Supplement

This supplement contains supporting material for the paper *Stein* Π*-Importance Sampling*. The mathematical background on Stein kernels is contained in Appendix A. The proof of Theorem 1 is contained in Appendix B. For implementation of Stein Π-Importance Sampling without the aid of automatic differentiation, various explicit derivatives are required; the relevant calculations can be found in Appendix C. The empirical protocols and additional empirical results are presented in Appendix D.

A Mathematical Background

This appendix contains mathematical background on reproducing kernels and Stein kernels, as used in the main text. Appendix A.1 introduces matrix-valued reproducing kernels, while Appendix A.2 specialises to Stein kernels by application of a Stein operator to a matrix-valued kernel. A selection of useful Stein kernels are presented in Appendix A.3.

A.1 Matrix-Valued Reproducing Kernels

A *matrix-valued kernel* is a function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$, that is both

1. symmetric; $K(x, y) = K(y, x)$ for all $x, y \in \mathbb{R}^d$, and
2. positive semi-definite; $\sum_{i=1}^n \sum_{j=1}^n \langle c_i, K(x_i, x_j) c_j \rangle \geq 0$ for all $x_1, \dots, x_n \in \mathbb{R}^d$ and $c_1, \dots, c_n \in \mathbb{R}^d$.

Let $K_x = K(\cdot, x)$. For vector-valued functions $g, g' : \mathbb{R}^d \to \mathbb{R}^d$, defined by $g = \sum_{i=1}^n K_{x_i} c_i$ and $g' = \sum_{j=1}^m K_{x_j'} c_i'$, define an inner product

$$\langle g, g' \rangle_{\mathcal{H}(K)} = \sum_{i=1}^n \sum_{j=1}^m \langle c_i, K(x_i, x_j') c_j' \rangle. \tag{9}$$

There is a unique Hilbert space of such vector-valued functions associated to K, denoted $\mathcal{H}(K)$; see Proposition 2.1 of Carmeli et al. (2006). This space is characterised as

$$\mathcal{H}(K) = \overline{\operatorname{span}}\{K_x c : x, c \in \mathbb{R}^d\}$$

where here the closure is taken with respect to the inner product in (9). It can be shown that $\mathcal{H}(K)$ is in fact a reproducing kernel Hilbert space (RKHS) which satisfies the *reproducing property*

$$\langle g, K_x c \rangle_{\mathcal{H}(K)} = \langle g(x), c \rangle$$

for all $g \in \mathcal{H}(K)$ and $x, c \in \mathbb{R}^d$. Matrix-valued kernels are the natural starting point for construction of KSDs, as described next.

A.2 Stein Kernels

A general construction for Stein kernels is to first identify a matrix-valued RKHS $\mathcal{H}(K)$ and an operator $S_P : \mathcal{H}(K) \to L^1(P)$ for which $\int S_p h \, \mathrm{d}P = 0$ for all $h \in \mathcal{H}(K)$. Such an operator will be called a *Stein operator*. The collection $\{S_p h : h \in \mathcal{H}(K)\}$ inherits the structure of an RKHS, whose reproducing kernel

$$k_P(x, y) = \langle S_P K_x, S_P K_y \rangle_{\mathcal{H}(K)} \tag{10}$$

is a Stein kernel, meaning that $\mu_P = 0$ where μ_P is the kernel mean embedding from (1); see Barp et al. (2022b). Explicit calculations for the Stein kernels considered in this work can be found in Appendix C.

For univariate distributions, Barbour (1988) proposed to obtain Stein operators from infinitessimal generators of P-invariant continuous-time Markov processes; see also Barbour (1990); Gotze (1991).

The approach was extended to multivariate distributions in Gorham and Mackey (2015). The starting point is the P-invariant Itô diffusion

$$\mathrm{d}X_t = \frac{1}{2}\frac{1}{p(X_t)}\nabla \cdot [p(X_t)M(X_t)]\mathrm{d}t + M(X_t)^{1/2}\mathrm{d}W_t, \tag{11}$$

where p is the density of P, assumed to be positive, $M : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a symmetric matrix called the *diffusion matrix*, and W_t is a standard Wiener process (Kent, 1978; Roberts and Stramer, 2002). Here the notation $[\nabla \cdot A]_i = \nabla \cdot (A_{i,\cdot}^\top)$ indicates the divergence operator applied to each row of the matrix $A(x) \in \mathbb{R}^{d\times d}$. The infinitessimal generator is

$$(A_P u)(x) = \frac{1}{2}\frac{1}{p(x)}\nabla \cdot [p(x)M(x)\nabla u(x)].$$

Substituting $h(x)$ for $\frac{1}{2}\nabla u(x)$, we obtain a Stein operator

$$(S_P h)(x) = \frac{1}{p(x)}\nabla \cdot [p(x)M(x)h(x)] \tag{12}$$

called the *diffusion Stein operator* (Gorham et al., 2019). This is indeed a Stein operator, since under mild integrability conditions on K, the divergence theorem gives that $\int S_p h \, \mathrm{d}P = 0$ for all $h \in \mathcal{H}(K)$; for full details and a proof see Barp et al. (2022b).

A.3 Selecting a Stein Kernel

There are several choices for a Stein kernel, and which we should use depends on what form of convergence we hope to control (Gorham and Mackey, 2017; Gorham et al., 2019; Hodgkinson et al., 2020; Barp et al., 2022b; Kanagawa et al., 2022). Appendix A.3.1 describes the Langevin–Stein kernel for weak convergence control, Appendix A.3.2 describes the KGM–Stein kernels for additional control over moments, and Appendix A.3.3 presents the Riemann–Stein kernel, whose convergence properties have to-date been less well-studied.

All of the kernels that we consider have length scale parameters that need to be specified, and some also have location parameters to be specified. As a reasonably automatic default we define

$$x_\star \in \arg\max p(x), \qquad \Sigma^{-1} = -\nabla^2 \log p(x_\star)$$

as a location and a matrix of characteristic length scales for P that will be used throughout. These values can typically be obtained using gradient-based optimisation, which is usually cheaper to perform compared to full approximation of P. It is assumed that $\nabla^2 \log p(x_\star)$ is positive definite in the sequel.

A.3.1 Weak Convergence Control with Langevin–Stein Kernels

The first kernel we consider, which we called the Langevin–Stein kernel in the main text, was introduced by Gorham and Mackey (2017). This Stein kernel was developed for the purpose of controlling the weak convergence of a sequence $(Q_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ to P. Recall that a sequence $(Q_n)_{n\in\mathbb{N}}$ is said to *converge weakly* (or *in distribution*) to P if $\int f \mathrm{d}Q_n \to \int f \mathrm{d}P$ for all continuous bounded functions $f : \mathbb{R}^d \to \mathbb{R}$. This convergence is denoted $Q_n \xrightarrow{\mathrm{d}} P$ in shorthand.

The problem considered in Gorham and Mackey (2017) was how to select a combination of matrix-valued kernel K (and, implicitly, a diffusion matrix M) such that the Stein kernel k_P in (10) generates a KSD $D_P(Q)$ in (4) for which $D_P(Q_n) \to 0$ implies $Q_n \xrightarrow{\mathrm{d}} P$. Their solution was to combine the inverse multi-quadric kernel with an identity diffusion matrix;

$$K(x, y) = (1 + \|x - y\|_\Sigma^2)^{-\beta} I, \qquad M(x) = I$$

for $\beta \in (0, 1)$. Provided that P has a density p for which $\nabla \log p(x)$ is Lipschitz, and that P is *distantly dissipative* (see Definition 4 of Gorham and Mackey, 2017), the associated KSD enjoys weak convergence control. Technically, the results in Gorham and Mackey (2017) apply only when $\Sigma = I$, but Theorem 4 in Chen et al. (2019) demonstrated that they hold also for any positive definite Σ. Following the recommendation of several previous authors, including Chen et al. (2018, 2019); Riabiz et al. (2022), we take $\beta = \frac{1}{2}$ throughout.

A.3.2 Moment Convergence Control with KGM–Stein Kernels

Despite its many elegant properties, weak convergence can be insufficient for applications where we are interested in integrals $\int f \, \mathrm{d}P$ for which the integrand $f : \mathbb{R}^d \to \mathbb{R}$ is unbounded. In particular, this is the case for moments of the form $f(x) = x_1^{\alpha_1} \dots x_d^{\alpha_d}$, $0 \neq \alpha \in \mathbb{N}_0^d$. In such situations, we may seek also the stronger property of *moment convergence control*. The development of KSDs for moment convergence control was recently considered by Kanagawa et al. (2022), and we refered to their construction as the KGM–Stein kernels in the main text. (For convenience, we have adopted the initials of the authors in naming the KGM–Stein kernel.)

A sequence $(Q_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ is said to converge to P *in the sth order moment* if $\int \|x\|^s \mathrm{d}Q_n(x) \to \int \|x\|^s \mathrm{d}P(x)$. To establish convergence of moments, we need an additional condition on top of weak convergence control: uniform integrability control. A sequence of measures $(Q_n)_{n\in\mathbb{N}}$ is said to have *uniformly integrable sth moments* if for any $\varepsilon > 0$, we can take $r > 0$ such that

$$\sup_{n\in\mathbb{N}} \int_{\|x\|>r} \|x\|^s \; \mathrm{d}Q_n(x) < \varepsilon.$$

This condition essentially states that the tail decay of the measures is well-controlled (so that it has a convergent moment). The KSD convergence $D_P(Q_n) \to 0$ implies uniform integrability if for any $\varepsilon > 0$, we can take $r_\varepsilon > 0$ and $f_\varepsilon \in \mathcal{H}(K)$ such that

$$S_P f_\varepsilon(x) \geq \|x\|^s 1\{\|x\| > r_\varepsilon\} - \varepsilon, \tag{13}$$

i.e., the Stein-modified RKHS can approximate the (norm-weighted) indicator function arbitrarily well. Such a function f_ε can be explicitly constructed (while not guaranteed to be a member of the RKHS). Specifically, the choice $f_\varepsilon = (1 - \iota_\varepsilon) g$ satisfies (13) under an appropriate dissipativity condition, where ι_ε is a differentiable indicator function vanishing outside a ball, and $g(x) = -x/\sqrt{1 + \|x\|^2}$. This motivated Kanagawa et al. (2022) to introduce the *sth order KGM–Stein kernel*, which is based on the matrix-valued kernel and diffusion matrix

$$K(x, y) = [\phi(\|x - y\|_\Sigma) + \kappa_{\mathrm{lin}}(x, y)]\, I, \qquad M(x) = (1 + \|x - x_\star\|_\Sigma^2)^{\frac{s-1}{2}} I,$$

where $(x, y) \mapsto \phi(\|x - y\|_\Sigma)$ is a C_0^1 universal kernel (see Barp et al., 2022b, Theorem 4.8). For comparability of our results, we take ϕ to be the inverse multi-quadric $\phi(r) = (1 + r^2)^{-1/2}$, and

$$\kappa_{\mathrm{lin}}(x, y) = \frac{1 + (x - x_\star)^\top \Sigma^{-1} (y - x_\star)}{\sqrt{1 + \|x - x_\star\|_\Sigma^2}\sqrt{1 + \|y - x_\star\|_\Sigma^2}}.$$

Here the normalised linear kernel κ_{lin} ensures $g \in \mathcal{H}(K)$, while the C_0^1 universal kernel ϕ allows approximation of $S_P \iota_\varepsilon g$; see Kanagawa et al. (2022).

A.3.3 Exploiting Geometry with Riemann–Langevin–Stein Kernels

For academic interest only, here we describe the *Riemann–Stein kernel* that featured in Figure 2 of the main text. This Stein kernel is motivated by the analysis of Gorham et al. (2019), who argued that the use of rapidly mixing Itô diffusions in Stein operators can lead to sharper convergence control. The Riemann–Stein kernel is based on the class of so-called *Riemannian* diffusions considered in Girolami and Calderhead (2011), who proposed to take the diffusion matrix M in (11) to be $M = (\mathcal{I}_{\text{prior}} + \mathcal{I}_{\text{Fisher}})^{-1}$, the inverse of the Fisher information matrix, $\mathcal{I}_{\text{Fisher}}$, regularised using the Hessian of the negative log-prior, $\mathcal{I}_{\text{prior}}$. For the two-dimensional illustration in Section 3.2, this leads to the diffusion matrix

$$M(x) = \left(I + \sum_{i=1}^{n} [\nabla f_i(x)][\nabla f_i(x)]^\top \right)^{-1},$$

where we recall that $y_i = f_i(x) + \epsilon_i$, where the ϵ_i are independent with $\epsilon_i \sim \mathcal{N}(0, 1)$, and the prior is $x \sim \mathcal{N}(0, 1)$. For the presented experiment we paired the above diffusion matrix with the inverse multi-quadric kernel $K(x, y) = (1 + \|x - y\|_\Sigma^2)^{-\beta}$ for $\beta = \frac{1}{2}$. The Riemann–Stein kernel extends naturally to distributions P defined on Riemannian manifolds $\mathcal{X}$; see Barp et al. (2022a) and Example 1 of Hodgkinson et al. (2020).

Unfortunately, the Riemann–Stein kernel is prohibitively expensive in most real applications, since each evaluation of M requires a full scan through the size-n dataset. The computational complexity of Stein Π-Thinning with the Riemann–Stein kernel is therefore $O(m^2n^2)$, which is unfavourable compared to the $O(m^2n)$ complexity in the case where the Stein kernel is not data-dependent. Furthermore, the convergence control properties of the Riemann–Stein kernel have yet to be established. For these reasons we included the Riemann–Stein kernel for illustration only; further groundwork will be required before the Riemann-Stein kernel can be practically used.

B Proof of Theorem 1

This appendix is devoted to the proof of Theorem 1. The proof is based on the recent work of Durmus and Moulines (2022), on the geometric convergence of MALA, and on the analysis of sparse (greedy) approximation of kernel discrepancies performed in Riabiz et al. (2022); these existing results are recalled in Appendix B.1. An additional technical result on preconditioned MALA is contained in Appendix B.2. The proof of Theorem 1 itself is contained in Appendix B.3.

B.1 Auxiliary Results

To precisely describe the results on which our analysis is based, we first need to introduce some notation and terminology. Let $V : \mathcal{X} \to [1, \infty)$ and, for a function $f : \mathcal{X} \to \mathbb{R}$ and a measure μ on $\mathcal{X}$, let

$$\|f\|_V := \sup_{x \in \mathcal{X}} \frac{|f(x)|}{V(x)}, \qquad \|\mu\|_V := \sup_{\|f\|_V \leq 1} \left| \int_{\mathcal{X}} f \mathrm{d}\mu \right|.$$

Recall that a Q-invariant Markov chain $(x_i)_{i \in \mathbb{N}} \subset \mathcal{X}$ with n^{th} step transition kernel Q^n is V*-uniformly ergodic* (see Theorem 16.0.1 of Meyn and Tweedie, 2012) if and only if $\exists R \in [0, \infty), \rho \in (0, 1)$ such that

$$\|\mathrm{Q}^n(x, \cdot) - Q\|_V \leq R\rho^n V(x) \tag{14}$$

for all initial states $x \in \mathcal{X}$ and all $n \in \mathbb{N}$.

Although MALA (Algorithm 1) is classical (Roberts and Stramer, 2002), until recently explicit sufficient conditions for ergodicity of MALA had not been obtained. The first result we will need is due Durmus and Moulines (2022), who presented the first explicit conditions for V-uniform convergence of MALA. It applies only to *standard* MALA, meaning that the preconditioning matrix M appearing in Algorithm 1 is the identity matrix. The extension of this result to preconditioned MALA will be handled in Appendix B.2.

Theorem 2. *Let $Q \in \mathcal{P}(\mathbb{R}^d)$ admit a density, q, such that*

(DM1) *there exists x_0 with $\nabla \log q(x_0) = 0$*

(DM2) *q is twice continuously differentiable with* $\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log q(x - x_0)\| < \infty$

(DM3) *there exists $b > 0$ and $B \geq 0$ such that $-\nabla^2 \log q(x - x_0) \succeq bI$ for all $\|x - x_0\| \geq B$.*

Then there exists $\epsilon_0 > 0$ such that for all step sizes $\epsilon \in (0, \epsilon_0)$, standard Q-invariant MALA *(i.e. with $M = I$) is V-uniformly ergodic for $V(x) = \exp\big(\frac{b}{16}\|x - x_0\|^2\big)$.*

Proof. This is Theorem 1 of Durmus and Moulines (2022). □

The next result that we will need establishes consistency of the greedy algorithm applied to samples from a Markov chain that is Q-invariant.

Theorem 3. *Let $P, Q \in \mathcal{P}(\mathcal{X})$ with $P \ll Q$. Let $k_P : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Stein kernel and let $D_P : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ denote the associated KSD. Consider a Q-invariant, time-homogeneous Markov chain $(x_i)_{i \in \mathbb{N}} \subset \mathcal{X}$ such that*

(R^+1) $(x_i)_{i \in \mathbb{N}}$ *is V-uniformly ergodic, such that* $V(x) \geq \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\sqrt{k_P(x)}$

(R^+2) $\sup_{i \in \mathbb{N}} \mathbb{E}\left[\frac{\mathrm{d}P}{\mathrm{d}Q}(x_i)\sqrt{k_P(x_i)}V(x_i)\right] < \infty$

(R$^+$3) *there exists* $\gamma > 0$ *such that* $\sup_{i\in\mathbb{N}} \mathbb{E}\left[\exp\left\{\gamma \max\left(1, \frac{\mathrm{d}P}{\mathrm{d}Q}(x_i)^2\right) k_P(x_i)\right\}\right] < \infty$.

Let $P_{n,m}$ *be the result of running the greedy algorithm in* (5). *If* $m \leq n$ *and* $\log(n) = O(m^{\gamma/2})$ *for some* $\gamma < 1$*, then* $D_P(P_{n,m}) \to 0$ *almost surely as* $m, n \to \infty$.

Proof. This is Theorem 3 of Riabiz et al. (2022). □

B.2 Preconditioned MALA

In addition to the auxiliary results in Appendix B.1, which concern standard MALA (i.e. with $M = I$), we require an elementary fact about MALA, namely that preconditioned MALA is equivalent to standard MALA under a linear transformation of the state variable. Recall that the M-preconditioned MALA algorithm is a Metropolis–Hastings algorithm whose proposal is the Euler–Maruyama discretisation of the Itô diffusion (11).

