of the KSD, we can write

$$\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_w) = \int k_p(\mathbf{x}, \mathbf{x}') q(\mathbf{x}) q(\mathbf{x}') d\mathbf{x} d\mathbf{x}'.$$

Additionally, given the above notations, q takes the form $q = w q_L + (1 - w) q_R$. Then, we can develop the KSD expression

to get

$$\begin{aligned}\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_w) =& \int k_p(\mathbf{x}, \mathbf{x}')(wq_L(\mathbf{x}) + (1-w)q_R(\mathbf{x}))(wq_L(\mathbf{x}') + (1-w)q_R(\mathbf{x}'))d\mathbf{x}d\mathbf{x}' \\ =& w^2 \int k_p(\mathbf{x}, \mathbf{x}')q_L(\mathbf{x})q_L(\mathbf{x}')d\mathbf{x}d\mathbf{x}' + (1-w)^2 \int k_p(\mathbf{x}, \mathbf{x}')q_R(\mathbf{x})q_R(\mathbf{x}')d\mathbf{x}d\mathbf{x}' \\ &+ 2w(1-w) \int k_p(\mathbf{x}, \mathbf{x}')q_L(\mathbf{x})q_R(\mathbf{x}')d\mathbf{x}d\mathbf{x}',\end{aligned}$$

where the last term follows from the symmetry of k_p. Finally, we have

$$\begin{aligned}\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_w) =& w^2\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_L) + (1-w)^2\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_R) \\ &+ 2w(1-w) \int k_p(\mathbf{x}, \mathbf{x}')q_L(\mathbf{x})q_R(\mathbf{x}')d\mathbf{x}d\mathbf{x}',\end{aligned}$$

and we denote by $\Delta_{L,R}$ the last term of this equation, which now writes

$$\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_w) = w^2\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_L) + (1-w)^2\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_R) + 2w(1-w)\Delta_{L,R}. \tag{4}$$

We first focus on the last term $\Delta_{L,R}$ of this expression, which can be shown to be arbitrarily small when μ gets large. According to Assumption 2.1, the distance between the centers of the two modes is 2μ, and both q_L and q_R have a compact support included in a ball of radius r. Consequently, for $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ such that $q_L(\mathbf{x}) > 0$ and $q_R(\mathbf{x}') > 0$, then $\|\mathbf{x} - \mathbf{x}'\|_2 > 2(\mu - r)$. Additionally, since the score s_p is continuous, s_p is bounded on a compact set, and it exists $s_0 > 0$ such that $s_p(\mathbf{x}) < s_0$ and $s_p(\mathbf{x}') < s_0$. Then, from Lemma B.1, we have

$$|k_p(\mathbf{x}, \mathbf{x}')| < \frac{s_0^2\ell}{2(\mu - r)} + \frac{2s_0\ell}{4(\mu - r)^2} + \frac{(d+3)\ell}{8(\mu - r)^3}.$$

Additionally, since ℓ is chosen with the median heuristic, and $w_p \neq 1/2$, then $\ell < r$, which gives

$$|k_p(\mathbf{x}, \mathbf{x}')| < \frac{s_0^2 r}{2(\mu - r)} + \frac{2s_0 r}{4(\mu - r)^2} + \frac{(d+3)r}{8(\mu - r)^3}.$$

Finally, we obtain the following upper bound

$$|\Delta_{L,R}| < \frac{s_0^2 r}{2(\mu - r)} + \frac{2s_0 r}{4(\mu - r)^2} + \frac{(d+3)r}{8(\mu - r)^3}.$$

Now, it is clear that

$$\lim_{\mu \to \infty} \Delta_{L,R} = 0.$$

Next, we reorder the terms of equation (4) to get a second-order polynomial in w as follows

$$\begin{aligned}\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_w) =& w^2\big[\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_L) + \mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - 2\Delta_{L,R}\big] \\ &- 2w\big[\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - \Delta_{L,R}\big] + \mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_R).\end{aligned}$$

Notice that the coefficient of w^2 is $\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_{1/2})/4$, and is therefore positive. Then, $\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_w)$ admits a unique minimum with respect to w, given by

$$w^\star = \frac{\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - \Delta_{L,R}}{\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_L) + \mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - 2\Delta_{L,R}}.$$

We rewrite $w^\star$ as follows,

$$\begin{aligned}
w^\star &= \frac{1/2\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) + 1/2\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L) - \Delta_{L,R} + 1/2\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) - 1/2\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)}{\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L) + \mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) - 2\Delta_{L,R}} \\
&= \frac{1}{2} + \frac{1}{2}\frac{\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) - \mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)}{\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L) + \mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) - 2\Delta_{L,R}} \\
&= \frac{1}{2} + \frac{1}{2}\frac{1 - \mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)}{1 + \mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) - 2\Delta_{L,R}/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)} \\
&= \frac{1}{2} + \frac{1}{2}\frac{1 - \mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)}{2(1 - \Delta_{L,R}/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)) + (\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) - 1)}.
\end{aligned}$$

We can deduce the following bound

$$\left| w^\star - \frac{1}{2} \right| \leq \frac{1}{2}\frac{|\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) - 1|}{|2(1 - \Delta_{L,R}/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)) + (\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) - 1)|}. \tag{5}$$

According to Assumption 2.2, with $0 < \eta < 1$,

$$\left| \frac{\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)}{\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)} - 1 \right| < \eta,$$

which gives an upper bound for the numerator of the right hand side of inequality (5). Additionally, for μ large enough, $\Delta_{R,L}$ is arbitrarily small, and in particular, we can have $\Delta_{R,L} < \mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)/2$, and then $2(1 - \Delta_{L,R}/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)) > 1$. Next, we use the triangle inequality to get

$$\begin{aligned}
&|2(1 - \Delta_{L,R}/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)) + (\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) - 1)| \\
&\qquad \geq 2(1 - \Delta_{L,R}/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)) - |\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_L)/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R) - 1| \\
&\qquad \geq 1 - \eta,
\end{aligned}$$

where the last inequality is obtained using $2(1 - \Delta_{L,R}/\mathrm{KSD}^2(\mathbb{P},\mathbb{Q}_R)) > 1$ for μ large enough, and Assumption 2.2 again. Finally, this lower bound on the denominator and the upper bound on the numerator combined with inequality (5) give

$$\left| w^\star - \frac{1}{2} \right| < \frac{\eta}{2(1-\eta)}.$$

□

Proof of Lemma B.1. We consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, such that $\mathbf{x} \neq \mathbf{y}$, $s_p(\mathbf{x}) < s_0$ and $s_p(\mathbf{y}) < s_0$. The Stein kernel obtained for the inverse multi-quadratic kernel function is defined by

$$\begin{aligned}
k_p(\mathbf{x},\mathbf{y}) = &-\frac{3}{\ell^4}\|\mathbf{x}-\mathbf{y}\|_2^2(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-5/2} \\
&+ \frac{1}{\ell^2}(d + (s_p(\mathbf{x}) - s_p(\mathbf{y})) \cdot (\mathbf{x}-\mathbf{y}))(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-3/2} \\
&+ (s_p(\mathbf{x}) \cdot s_p(\mathbf{y}))(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-1/2}.
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
\frac{3}{\ell^4}\|\mathbf{x}-\mathbf{y}\|_2^2(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-5/2} &< \frac{3}{\ell^4}\|\mathbf{x}-\mathbf{y}\|_2^2(\|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-5/2} \\
&< 3\ell\|\mathbf{x}-\mathbf{y}\|_2^{-3}.
\end{aligned}$$

For the second term, we first have

$$d/\ell^2(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-3/2} < d/\ell^2(\|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-3/2} < d\ell\|\mathbf{x}-\mathbf{y}\|_2^{-3}.$$

