Supplement

This electronic supplement is organised as follows: In Section A proofs for the theoretical results in the main text are provided. In Section B we provide details for the two existing methods (MED, SVGD) that formed our experimental benchmark. Then, in Section C, we provide additional numerical results that elaborate on those reported in the main text.

Code Code to reproduce these experiments is available from:

`github.com/wilson-ye-chen/stein_points`

A. Proof of Theoretical Results in the Main Text

A.1. Proofs of Theorems 1 and 2: Stein Herding and Stein Greedy Convergence

We will show that both Theorem 1 and Theorem 2 follow from the following unified Stein Point convergence result, proved in Section A.1.3.

Theorem 5 (Stein Point Convergence)**.** *Suppose k_0 with $k_{0,P} = 0$ is a P-sub-exponential reproducing kernel. Then there exist constants $c_1, c_2 > 0$ depending only on k and P such that any point sequence $\{x_i\}_{i=1}^n$ satisfying*

$$\frac{k_0(x_j,x_j)}{2} + \sum_{i=1}^{j-1} k_0(x_i,x_j) \leq \frac{\delta}{2} + \frac{S_j^2}{2} + \min_{x\in X: k_0(x,x)\leq S_j^2} \sum_{i=1}^{j-1} k_0(x_i,x)$$

with $S_j \in [\sqrt{2\log(j)/c_2}, \sqrt{2\log(n)/c_2}]$ for each $1 \leq j \leq n$ and $\delta \geq 0$ also satisfies

$$D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n) \leq e^{\pi/2}\sqrt{\tfrac{2\log(n)}{c_2 n} + \tfrac{c_1}{n} + \tfrac{\delta}{n}}.$$

A.1.1. Proof of Theorem 1: Stein Herding Convergence

Instantiate the constants $c_1, c_2 > 0$ from Theorem 5, and consider any point sequence $\{x_i\}_{i=1}^n$ satisfying

$$\sum_{i=1}^{j-1} k_0(x_i,x_j) \leq \frac{\delta}{2} + \min_{x\in X: k_0(x,x)\leq R_j^2} \textstyle\sum_{i=1}^{j-1} k_0(x_i,x)$$

with $k_0(x_j,x_j) \leq R_j^2 \in [2\log(j)/c_2, 2\log(n)/c_2]$. We immediately have

$$\frac{k_0(x_j,x_j)}{2} + \sum_{i=1}^{j-1} k_0(x_i,x_j) \leq \frac{\delta}{2} + \frac{R_j^2}{2} + \min_{x\in X: k_0(x,x)\leq R_j^2} \sum_{i=1}^{j-1} k_0(x_i,x)$$

so the desired conclusion follows from Theorem 5.

A.1.2. Proof of Theorem 2: Stein Greedy Convergence

Instantiate the constants $c_1, c_2 > 0$ from Theorem 5, and consider any point sequence $\{x_i\}_{i=1}^n$ satisfying

$$\frac{k_0(x_j,x_j)}{2} + \sum_{i=1}^{j-1} k_0(x_i,x_j) \leq \frac{\delta}{2} + \min_{x\in X: k_0(x,x)\leq R_j^2} \frac{k_0(x,x)}{2} + \sum_{i=1}^{j-1} k_0(x_i,x)$$

with $S_j = \sqrt{2\log(j)/c_2} \leq R_j \leq \infty$ for each $1 \leq j \leq n$. We immediately have

$$\frac{k_0(x_j,x_j)}{2} + \sum_{i=1}^{j-1} k_0(x_i,x_j) \leq \frac{\delta}{2} + \min_{x\in X: k_0(x,x)\leq S_j^2} \frac{k_0(x,x)}{2} + \sum_{i=1}^{j-1} k_0(x_i,x) \leq \frac{\delta}{2} + \frac{S_j^2}{2} + \min_{x\in X: k_0(x,x)\leq S_j^2} \sum_{i=1}^{j-1} k_0(x_i,x),$$

so the desired conclusion follows from Theorem 5.