Proposition 1. *Let* $M(x) \equiv M$ *for a symmetric positive definite and position-independent matrix* $M \in \mathbb{R}^{d\times d}$*. Let* $Q \in \mathcal{P}(\mathbb{R}^d)$ *admit a probability density function (PDF)* q *for which the* Q*-invariant diffusion* $(X_t)_{t\geq 0}$*, given by setting* $p = q$ *in* (11)*, is well-defined. Then under the change of variables* $Y_t := M^{1/2}X_t$,

$$\mathrm{d}Y_t = \frac{1}{2}(\nabla \log \tilde{q})(Y_t)\mathrm{d}t + \mathrm{d}W_t, \tag{15}$$

where $\tilde{q}(x) \propto q(M^{-1/2}x)$ *for all* $x \in \mathbb{R}^d$.

Proof. From the chain rule,

$$(\nabla \log \tilde{q})(y) = \nabla_y \log q(M^{-1/2}y) = M^{-1/2}(\nabla \log q)(M^{-1/2}y),$$

and thus, substituting $Y_t = M^{1/2}X_t$, (15) is equal to

$$\begin{aligned}\mathrm{d}X_t &= M^{-1/2}\left[\frac{1}{2}M^{-1/2}(\nabla \log q)(M^{-1/2}M^{1/2}X_t) + \mathrm{d}W_t\right] \\ &= \frac{1}{2}M^{-1}(\nabla \log q)(X_t) + M^{-1/2}\mathrm{d}W_t,\end{aligned}$$

which is identical to (11) in the case where $M(x) = M$ is constant. □

Let Q and $\tilde{Q}$ be the distributions referred to in Proposition 1, whose PDFs are respectively $q(x)$ and $\tilde{q}(x) \propto q(M^{-1/2}x)$. Proposition 1 then implies that the M-preconditioned MALA algorithm applied to Q (i.e. Algorithm 1 for $\Pi = Q$) is equivalent to the standard MALA algorithm (i.e. $M = I$) applied to $\tilde{Q}$. This fact allows us to generalise the result of Theorem 2 as follows:

Corollary 1. *Consider a symmetric positive definite matrix* $M \in \mathbb{R}^{d\times d}$*. Assume that conditions* (DM1-3) *in Theorem 2 are satisfied. Then there exists* $\epsilon_0' > 0$ *and* $b' > 0$ *such that for all step sizes* $\epsilon \in (0, \epsilon_0')$*, the* M*-preconditioned* Q*-invariant* MALA *is* V*-uniformly ergodic for* $V(x) = \exp\left(\frac{b'}{16}\|x - x_0\|^2\right)$.

Proof. From Theorem 2 and Proposition 1, the result follows if we can establish (DM1-3) for $\tilde{Q}$, since M-preconditioned MALA is equivalent to standard MALA applied to $\tilde{Q}$. For a matrix $A \in \mathbb{R}^{d\times d}$, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ respectively denote the minimum and maximum eigenvalues of A. For (DM1) we set $y_0 = M^{1/2}x_0$ and observe that

$$(\nabla \log \tilde{q})(y_0) = M^{-1/2}(\nabla \log q)(x_0) = 0.$$

For (DM2) we have that

$$\begin{aligned}\sup_{y\in\mathbb{R}^d} \|\nabla^2(\log \tilde{q})(y - y_0)\| &= \sup_{y\in\mathbb{R}^d} \|M^{-1/2}(\nabla^2 \log q)(M^{-1/2}(y - y_0))M^{-1/2}\| \\ &\leq \lambda_{\min}(M)^{-1} \sup_{x\in\mathbb{R}^d} \|(\nabla^2 \log q)(x - x_0)\| < \infty.\end{aligned}$$

For (DM3) we have that

$$\begin{aligned}-(\nabla^2 \log \tilde{q})(y - y_0) &= -M^{-1/2}(\nabla^2 \log q)(M^{-1/2}(y - y_0))M^{-1/2} \\ &= -M^{-1/2}(\nabla^2 \log q)(x - x_0)M^{-1/2} \succeq M^{1/2}(bI)M^{1/2} = bM^{-1} \succeq b'I\end{aligned}$$

where $b' = b\lambda_{\max}(M)^{-1}$, which holds for all $\|x - x_0\| \geq B$, and in particular for all $\|y - y_0\| \geq B'$ where $B' = B\lambda_{\max}(M)^{1/2}$. Thus (DM1-3) are established for $\tilde{Q}$. ☐

Remark 1. *The choice $M = \Sigma^{-1}$, which sets the preconditioner matrix M equal to the inverse of the length scale matrix Σ used in the specification of the kernel K (c.f. Appendix A.3), leads to the elegant interpretation that Stein Π-Importance Sampling applied to M-preconditioned* MALA *is equivalent to the Stein Π-Importance Sampling applied to standard* MALA *(i.e. with $M = I$) for the whitened target $\tilde{P}$ with PDF $\tilde{p}(x) \propto p(M^{-1/2}x)$. For our experiments, however, the preconditioner matrix M was learned during a warm-up phase of* MALA*, since in general the curvature of P (captured by Σ) and the curvature of Π (captured by M^{-1}) may be different.*

B.3 Proof of Theorem 1

The route to establishing Theorem 1 has three parts. First, we establish (DM1-3) of Theorem 2 with $Q = \Pi$, to deduce from Corollary 1 that Π-invariant M-preconditioned MALA is V-uniformly ergodic. This in turn enables us to establish conditions (R$^+$1-3) of Theorem 3, again for $Q = \Pi$, from which the strong consistency $D_P(P_{n,m}) \overset{\text{a.s.}}{\to} 0$ of SΠT-MALA is established. Finally, we note that $0 \leq D_P(P_n^\star) \leq D_P(P_{n,m})$, since the support of $P_{n,m}$ is contained in the support of $P_n^\star$, and the latter is optimally weighted, whence also the strong consistency of SΠIS-MALA.

Establish (DM1-3) First we establish (DM1-3) for $Q = \Pi$. Fix $x_0 \in \mathbb{R}^d$. For (DM2), first recall that the range of k_P is $[C_1^2, \infty)$ where $C_1 > 0$, from Assumption 1. Since $\log(\cdot)$ has bounded second derivatives on $[C_1^2, \infty)$, there is a constant $C > 0$ such that

$$\forall x \in \mathbb{R}^d, \qquad \|\nabla^2 \log k_P(x)\| \leq C\|\nabla^2 k_P(x)\|.$$

Thus, using compactness of the set $\{x : \|x - x_0\| \leq B_2\}$,

$$\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log k_P(x)\| \leq C \max\left(\underbrace{\sup_{\|x - x_0\| \leq B_2} \|\nabla^2 k_P(x)\|}_{< \infty \text{ by (A3)}}, \underbrace{\sup_{\|x - x_0\| \geq B_2} \|\nabla^2 k_P(x)\|}_{< b_2 \|I\| \text{ by (A4)}} \right) < \infty. \tag{16}$$

Now, π is twice differentiable as it is the product of twice differentiable functions p and $k_P^{1/2}$ from (A1) and (A3), and moreover

$$\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log \pi(x - x_0)\| \leq \underbrace{\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log p(x)\|}_{< \infty \text{ by (A1)}} + \frac{1}{2} \underbrace{\sup_{x \in \mathbb{R}^d} \|\nabla^2 \log k_P(x)\|}_{< \infty \text{ by (16)}} < \infty,$$

so (DM2) is satisfied. For (DM3), first note from the chain and product rules that for all $\|x\| \geq B_2$

$$\nabla^2 \log k_P(x - x_0) = \underbrace{\frac{\nabla^2 k_P(x - x_0)}{k_P(x - x_0)}}_{\preceq (b_2/C_1^2)I \text{ by (A4)}} - \underbrace{\frac{[\nabla k_P(x - x_0)][\nabla k_P(x - x_0)]^\top}{k_P(x - x_0)^2}}_{\succeq 0} \preceq \frac{b_2}{C_1^2} I. \tag{17}$$

Thus, for all $\|x - x_0\| \geq B := \|x_0\| + \max(B_1, B_2)$,

$$-\nabla^2 \log \pi(x - x_0) = \underbrace{-\nabla^2 \log p(x - x_0)}_{\succeq b_1 I \text{ by (A2)}} - \frac{1}{2} \underbrace{\nabla^2 \log k_P(x - x_0)}_{\preceq (b_2/C_1^2)I \text{ by (17)}} \succeq \underbrace{\left(b_1 - \frac{b_2}{2C_1^2} \right)}_{=: b > 0} I \tag{18}$$

as required. The same argument establishes (DM1); from (18) we have $\lim_{\|x\| \to \infty} \pi(x) = 0$, and since π is a continuously differentiable density there must exist an x_0 at which π is locally minimised. Thus we have established (DM1-3) for $Q = \Pi$ and we may conclude from Corollary 1 that there is an $\epsilon_0' > 0$ and $b' > 0$ such that, for all $\epsilon \in (0, \epsilon_0')$, the Π-invariant M-preconditioned MALA chain $(x_i)_{i \in \mathbb{N}}$ is V-uniformly ergodic for $V(x) = C_2 \exp\left(\frac{b'}{16} \|x - x_0\|^2 \right)$ (since if a Markov chain is V-uniformly ergodic, then it is also CV-uniformly ergodic).

Establish (R^+1-3) The aim is now to establish conditions (R^+1-3) of Theorem 3 for $Q = \Pi$. By construction $\mathrm{d}P/\mathrm{d}\Pi = C_2/\sqrt{k_P(x)} < C_2/C_1 < \infty$, where C_1 and C_2 were defined in Assumption 1, so that $P \ll \Pi$. It has already been established that $(x_i)_{i \in \mathbb{N}}$ is V-uniformly ergodic, and further

$$V(x) = C_2 \exp\left(\frac{b'}{16}\|x - x_0\|^2\right) \geq C_2 = \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x)\sqrt{k_P(x)}$$

for all x, which establishes (R^+1). Let R and ρ denote constants for which the V-uniform ergodicity property (14) is satisfied. From V-uniform ergodicity, the integral $\int V \, \mathrm{d}\Pi$ exists and

$$\begin{aligned} \left|\mathbb{E}\left[\frac{\mathrm{d}P}{\mathrm{d}\Pi}(x_i)\sqrt{k_P(x_i)}V(x_i)\right] - C_2 \int V \, \mathrm{d}\Pi\right| &= C_2 \left|\mathbb{E}[V(x_i)] - \int V \, \mathrm{d}\Pi\right| \\ &\leq C_2 R \rho^n V(x_0) \to 0 \end{aligned}$$

which establishes (R^+2). Fix $\gamma > 0$. By construction $\mathrm{d}P/\mathrm{d}\Pi \leq C_2/C_1$, and thus

$$\exp\left\{\gamma \max\left(1, \frac{\mathrm{d}P}{\mathrm{d}\Pi}(x)^2\right) k_P(x)\right\} < \exp\left\{\tilde{\gamma} k_P(x)\right\}$$

where $\tilde{\gamma} = \max(1, C_2/C_1)\gamma$. Since we have assumed that k_P is continuous with, from (A4),

$$b_3 := \limsup_{\|x\| \to \infty} \frac{k_P(x)}{\|x\|^2} < \infty,$$

we may take γ such that $\tilde{\gamma} b_3 < b'/16$, so that $\|x \mapsto \exp\{\tilde{\gamma} k_P(x)\}\|_V < \infty$ and in particular

$$\left|\mathbb{E}\left[\exp\{\tilde{\gamma} k_P(x_i)\}\right] - \int \exp\{\tilde{\gamma} k_P(x)\} \, \mathrm{d}\Pi(x)\right| \leq \|x \mapsto \exp\{\tilde{\gamma} k_P(x)\}\|_V \times R \rho^n V(x_0) \to 0$$

which establishes (R^+3). Thus we have established (R^+1-3) for $Q = \Pi$, so from Theorem 3 we have strong consistency of SΠT-MALA (i.e. $D_P(P_{n,m}) \overset{\text{a.s.}}{\to} 0$) provided that $m \leq n$ with $\log(n) = O(m^{\gamma/2})$ for some $\gamma < 1$. The latter condition is equivalent to $m = \Omega((\log n)^\delta)$ for some $\delta > 2$, which we used for the statement. Since $0 \leq D_P(P_n^\star) \leq D_P(P_{n,m})$, the strong consistency of SΠIS-MALA is also established.

C Explicit Calculation of Stein Kernels

This appendix contains explicit calculations for the Langevin–Stein and KGM–Stein kernels k_P, which are sufficient to implement Stein Π-Importance Sampling and Stein Π-Thinning. These calculations can also be performed using automatic differentiation, but comparison to the analytic expressions is an important step in validation of computer code.

To proceed, we observe that the diffusion Stein operator S_P in (12) applied to a matrix-valued kernel K is equivalent to the Langevin–Stein operator applied to the kernel $C(x,y) = M(x)K(x,y)M(y)^\top$. In the case of the Langevin–Stein and KGM–Stein kernels we have $K(x,y) = \kappa(x,y)I$ for some $\kappa(x,y)$ and $M(x) = (1 + \|x - x_\star\|_\Sigma^2)^{(s-1)/2} I$ for some $s \in \{0, 1, 2, \dots\}$. Thus $C(x,y) = c(x,y)I$ where

$$c(x,y) := (1 + \|x - x_\star\|_\Sigma^2)^{(s-1)/2}(1 + \|y - x_\star\|_\Sigma^2)^{(s-1)/2}\kappa(x,y)$$

and

$$\begin{aligned} k_P(x,y) = \nabla_x \cdot \nabla_y c(x,y) + [\nabla_x c(x,y)] \cdot [\nabla_y \log p(y)] + [\nabla_y c(x,y)] \cdot [\nabla_x \log p(x)] \\ + c(x,y)[\nabla_x \log p(x)] \cdot [\nabla_y \log p(y)], \end{aligned}$$

following the calculations in Oates et al. (2017). To evaluate the terms in this formula we start by differentiating $c(x, y)$, to obtain

$$
\begin{aligned}
\nabla_x c(x,y) &= (1+\|x-x_\star\|_\Sigma^2)^{(s-1)/2}(1+\|y-x_\star\|_\Sigma^2)^{(s-1)/2} \\
&\qquad \times \left[\frac{(s-1)\kappa(x,y)\Sigma^{-1}(x-x_\star)}{1+\|x-x_\star\|_\Sigma^2} + \nabla_x \kappa(x,y)\right] \\
\nabla_y c(x,y) &= (1+\|x-x_\star\|_\Sigma^2)^{(s-1)/2}(1+\|y-x_\star\|_\Sigma^2)^{(s-1)/2} \\
&\qquad \times \left[\frac{(s-1)\kappa(x,y)\Sigma^{-1}(y-x_\star)}{1+\|y-x_\star\|_\Sigma^2} + \nabla_y \kappa(x,y)\right] \\
\nabla_x \cdot \nabla_y c(x,y) &= (1+\|x-x_\star\|_\Sigma^2)^{(s-1)/2}(1+\|y-x_\star\|_\Sigma^2)^{(s-1)/2} \\
&\qquad \times \left[\frac{(s-1)^2\kappa(x,y)(x-x_\star)^\top\Sigma^{-2}(y-x_\star)}{(1+\|x-x_\star\|_\Sigma^2)(1+\|y-x_\star\|_\Sigma^2)} + \frac{(s-1)(y-x_\star)^\top\Sigma^{-1}\nabla_x\kappa(x,y)}{(1+\|y-x_\star\|_\Sigma^2)}\right. \\
&\qquad\qquad \left. + \frac{(s-1)(x-x_\star)^\top\Sigma^{-1}\nabla_y\kappa(x,y)}{(1+\|x-x_\star\|_\Sigma^2)} + \nabla_x \cdot \nabla_y \kappa(x,y)\right].
\end{aligned}
$$

These expressions involve gradients of $\kappa(x, y)$, and explicit formulae for these are presented for the choice of $\kappa(x, y)$ corresponding to the Langevin–Stein kernel in Appendix C.1, and to the KGM–Stein kernel in Appendix C.2.

To implement Stein Π-Thinning we require access to both $k_P(x)$ and $\nabla k_P(x)$, the latter for use in the proposal distribution and acceptance probability in MALA. These quantities will now be calculated. In what follows we assume that $\kappa(x, y)$ is continuously differentiable, so that partial derivatives with respect to x and y can be interchanged. Then

$$
\begin{aligned}
c_0(x) &:= c(x,x) \\
&= (1+\|x-x_\star\|_\Sigma^2)^{s-1}\kappa(x,x) \\
c_1(x) &:= \left.\nabla_x c(x,y)\right|_{y\to x} \\
&= (1+\|x-x_\star\|_\Sigma^2)^{s-1}\left[\frac{(s-1)\kappa(x,x)\Sigma^{-1}(x-x_\star)}{(1+\|x-x_\star\|_\Sigma^2)} + \left.\nabla_x\kappa(x,y)\right|_{y\to x}\right] \\
c_2(x) &:= \left.\nabla_x \cdot \nabla_y c(x,y)\right|_{y\to x} \\
&= (1+\|x-x_\star\|_\Sigma^2)^{s-1}\left[\frac{(s-1)^2\kappa(x,x)(x-x_\star)^\top\Sigma^{-2}(x-x_\star)}{(1+\|x-x_\star\|_\Sigma^2)^2}\right. \\
&\qquad\qquad \left. + \frac{2(s-1)(x-x_\star)^\top\Sigma^{-1}\ \left.\nabla_x\kappa(x,y)\right|_{y\to x}}{(1+\|x-x_\star\|_\Sigma^2)} + \left.\nabla_x\cdot\nabla_y\kappa(x,y)\right|_{y\to x}\right]
\end{aligned}
$$

so that

$$
k_P(x) := k_P(x,x) = c_2(x) + 2c_1(x)\cdot\nabla_x \log p(x) + c_0(x)\|\nabla_x \log p(x)\|^2. \tag{19}
$$

Let $[\nabla_x c_1(x)]_{i,j} = \partial_{x_i}[c_1(x)]_j$ and $[\nabla_x^2 \log p(x)]_{i,j} = \partial_{x_i}\partial_{x_j}\log p(x)$. Now we can differentiate (19) to get

$$
\begin{aligned}
\nabla_x k_P(x) &= \nabla_x c_2(x) + 2[\nabla_x c_1(x)][\nabla_x \log p(x)] + 2[\nabla_x^2 \log p(x)]c_1(x) \\
&\qquad + [\nabla_x c_0(x)]\|\nabla_x \log p(x)\|^2 + 2c_0(x)[\nabla_x^2 \log p(x)][\nabla_x \log p(x)].
\end{aligned} \tag{20}
$$

In what follows we also derive explicit formulae for $c_0(x)$, $c_1(x)$ and $c_2(x)$, and hence for $\nabla_x c_0(x)$, $\nabla_x c_1(x)$ and $\nabla_x c_2(x)$, for the case of the Langevin–Stein kernel in Appendix C.1, and the KGM–Stein kernel in Appendix C.2.