For the second part of the second term, we use Cauchy-Schwartz inequality to get

$$\big|(s_p(\mathbf{x}) - s_p(\mathbf{y})) \cdot (\mathbf{x} - \mathbf{y})\big| \leq \|(s_p(\mathbf{x}) - s_p(\mathbf{y}))\|_2 \|\mathbf{x} - \mathbf{y}\|_2 \leq 2s_0 \|\mathbf{x} - \mathbf{y}\|_2,$$

and then

$$\begin{aligned}\frac{1}{\ell^2}|(s_p(\mathbf{x}) - s_p(\mathbf{y})) \cdot (\mathbf{x} - \mathbf{y})|(1 + \|\mathbf{x} - \mathbf{y}\|_2^2/\ell^2)^{-3/2} \\ < 2s_0/\ell^2 \|\mathbf{x} - \mathbf{y}\|_2 (\|\mathbf{x} - \mathbf{y}\|_2^2/\ell^2)^{-3/2} \\ < 2s_0 \ell \|\mathbf{x} - \mathbf{y}\|_2^{-2}.\end{aligned}$$

Finally, for the last term, we have

$$|s_p(\mathbf{x}) \cdot s_p(\mathbf{y})|(1 + \|\mathbf{x} - \mathbf{y}\|_2^2/\ell^2)^{-1/2} < s_0^2 \ell \|\mathbf{x} - \mathbf{y}\|_2^{-1}.$$

Overall, we use the triangle inequality and all the previous inequalities to upper k_p as follows

$$|k_p(\mathbf{x}, \mathbf{y})| < 3\ell \|\mathbf{x} - \mathbf{y}\|_2^{-3} + d\ell \|\mathbf{x} - \mathbf{y}\|_2^{-3} + 2s_0 \ell \|\mathbf{x} - \mathbf{y}\|_2^{-2} + s_0^2 \ell \|\mathbf{x} - \mathbf{y}\|_2^{-1}.$$

Finally, if $\|\mathbf{x} - \mathbf{y}\|_2 > \delta$, we have

$$|k_p(\mathbf{x}, \mathbf{y})| < \frac{s_0^2 \ell}{\delta} + \frac{2s_0 \ell}{\delta^2} + \frac{(d+3)\ell}{\delta^3}.$$

□

C. Proofs of Theorem 2.4, Corollary 2.5, and Corollary 2.6

Lemma C.1. *Let k_p be the Stein kernel associated to the IMQ kernel, with parameters $c = 1$, $\beta = 1/2$, and $\ell > 0$. For $s_0 \geq \min_{\mathbf{x} \in \mathbb{R}^d} \|s(\mathbf{x})\|_2$, and $\mathbf{x}, \mathbf{y} \in \mathcal{M}_{s_0}$, we have*

$$-\frac{c_0}{\ell^2} - \frac{c_1 s_0}{\ell} - s_0^2 \leq k_p(\mathbf{x}, \mathbf{y}) \leq \frac{d}{\ell^2} + \frac{c_1 s_0}{\ell} + s_0^2,$$

where $c_0 = 2\left(\frac{3}{5}\right)^{5/2} \approx 0.56$, and $c_1 = \frac{4}{3^{3/2}} \approx 0.77$.

Proof of Theorem 2.4. By definition of the kernelized Stein discrepancy between the target distribution $\mathbb{P}$ and the empirical measure $\mathbb{P}_m = \frac{1}{m}\sum_{i=1}^m \delta(\mathbf{X}_i)$, we have

$$\mathbb{E}[\mathrm{KSD}^2(\mathbb{P}, \mathbb{P}_m)] = \frac{1}{m^2} \sum_{i,j=1}^m \mathbb{E}[k_p(\mathbf{X}_i, \mathbf{X}_j)]$$

In what follows, k_p denotes the Stein kernel obtained for the inverse multi-quadratric kernel function, *i.e.*,

$$\begin{aligned}k_p(\mathbf{x}, \mathbf{y}) = &-\frac{3}{\ell^4}\|\mathbf{x} - \mathbf{y}\|_2^2 (1 + \|\mathbf{x} - \mathbf{y}\|_2^2/\ell^2)^{-5/2} \\ &+ \frac{1}{\ell^2}(d + (s_p(\mathbf{x}) - s_p(\mathbf{y})) \cdot (\mathbf{x} - \mathbf{y}))(1 + \|\mathbf{x} - \mathbf{y}\|_2^2/\ell^2)^{-3/2} \\ &+ (s_p(\mathbf{x}) \cdot s_p(\mathbf{y}))(1 + \|\mathbf{x} - \mathbf{y}\|_2^2/\ell^2)^{-1/2}.\end{aligned}$$

Given that $\mathbf{X}_i$ and $\mathbf{X}_j$ are independent random variables that follow the distribution $\mathbb{P}$, one has $\mathbb{E}[k_p(\mathbf{X}_i, \mathbf{X}_j)] = 0$ for any $i \neq j$. Using this property and the closed-form expression of the Stein kernel k_p, it is found that

$$\begin{aligned}\mathbb{E}[\mathrm{KSD}^2(\mathbb{P}, \mathbb{P}_m)] &= \frac{1}{m^2} \sum_{i=1}^m \mathbb{E}[k_p(\mathbf{X}_i, \mathbf{X}_i)] \\ &= \frac{1}{m} \mathbb{E}[k_p(\mathbf{X}, \mathbf{X})] \\ &= \frac{d}{m\ell^2} + \frac{\mathbb{E}[\|s_p(\mathbf{X})\|_2^2]}{m},\end{aligned}$$

where $\mathbf{X} \sim \mathbb{P}$. On the other hand, the kernelized Stein discrepancy between the target $\mathbb{P}$ and the empirical distribution $\mathbb{Q}_m = \frac{1}{m}\sum_{i=1}^m \delta(\mathbf{x}_i)$ is given by

$$\begin{aligned}\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_m) &= \frac{1}{m^2}\sum_{i,j=1}^m k_p(\mathbf{x}_i, \mathbf{x}_j) \\ &= \frac{1}{m^2}\sum_{i=1}^m k_p(\mathbf{x}_i, \mathbf{x}_i) + \frac{1}{m^2}\sum_{i\neq j}^m k_p(\mathbf{x}_i, \mathbf{x}_j) \\ &= \frac{d}{m\ell^2} + \frac{1}{m^2}\sum_{i=1}^m \|s_p(\mathbf{x}_i)\|_2^2 + \frac{1}{m^2}\sum_{i\neq j}^m k_p(\mathbf{x}_i, \mathbf{x}_j).\end{aligned}$$

Next, it can be shown that the difference $m\big(\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_m) - \mathbb{E}[\mathrm{KSD}^2(\mathbb{P}, \mathbb{P}_m)]\big)$ takes the form

$$m\big(\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_m) - \mathbb{E}[\mathrm{KSD}^2(\mathbb{P}, \mathbb{P}_m)]\big) = \frac{1}{m}\sum_{i=1}^m \|s_p(\mathbf{x}_i)\|_2^2 + \frac{1}{m}\sum_{i\neq j}^m k_p(\mathbf{x}_i, \mathbf{x}_j) - \mathbb{E}[\|s_p(\mathbf{X})\|_2^2].$$

Since $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M}_{s_0}$, we can use Lemma C.1 to bound the terms $k_p(\mathbf{x}_i, \mathbf{x}_j)$, and then obtain

$$m\big(\mathrm{KSD}^2(\mathbb{P}, \mathbb{Q}_m) - \mathbb{E}[\mathrm{KSD}^2(\mathbb{P}, \mathbb{P}_m)]\big) \leq s_0^2 + (m-1)\Big(\frac{s_0 c_1}{\ell} + \frac{d}{\ell^2} + s_0^2\Big) - \mathbb{E}[\|s_p(\mathbf{X})\|_2^2],$$

where the right hand side is always negative if

$$m < 1 + \frac{\mathbb{E}[\|s_p(\mathbf{X})\|_2^2] - s_0^2}{d/\ell^2 + c_1 s_0/\ell + s_0^2},$$

which provides the advertised result, since $c_1 < 1$. □

Proof of Lemma C.1. The Stein kernel k_p obtained for the inverse multi-quadratric kernel function, with parameters $c = 1$, $\beta = 1/2$, and $\ell > 0$ is given by

$$\begin{aligned}k_p(\mathbf{x}, \mathbf{y}) = &-\frac{3}{\ell^4}\|\mathbf{x}-\mathbf{y}\|_2^2(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-5/2} \\ &+ \frac{1}{\ell^2}(d + (s_p(\mathbf{x}) - s_p(\mathbf{y}))\cdot(\mathbf{x}-\mathbf{y}))(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-3/2} \\ &+ (s_p(\mathbf{x})\cdot s_p(\mathbf{y}))(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-1/2}.\end{aligned}$$