A.1.3. Proof of Theorem 5: Stein Point Convergence

Our high-level strategy is to show that, when k_0 is P-sub-exponential, optimizing over a suitably truncated search space on each step is sufficient to optimize the discrepancy globally. To obtain an explicit rate of convergence, we adapt the greedy approximation error analysis of Jones (1992), which applies to uniformly bounded kernels. We begin by fixing any sequence of truncation levels $(S_j)_{j=1}^{\infty}$ with each $S_j \in [0, \infty)$, defining the truncation sets $B_j = \{x \in X : k_0(x,x) \leq S_j^2\}$, and letting $\mathcal{M}_j$ denote the convex hull of $\{k_0(x,\cdot)\}_{x \in B_j}$. Next we identify a truncation-optimal $h_j \in \arg\min_{f \in \mathcal{M}_j} J(f)$. Now, fix any point sequence $\{x_i\}_{i=1}^n$ satisfying

$$\frac{k_0(x_j,x_j)}{2} + \sum_{i=1}^{j-1} k_0(x_i,x_j) \leq \frac{\delta}{2} + \frac{S_j^2}{2} + \min_{x \in X : k_0(x,x) \leq S_j^2} \sum_{i=1}^{j-1} k_0(x_i,x)$$

for some approximation level $\delta \geq 0$ and each $1 \leq j \leq n$. In the remainder, we will recursively bound the discrepancy of this point sequence in terms of each S_j and $\|h_j\|_{\mathcal{K}_0}$, bound each $\|h_j\|_{\mathcal{K}_0}$ in terms of S_j using the P-sub-exponential tails of k_0, and show that an appropriate setting of each S_j delivers the advertised claim.

Bounding discrepancy For each j, let $f_j = \frac{1}{j}\sum_{i=1}^j k_0(x_i,\cdot)$ and $\epsilon_j = D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^j) = \|f_j\|_{\mathcal{K}_0}$. By Cauchy-Schwarz and the arithmetic-geometric mean inequality, we have the estimates

$$\begin{aligned}
n^2\epsilon_n^2 - \delta &= k_0(x_n,x_n) + (n-1)^2\epsilon_{n-1}^2 + 2(n-1)f_{n-1}(x_n) - \delta \\
&\leq S_n^2 + (n-1)^2\epsilon_{n-1}^2 + 2(n-1)\min_{x \in B_n} f_{n-1}(x) \\
&= S_n^2 + (n-1)^2\epsilon_{n-1}^2 + 2(n-1)\inf_{f \in \mathcal{M}_n} \langle f, f_{n-1}\rangle_{\mathcal{K}_0} \\
&\leq S_n^2 + (n-1)^2\epsilon_{n-1}^2 + 2(n-1)\langle h_n, f_{n-1}\rangle_{\mathcal{K}_0} \\
&\leq S_n^2 + (n-1)^2\epsilon_{n-1}^2 + n^2\|h_n\|_{\mathcal{K}_0}^2 + \frac{(n-1)^2}{n^2}\epsilon_{n-1}^2.
\end{aligned}$$

Unrolling the recursion, we obtain

$$n^2\epsilon_n^2 \leq \sum_{i=0}^{n-1}(S_{n-i}^2 + \delta + \|h_{n-i}\|_{\mathcal{K}_0}^2 (n-i)^2)\prod_{j=1}^{i}(1 + 1/(n-j+1)^2).$$

Moreover, the products in this expression are uniformly bounded in i as

$$\log(\prod_{j=1}^{i}(1 + 1/(n-j+1)^2)) = \sum_{j=1}^{i}\log((1 + 1/(n-j+1)^2)) \leq \int_0^{\infty}\log(1 + 1/x^2)\,dx = \pi.$$

Therefore,

$$n^2\epsilon_n^2 \leq e^{\pi}\sum_{i=1}^{n} S_i^2 + \delta + i^2\|h_i\|_{\mathcal{K}_0}^2.$$

Bounding $\|h_i\|_{\mathcal{K}_0}$ To bound each $\|h_i\|_{\mathcal{K}_0}$, we consider the truncated mean embeddings

$$\begin{aligned}
k_i^- &:= \int k_0(x,\cdot)\mathbb{I}\left[k_0(x,x) \leq S_i^2\right]\mathrm{d}P(x) \quad \text{and} \\
k_i^+ &:= \int k_0(x,\cdot)\mathbb{I}\left[k_0(x,x) > S_i^2\right]\mathrm{d}P(x) = k_P - k_i^-.
\end{aligned}$$