C.1 Explicit Formulae for the Langevin–Stein Kernel

The Langevin–Stein kernel from Appendix A.3.1 corresponds to the choice $s = 1$ and $\kappa(x, y)$ the inverse multi-quadric kernel, so that

$$\begin{aligned}
\kappa(x,y) &= (1+\|x-y\|_\Sigma^2)^{-\beta} \\
\nabla_x \kappa(x,y) &= -2\beta(1+\|x-y\|_\Sigma^2)^{-\beta-1}\Sigma^{-1}(x-y) \\
\nabla_y \kappa(x,y) &= 2\beta(1+\|x-y\|_\Sigma^2)^{-\beta-1}\Sigma^{-1}(x-y) \\
\nabla_x \cdot \nabla_y \kappa(x,y) &= -4\beta(\beta+1)(1+\|x-y\|_\Sigma^2)^{-\beta-2}(x-y)^\top\Sigma^{-2}(x-y) \\
&\qquad + 2\beta \mathrm{tr}(\Sigma^{-1})(1+\|x-y\|_\Sigma^2)^{-\beta-1}.
\end{aligned}$$

Evaluating on the diagonal:

$$\begin{aligned}
\kappa(x,x) &= 1 \\
\nabla_x \kappa(x,y)|_{y\to x} = \nabla_y \kappa(x,y)|_{y\to x} &= 0 \\
\nabla_x \cdot \nabla_y \kappa(x,y)|_{y\to x} &= 2\beta \mathrm{tr}(\Sigma^{-1}),
\end{aligned}$$

so that $c_0(x) = 1$, $c_1(x) = 0$, $c_2(x) = 2\beta\mathrm{tr}(\Sigma^{-1})$. Differentiating these formulae, $\nabla_x c_0(x) = 0$, $\nabla_x c_1(x) = 0$, $\nabla_x c_2(x) = 0$.

C.2 Explicit Formulae for the KGM–Stein Kernel

The KGM kernel of order s from Appendix A.3.2 corresponds to the choice

$$\kappa(x,y) = (1+\|x-y\|_\Sigma^2)^{-\beta} + \frac{1+(x-x_\star)^\top\Sigma^{-1}(y-x_\star)}{(1+\|x-x_\star\|_\Sigma^2)^{s/2}(1+\|y-x_\star\|_\Sigma^2)^{s/2}},$$

for which we have

$$\begin{aligned}
\nabla_x \kappa(x,y) &= -2\beta(1+\|x-y\|_\Sigma^2)^{-\beta-1}\Sigma^{-1}(x-y) \\
&\quad + \frac{\Sigma^{-1}(y-x_\star) - s[1+(x-x_\star)^\top\Sigma^{-1}(y-x_\star)]\Sigma^{-1}(x-x_\star)(1+\|x-x_\star\|_\Sigma^2)^{-1}}{(1+\|x-x_\star\|_\Sigma^2)^{s/2}(1+\|y-x_\star\|_\Sigma^2)^{s/2}} \\
\nabla_y \kappa(x,y) &= 2\beta(1+\|x-y\|_\Sigma^2)^{-\beta-1}\Sigma^{-1}(x-y) \\
&\quad + \frac{\Sigma^{-1}(x-x_\star) - s[1+(x-x_\star)^\top\Sigma^{-1}(y-x_\star)]\Sigma^{-1}(y-x_\star)(1+\|y-x_\star\|_\Sigma^2)^{-1}}{(1+\|x-x_\star\|_\Sigma^2)^{s/2}(1+\|y-x_\star\|_\Sigma^2)^{s/2}} \\
\nabla_x \cdot \nabla_y \kappa(x,y) &= -4\beta(\beta+1)(1+\|x-y\|_\Sigma^2)^{-\beta-2}(x-y)^\top\Sigma^{-2}(x-y) + 2\beta\mathrm{tr}(\Sigma^{-1})(1+\|x-y\|_\Sigma^2)^{-\beta-1} \\
&\quad + \frac{\left[\begin{array}{l}
\mathrm{tr}(\Sigma^{-1}) - s(1+\|x-x_\star\|_\Sigma^2)^{-1}(x-x_\star)^\top\Sigma^{-2}(x-x_\star) \\
\qquad -s(1+\|y-x_\star\|_\Sigma^2)^{-1}(y-x_\star)^\top\Sigma^{-2}(y-x_\star) \\
\qquad +s^2[1+(x-x_\star)^\top\Sigma^{-1}(y-x_\star)](1+\|x-x_\star\|_\Sigma^2)^{-1}(1+\|y-x_\star\|_\Sigma^2)^{-1} \\
\qquad\qquad \times(x-x_\star)^\top\Sigma^{-2}(y-x_\star)
\end{array}\right]}{(1+\|x-x_\star\|_\Sigma^2)^{s/2}(1+\|y-x_\star\|_\Sigma^2)^{s/2}}.
\end{aligned}$$

Evaluating on the diagonal:

$$\begin{aligned}
\kappa(x,x) &= 1 + (1+\|x-x_\star\|_\Sigma^2)^{-s+1} \\
\nabla_x \kappa(x,y)|_{y\to x} = \nabla_y \kappa(x,y)|_{y\to x} &= -(s-1)\Sigma^{-1}(x-x_\star)(1+\|x-x_\star\|_\Sigma^2)^{-s} \\
\nabla_x \cdot \nabla_y \kappa(x,y)|_{y\to x} &= 2\beta\mathrm{tr}(\Sigma^{-1}) + \mathrm{tr}(\Sigma^{-1})(1+\|x-x_\star\|_\Sigma^2)^{-s} \\
&\qquad + s(s-2)(1+\|x-x_\star\|_\Sigma^2)^{-s-1}(x-x_\star)^\top\Sigma^{-2}(x-x_\star)
\end{aligned}$$

so that

$$\begin{aligned}
c_0(x) &= 1 + (1+\|x-x_\star\|_\Sigma^2)^{s-1} \\
c_1(x) &= (s-1)(1+\|x-x_\star\|_\Sigma^2)^{s-2}\Sigma^{-1}(x-x_\star) \\
c_2(x) &= \frac{[(s-1)^2(1+\|x-x_\star\|_\Sigma^2)^{s-1}-1](x-x_\star)^\top\Sigma^{-2}(x-x_\star)}{(1+\|x-x_\star\|_\Sigma^2)^2} + \frac{\mathrm{tr}(\Sigma^{-1})[1+2\beta(1+\|x-x_\star\|_\Sigma^2)^s]}{(1+\|x-x_\star\|_\Sigma^2)}.
\end{aligned}$$

Differentiating these formulae:

$$\begin{aligned}
\nabla_x c_0(x) &= 2(s-1)(1+\|x-x_\star\|_\Sigma^2)^{s-2}\Sigma^{-1}(x-x_\star) \\
\nabla_x c_1(x) &= 2(s-1)(s-2)(1+\|x-x_\star\|_\Sigma^2)^{s-3}[\Sigma^{-1}(x-x_\star)][\Sigma^{-1}(x-x_\star)]^\top \\
&\qquad + (s-1)(1+\|x-x_\star\|_\Sigma^2)^{s-2}\Sigma^{-1} \\
\nabla_x c_2(x) &= 2(s-1)^2(s-3)(1+\|x-x_\star\|_\Sigma^2)^{s-4}[(x-x_\star)^\top\Sigma^{-2}(x-x_\star)]\Sigma^{-1}(x-x_\star) \\
&\qquad + 2(s-1)^2(1+\|x-x_\star\|_\Sigma^2)^{s-3}\Sigma^{-2}(x-x_\star) \\
&\qquad + 4\beta\mathrm{tr}(\Sigma^{-1})(s-1)(1+\|x-x_\star\|_\Sigma^2)^{s-2}\Sigma^{-1}(x-x_\star) \\
&\qquad - 2(1+\|x-x_\star\|_\Sigma^2)^{-2}[\Sigma^{-2}(x-x_\star)+\mathrm{tr}(\Sigma^{-1})\Sigma^{-1}(x-x_\star)] \\
&\qquad + 4(1+\|x-x_\star\|_\Sigma^2)^{-3}[(x-x_\star)^\top\Sigma^{-2}(x-x_\star)]\Sigma^{-1}(x-x_\star).
\end{aligned}$$

These complete the analytic calculations necessary to compute the Stein kernel k_P and its gradient.

D Empirical Assessment

This appendix contains full details of the empirical protocols that were employed and the additional empirical results described in the main text. Appendix D.1 discusses the effect of dimension on our proposed Π. Additional illuatrative results from Section 3.2 are contained in Appendix D.2. The full details for how MALA was implemented are contained in Appendix D.3. An additional illustration using a generalised auto-regressive moving average (GARCH) model is presented in Appendix D.4. The full results for SΠIS-MALA are contained in Appendix D.5, and in Appendix D.6 the convergence of the sparse approximation provided by SΠT-MALA to the optimal weighted approximation is investigated. Finally, the performance of KSDs is quantified using the 1-Wasserstein divergence in Appendix D.7.

D.1 The Effect of Dimension on Π

The improvement of Stein Π-Importance Sampling over the default Stein importance sampling algorithm (i.e. $\Pi = P$) can be expected to reduce as the dimension d of the target P is increased. To see this, consider the Langevin–Stein kernel

$$k_P(x) = c_1 + c_2\|\nabla \log p(x)\|_\Sigma^2 \tag{21}$$

for some $c_1, c_2 > 0$; see Appendix C. Taking $P = \mathcal{N}(0, I_{d\times d})$, for which the length scale matrix Σ appearing in Appendix A.3 is $\Sigma = I_{d\times d}$, we obtain

$$k_P(x) = c_1 + c_2\|x\|^2.$$

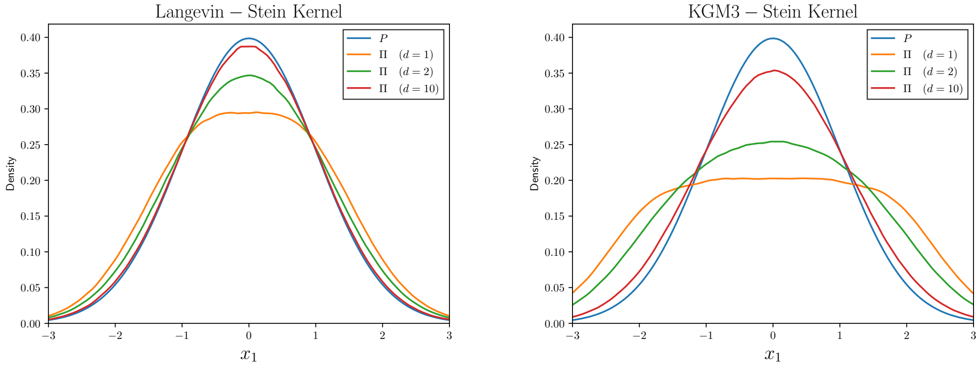
However, the sampling distribution Π defined in (8) depends on k_P only up to an unspecified normalisation constant; we may therefore equally consider the asymptotic behaviour of $\tilde{k}_P(x) := k_P(x)/d$. Let $X \sim P$. Then $\mathbb{E}[\tilde{k}_P(X)] = c_2$ is a d-independent constant, and

$$\left\|\tilde{k}_P - \mathbb{E}[\tilde{k}_P(X)]\right\|_{L^2(P)}^2 = \int \left[\frac{k_P(x) - (c_1 + c_2 d)}{d}\right]^2 \mathrm{d}P(x) = \frac{2c_2^2}{d} \to 0$$

as $d \to \infty$. This shows that $\tilde{k}_P$ converges to a constant function in $L^2(P)$, and thus for "typical" values of x in the effective support of P,

$$\pi(x) \propto p(x)\sqrt{\tilde{k}_P(x)} \overset{\approx}{\propto} p(x),$$

so that $\Pi \approx P$ in the $d \to \infty$ limit. This intuition is borne out in simulations involving both the Langevin–Stein kernel (as just discussed) and also the KGM3–Stein kernel. Indeed, Figure S1 shows that as the dimension d is increased, the marginal distributions of Π become increasingly similar to those of P.



(S1.1) (S1.2)

Figure S1: The effect of dimension on Π: Here P was taken to be the standard Gaussian distribution $\mathcal{N}(0, I_{d \times d})$ in $\mathbb{R}^d$ and the proposed distribution Π was computed. The marginal distribution of the first component of Π is plotted for $d \in \{1, 2, 10\}$, for both (a) the Langevin–Stein kernel and (b) the KGM3–Stein kernel.

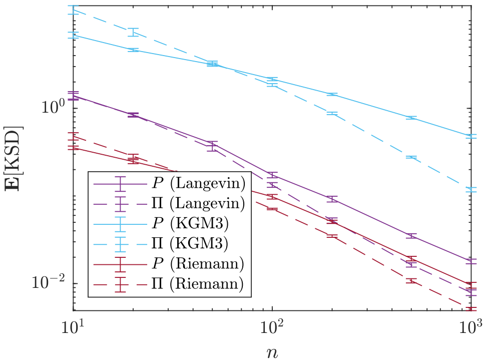


Figure S2: Assessing the performance of the sampling distributions Π shown in Figure 2. The mean kernel Stein discrepancy (KSD) for computation performed using the Langevin–Stein kernel (purple), the KGM3–Stein kernel (blue), and the Riemann–Stein kernel (red); in each case, KSD was computed using the same Stein kernel used to construct Π. Solid lines indicate the baseline case of sampling from P, while dashed lines indicate the proposed approach of sampling from Π. (The experiment was repeated 10 times and standard error bars are plotted.)

D.2 2D Illustration from the Main Text

Section 3.2 of the main text contained a 2-dimensional illustration of Stein Π-Importance Sampling and presented the distributions Π corresponding to different choices of Stein kernel. Here, in Figure S2, we present the mean KSDs for Stein Π-Importance Sampling performed using the Langevin–Stein kernel (purple), the KGM3–Stein kernel (blue), and the Riemann–Stein kernel (red), corresponding to the sampling distributions Π displayed in Figure 2 of the main text.

For this experiment, exact sampling from both P and Π was performed using a fine grid on which all probabilities were calculated and appropriately normalised. Results are in broad agreement with the 1-dimensional illustration contained in the main text, in the sense that in all cases Stein Π-Importance Sampling provides a significant improvement over the default Stein importance sampling method with Π equal to P.

Algorithm 4 Adaptive MALA

```
Require: $x_{0,0}$ (initial state), $\epsilon_0$ (initial step size), $M_0$ (initial preconditioner matrix), $\{n_i\}_{i=0}^{h-1}$ (epoch lengths),
    $\{\alpha_i\}_{i=1}^{h-1}$ (learning schedule), $h$ (number of epochs), $k_P$ (Stein kernel)
$\{x_{0,1} \dots, x_{0,n_0}\} \leftarrow \texttt{MALA}(x_{0,0}, \epsilon_0, M_0, n_0, k_P)$
for $i = 1, \dots, h-1$ do
    $x_{i,0} \leftarrow x_{i-1,n_{i-1}}$
    $\rho_{i-1} \leftarrow \frac{1}{n_{i-1}} \sum_{j=1}^{n_{i-1}} 1_{x_{i-1,j} \neq x_{i-1,j-1}}$        ▷ Average acceptance rate for chain $i$
    $\epsilon_i \leftarrow \epsilon_{i-1} \exp(\rho_{i-1} - 0.57)$        ▷ Update step size
    $M_i \leftarrow \alpha_i M_i + (1-\alpha_i)\mathrm{cov}(\{x_{i-1,1} \dots, x_{i-1,n_{i-1}}\})$        ▷ Update preconditioner matrix
    $\{x_{i,1} \dots, x_{i,n_i}\} \leftarrow \texttt{MALA}(x_{i,0}, \epsilon_i, M_i, n_i, k_P)$
end for
```

D.3 Implementation of MALA

For implementation of MALA in Algorithm 4 we are required to specify a step size ϵ and a preconditioner matrix M. In general, suitable values for both of these parameters will be problem-dependent. Standard practice is to perform some form of manual or automated tuning to arrive at parameter values for which the average acceptance rate is close to 0.57, motivated by the asymptotic analysis of Roberts and Rosenthal (1998). Adaptive MCMC algorithms, which seek to optimise the parameters of MCMC algorithms such as MALA during the warm-up period, provide an appealing solution, and was the approach taken in this work.

The adaptive MALA algorithm which we used is contained in Algorithm 4, where we have let $\texttt{MALA}(x, \epsilon, M, n, k_P)$ denote the output from the preconditioned MALA with initial state x, step size ϵ, preconditioner matrix M, and chain length n, described in Algorithm 1. In Algorithm 4, we use $\mathrm{cov}(\cdot)$ to denote the sample covariance matrix. The algorithm monitors the average acceptance rate and increases or decreases it according to whether it is below or above, respectively, the 0.57 target. For the preconditioner matrix, the sample covariance matrix of samples obtained from the penultimate tuning run of MALA is used. For all experiments that we report using MALA, we set $\epsilon_0 = 1$, $M_0 = I_d$, $h = 10$, and $\alpha_1 = \dots = \alpha_9 = 0.3$. The warm-up epoch lengths were $n_0 = \dots = n_8 = 1,000$ and the final epoch length was $n_9 = 10^5$. The samples $\{x_{h-1,1}, \dots, x_{h-1,n_{i-1}}\}$ from the final epoch are returned, and constituted output from MALA for our experimental assessment.

To sample from P instead of Π, we used Algorithm 4 we formally set $k_P(x) = 1$ for all $x \in \mathbb{R}^d$, which recovers $\Pi = P$ as the target.

D.4 Illustration on a GARCH Model

This appendix contains an additional illustrative experiment, concerning a GARCH model that is a particular instance of a model from the `PosteriorDB` database discussed in Section 4. The purpose of this illustration is to facilitate an empirical investigation in a slightly higher dimension ($d = 4$) and to explore the effect of changing the order s of the KGM–Stein kernel defined in Appendix A.3.2.