Since the first term is always negative, we obtain

$$\begin{aligned}k_p(\mathbf{x}, \mathbf{y}) \leq &\frac{d}{\ell^2} + \frac{1}{\ell^2}|(s_p(\mathbf{x}) - s_p(\mathbf{y}))\cdot(\mathbf{x}-\mathbf{y})|(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-3/2} \\ &+ |(s_p(\mathbf{x})\cdot s_p(\mathbf{y}))|(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-1/2}.\end{aligned}$$

We define the function g_1 for $z \geq 0$ as

$$g_1(z) = \frac{z}{(1+z^2)^{3/2}},$$

and a simple function analysis shows that

$$\frac{c_1}{2} \overset{\text{def}}{=} \sup_{z\geq 0} g_1(z) = \frac{2}{3^{3/2}} \approx 0.385.$$

Additionally, we clearly have $(1 + \|\mathbf{x}-\mathbf{y}\|_2^2/\ell^2)^{-1/2} \leq 1$, and combining these last two results, we get

$$k_p(\mathbf{x}, \mathbf{y}) \leq \frac{d}{\ell^2} + \frac{c_1}{2\ell}\Big|(s_p(\mathbf{x}) - s_p(\mathbf{y}))\cdot\frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|_2}\Big| + |(s_p(\mathbf{x})\cdot s_p(\mathbf{y}))|.$$

We can apply Cauchy-Schwartz inequality, and since $\mathbf{x}, \mathbf{y} \in \mathcal{M}_{s_0}$, we get

$$\left|(s_p(\mathbf{x}) - s_p(\mathbf{y})) \cdot \frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|_2}\right| \leq \|(s_p(\mathbf{x}) - s_p(\mathbf{y}))\|_2 \frac{\|\mathbf{x}-\mathbf{y}\|_2}{\|\mathbf{x}-\mathbf{y}\|_2} \leq 2s_0,$$

and also

$$|(s_p(\mathbf{x}) \cdot s_p(\mathbf{y}))| \leq s_0^2.$$

Overall, we obtain

$$k_p(\mathbf{x}, \mathbf{y}) \leq \frac{d}{\ell^2} + \frac{c_1 s_0}{\ell} + s_0^2.$$

For the lower bound, we first define

$$g_0(z) = \frac{3z^2}{(1+z^2)^{5/2}},$$

and a simple function analysis shows that

$$c_0 \stackrel{\text{def}}{=} \sup_{z \geq 0} g_0(z) = 2\Big(\frac{3}{5}\Big)^{5/2} \approx 0.558.$$

Similarly to the upper bound case, we finally get

$$k_p(\mathbf{x}, \mathbf{y}) \geq -\frac{c_0}{\ell^2} - \frac{c_1 s_0}{\ell} - s_0^2.$$

□

Proof of Corollary 2.5. As minimum and saddle points are stationary points of p, we have

$$\{\mathbf{x}_i\}_{i=1}^m \subset \mathcal{M}_0,$$

and we can apply Theorem 2.4 for $s_0 = 0$ to get the final result. □

Proof of Corollary 2.6. Let the density p be a Gaussian mixture model of two components with equal weights, respectively centered in $(-\mu, \mathbf{0}_{d-1})$ and $(\mu, \mathbf{0}_{d-1})$, and of variance $\sigma^2 \mathbf{I_d}$, and let $\nu = \mu/\sigma$. We assume that $\nu > 1$ and $0 \leq s_0 < \big[\nu\sqrt{\nu^2-1} - \ln(\nu + \sqrt{\nu^2-1})\big]/\mu$. Then, according to Theorem 2.4, for any $\{\mathbf{x}_i\}_{i=1}^m \subset \mathcal{M}_{s_0}$ of empirical measure $Q_m = \frac{1}{m}\sum_{i=1}^m \delta(\mathbf{x}_i)$, we have

(i) $\mathrm{KSD}^2\big(\mathbb{P}, \mathbb{Q}_m\big) < \mathbb{E}[\mathrm{KSD}^2\big(\mathbb{P}, \mathbb{P}_m\big)]$ if m and s_0 satisfy $m < 1 + \frac{\mathbb{E}[\|s_p(\mathbf{X})\|_2^2] - s_0^2}{d/\ell^2 + s_0/\ell + s_0^2}$.

To prove statement (ii), we need to characterize the shape of the set $\mathcal{M}_{s_0} \subset \mathbb{R}^d$, given by the level lines of the squared score norm $\|s_p(\mathbf{x})\|_2^2$. The density p is a Gaussian mixture, i.e.,

$$p(\mathbf{x}) = \frac{1}{2(2\pi)^{d/2}\sigma^d} e^{-\|\mathbf{x}^{(-1)}\|_2^2/2\sigma^2}\big(e^{-(x^{(1)}+\mu)^2/2\sigma^2} + e^{-(x^{(1)}-\mu)^2/2\sigma^2}\big),$$

where $\mathbf{x}^{(-1)}$ is the vector $\mathbf{x}$ without the first component. Then, the score is also given by an explicit formula,

$$s_p(\mathbf{x}) = \begin{pmatrix} -\frac{x^{(1)}}{\sigma^2} + \frac{\mu}{\sigma^2}\tanh(\frac{\mu}{\sigma^2}x^{(1)}) \\ -\frac{\mathbf{x}^{(-1)}}{\sigma^2} \end{pmatrix},$$

where $\tanh$ is the standard hyperbolic tangent function. An important property of this score function is that the j-th component of s_p only depends on $x^{(j)}$, which makes $s_p^{(j)}(\mathbf{x})$ invariant by any translation orthogonal to the j-th axis. Then, we can compute the squared score norm

$$\|s_p(\mathbf{x})\|_2^2 = s_p^{(1)}(x^{(1)})^2 + \frac{\|\mathbf{x}^{(-1)}\|_2^2}{\sigma^4},$$

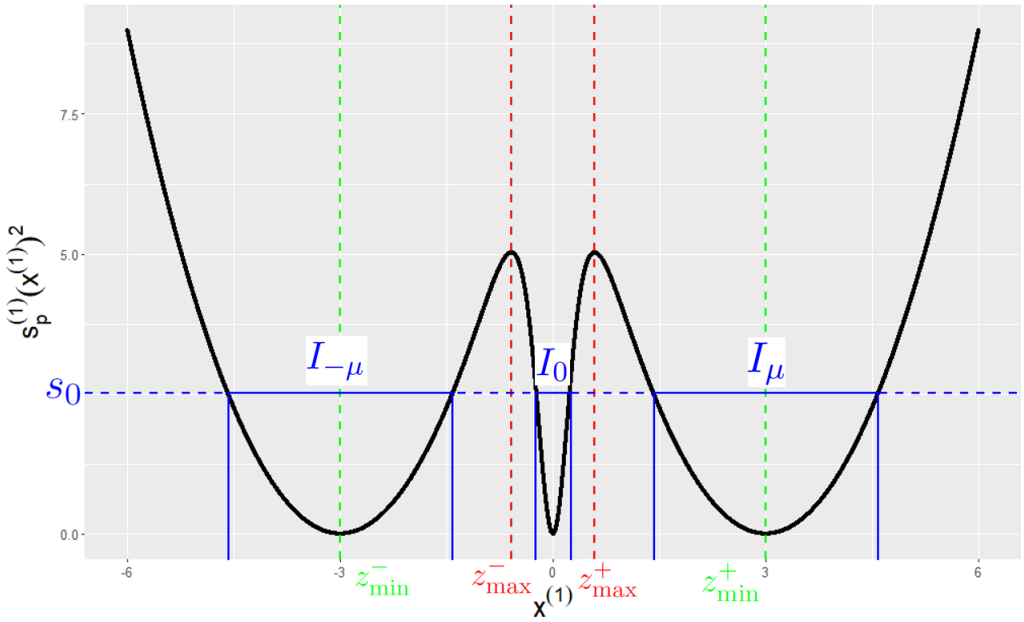

Figure 15. Squared first component of the score for a Gaussian mixture with $\mu = 3$ and $\sigma = 1$.