Since $k_P = 0$, we have $\|k_i^+\|_{\mathcal{K}_0} = \|k_i^-\|_{\mathcal{K}_0}$. Moreover, since $k_i^- \in \mathcal{M}_i$, we deduce that

$$\begin{aligned}
\|h_i\|_{\mathcal{K}_0}^2 &\leq \|k_i^-\|_{\mathcal{K}_0}^2 = \|k_i^+\|_{\mathcal{K}_0}^2 \\
&= \iint k_0(x,y)\mathbb{I}\left[k_0(x,x) > S_i^2\right]\mathrm{d}P(x)\mathbb{I}\left[k_0(y,y) > S_i^2\right]\mathrm{d}P(y) \\
&\leq \left(\int \sqrt{k_0(x,x)}\mathbb{I}\left[k_0(x,x) > S_i^2\right]\mathrm{d}P(x)\right)^2 \\
&\leq \int k_0(x,x)\mathbb{I}\left[k_0(x,x) > S_i^2\right]\mathrm{d}P(x)
\end{aligned}$$

where the final two inequalities follow by Cauchy-Schwarz and Jensen's inequality.

Let $Y = k_0(Z, Z)$ for $Z \sim P$. We will bound the tail expectation in the final display by considering the biased random variable $Y^* = k_0(Z^*, Z^*)$ for Z^* with density $\rho(z^*) = \frac{k_0(z^*,z^*)p(z^*)}{\mathbb{E}[Y]}$. By (Wainwright, 2017, Thm. 2.2), since Y is sub-exponential, there exists $c_0 > 0$ such that $\mathbb{E}[e^{\lambda Y}] < \infty$ for all $|\lambda| \leq c_0$. For any $\lambda \neq 0$ with $|\lambda| \leq c_0/2$, we have, by the relation $x \leq e^x$,

$$\mathbb{E}[e^{\lambda Y^*}] = \mathbb{E}[e^{\lambda k_0(Z^*,Z^*)}] = \frac{\mathbb{E}[k_0(Z,Z)e^{\lambda k_0(Z,Z)}]}{\mathbb{E}[Y]} = \frac{\mathbb{E}[\lambda Y e^{\lambda Y}]}{\lambda \mathbb{E}[Y]} \leq \frac{\mathbb{E}[e^{2\lambda Y}]}{\lambda \mathbb{E}[Y]} < \infty.$$

Hence, by (Wainwright, 2017, Thm. 2.2), Y^* is also sub-exponential and satisfies, for some $\tilde{c}_1, c_2 > 0$, $\mathbb{P}(Y^* \geq t) \leq \tilde{c}_1 e^{-c_2 t}$ for all $t > 0$.

Applying this finding to the bounding of h_i, we obtain

$$\|h_i\|^2_{\mathcal{K}_0} \leq \int k_0(x,x)\mathbb{I}\left[k_0(x,x) > S_i^2\right] \mathrm{d}P(x) = \mathbb{E}[Y] \int \mathbb{I}\left[k_0(x,x) > S_i^2\right] \rho(x)\, dx = \mathbb{E}[Y]\mathbb{P}\left(Y^* \geq S_i^2\right) \leq c_1 e^{-c_2 S_i^2}$$

where $c_1 = \tilde{c}_1 \mathbb{E}[Y]$. Hence

$$D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n) \leq e^{\pi/2} \sqrt{\frac{1}{n^2} \sum_{i=1}^n S_i^2 + \delta + i^2 c_1 e^{-c_2 S_i^2}}.$$

Setting each S_i By choosing $S_i \in [\sqrt{2\log(i)/c_2}, \sqrt{2\log(n)/c_2}]$ for each i we obtain

$$D_{\mathcal{K}_0,P}(\{x_i\}_{i=1}^n) \leq e^{\pi/2} \sqrt{\frac{1}{n^2} \sum_{i=1}^n \frac{2\log(n)}{c_2} + \delta + c_1} \leq e^{\pi/2} \sqrt{\frac{2\log(n)}{c_2 n} + \frac{\delta}{n} + \frac{c_1}{n}}.$$