First we describe the GARCH model that was used. These models are widely-used in econometrics to describe time series data $\{y_t\}_{t=1}^n$ in settings where the volatility process is assumed to be time-varying (but stationary). In particular, we consider the GARCH(1,1) model

$$\begin{aligned} y_t &= \phi_1 + a_t, \\ a_t &= \sigma_t \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0,1), \\ \sigma_t^2 &= \phi_2 + \phi_3 a_{t-1}^2 + \phi_4 \sigma_{t-1}^2, \end{aligned}$$

where $\phi_2 > 0$, $\phi_3 > 0$, $\phi_4 > 0$, and $\phi_3 + \phi_4 < 1$ are the model parameters, constrained to a subset of $\mathbb{R}^4$. For ease of sampling, a change of variables $\tau : (\phi_1, \phi_2, \phi_3, \phi_4) \mapsto \theta$ is performed in such a way that the parameter $\theta \in \mathbb{R}^4$ is unconstrained. Assuming an improper flat prior on θ, the log-posterior density for θ is given up to an additive constant by

$$\log p(\theta \mid y_1, \dots, y_n) \stackrel{+C}{=} \sum_{t=1}^n \left[-\frac{1}{2} \log\bigl(\sigma_t^2\bigr) - \frac{y_t^2}{2\sigma_t^2} \right] + \log |J_{\tau^{-1}}(\theta)|,$$

where $|J_{\tau^{-1}}(\theta)|$ is the Jacobian determinant of τ^{-1}.

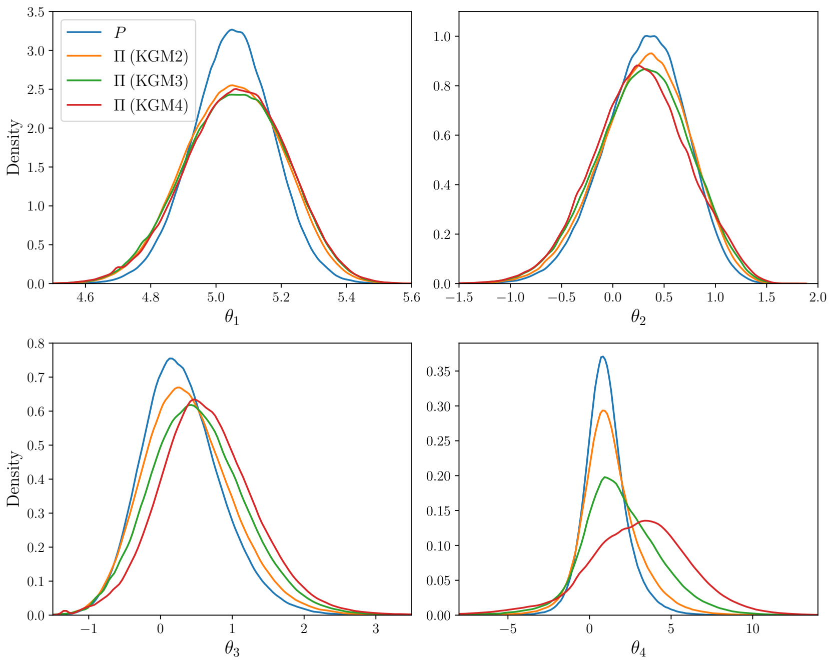


Figure S3: Illustrating the shape of Π based on the KGMs–Stein kernel for a GARCH(1,1) model, controlling convergence of moments up to order $s \in \{2, 3, 4\}$. The marginal density functions of each distribution were approximated using one-million samples obtained using MCMC.

For this illustration, real data were provided within the model description of `PosteriorDB`, for which the estimated *maximum a posteriori* parameter is $\hat{\phi} = (5.04, 1.36, 0.53, 0.31)$. The marginal distributions of Π corresponding to the KGM–Stein kernels of orders $s \in \{2, 3, 4\}$ are compared to the marginals of P in Figure S3. It can be seen that higher orders s correspond to greater over-dispersion of Π; this makes intuitive sense since larger s corresponds to a more stringent KSD (controlling the convergence of moments up to order s) which places greater emphasis on how the tails of P are approximated. Further, for the final skewed marginal of P, we note that the distribution Π exaggerates the skew, placing more of its mass in the tail of the direction which is positively skewed. Further discussion of skewed targets is contained in Appendix D.8.

D.5 Stein Π-Importance Sampling for `PosteriorDB`

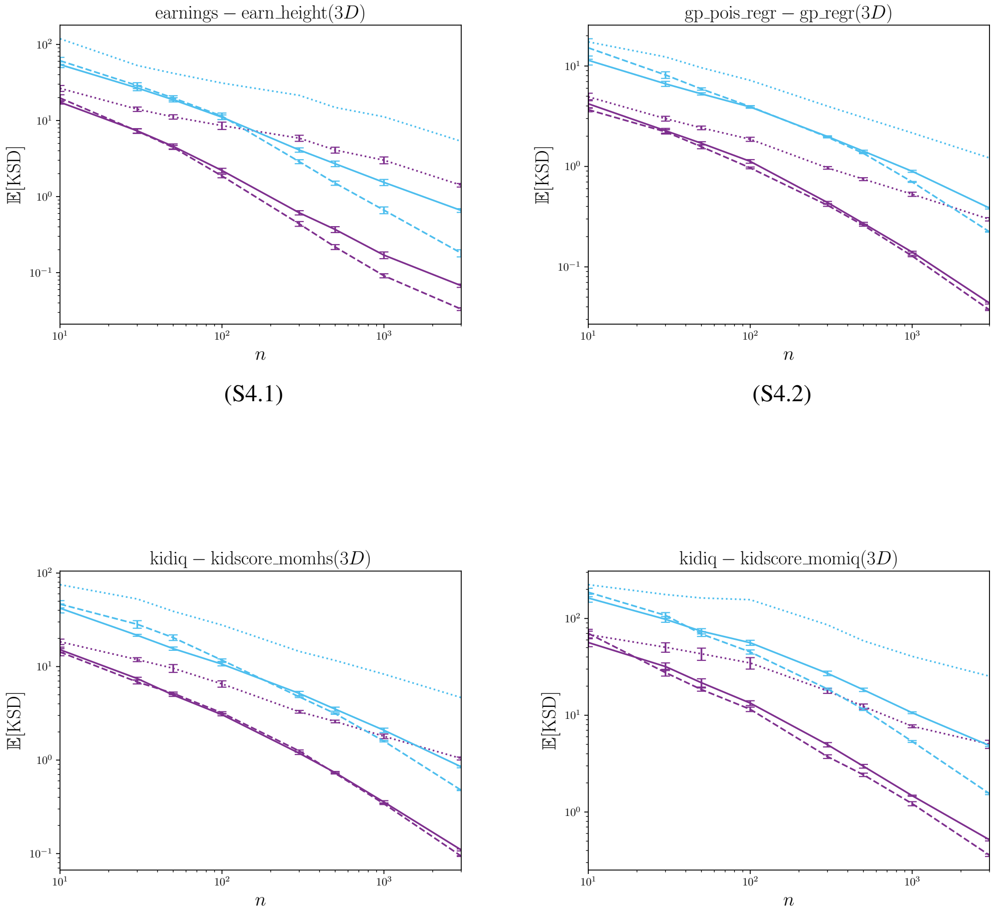
To introduce objectivity into our assessment, we exploited the `PosteriorDB` benchmark (Magnusson et al., 2022). This ongoing project is an attempt toward standardised benchmarking, consisting of a collection of posteriors to be numerically approximated. The test problems in `PosteriorDB` are defined in the `Stan` probabilistic programming language, and so `BridgeStan` (Roualdes et al., 2023) was used to directly access posterior densities and their gradients as required. The ambition of `PosteriorDB` is to provide an extensive set of benchmark tasks; at the time we conducted our research, `PosteriorDB` was at Version 0.4.0 and contained 149 models, of which 47 came equipped with a gold-standard sample of size $n = 10^3$, generated from a long run of Hamiltonian Monte Carlo (the No-U-Turn sampler in `Stan`). Of these 47 models, a subset of 40 were found to be compatible with `BridgeStan`, which was at Version 1.0.2 at the time this research was performed. The version of `Stan` that we used was `Stanc3` Version 2.31.0 (Unix). Thus we used a total of 40 test problems for our empirical assessment.

For each test problem, a total of 10 replicate experiments were performed and standard errors were computed. A sampling method was defined as being *significantly better* for approximation of a given target, compared to all other methods considered, if had lower mean KSD *and* the standard error bar did not overlap with the standard error bar of any other method. Table 1 in the main text summarises the performance of SΠIS-MALA, fixing the number of samples to be $n = 3 \times 10^3$. In this appendix, full empirical results are provided.

For sampling from MALA, we used the adaptive algorithm described in Appendix D.3 with a final epoch of length $n_{\max} = 10^5$. Then, whenever a set of $n \ll n_{\max}$ consecutive samples from MALA are required for our experimental assessment, these were obtained by selecting at random a consecutive sequence of length n from the total chain of length 10^5. This ensures that the performance of unprocessed MALA that we report is not negatively affected by burn-in, in so far as is practical to control.

Full results are presented in Figure S4. These results broadly support the interpretation that SΠIS-MALA usually outperforms SIS-MALA, or otherwise both methods provide a similar level of performance, for the sufficiently large sample sizes n considered. The sample size threshold at which SΠIS-MALA outperforms SIS-MALA appears to be dimension-dependent. A notable exception is panel 29 of Figure S4, a $d = 10$ dimensional task for which SΠIS-MALA provided a substantially worse approximation in KSD for the range of values of n considered.

Figure S4: Benchmarking on `PosteriorDB`. Here we compared raw output from MALA (dotted lines) with the post-processed output provided by the default Stein importance sampling method of Liu and Lee (2017) (SIS-MALA; solid lines) and the proposed Stein Π-Importance Sampling method (SΠIS-MALA; dashed lines). The Langevin (purple) and KGM3–Stein kernels (blue) were used for SIS-MALA and SΠIS-MALA and the associated KSDs are reported as the number n of iterations of MALA is varied. Ten replicates were computed and standard errors were plotted. The name of each model is shown in the title of the corresponding panel, and the dimension d of the parameter vector is given in parentheses. [Langevin–Stein kernel: ········MALA, ——SIS-MALA, - - - -SΠIS-MALA. KGM3–Stein kernel: ········MALA, ——SIS-MALA, - - - -SΠIS-MALA.]

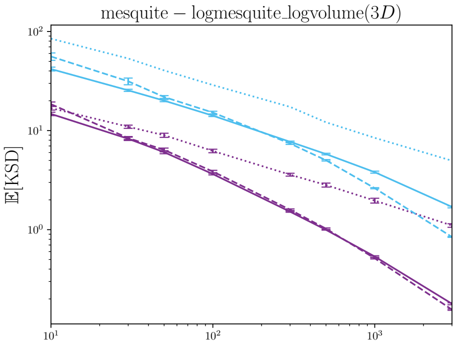


(S4.1)

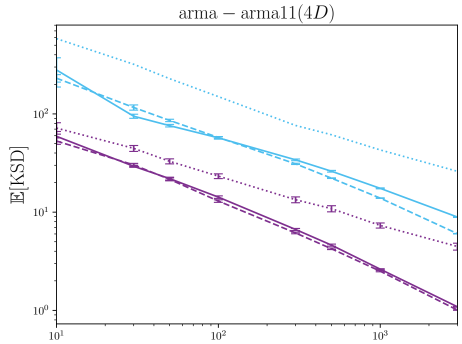
(S4.2)

(S4.3)

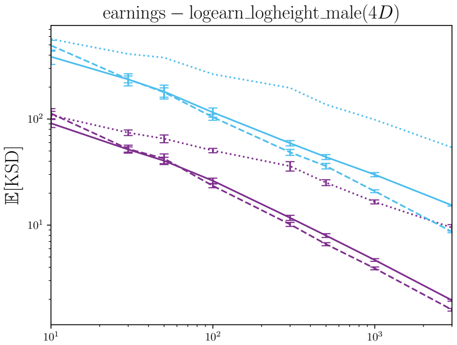
(S4.4)



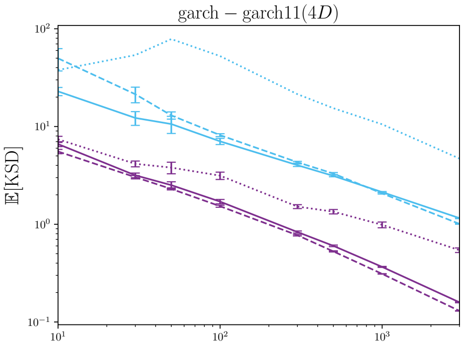
(S4.5)



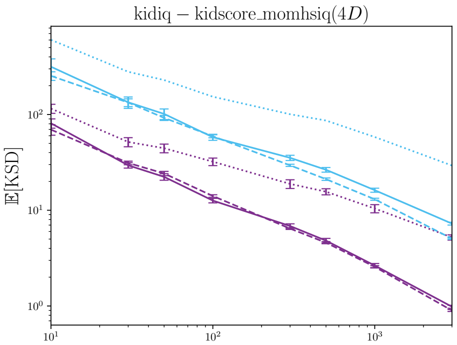
(S4.6)



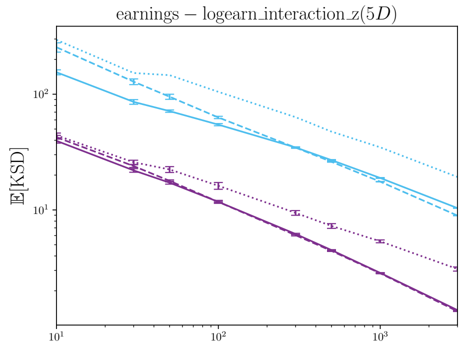
(S4.7)



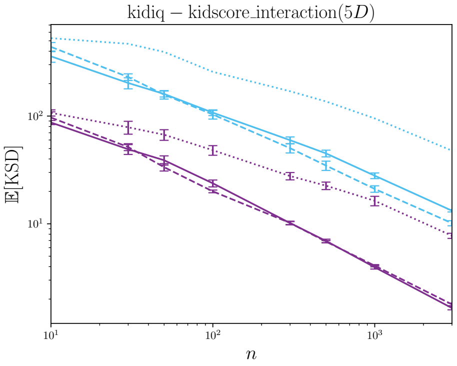
(S4.8)



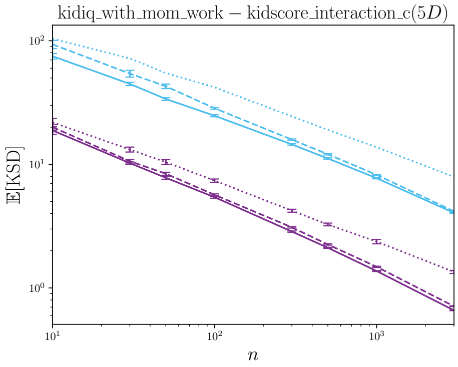
(S4.9)



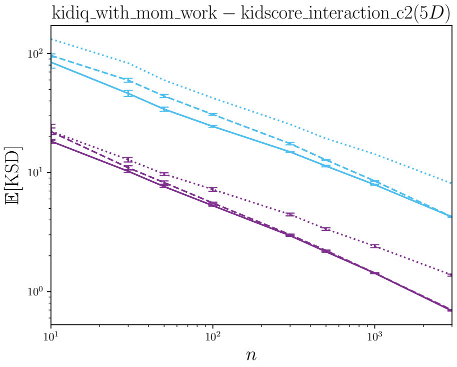
(S4.10)



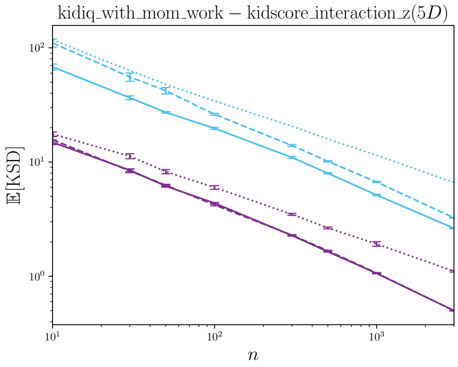
(S4.11)



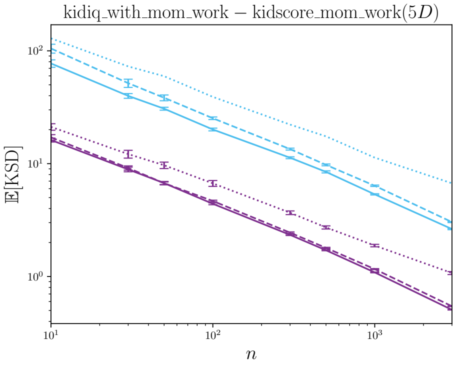
(S4.12)



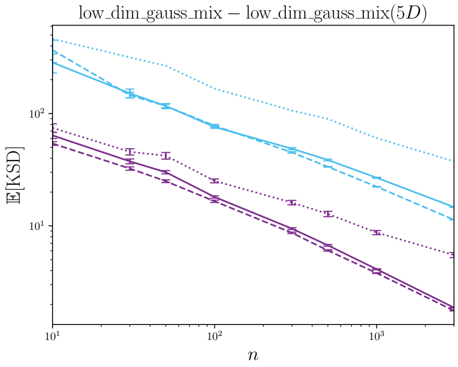
(S4.13)



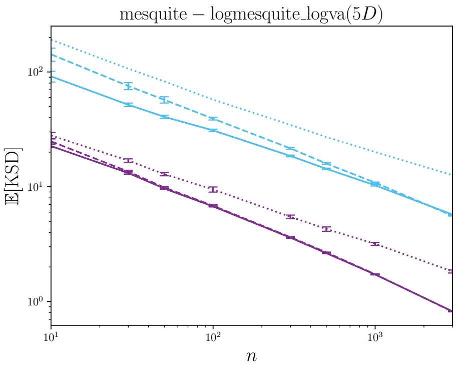
(S4.14)



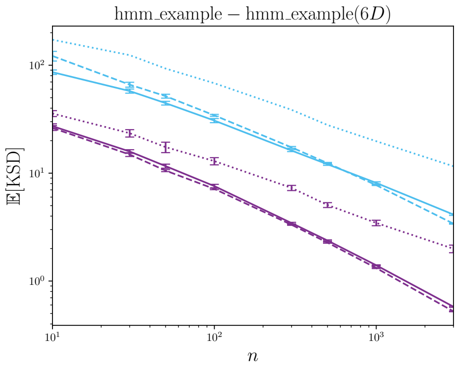
(S4.15)



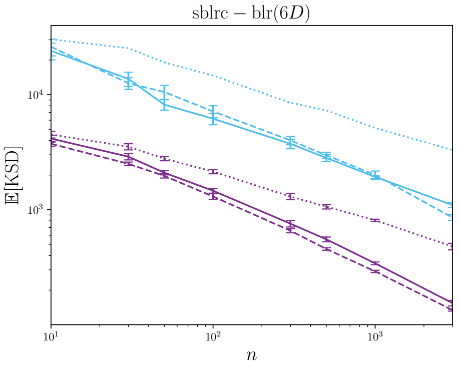
(S4.16)