where $s_p^{(1)}(z)^2 = \left(\frac{z}{\sigma^2} - \frac{\mu}{\sigma^2}\tanh(\frac{\mu}{\sigma^2}z)\right)^2$. A simple function analysis of this univariate function, illustrated in Figure 15, shows that $s_p^{(1)}(z)^2$ has two local maximum in $z^-_{\max}$ and $z^+_{\max}$, and three local minimum in $z^-_{\min}$, 0, and $z^+_{\min}$, provided that $\nu = \mu/\sigma > 1$. We also get that $s_p^{(1)}(z)^2$ grows to $+\infty$ when $x^{(1)} \to +/-\infty$. The extreme values are ordered as follows

$$-\mu < z^-_{\min} < z^-_{\max} < 0 < z^+_{\max} < z^+_{\min} < \mu.$$

The values of $z^-_{\min}$, $z^-_{\max}$, $z^+_{\max}$, and $z^+_{\min}$ are given by the zeros of the first derivative of $s_p^{(1)}(z)^2$, defined by

$$\frac{ds_p^{(1)}(z)^2}{dz} = 2\Big(-\frac{1}{\sigma^2} + (\frac{\mu}{\sigma^2})^2\frac{1}{\cosh(\frac{\mu}{\sigma^2}z)^2}\Big)\Big(-\frac{z}{\sigma^2} + \frac{\mu}{\sigma^2}\tanh\left(\frac{\mu}{\sigma^2}z\right)\Big).$$

This derivative vanishes when one of the two factors is null. Since $\mu/\sigma > 1$, the first term is null when

$$\frac{\mu^2}{\sigma^2}\frac{1}{\cosh(\frac{\mu}{\sigma^2}z)^2} = 1,$$

which leads to

$$z^-_{\max} = -\frac{\sigma^2}{\mu}\text{arcosh}(\frac{\mu}{\sigma}) \quad \text{and} \quad z^+_{\max} = \frac{\sigma^2}{\mu}\text{arcosh}(\frac{\mu}{\sigma}).$$

The second factor is null when

$$\tanh\left(\frac{\mu}{\sigma^2}z\right) - \frac{z}{\mu} = 0. \tag{6}$$

Obviously, $z = 0$ is solution. Since $\mu/\sigma > 1$, equation (6) has two additional solutions. Although they do not have a closed form, we have

$$z^-_{\min} \in (-\mu, z^-_{\max})$$
$$z^+_{\min} \in (z^+_{\max}, \mu).$$

Also notice that, as μ/σ gets larger, $z^-_{\min}$ is closer to $-\mu$, and $z^+_{\min}$ to μ. For example in Figure 15, we set $\mu/\sigma = 3$, and we hardly see a gap between $-\mu$ and $z^-_{\min}$, or μ and $z^+_{\min}$.

By definition, for $\mathbf{x} \in \mathcal{M}_{s_0}$, we have

$$s_p^{(1)}(x^{(1)})^2 \leq \|s_p(\mathbf{x})\|_2^2 \leq s_0^2.$$

Therefore, given the variations of $s_p^{(1)}(x^{(1)})^2$ detailed above and illustrated in Figure 15, if $s_0^2 < s_p^{(1)}(z_{\max}^-)^2 = s_p^{(1)}(z_{\max}^+)^2$, it exists three disjoint intervals $I_{-\mu}, I_0, I_\mu \subset \mathbb{R}$, respectively centered around $-\mu$, 0, and μ, such that

$$\mathbf{x} \in \mathcal{M}_{s_0} \implies x^{(1)} \in I_{-\mu} \cup I_0 \cup I_\mu.$$

To conclude, we compute the value of $s_p^{(1)}(z_{\max}^-)$, that is

$$s_p^{(1)}(z_{\max}^+) = -\frac{1}{\mu}\text{arcosh}\big(\frac{\mu}{\sigma}\big) + \frac{\mu}{\sigma^2}\tanh\big(\text{arcosh}\big(\frac{\mu}{\sigma}\big)\big).$$

Using the formulas $\tanh(\text{arcosh}(x)) = \frac{\sqrt{x^2-1}}{x}$ for $|x| > 1$, $\text{arcosh}(x) = \ln(x + \sqrt{x^2-1})$, and with $\nu = \mu/\sigma$, we get

$$\begin{aligned} s_p^{(1)}(z_{\max}^+) &= \sqrt{(\mu/\sigma)^2 - 1}/\sigma - \ln(\mu/\sigma + \sqrt{(\mu/\sigma)^2 - 1})/\mu \\ &= \big[\nu\sqrt{\nu^2-1} - \ln(\nu + \sqrt{\nu^2-1})\big]/\mu, \end{aligned}$$

which is always strictly positive since $\nu > 1$. By assumption, $0 \leq s_0 < \big[\nu\sqrt{\nu^2-1} - \ln(\nu + \sqrt{\nu^2-1})\big]/\mu$, and therefore, we have $s_0 < s_p^{(1)}(z_{\max}^+) = s_p^{(1)}(z_{\max}^-)$, which concludes the proof of statement (ii).

□

D. Proof of Theorem 3.2

Proof of Theorem 3.2. We consider the mixture distributions p and q satisfying Assumption 2.1, for $\mu > 0$ and $r > 0$, and denote by q_L the distribution of the left mode of the mixture q, and similarly, q_R is the distribution of the right mode of q. The probability measures $\mathbb{Q}_L$ and $\mathbb{Q}_R$ respectively admits the densities q_L and q_R. As in the proof of Theorem 2.3, we get that

$$\begin{aligned} \text{KSD}^2(\mathbb{P}, \mathbb{Q}_w) =& w^2\big[\text{KSD}^2(\mathbb{P}, \mathbb{Q}_L) + \text{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - 2\Delta_{L,R}\big] \\ &- 2w\big[\text{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - \Delta_{L,R}\big] + \text{KSD}^2(\mathbb{P}, \mathbb{Q}_R). \end{aligned}$$

Next, we define $\mathbf{Z} \sim \mathbb{Q}_w$, $\mathbf{Z}_L \sim \mathbb{Q}_L$, and $\mathbf{Z}_R \sim \mathbb{Q}_R$, and develop the entropic regularization term,

$$\begin{aligned} \mathbb{E}[\log(p(\mathbf{Z}))] &= \int \log(p(\mathbf{x}))(wq_L(\mathbf{x}) + (1-w)q_R(\mathbf{x}))d\mathbf{x} \\ &= w\int \log(p(\mathbf{x}))q_L(\mathbf{x})d\mathbf{x} + (1-w)\int \log(p(\mathbf{x}))q_R(\mathbf{x}))d\mathbf{x} \\ &= w\mathbb{E}[\log(p(\mathbf{Z}_L))] + (1-w)\mathbb{E}[\log(p(\mathbf{Z}_R))]. \end{aligned}$$

Combining these two results, we have

$$\begin{aligned} \text{KSD}_\lambda^2(\mathbb{P}, \mathbb{Q}_w) =& w^2\big[\text{KSD}^2(\mathbb{P}, \mathbb{Q}_L) + \text{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - 2\Delta_{L,R}\big] \\ &- 2w\big[\text{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - \Delta_{L,R} + \lambda/2(\mathbb{E}[\log(p(\mathbf{Z}_L))] - \mathbb{E}[\log(p(\mathbf{Z}_R))])\big] \\ &+ \text{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - \lambda\mathbb{E}[\log(p(\mathbf{Z}_R))]. \end{aligned}$$

We recall that p does not depend on w, but only on the fixed weight w_p and the two mode distributions. Consequently, $\text{KSD}_\lambda^2(\mathbb{P}, \mathbb{Q}_w)$ is a second-order polynomial with respect to w. As for Theorem 2.3, the coefficient of w^2 is $\text{KSD}^2(\mathbb{P}, \mathbb{Q}_{1/2})/4$, and is therefore positive. Then, the polynomial is minimized with respect to w, at $w^\star$ defined by

$$w_\lambda^\star = \frac{\text{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - \Delta_{L,R} + \lambda/2(\mathbb{E}[\log(p(\mathbf{Z}_L))] - \mathbb{E}[\log(p(\mathbf{Z}_R))])}{\text{KSD}^2(\mathbb{P}, \mathbb{Q}_L) + \text{KSD}^2(\mathbb{P}, \mathbb{Q}_R) - 2\Delta_{L,R}}.$$