A.2. Proof of Theorem 3: Log Inverse KSD Controls Convergence

Fix any $\alpha > 0$ and $\beta < 0$. Our proof will leverage (Gorham & Mackey, 2017, Thm. 7). This requires demonstrating two separate properties for the log inverse kernel: first, the log inverse function $\Phi(z) \triangleq (\alpha + \log(1 + \|z\|_2^2))^\beta$ has a nonvanishing generalized Fourier transform, and second, whenever $D_{\mathcal{K}_0,P}(\mu_m) \to 0$, the measures μ_m are uniformly tight. We will repeatedly use the notation $\gamma(r) \triangleq (\alpha + \log(1 + r))^\beta$ and $\phi(r) \triangleq \gamma(r^2)$ throughout the proof. Moveover, we will use $\hat{f}$ to denote the (generalized) Fourier transform of a function f, and V_d will represent the volume of the unit Euclidean ball in d dimensions. Finally, we write $f^{(m)}$ for the m-th derivative of any sufficiently differentiable function $f : \mathbb{R} \to \mathbb{R}$.

To demonstrate the first property, we begin with the following lemma.

Lemma 6 (Log Inverse Function Is Completely Monotone)**.** *Fix any $\alpha > 0$ and $\beta < 0$. The function $\gamma(r) \triangleq (\alpha + \log(1 + r))^\beta$ is* completely monotone*, i.e., $\gamma \in C^\infty$ and $(-1)^m \gamma^{(m)}(r) \geq 0$ for all $m \in \mathbb{N}_0$ and all $r \geq 0$, and hence the function $k_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by $k_2(x, x') \triangleq \gamma(\|x - x'\|_2^2)$ is a kernel function for all dimensions $d \in \mathbb{N}$.*

Proof. By (Wendland, 2004, Theorem 7.13) we know that Φ is positive semidefinite for all dimensions $d \in \mathbb{N}$ if and only if γ is completely monotone. Thus it remains to show that γ is completely monotone.

Since $\alpha > 0$, $\gamma(r) > 0$ for all $r \geq 0$. To verify $(-1)^m \gamma^{(m)}(r) \geq 0$ for all $m \geq 1$, we will proceed by induction. Let us suppose that for some $m \geq 1$,

$$\gamma^{(m)}(r) = (-1)^m \sum_{l=1}^m c_{l,m} (\alpha + \log(1 + r))^{\beta - l} (1 + r)^{-m} \tag{10}$$

where each $c_{l,m} \in \mathbb{R}$ is positive. Taking another derivative yields

$$\gamma^{(m+1)}(r) = (-1)^{m+1} \sum_{l=1}^{m+1} c_{l,m+1} (\alpha + \log(1 + r))^{\beta - l} (1 + r)^{-m-1},$$

where $c_{1,m+1} \triangleq m\, c_{1,m}$, $c_{l,m+1} \triangleq m\, c_{l,m} + (l - \beta - 1)\, c_{l-1,m}$ for $l > 1$ and $c_{l,m} \triangleq 0$ for all $l > m$, completing the induction step.

As for the base case, notice $\gamma'(r) = \beta(\alpha + \log(1+r))^{\beta-1}(1+r)^{-1}$, which establishes the identity for $l = 1$ by setting $c_{1,1} \triangleq -\beta$. The conclusion of this proof by induction implies $(-1)^m \gamma^{(m)}(r) \geq 0$ for all m and all $r \geq 0$. By (10), $\gamma \in C^\infty$, establishing the lemma. □

Knowing that γ is a completely monotone function, we can now demonstrate $\hat{\Phi}$ has a nonvanishing generalized Fourier transform.