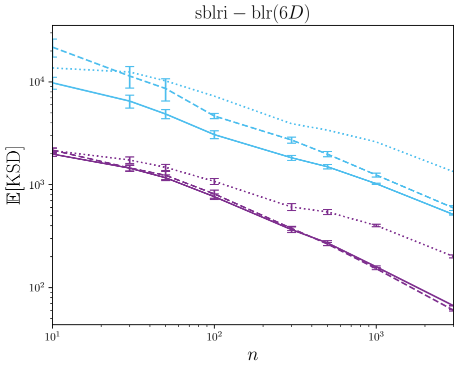
(S4.17)



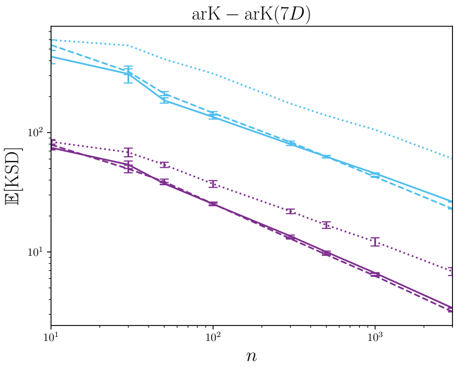
(S4.18)



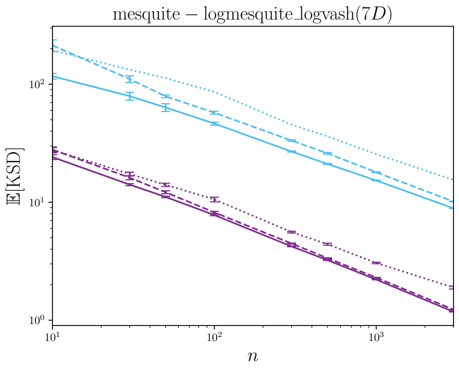
(S4.19)



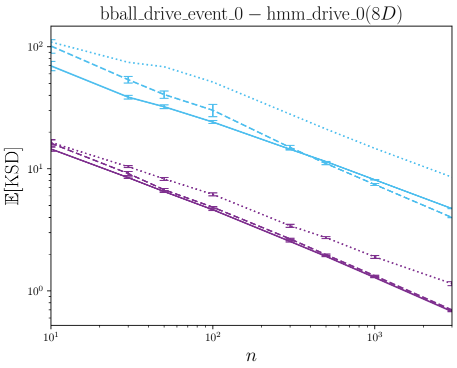
(S4.20)



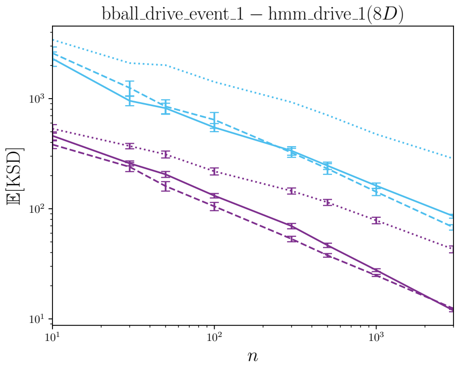
(S4.21)



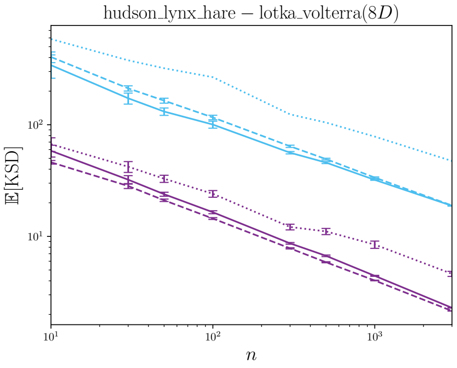
(S4.22)


bball_drive_event_0 − hmm_drive_0(8D)

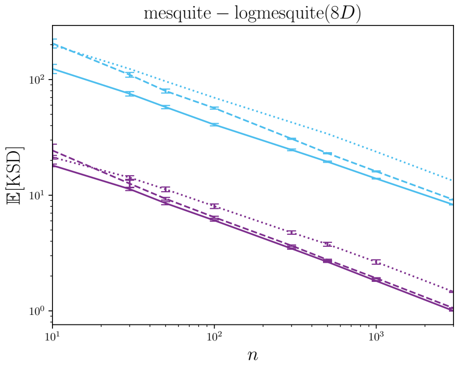
(S4.23)


bball_drive_event_1 − hmm_drive_1(8D)

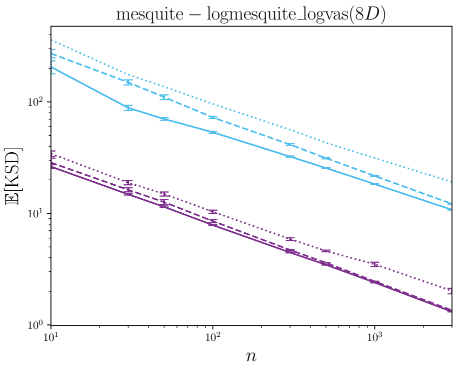
(S4.24)


hudson_lynx_hare − lotka_volterra(8D)

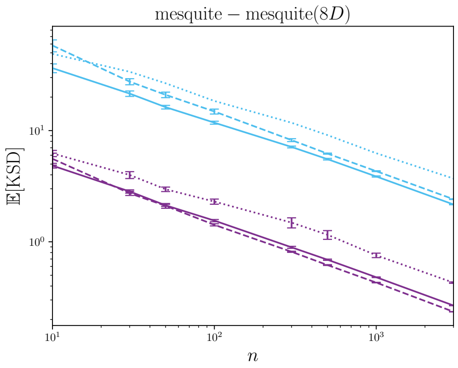
(S4.25)


mesquite − logmesquite(8D)

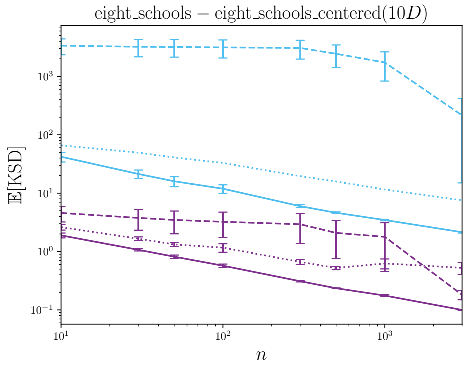
(S4.26)


mesquite − logmesquite_logvas(8D)

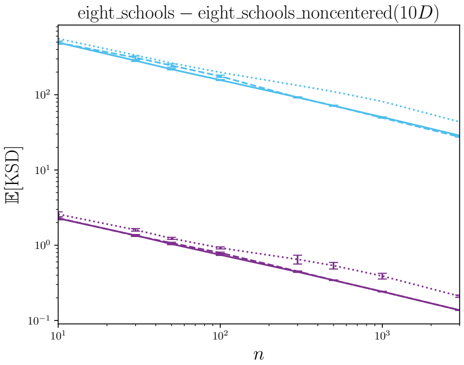
(S4.27)


mesquite − mesquite(8D)

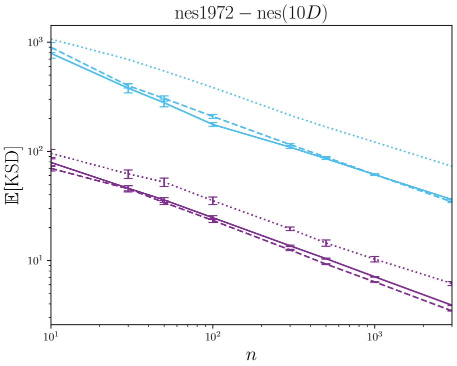
(S4.28)



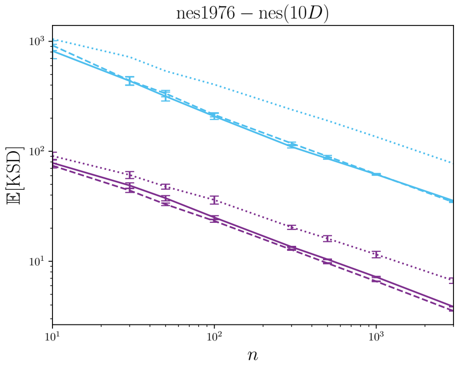
(S4.29)



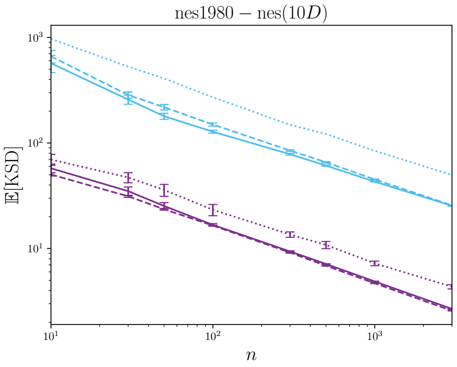
(S4.30)



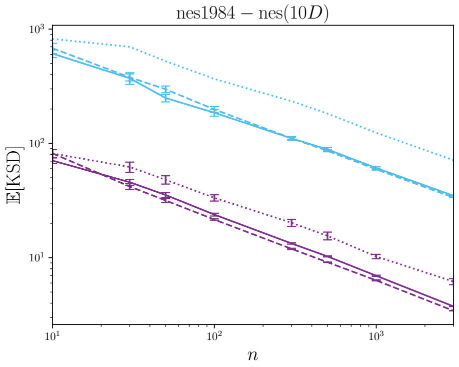
(S4.31)



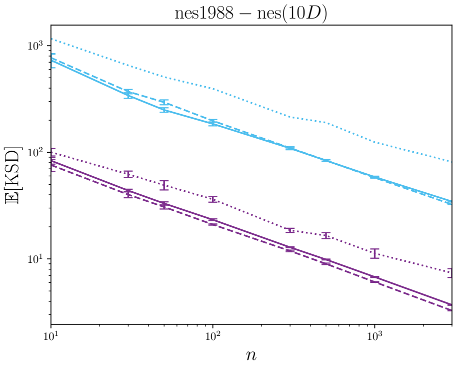
(S4.32)



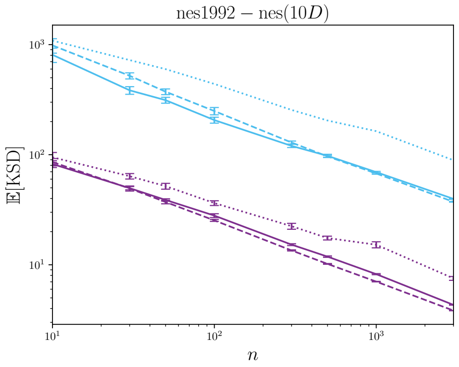
(S4.33)



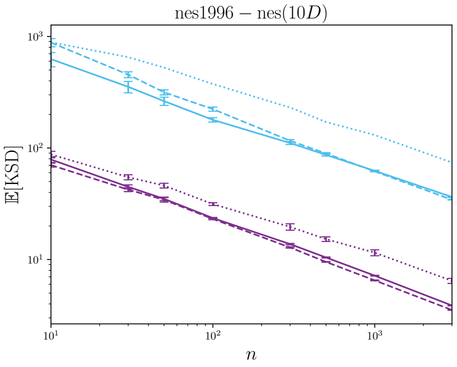
(S4.34)



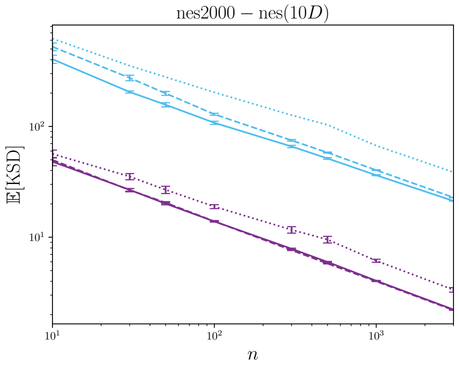
(S4.35)



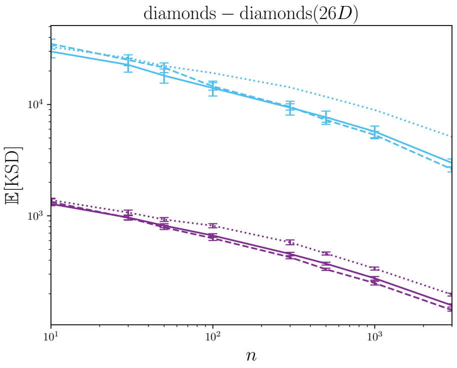
(S4.36)



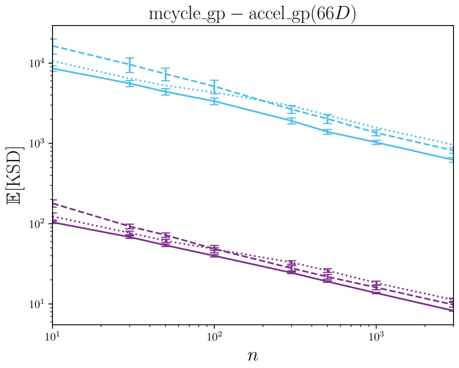
(S4.37)



(S4.38)



(S4.39)



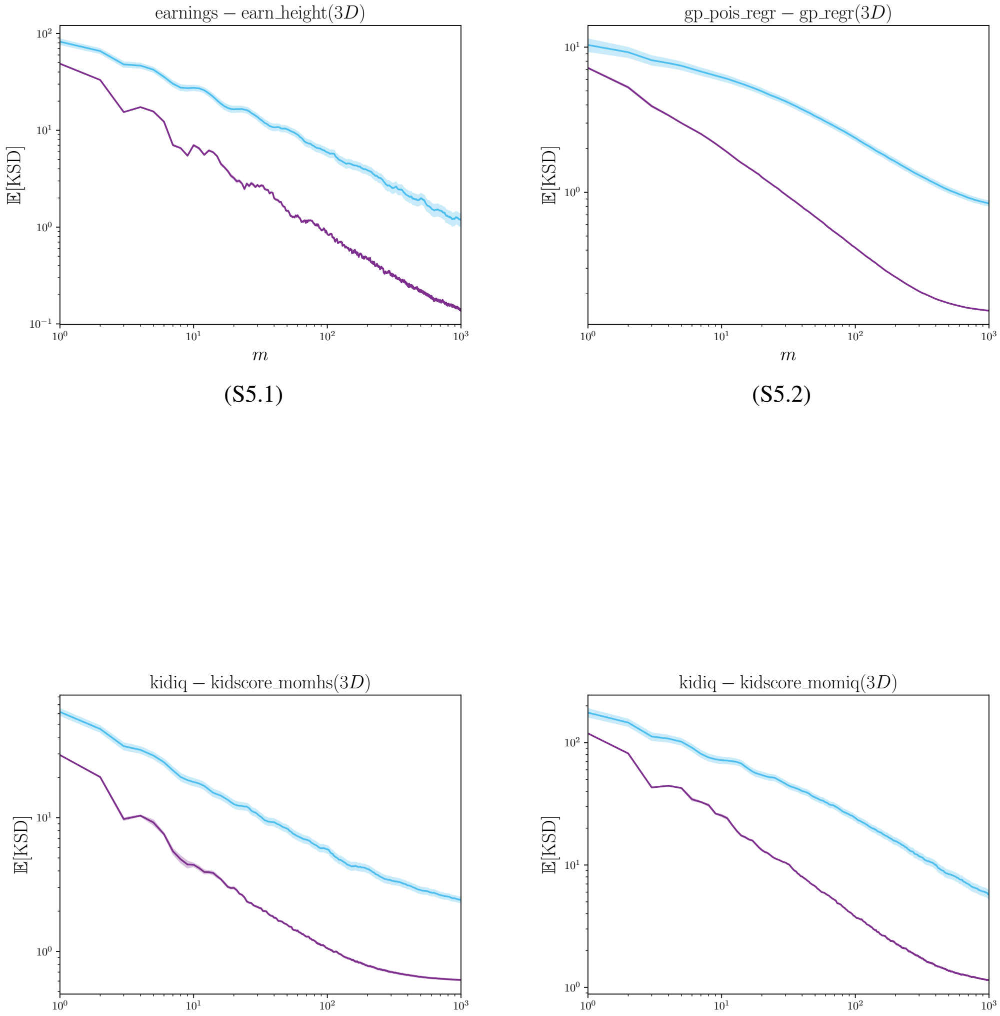
(S4.40)

D.6 Stein Π-Thinning for `PosteriorDB`

The results presented in the main text concerned $n = 3 \times 10^3$ samples from MALA, which is near the limit at which the optimal weights $w^\star$ can be computed in a few seconds on a laptop PC. For larger values of n, sparse approximation methods are likely to required. In the main text we presented *Stein* Π-*Thinning*, which employs a greedy optimisation perspective to obtain a sparse approximation to the optimal weights at cost $O(m^2 n)$, where m are the number of greedy iterations performed. Explicit and verifiable conditions for the strong consistency of the resulting SΠT-MALA algorithm were established in Section 3.3. The purpose of this appendix is to empirically explore the convergence of SΠT-MALA using the `PosteriorDB` test bed.

In the experiments we report the number of MALA samples was fixed to $n = 10^3$ and the number of greedy iterations was varied from $m = 1$ to $m = 10^3$. The results, in Figure S5, indicate that for most models in `PosteriorDB` the minimum value of KSD is approximately reached when m is anywhere from $\frac{n}{10}$ to $\frac{n}{2}$, representing a modest but practically significant reduction in computational cost compared to SΠIS-MALA. This agrees with the qualitative findings reported in the original Stein thinning paper of Riabiz et al. (2022).

Figure S5: Benchmarking on `PosteriorDB`. Here we investigate the convergence of the sparse approximation provided by the proposed Stein Π-Thinning method (SΠT-MALA). The Langevin (purple) and KGM3–Stein kernels (blue) were used for SΠT-MALA and the associated KSDs are reported as the number m of iterations of Stein thinning is varied. Ten replicates were computed and standard errors were plotted. The name of each model is shown in the title of the corresponding panel, and the dimension d of the parameter vector is given in parentheses. [Langevin–Stein kernel: — SΠT-MALA. KGM3–Stein kernel: — SΠT-MALA.]