Since $\mathbb{E}[\log(p(\mathbf{Z}_L))] - \mathbb{E}[\log(p(\mathbf{Z}_R))] \neq 0$ by assumption, we get that $w_\lambda^\star = w_p$ for lambda defined by

$$\lambda = 2\frac{w_p\text{KSD}^2(\mathbb{P}, \mathbb{Q}_L) - (1-w_p)\text{KSD}^2(\mathbb{P}, \mathbb{Q}_R) + (1-2w_p)\Delta_{L,R}}{\mathbb{E}[\log(p(\mathbf{Z}_L))] - \mathbb{E}[\log(p(\mathbf{Z}_R))]}.$$

□

E. Proof of Theorem 3.3

Proof of Theorem 3.3. Following the same approach as in the proof of Theorem 2.4, we have

$$\mathbb{E}[\text{L-KSD}^2(\mathbb{P}, \mathbb{P}_m)] = \frac{d}{m\ell^2} + \frac{\mathbb{E}[\|s_p(\mathbf{X})\|_2^2]}{m} + \frac{\mathbb{E}[\Delta^+ \log p(\mathbf{X})]}{m},$$

where $\mathbf{X} \sim \mathbb{P}$, and, with the empirical distribution $\mathbb{Q}_m = \frac{1}{m}\sum_{i=1}^m \delta(\mathbf{x}_i)$,

$$\text{L-KSD}^2(\mathbb{P}, \mathbb{Q}_m) = \frac{d}{m\ell^2} + \frac{1}{m^2}\sum_{i=1}^m \|s_p(\mathbf{x}_i)\|_2^2 + \frac{1}{m^2}\sum_{i\neq j}^m k_p(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{m^2}\sum_{i=1}^m \Delta^+ \log(p(\mathbf{x}_i)).$$

As $\{\mathbf{x}_i\}_{i=1}^m$ are concentrated at a local minimum or saddle point $\mathbf{x}_0$, the score is null for all particles, as well as the distances between them, and we get

$$\text{L-KSD}^2(\mathbb{P}, \mathbb{Q}_m) = \frac{d}{m\ell^2} + \frac{(m-1)d}{m\ell^2} + \frac{\Delta^+ \log(p(\mathbf{x}_0))}{m}.$$

Next, the difference $m\big(\text{L-KSD}^2(\mathbb{P}, \mathbb{Q}_m) - \mathbb{E}[\text{L-KSD}^2(\mathbb{P}, \mathbb{P}_m)]\big)$ writes

$$m\big(\text{L-KSD}^2(\mathbb{P}, \mathbb{Q}_m) - \mathbb{E}[\text{L-KSD}^2(\mathbb{P}, \mathbb{P}_m)]\big) = \Delta^+ \log(p(\mathbf{x}_0)) + (m-1)\frac{d}{\ell^2} - \big(\mathbb{E}[\|s_p(\mathbf{X})\|_2^2] + \mathbb{E}[\Delta^+ \log p(\mathbf{X})]\big).$$

By definition,

$$\Delta^+ \log p(\mathbf{x}) \stackrel{def}{=} \sum_{j=1}^d \left(\frac{\partial^2 \log p(\mathbf{x})}{\partial x^{(j)2}}\right)^+, \tag{7}$$

with

$$\frac{\partial^2 \log p(\mathbf{x})}{\partial x^{(j)2}} = \frac{1}{p(\mathbf{x})}\frac{\partial^2 p(\mathbf{x})}{\partial x^{(j)2}} - \left(\frac{1}{p(\mathbf{x})}\frac{\partial p(\mathbf{x})}{\partial x^{(j)}}\right)^2.$$

As $\mathbf{x}_0$ is a stationary point of p, $\partial p(\mathbf{x}_0)/\partial x^{(j)} = 0$, and we obtain

$$\frac{\partial^2 \log p(\mathbf{x}_0)}{\partial x^{(j)2}} = \frac{1}{p(\mathbf{x}_0)}\frac{\partial^2 p(\mathbf{x}_0)}{\partial x^{(j)2}},$$

leading to

$$\Delta^+ \log p(\mathbf{x}_0) = \sum_{j=1}^d \frac{1}{p(\mathbf{x}_0)}\left(\frac{\partial^2 p(\mathbf{x}_0)}{\partial x^{(j)2}}\right)^+ = \frac{\Delta^+ p(\mathbf{x}_0)}{p(\mathbf{x}_0)}. \tag{8}$$

Finally, we get

$$m\big(\text{L-KSD}^2(\mathbb{P}, \mathbb{Q}_m) - \mathbb{E}[\text{L-KSD}^2(\mathbb{P}, \mathbb{P}_m)]\big) = \frac{\Delta^+ p(\mathbf{x}_0)}{p(\mathbf{x}_0)} + (m-1)\frac{d}{\ell^2} - \big(\mathbb{E}[\|s_p(\mathbf{X})\|_2^2] + \mathbb{E}[\Delta^+ \log p(\mathbf{X})]\big),$$

which ensures that $\text{L-KSD}^2(\mathbb{P}, \mathbb{Q}_m) - \mathbb{E}[\text{L-KSD}^2(\mathbb{P}, \mathbb{P}_m)] > 0$ for $m \geq 2$, provided that

$$p(\mathbf{x}_0) < \frac{\Delta^+ p(\mathbf{x}_0)}{\mathbb{E}[\|s_p(\mathbf{X})\|_2^2] + \mathbb{E}[\Delta^+ \log p(\mathbf{X})]}.$$

□

F. Proof of Theorem 3.5

Theorem 3.5 extends Theorem 3 from Riabiz et al. (2022) to regularized Stein Thinning, using Lemmas F.1-F.2 and assuming the following convergence rate of the regularization parameter $\lambda_{m_n} = o(\log(m_n)/m_n)$. Lemma F.1 also extends Theorem 1 from Riabiz et al. (2022)[Theorem 1], whereas Lemma F.2 is a slight modification of Lemma 5 from Riabiz et al. (2022).

Lemma F.1. *Let $\mathbb{P}$ be a probability measure on $\mathbb{R}^d$ that admits density $p \in \mathcal{C}^2(\mathbb{R}^d)$, k_p be a reproducing Stein kernel, and $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ a fixed set of points. If π is an index sequence of length m produced by regularized Stein thinning, then we have for $\lambda > 0$,*

$$\mathrm{KSD}^2\big(\frac{1}{m}\sum_{j=1}^{m}\delta(\mathbf{x}_{\pi_j})\big) \leq \mathrm{KSD}^2\big(\sum_{i=1}^{n} w_i^\star \delta(\mathbf{x}_i)\big) + \frac{1+\log(m)}{m}\max_{i=1,\dots,n} k_p(\mathbf{x}_i,\mathbf{x}_i) + \frac{1+\log(m)}{m}\max_{i=1,\dots,n}\Delta^+\log(p(\mathbf{x}_i)) + 2\lambda \max_{i=1,\dots,n}|\log(p(\mathbf{x}_i))|,$$

where the weights $w^\star$ are defined as

$$w^\star \in \arg\min_{\substack{\sum_i w_i = 1 \\ w_i \geq 0}} \mathrm{KSD}^2\big(\sum_{i=1}^{n} w_i \delta(\mathbf{x}_i)\big).$$

Lemma F.2. *Let f be a non-negative function on $\mathbb{R}^d$. Consider a sequence of random variables $(\mathbf{X}_i)_{i\in\mathbb{N}} \subset \mathbb{R}^d$ such that, for some $\gamma > 0$,*

$$b \overset{def}{=} \sup_{i\in\mathbb{N}} \mathbb{E}[e^{\gamma f(\mathbf{X}_i)}] < \infty.$$

If $\log(n)^\beta < m_n < n$, with any $\beta > 1$, then we have almost surely,

$$\lim_{n\to\infty} \frac{\log(m_n)}{m_n} \max_{i=1,\dots,n} f(\mathbf{X}_i) = 0.$$

The proofs of Lemmas F.1 and F.2 are reported at the end of this section. We first proceed with the proof of Theorem 3.5.

Proof of Theorem 3.5. From Lemma F.1, we have

$$\mathrm{KSD}^2\big(\frac{1}{m_n}\sum_{j=1}^{m_n}\delta(\mathbf{Z}_{\pi_j})\big) \leq \underbrace{\mathrm{KSD}^2\big(\sum_{i=1}^{n} w_i^\star \delta(\mathbf{Z}_i)\big)}_{(\star)} + \underbrace{\frac{1+\log(m_n)}{m_n}\max_{i=1,\dots,n} k_p(\mathbf{Z}_i,\mathbf{Z}_i)}_{(\star\star)} + \underbrace{\frac{1+\log(m_n)}{m_n}\max_{i=1,\dots,n}\Delta^+\log(p(\mathbf{Z}_i))}_{(\star\star\star)} + \underbrace{2\lambda_{m_n}\max_{i=1,\dots,n}|\log(p(\mathbf{Z}_i))|}_{(\diamond)}.$$

Riabiz et al. (2022)[Proof of Theorem 3, p.12] showed that the term $(\star)$ converges towards 0 almost surely as $n \to \infty$. For the remaining terms, from Assumption 3.4, we have

$$\sup_{i\in\mathbb{N}} \mathbb{E}[e^{\gamma(d/\ell^2 + \|s_p(\mathbf{Z}_i)\|_2^2)}] < \infty\,, \quad \sup_{i\in\mathbb{N}} \mathbb{E}[e^{\gamma|\log(p(\mathbf{Z}_i))|}] < \infty\,,$$

and

$$\sup_{i\in\mathbb{N}} \mathbb{E}[e^{\gamma\Delta^+\log(p(\mathbf{Z}_i))}] < \infty\,.$$

We can use Lemma F.2 with $f(\mathbf{x}) = k_p(\mathbf{x},\mathbf{x})$ and $f(\mathbf{x}) = \Delta^+\log(\mathbf{x})$ to deduce that $(\star\star) \to 0$ and $(\star\star\star) \to 0$, respectively. The remaining term $(\diamond)$ can be rewritten as

$$2\lambda_{m_n}\max_{i=1,\dots,n}|\log(p(\mathbf{Z}_i))| = \frac{2m_n\lambda_{m_n}}{\log(m_n)} \times \frac{\log(m_n)}{m_n}\max_{i=1,\dots,n}|\log(p(\mathbf{Z}_i))|.$$

Using the assumption that $\lambda_{m_n} = o(\log(m_n)/m_n)$ and Lemma F.2 with $f(\mathbf{x}) = |\log(p(\mathbf{x}))|$, we conclude that $(\diamond) \to 0$. It follows that $\mathrm{KSD}^2\big(\frac{1}{m_n}\sum_{j=1}^{m_n}\delta(\mathbf{x}_{\pi_j})\big) \to 0$ almost surely as $n \to \infty$. Given that p is assumed to be distantly dissipative, we apply Theorem 4 from Chen et al. (2019) to obtain that $\mathbb{Q}_{m_n} \Rightarrow \mathbb{P}$ almost surely, as $n \to \infty$. □

Proof of Lemma F.1. We consider an iteration $t \in \{2, \dots, m\}$ of regularized Stein thinning, where $m \geq 2$ is the final length of the thinned sample. We define $a_t = t^2\mathrm{KSD}^2(\mathbb{P}, \frac{1}{t}\sum_{j=1}^{t}\delta(\mathbf{x}_{\pi_j}))$ and $f_t = \sum_{j=1}^{t} k_p(\mathbf{x}_{\pi_j}, \cdot)$. We also denote $S_1^2 = \max_{i=1,\dots,n} k_p(\mathbf{x}_i, \mathbf{x}_i) + \max_{i=1,\dots,n} \Delta^+ \log(p(\mathbf{x}_i))$, and $S_2 = 2\max_{i=1,\dots,n} |\log(p(\mathbf{x}_i))|$. Using the definition of the squared KSD of an empirical measure, we have

$$a_t = a_{t-1} + k_p(\mathbf{x}_{\pi_t}, \mathbf{x}_{\pi_t}) + 2\sum_{j=1}^{t-1} k_p(\mathbf{x}_{\pi_j}, \mathbf{x}_{\pi_t}).$$

Let $\mathbf{x}_t^\star = \arg\min_{\mathbf{y}\in\{\mathbf{x}_i\}_{i=1}^n} f_{t-1}(\mathbf{y})$. By definition, $\mathbf{x}_{\pi_t}$ minimizes the cost function of the regularized Stein thinning algorithm at iteration t, and we have

$$\begin{aligned} k_p(\mathbf{x}_{\pi_t}, \mathbf{x}_{\pi_t}) + 2\sum_{j=1}^{t-1} k_p(\mathbf{x}_{\pi_j}, \mathbf{x}_{\pi_t}) + \Delta^+ \log(p(\mathbf{x}_{\pi_t})) - \lambda t \log(p(\mathbf{x}_{\pi_t})) \\ \leq k_p(\mathbf{x}_t^\star, \mathbf{x}_t^\star) + 2\sum_{j=1}^{t-1} k_p(\mathbf{x}_{\pi_j}, \mathbf{x}_t^\star) + \Delta^+ \log(p(\mathbf{x}_t^\star)) - \lambda t \log(p(\mathbf{x}_t^\star)). \end{aligned}$$

We combine this last inequality with the first equation to obtain

$$\begin{aligned} a_t \leq a_{t-1} + k_p(\mathbf{x}_t^\star, \mathbf{x}_t^\star) + 2\sum_{j=1}^{t-1} k_p(\mathbf{x}_{\pi_j}, \mathbf{x}_t^\star) + \Delta^+ \log(p(\mathbf{x}_t^\star)) - \Delta^+ \log(p(\mathbf{x}_{\pi_t})) \\ - \lambda t(\log(p(\mathbf{x}_t^\star)) - \log(p(\mathbf{x}_{\pi_t}))), \end{aligned}$$

and then,

$$\begin{aligned} a_t \leq a_{t-1} + k_p(\mathbf{x}_t^\star, \mathbf{x}_t^\star) + 2\sum_{j=1}^{t-1} k_p(\mathbf{x}_{\pi_j}, \mathbf{x}_t^\star) + \Delta^+ \log(p(\mathbf{x}_t^\star)) - \Delta^+ \log(p(\mathbf{x}_{\pi_t})) \\ + \lambda t(|\log(p(\mathbf{x}_t^\star))| + |\log(p(\mathbf{x}_{\pi_t}))|). \end{aligned}$$

By definition, $0 \leq |\log(p(\mathbf{x}_i))| \leq S_2$ and $k_p(\mathbf{x}_t^\star, \mathbf{x}_t^\star) + \Delta^+ \log(p(\mathbf{x}_t^\star)) - \Delta^+ \log(p(\mathbf{x}_{\pi_t})) \leq S_1^2$, hence, we have

$$a_t \leq a_{t-1} + S_1^2 + t\lambda S_2 + 2\min_{\mathbf{y}\in\{\mathbf{x}_i\}_{i=1}^n} f_{t-1}(\mathbf{y}).$$

As in the proof of (Riabiz et al., 2022), we have $\min_{\mathbf{y}\in\{\mathbf{x}_i\}_{i=1}^n} f_{t-1}(\mathbf{y}) \leq \sqrt{a_{t-1}}\|h^\star\|_{\mathcal{H}(k_p)}$ where $h^\star$ is the element in the RKHS $\mathcal{H}(k_p)$ of the form $h^\star = \sum_{i=1}^n w_i^\star k_p(\mathbf{x}_i, \cdot)$. As a result, a_t is bounded as follows:

$$a_t \leq a_{t-1} + S_1^2 + t\lambda S_2 + 2\sqrt{a_{t-1}}\|h^\star\|_{\mathcal{H}(k_p)}.$$

We then show by induction that

$$a_t \leq t^2(\|h^\star\|_{\mathcal{H}(k_p)}^2 + C_t + \lambda S_2),$$

where

$$C_t \stackrel{def}{=} \frac{1}{t}\left(S_1^2 - \|h^\star\|_{\mathcal{H}(k_p)}^2\right)\sum_{j=1}^{t}\frac{1}{j}.$$

With such as a result, we will have $\mathrm{KSD}^2(\mathbb{P}, \frac{1}{t}\sum_{j=1}^{t}\delta(\mathbf{x}_{\pi_j})) \leq \mathrm{KSD}^2(\mathbb{P}, \sum_{i=1}^n w_i^\star \delta(\mathbf{x}_{\pi_j})) + C_t + \lambda S_2$ and obtain the advertised result in Theorem F.1 for the last iteration $t = m$.