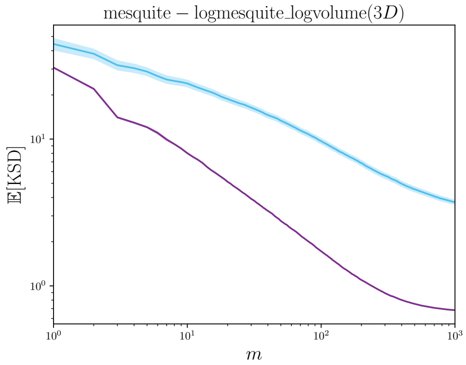


(S5.1)

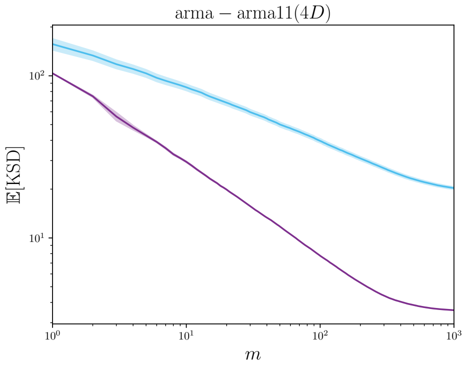
(S5.2)

(S5.3)

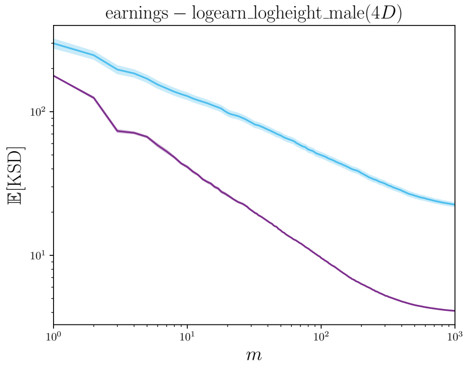
(S5.4)



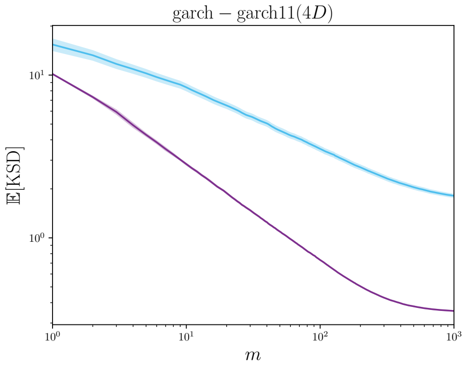
(S5.5)



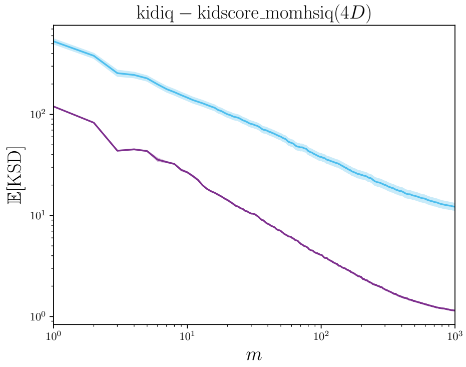
(S5.6)



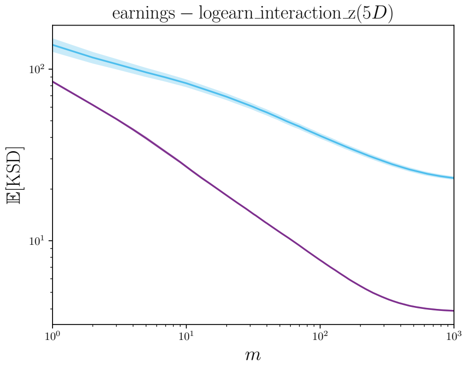
(S5.7)



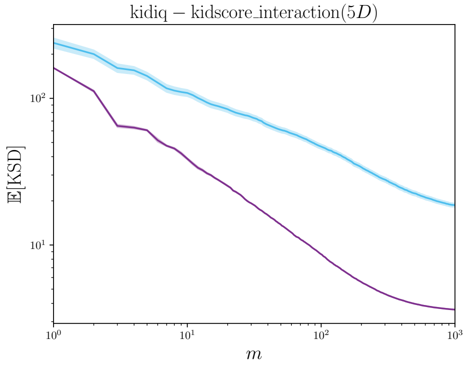
(S5.8)



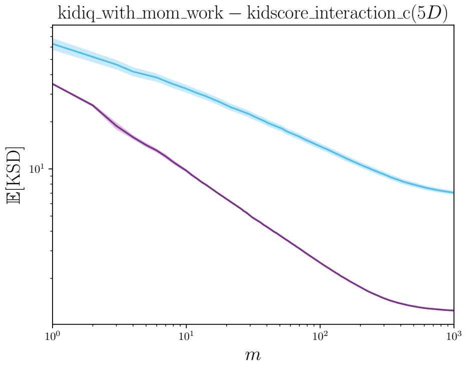
(S5.9)



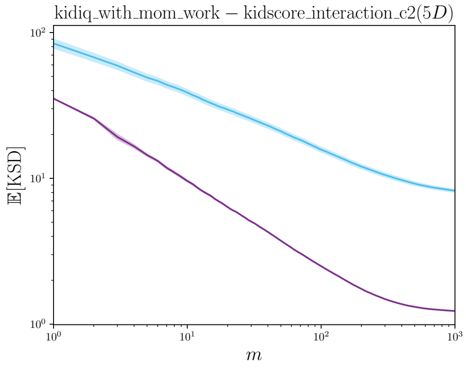
(S5.10)


kidiq − kidscore_interaction($5D$)

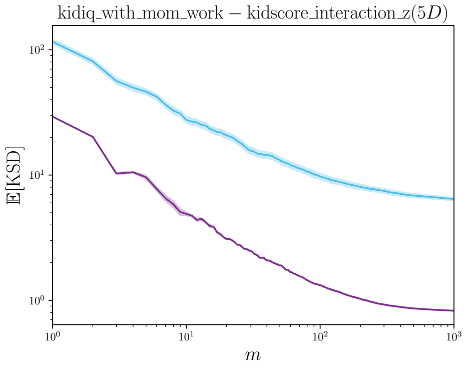
(S5.11)


kidiq_with_mom_work − kidscore_interaction_c($5D$)

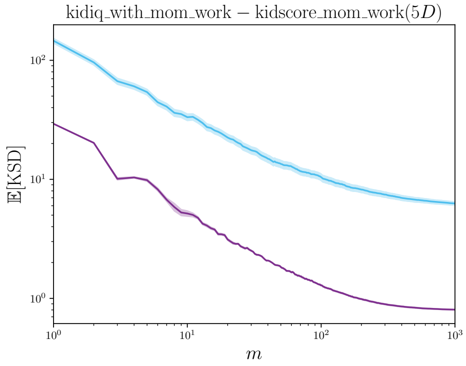
(S5.12)


kidiq_with_mom_work − kidscore_interaction_c2($5D$)

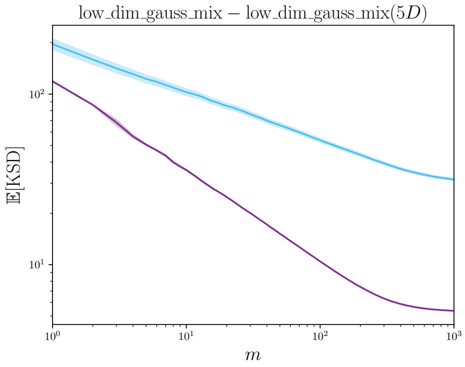
(S5.13)


kidiq_with_mom_work − kidscore_interaction_z($5D$)

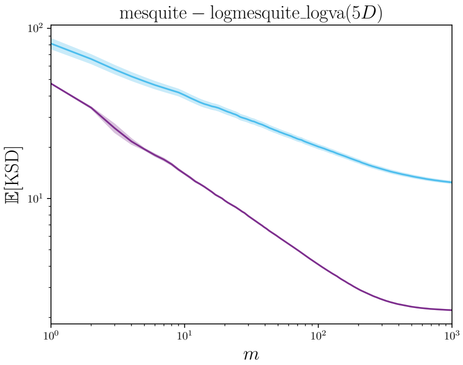
(S5.14)


kidiq_with_mom_work − kidscore_mom_work($5D$)

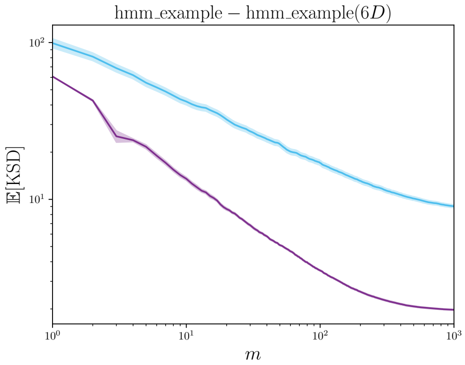
(S5.15)


low_dim_gauss_mix − low_dim_gauss_mix($5D$)

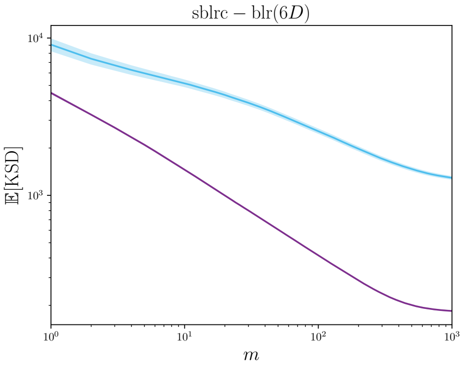
(S5.16)



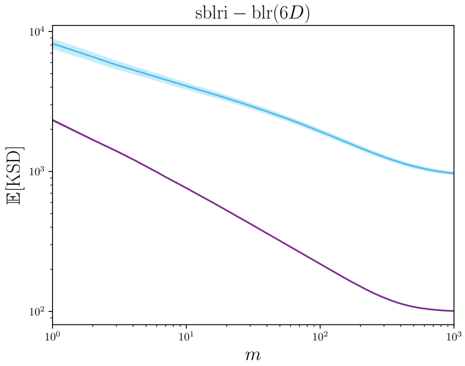
(S5.17)



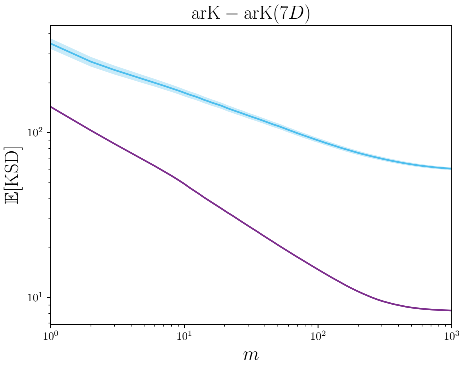
(S5.18)



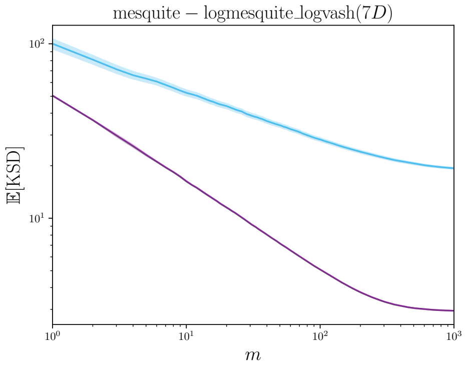
(S5.19)



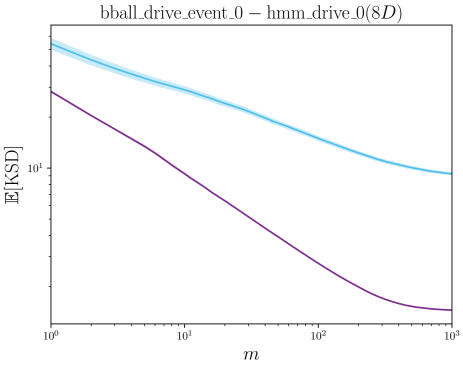
(S5.20)



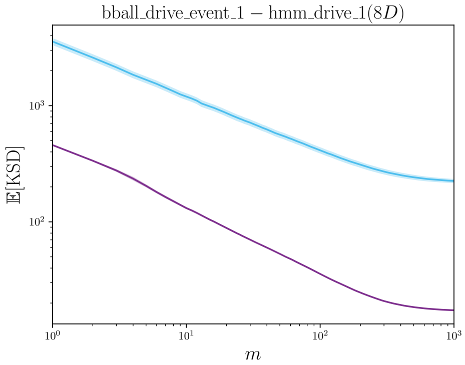
(S5.21)



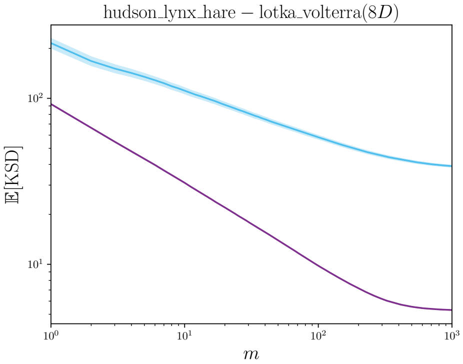
(S5.22)



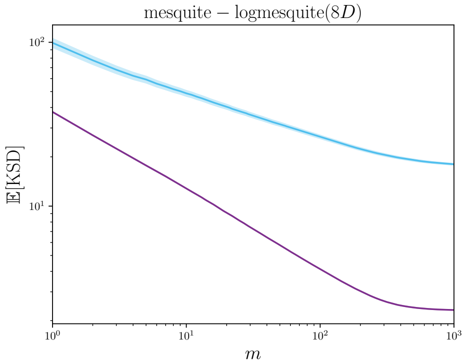
(S5.23)



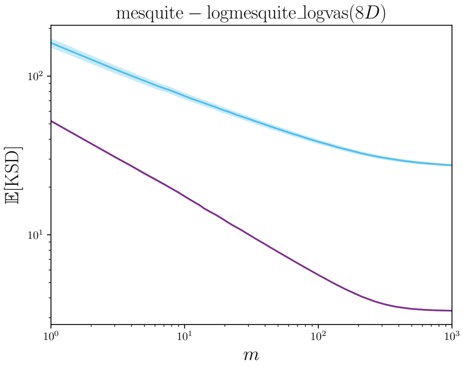
(S5.24)



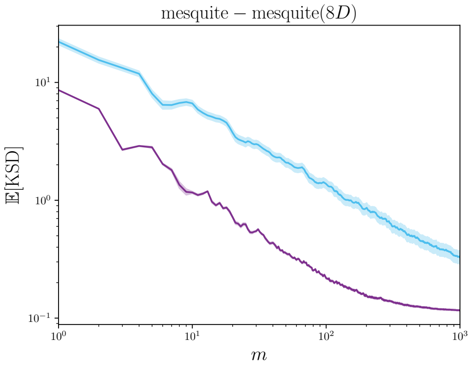
(S5.25)



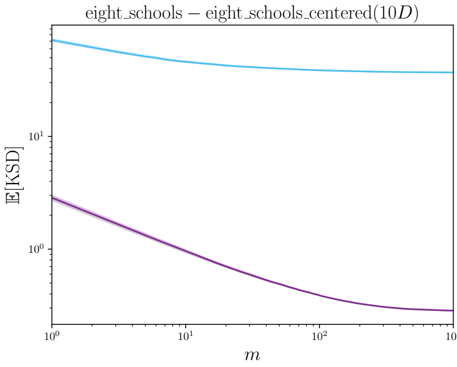
(S5.26)



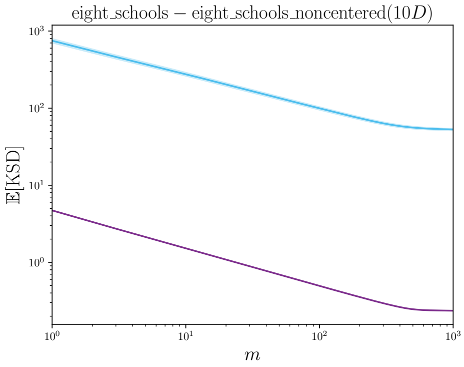
(S5.27)



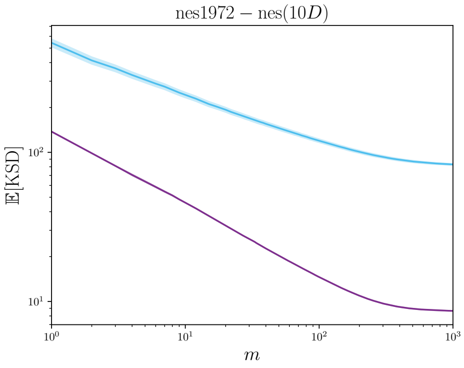
(S5.28)



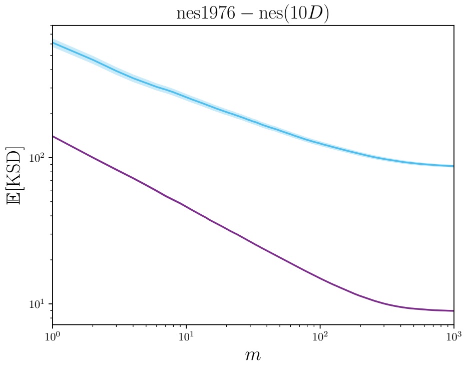
(S5.29)



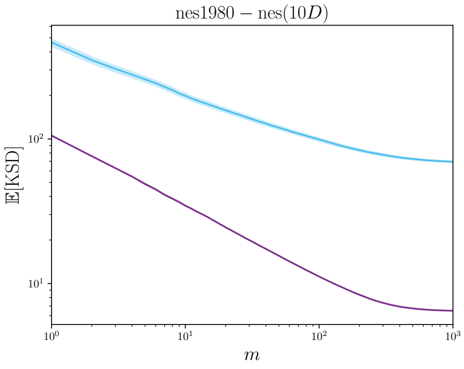
(S5.30)



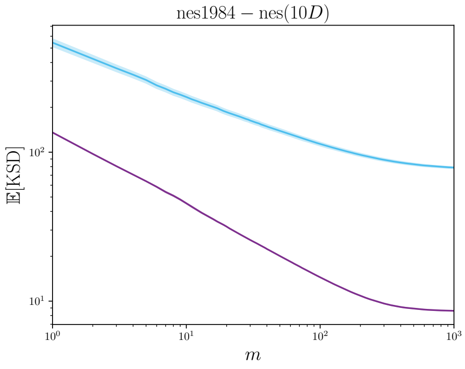
(S5.31)



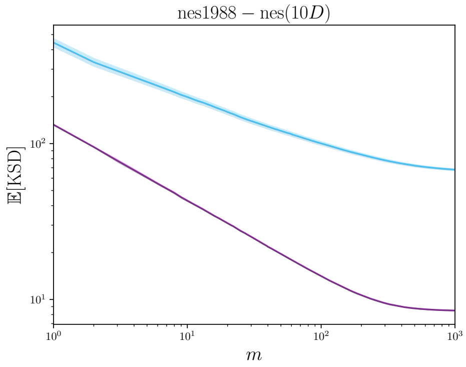
(S5.32)