For $t = 1$, we have $a_1 = k_p(\mathbf{x}_{\pi_1}, \mathbf{x}_{\pi_1}) \leq S_1^2$ and thus $a_1 \leq \|h^\star\|^2_{\mathcal{H}(k_p)} + C_1 + \lambda S_2$. For a fixed $t \geq 2$, assume that $a_{t-1} \leq (t-1)^2(\|h^\star\|^2_{\mathcal{H}(k_p)} + C_{t-1} + \lambda S_2)$ where $C_{t-1} = \frac{1}{t-1}(S_1^2 - \|h^\star\|^2_{\mathcal{H}(k_p)}) \sum_{j=1}^{t-1} \frac{1}{j}$. We then have

$$\begin{aligned} a_t \leq & a_{t-1} + S_1^2 + t\lambda S_2 + 2\sqrt{a_{t-1}}\|h^\star\|_{\mathcal{H}(k_p)} \\ \leq & (t-1)^2(\|h^\star\|^2_{\mathcal{H}(k_p)} + C_{t-1} + \lambda S_2) + S_1^2 + t\lambda S_2 \\ & + 2(t-1)\sqrt{\|h^\star\|^2_{\mathcal{H}(k_p)} + C_{t-1} + \lambda S_2}\|h^\star\|_{\mathcal{H}(k_p)} \\ = & t^2(\|h^\star\|^2_{\mathcal{H}(k_p)} + C_t + \lambda S_2) + R_t \end{aligned} \tag{9}$$

where

$$\begin{aligned} R_t = & (t-1)^2 C_{t-1} - t^2 C_t + (1-2t)(\|h^\star\|^2_{\mathcal{H}(k_p)} + \lambda S_2) + S_1^2 \\ & + t\lambda S_2 + 2(t-1)\sqrt{\|h^\star\|^2_{\mathcal{H}(k_p)} + C_{t-1} + \lambda S_2}\|h^\star\|_{\mathcal{H}(k_p)} \\ = & (t-1)^2 C_{t-1} - t^2 C_t + (1-2t)\|h^\star\|^2_{\mathcal{H}(k_p)} + S_1^2 \\ & + \lambda S_2(1-t) + 2(t-1)\sqrt{\|h^\star\|^2_{\mathcal{H}(k_p)} + C_{t-1} + \lambda S_2}\|h^\star\|_{\mathcal{H}(k_p)} \end{aligned}$$

Using (Riabiz et al., 2022)[Lemma 1], we have

$$2\|h^\star\|_{\mathcal{H}(k_p)}\sqrt{\|h^\star\|^2_{\mathcal{H}(k_p)} + C_{t-1} + \lambda S_2} \leq 2\|h^\star\|^2_{\mathcal{H}(k_p)} + C_{t-1} + \lambda S_2$$

It follows from Equation (9) that we need $R_t \leq 0$, *i.e.*,

$$2\|h^\star\|^2_{\mathcal{H}(k_p)} + C_{t-1} + \lambda S_2 \leq \frac{t^2 C_t - (t-1)^2 C_{t-1}}{t-1} - \frac{S_1^2 - \|h^\star\|^2_{\mathcal{H}(k_p)}}{t-1} + \lambda S_2 + 2\|h^\star\|^2_{\mathcal{H}(k_p)}.$$

The above inequality is always satisfied as long as

$$2\|h^\star\|^2_{\mathcal{H}(k_p)} + C_{t-1} + \leq \frac{t^2 C_t - (t-1)^2 C_{t-1}}{t-1} - \frac{S_1^2 - \|h^\star\|^2_{\mathcal{H}(k_p)}}{t-1} + 2\|h^\star\|^2_{\mathcal{H}(k_p)},$$

which is equivalent to

$$tC_t - (t-1)C_{t-1} \geq \frac{1}{t}(S_1^2 - \|h^\star\|^2_{\mathcal{H}(k_p)}),$$

and always true by definition of C_t. Hence we have shown that $a_t \leq t^2(\|h^\star\|^2_{\mathcal{H}(k_p)} + C_t + \lambda S_2)$. Given that $\|h^\star\|^2_{\mathcal{H}(k_p)} = \mathrm{KSD}^2(\mathbb{P}, \sum_{i=1}^n w_i^\star \delta(\mathbf{x}_i))$, we have

$$\mathrm{KSD}^2(\mathbb{P}, \frac{1}{t}\sum_{j=1}^{t} \delta(\mathbf{x}_{\pi_j})) \leq \mathrm{KSD}^2(\mathbb{P}, \sum_{i=1}^{n} w_i^\star \delta(\mathbf{x}_i)) + C_t + \lambda S_2\,,$$

where $C_t \leq \frac{1+\log(t)}{t}\big(\max_{i=1,\ldots,n} k_p(\mathbf{x}_i, \mathbf{x}_i) + \max_{i=1,\ldots,n} \Delta^+ \log(p(\mathbf{x}_i))\big)$ (see (Riabiz et al., 2022)[Lemma 2]). □

Proof of Lemma F.2. We follow the proof of Lemma 5 from (Riabiz et al., 2022). In the last step of the proof, we essentially need to show that

$$\sum_{m_n=1}^{\infty} c_1(m_n^2) < \infty \quad \text{and} \quad \sum_{m_n=1}^{\infty} c_2(m_n) < \infty$$

where

$$c_1(m_n^2) \stackrel{def}{=} \frac{2\log(m_n)}{m_n^2}\frac{\log(nb)}{\gamma} \quad \text{and} \quad c_2(m_n) \stackrel{def}{=} 4\frac{\log(m_n)}{m_n^2}\frac{\log(n((m_n+1)^2)b)}{\gamma}\,.$$

With the assumption that $\log(n)^\beta \leq m_n < n$ with $\beta > 1$, it is deduced that $c_1(m_n^2) \to 0$ and $c_2(m_n) \to 0$ as $n \to \infty$. □

G. Laplacian Stein Operator

Instead of the standard Langevin operator, we can use $\mathcal{T}_p(g) = \Delta(pg)/p$, mentioned in Oates et al. (2017, Appendix A.2). However, such Stein operator introduces similar problems as Pathology I, since the Stein kernel associated with $\mathcal{T}_p'(g)$ also has spurious minimum in regions where second derivatives of p vanish, as illustrated below. Therefore, it is more appropriate to taylor a specific Laplacian correction as proposed in this paper, which cannot directly be derived from a Stein kernel, since it is not differentiable. In this appendix, we study the operator $\mathcal{T}_p$ defined as $\mathcal{T}_p g = \Delta(pg)/p$ and such that (Oates et al., 2017, Appendix A.2)

$$\mathbb{E}[(\mathcal{T}_p g)(\mathbf{Z})] = 0\,,$$

for all g belonging to $\mathcal{G}$, and with $\mathbf{Z} \sim \mathbb{P}$. For each dimension $j \in \{1, \dots, d\}$, let $\mathcal{T}_p^j$ be the operator defined as

$$(\mathcal{T}_p^j g)(\mathbf{x}) = \frac{1}{p(\mathbf{x})}\Big(\nabla^2_{x_j} p(\mathbf{x}) g(\mathbf{x}) + 2\nabla_{x_j} p(\mathbf{x}) \nabla_{x_j} g(\mathbf{x}) + p(\mathbf{x})\nabla^2_{x_i} g(\mathbf{x})\Big)$$

for $g : \mathbb{R}^d \to \mathbb{R}$, and where x_j is the j-th coordinate of $\mathbf{x}$ to simplify notations. The operator $\mathcal{T}_p g$ can then be rewritten as $(\mathcal{T}_p g)(\mathbf{x}) = \sum_{j=1}^d (\mathcal{T}_p^j g)(\mathbf{x})$. Gorham & Mackey (2017)[Proposition 2] establishes the closed-form expression of the KSD in the case of the multidimensional Langevin operator. We generalize the proof of Gorham & Mackey (2017)[Proposition 2] for the Laplacian operator $(\mathcal{T}_p g)(\mathbf{x}) = \Delta(p(\mathbf{x})g(\mathbf{x}))/p(\mathbf{x})$ and establish a closed-form expression of the Stein kernel $k_p : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. For a given kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $k \in \mathcal{C}^{2,2}$, the Stein kernel k_p is given by (Gorham & Mackey, 2017)

$$k_p(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^d k_p^j(\mathbf{x}, \mathbf{y}), \tag{10}$$

where

$$k_p^j(\mathbf{x}, \mathbf{y}) = \langle \mathcal{T}_p^j(k(\mathbf{x}, \cdot)), \mathcal{T}_p^j(k(\cdot, \mathbf{y}))\rangle_{\mathcal{H}_k}\,. \tag{11}$$

After a few developments, it is found that

$$\begin{aligned}
p(\mathbf{x})p(\mathbf{y})k_p^j(\mathbf{x}, \mathbf{y}) &= \nabla^2_{x_j} p(\mathbf{x}) \nabla^2_{y_j} p(\mathbf{y}) k(\mathbf{x}, \mathbf{y}) \\
&+ 2\nabla^2_{x_j} p(\mathbf{x}) \nabla_{y_j} p(\mathbf{y}) \nabla_{y_j} k(\mathbf{x}, \mathbf{y}) + p(\mathbf{y})\nabla^2_{x_j} p(\mathbf{x}) \nabla^2_{y_j} k(\mathbf{x}, \mathbf{y}) \\
&+ 2\nabla_{x_j} p(\mathbf{x}) \nabla^2_{y_j} p(\mathbf{y}) \nabla_{x_j} k(\mathbf{x}, \mathbf{y}) + 4\nabla_{x_j} p(\mathbf{x}) \nabla_{y_j} p(\mathbf{y}) \nabla_{x_i} \nabla_{y_j} k(\mathbf{x}, \mathbf{y}) \\
&+ 2p(\mathbf{y}) \nabla_{x_j} p(\mathbf{x}) \nabla_{x_j} \nabla^2_{y_j} k(\mathbf{x}, \mathbf{y}) + p(\mathbf{x}) \nabla^2_{y_j} p(\mathbf{y}) \nabla^2_{x_j} k(\mathbf{x}, \mathbf{y}) \\
&+ 2p(\mathbf{x}) \nabla_{y_j} p(\mathbf{y}) \nabla^2_{x_j} \nabla_{y_j} k(\mathbf{x}, \mathbf{y}) + p(\mathbf{x})p(\mathbf{y}) \nabla^2_{x_j} \nabla^2_{y_j} k(\mathbf{x}, \mathbf{y})
\end{aligned} \tag{12}$$

A closed-form expression can then be obtained in the case of, *e.g.*, an inverse multiquadratic kernel of the form $k(\mathbf{x}, \mathbf{y}) = (1 + \|\mathbf{x} - \mathbf{y}\|_2^2/\ell^2)^{-1/2}$ where ℓ denotes the bandwidth. In this case, one has

$$\begin{aligned}
\nabla_{y_j} k(\mathbf{x}, \mathbf{y}) &= \frac{1}{\ell^2} k(\mathbf{x}, \mathbf{y})^3 (x_j - y_j)\,, \quad \nabla_{x_j} k(\mathbf{x}, \mathbf{y}) = -\nabla_{y_j} k(\mathbf{x}, \mathbf{y}) \\
\nabla^2_{y_j} k(\mathbf{x}, \mathbf{y}) &= \nabla^2_{x_j} k(\mathbf{x}, \mathbf{y}) = -\frac{1}{\ell^2} k(\mathbf{x}, \mathbf{y})^3 + \frac{3}{\ell^4} k(\mathbf{x}, \mathbf{y})^5 (x_j - y_j)^2 \\
\nabla_{x_j} \nabla_{y_j} k(\mathbf{x}, \mathbf{y}) &= \frac{1}{\ell^2} k(\mathbf{x}, \mathbf{y})^3 - \frac{3}{\ell^4} k(\mathbf{x}, \mathbf{y})^5 (x_j - y_j)^2 \\
\nabla^2_{x_j} \nabla_{y_j} k(\mathbf{x}, \mathbf{y}) &= \frac{-9k(\mathbf{x}, \mathbf{y})^5}{\ell^4}(x_j - y_j) + \frac{15k(\mathbf{x}, \mathbf{y})^7}{\ell^6}(x_j - y_j)^3 \\
\nabla_{x_j} \nabla^2_{y_j} k(\mathbf{x}, \mathbf{y}) &= -\nabla^2_{x_j} \nabla_{y_j} k(\mathbf{x}, \mathbf{y}) \\
\nabla^2_{x_j} \nabla^2_{y_j} k(\mathbf{x}, \mathbf{y}) &= \frac{9k(\mathbf{x}, \mathbf{y})^5}{\ell^4} - \frac{90k(\mathbf{x}, \mathbf{y})^7}{\ell^6}(x_j - y_j)^2 + \frac{105k(\mathbf{x}, \mathbf{y})^9}{\ell^8}(x_j - y_j)^4
\end{aligned} \tag{13}$$

A closed-form expression of k_p can then be obtained by combining Equations (10)-(13). In contrast to the Langevin Stein kernel (see Equation 2), the above Stein kernel involves second-order derivatives of the density p.

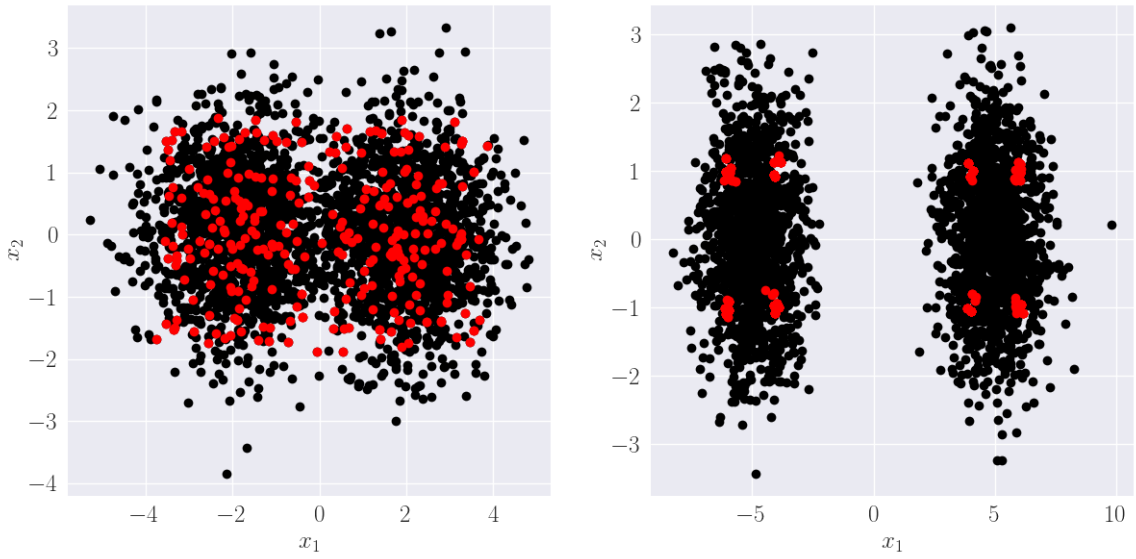

Figure 16. Stein thinning with the Stein operator $\mathcal{T}_p g = \Delta(pg)/p$, for Gaussian mixtures of Example 1, with $\mu = 2$ (left panel), and $\mu = 5$ (right panel), $\sigma = 1$, and $w = 0.5$.

We run experiments based on our Example 1 of Gaussian mixtures for this new Stein operator $\mathcal{T}_p g = \Delta(pg)/p$. We sequentially set $\mu = 2$ and $\mu = 5$, with $\sigma = 1$ and $w = 0.5$. Results are displayed in Figure 16. In the left panel with $\mu = 2$, we see that Pathology II does not occur, as opposed to Figure 2 with the Langevin Stein operator. However, when μ is set to 5 in the right panel, all particles are concentrated in spurious minimum again. Therefore, introducing higher-order derivatives of the target density through the Stein operator does not seem to be a promising route.