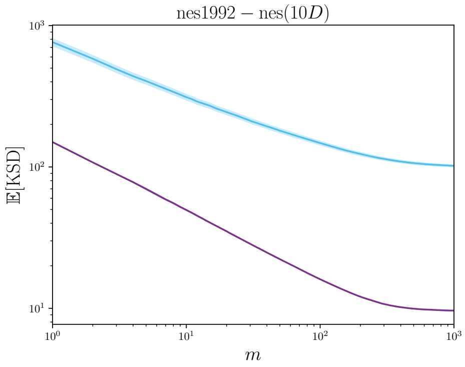
(S5.33)



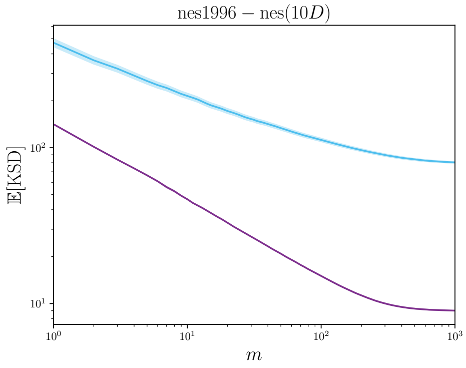
(S5.34)



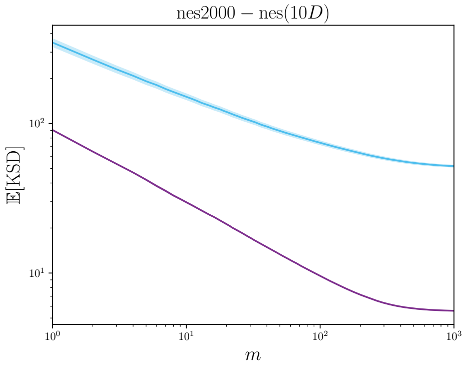
(S5.35)



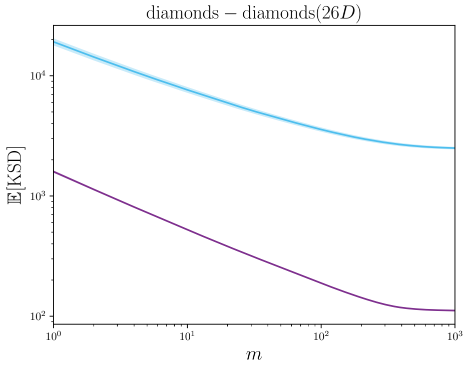
(S5.36)



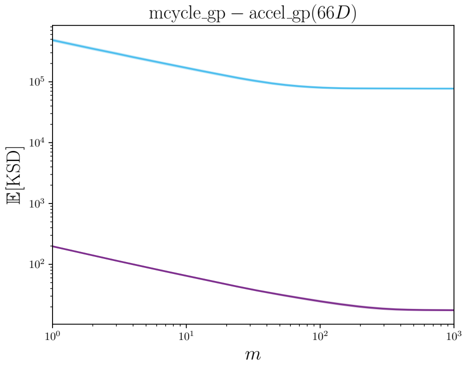
(S5.37)



(S5.38)



(S5.39)



(S5.40)

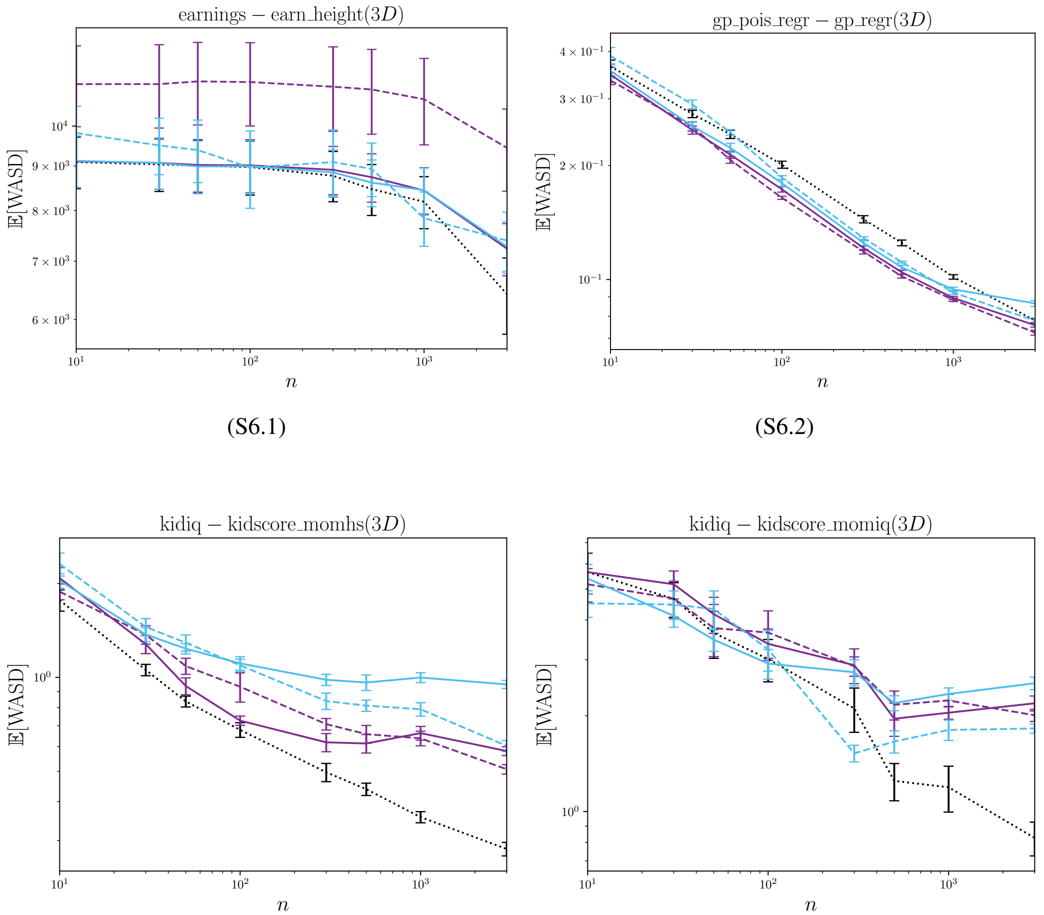
D.7 Performance of Stein Discrepancies

The properties of Stein discrepancies was out of scope for this work. Nonetheless, there is much interest in better understanding the properties of KSDs, and in this appendix the performance of SΠIS-MALA in terms of 1-Wasserstein divergence is reported. This was made possible since `PosteriorDB` supplies a set of posterior samples obtained from a long run of Hamiltonian Monte Carlo (the No-U-Turn sampler in `Stan`) which we treat as a gold standard.

Full results are presented in S6. Broadly speaking, for most models the minimisation of KSD seems to be associated with minimisation of 1-Wasserstein distance, however there are some models for which minimisation of KSD is loosely, if at all, related to minimisation of 1-Wasserstein divergence. In these cases, we attribute this performance to the *blindness to mixing proportions* phenomena, described in Wenliang and Kanagawa (2021); Koehler et al. (2022); Liu et al. (2023). Convergence in 1-Wasserstein is equivalent to weak convergence plus convergence of the first moment, so the KGM–Stein kernels of order $s \geq 1$ control convergence in 1-Wasserstein. In Section 2.3 we proved that SΠIS-MALA is strongly consistent in KSD for the KGM–Stein kernel in the case $s = 1$, so we can expect strong consistency in 1-Wasserstein divergence for SΠIS-MALA in this case as well. It is interesting to observe that better 1-Wasserstein quantisations tend to be provided by SΠIS-MALA compared to SIS-MALA when either the Langevin–Stein or KGM–Stein kernel are used.

The development of improved Stein discrepancies is an active area of research, and we emphasise that the methodology developed in this work can be applied to *any* KSDs, including potentially KSDs with better or more direct control over standard notions of convergence (such as 1-Wasserstein) that in the future may be developed.

Figure S6: Performance of Stein discrepancies on `PosteriorDB`. Here we compared raw output from MALA (dotted lines) with the post-processed output provided by the default Stein importance sampling method of Liu and Lee (2017) (SIS-MALA; solid lines) and the proposed Stein Π-Importance Sampling method (SΠIS-MALA; dashed lines). The Langevin (purple) and KGM3–Stein kernels (blue) were used for SIS-MALA and SΠIS-MALA, and the 1-Wasserstein divergence is reported as the number n of iterations of MALA is varied. Ten replicates were computed and standard errors were plotted. The name of each model is shown in the title of the corresponding panel, and the dimension d of the parameter vector is given in parentheses. [Legend: ⋯⋯ Raw MALA. Langevin–Stein kernel: ━━ SIS-MALA, - - - SΠIS-MALA. KGM3–Stein kernel: ━━ SIS-MALA, - - - SΠIS-MALA.]

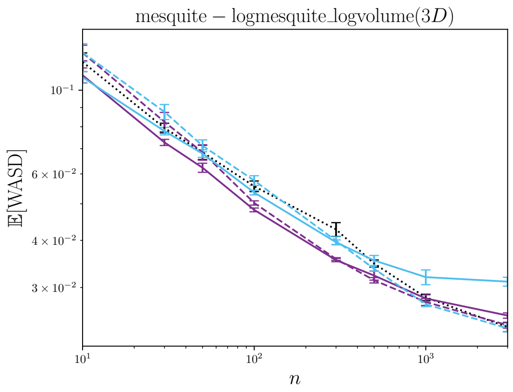


(S6.1)

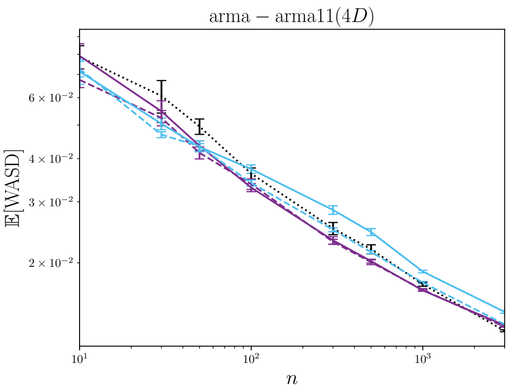
(S6.2)

(S6.3)

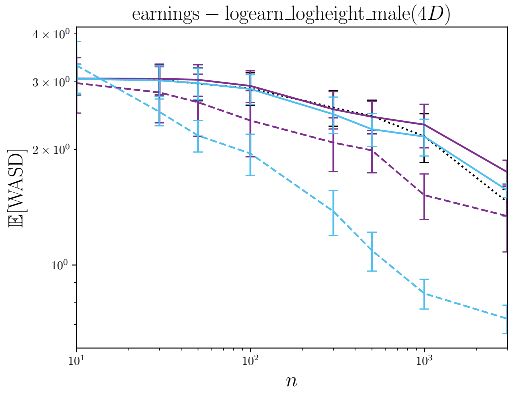
(S6.4)



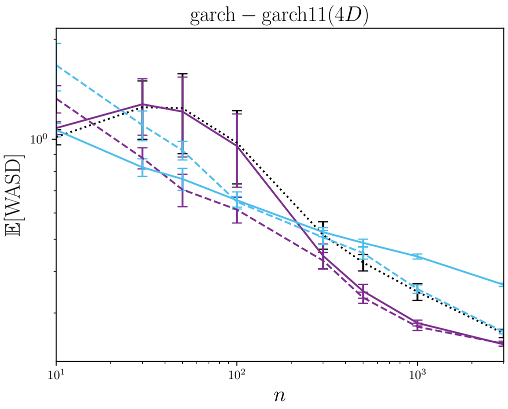
(S6.5)



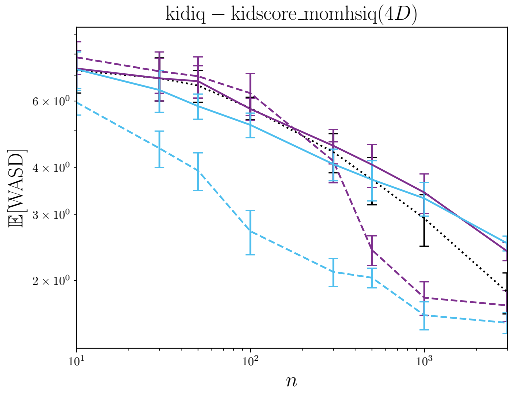
(S6.6)



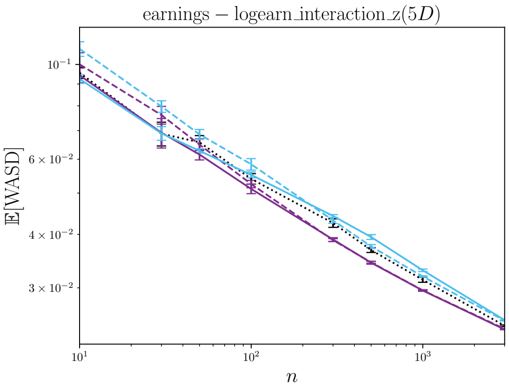
(S6.7)



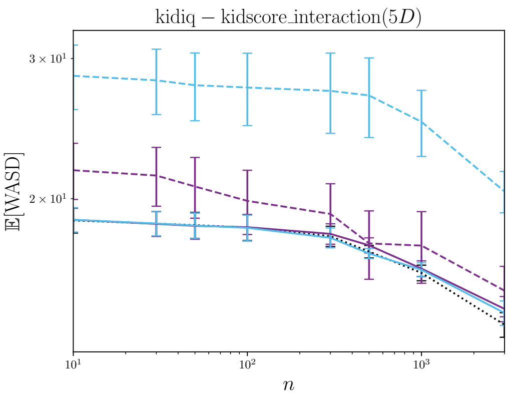
(S6.8)



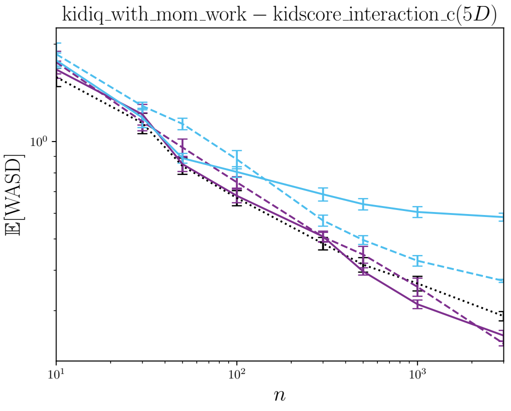
(S6.9)



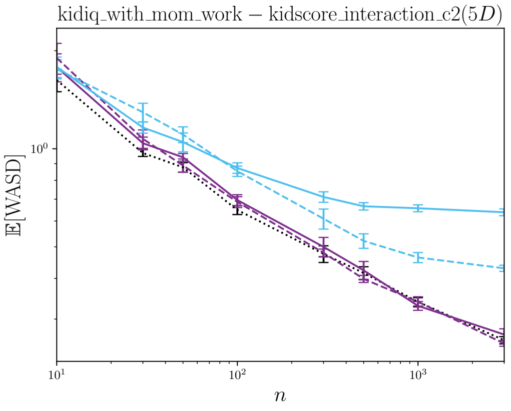
(S6.10)



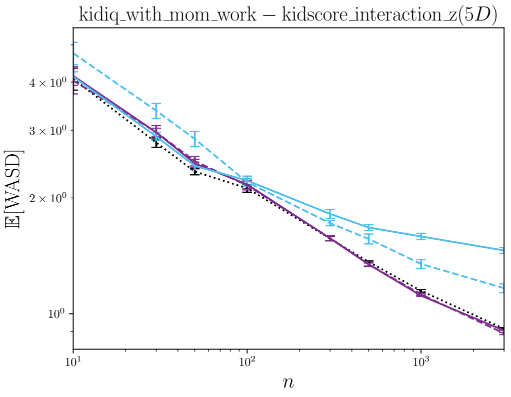
(S6.11)



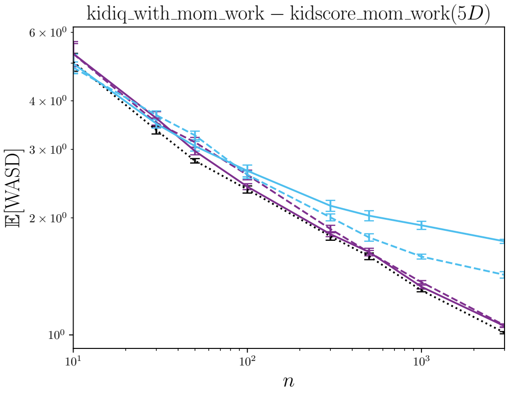
(S6.12)



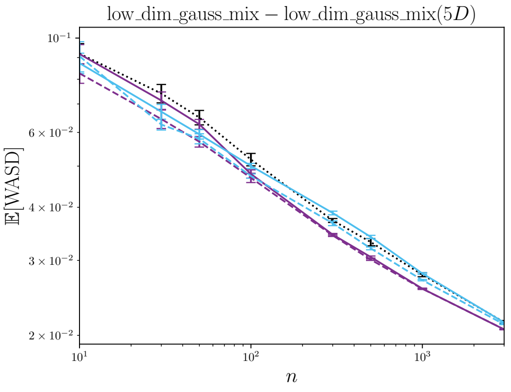
(S6.13)



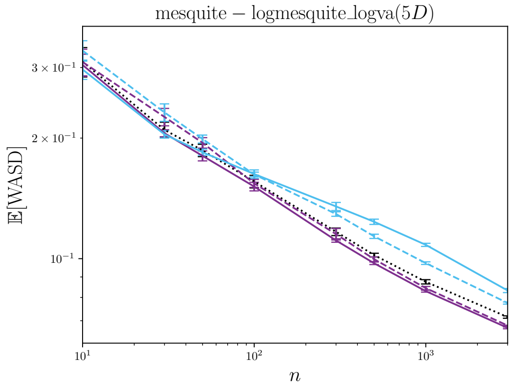
(S6.14)

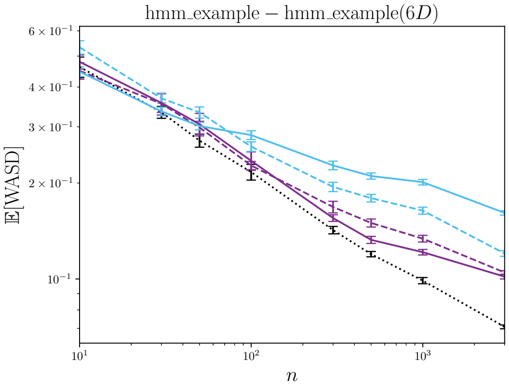


(S6.15)



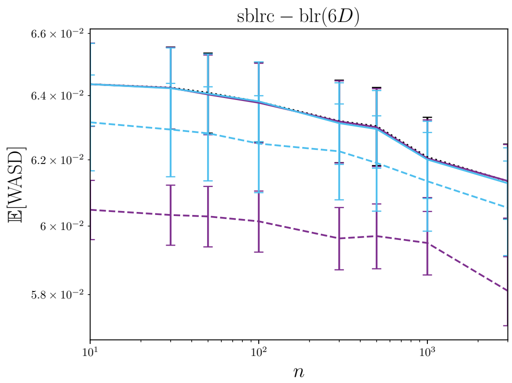
(S6.16)

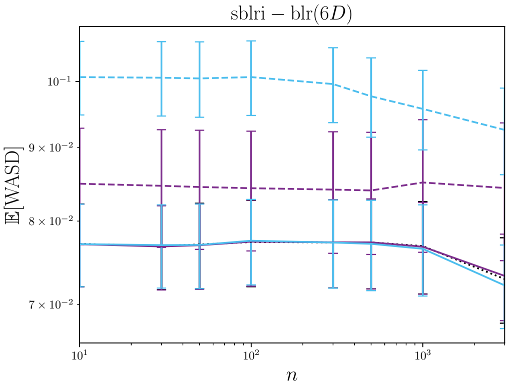




(S6.17)

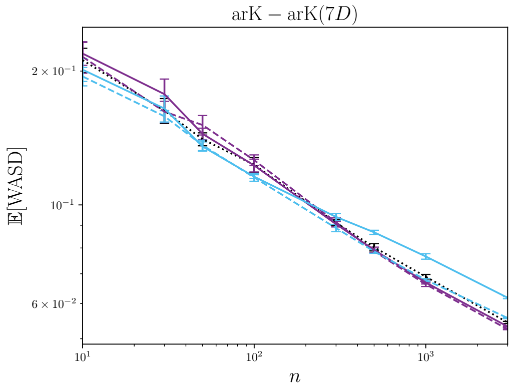
(S6.18)

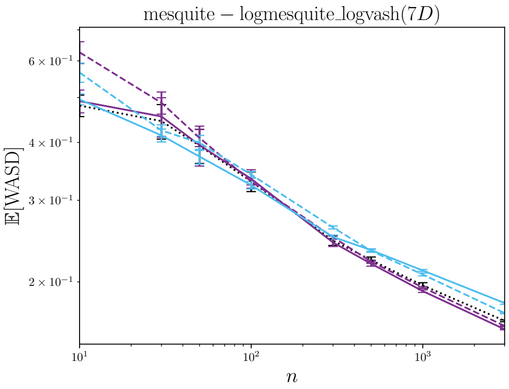




(S6.19)

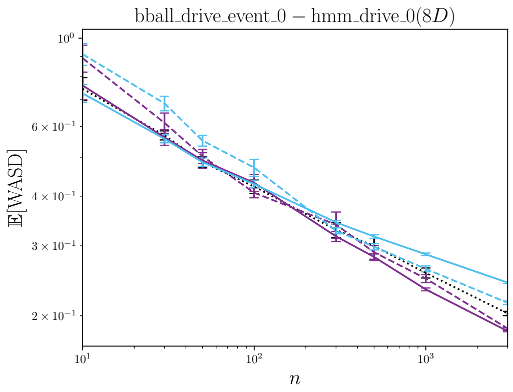
(S6.20)



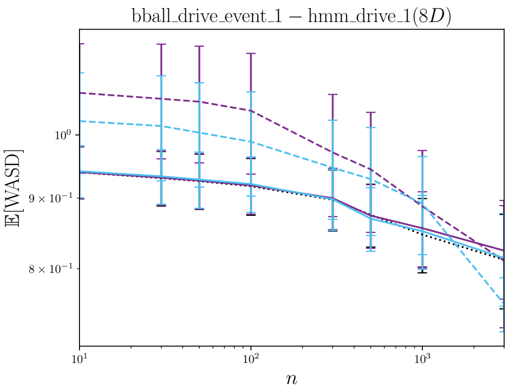


(S6.21)

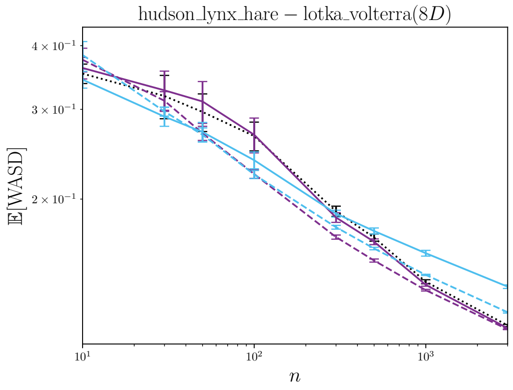
(S6.22)



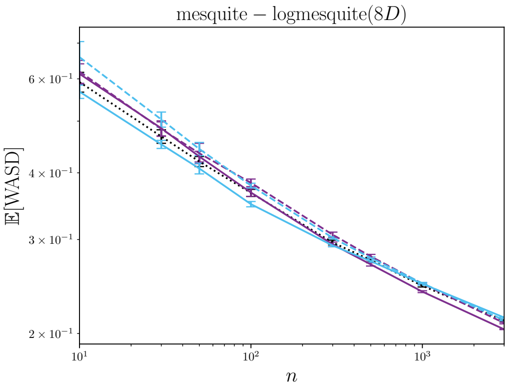
(S6.23)



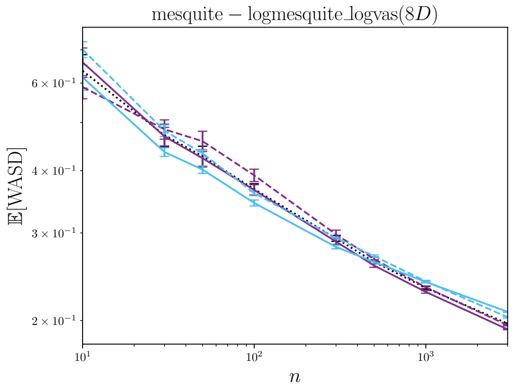
(S6.24)



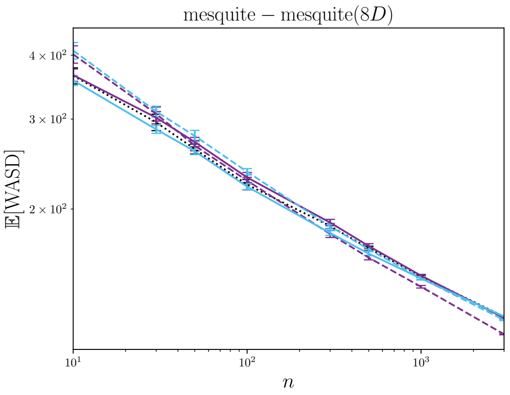
(S6.25)



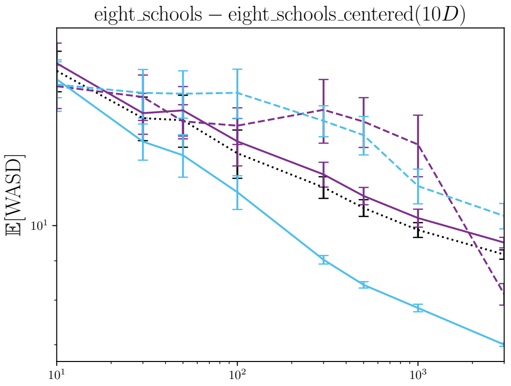
(S6.26)



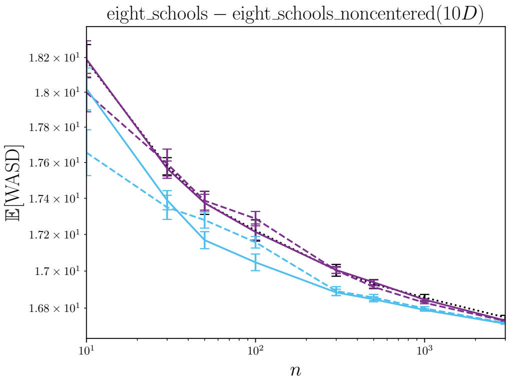
(S6.27)



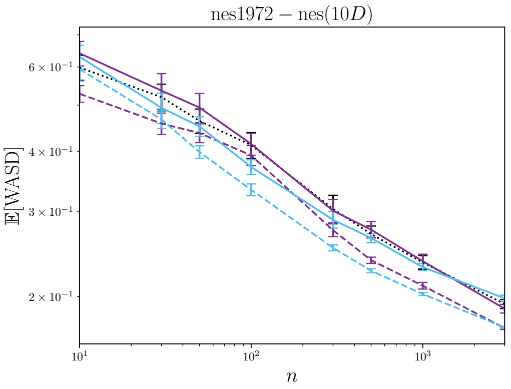
(S6.28)



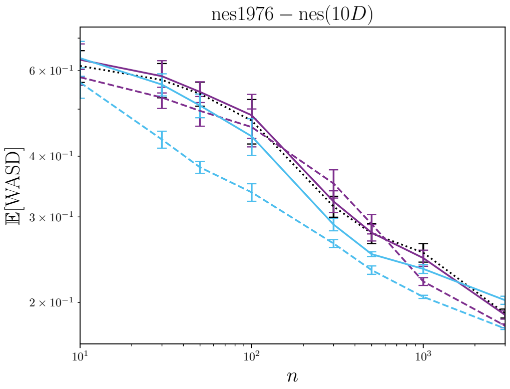
(S6.29)



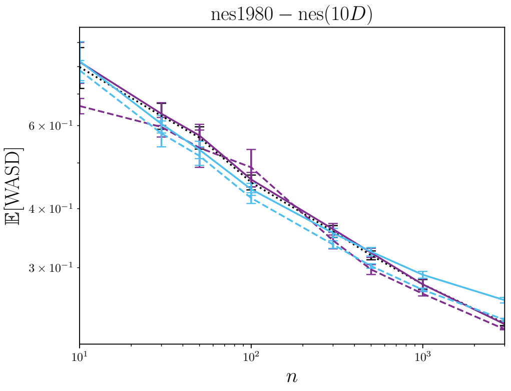
(S6.30)



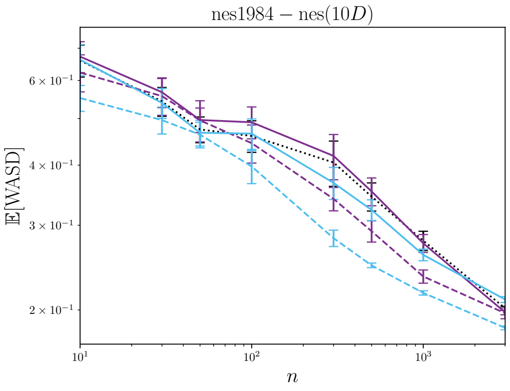
(S6.31)



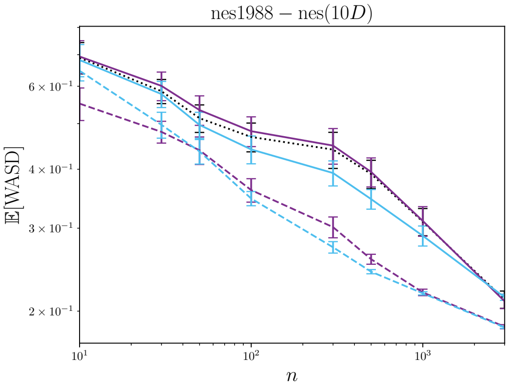
(S6.32)



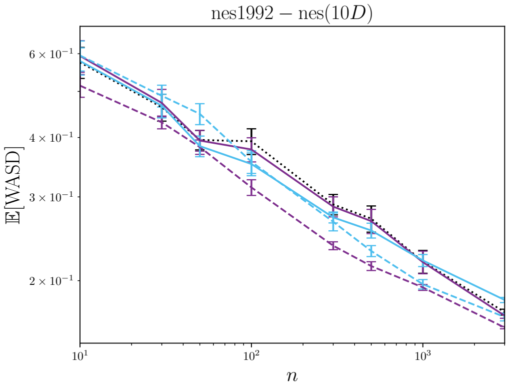
(S6.33)



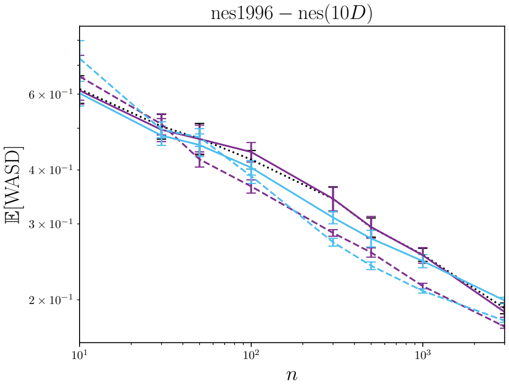
(S6.34)



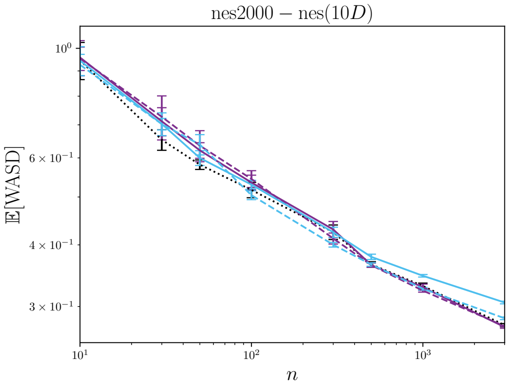
(S6.35)



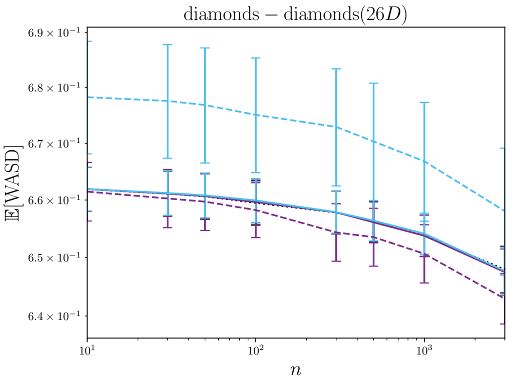
(S6.36)



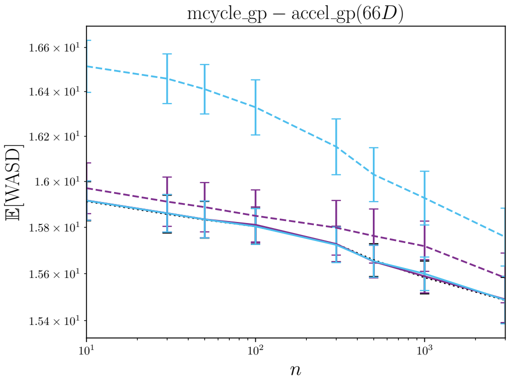
(S6.37)



(S6.38)



(S6.39)



(S6.40)

D.8 Investigation for a Skewed Target

This final appendix contrasts the 1-Wasserstein optimal sampling distribution Π_1 (c.f. Section 2.1), with the choice of Π that we recommended in (8). In particular, we focus on the KGM3–Stein kernel under a heavily skewed P, for which Π_1 and Π can be markedly different.

For this investigation a bivariate skew-normal target was constructed, where the density is given by $p(x_1, x_2) = 4\phi(x_1)\Phi(6x_1)\phi(x_2)\Phi(-3x_2)$, with ϕ and Φ respectively denoting the density and distribution functions of a standard Gaussian. The density p of P, together with the marginal densities of Π_1 and Π, are plotted in Figure S7. It can be seen that, while both Π_1 and Π are over-dispersed

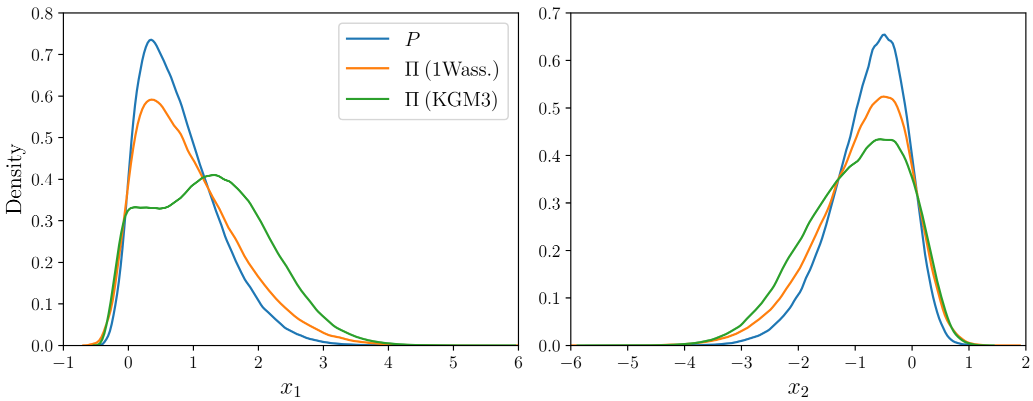


Figure S7: Comparing the proposed distribution Π (KGM3; based on the KGM3–Stein kernel) to Π_1 (1Wass.; the optimal choice for 1-Wasserstein quantisation from Section 2.1) for a bivariate skew-normal target ($d = 2$). The marginal density functions of each distribution were approximated using 10^6 samples from MCMC.

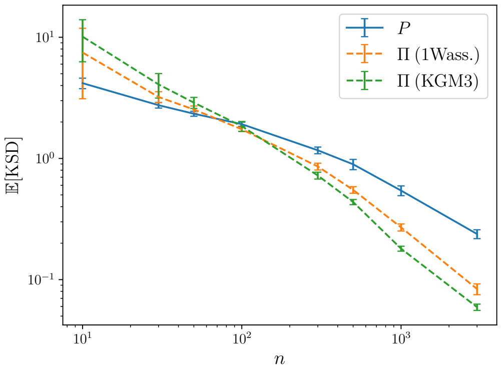


Figure S8: Comparing the performance of using the proposed distribution Π (KGM3; based on the KGM3–Stein kernel) to Π_1 (1Wass.; the optimal choice for 1-Wasserstein quantisation from Section 2.1) for a bivariate skew-normal target ($d = 2$). The mean kernel Stein discrepancy (KSD) for Stein Π-Importance Sampling was estimated; in each case, the KSD based on the KGM3–Stein kernel was computed. Solid lines indicate the baseline case of sampling from P, while dashed lines indicate sampling from Π. (The experiment was repeated 10 times and standard error bars are plotted.)

with respect to P, our recommended Π assigns proportionally more mass to the tail that is positively skewed.

The performance of Stein Π-Importance Sampling based on Π_1 and Π is compared in Figure S8. Though both choices lead to an improvement relative to Stein importance sampling algorithm with $\Pi = P$, the use of Π leads to a significant further reduction (on average) in KSD compared to Π_1. Based on our investigations, this finding seems general; the use of Π_1 does not realise the full potential of Stein Π-Imporance sampling when the target is skewed.