Lemma 7 (Log Inverse Function Has Nonvanishing GFT)**.** *Consider the function $\Phi : \mathbb{R}^d \to \mathbb{R}$ given by $\Phi(z) = (\alpha + \log(1 + \|z\|_2^2))^\beta$ for some $\alpha > 0$ and $\beta < 0$. Its generalized Fourier transform $\hat{\Phi}(w)$ is radial, nonvanishing, and continuous for $w \neq 0$. Moreover, $\hat{\Phi}(w) \to 0$ as $\|w\|_2 \to \infty$.*

Proof. We will first use induction to prove an intermediate result that states for any $m \in \mathbb{N}_0$,

$$\Delta^m \Phi(z) = \sum_{(u,v) \in S_m} \tau_{u,v} \|z\|_2^{2v} \gamma^{(u)}(\|z\|_2^2) \tag{11}$$

where $\tau_{u,v} > 0$ are positive reals, $S_m = \{(u,v) \in \mathbb{N}_0^2 \mid v \leq u - m, u \leq 2m\}$ and $\gamma(r) \triangleq (\alpha + \log(1+r))^\beta$.

Note for the base case $m = 0$, the claim above for $\Delta^0 \Phi = \Phi$ clearly holds. Now suppose it holds from some $m \in \mathbb{N}_0$. If $A : \mathbb{R}^d \to \mathbb{R}$ is a function that can decomposed as $A(z) \triangleq f(\|z\|_2^2)\, g(\|z\|_2^2)$ where $f, g \in C^\infty([0,\infty))$, then we have

$$\Delta A(z) = \left[2d g'(\|z\|_2^2) + 4\|z\|_2^2 g''(\|z\|_2^2)\right] f(\|z\|_2^2) + \left[2d g(\|z\|_2^2) + 4\|z\|_2^2 g'(\|z\|_2^2)\right] f'(\|z\|_2^2) + 4\|z\|_2^2 g(\|z\|_2^2) f''(\|z\|_2^2). \tag{12}$$

Consider each term in the decomposition of $\Delta^m \Phi(z)$ from the induction hypothesis. If we let $g(r) = r^v$ and $f(r) = \phi^{(u)}(r)$, we see that each term from (12) is of the form $\tau'_{u,v} \|z\|_2^{2v'} \phi^{(u')}(\|z\|_2^2)$ where the values for (u', v') are $(u, v-1), (u, v-1), (u+1, v), (u+1, v), (u+2, v+1)$ respectively. Notice that when $v = 0$ or $v = 1$, the first or second derivative of g will be zero and these terms may disappear altogether. Thus all these tuples will lie in S_{m+1} for any $(u,v) \in S_m$, and so we must have $\Delta^{m+1}\Phi(z)$ satisfies the induction hypothesis as well, completing the proof by induction.

Now we can prove the lemma. Suppose $2m \geq d$. Then by the triangle inequality and a radial substitution (Baker, 1999),

$$\int_{\mathbb{R}^d} |\Delta^m \Phi(z)|\, dz \leq \sum_{(u,v) \in S_m} \int_{\mathbb{R}^d} \tau_{u,v} \|z\|_2^{2v} |\phi^{(u)}(\|z\|_2^2)|\, dz = d\, V_d \sum_{(u,v) \in S_m} \int_0^\infty \tau_{u,v} r^{2v+d-1} |\phi^{(u)}(r^2)|\, dr.$$

Because $|\phi^{(u)}(r)| = O(r^{-u} \log^{\beta-1}(r))$ as $r \to \infty$ for $u \in \mathbb{N}$ by (10), we see that each integrand above is $O(r^{2(v-u)+d-1} \log^{\beta-1}(r))$. But since $v \leq u - m$, this will imply that each integrand is $O(r^{-2m+d-1} \log^{\beta-1}(r))$, which is integrable for large r yielding $\Delta^m \Phi \in L^1(\mathbb{R}^d)$.

By (Steinwart & Christmann, 2008, Lemma 4.34) and the fact that positive definiteness is preserved by summation, we have $\Delta^m \Phi$ is a positive definite function. This along with the fact that $\Delta^m \Phi \in L^1(\mathbb{R}^d)$ allows us to invoke (Wendland, 2004, Theorem 6.11) and (Wendland, 2004, Theorem 6.18) to obtain $\widehat{\Delta^m \Phi}$ is continuous, radial and nonvanishing. Moreover, $\Delta^m \Phi$ belonging to $L^1(\mathbb{R}^d)$ implies its Fourier transform belongs to $L^\infty(\mathbb{R}^d)$. The lemma follows by noticing $\widehat{\Delta^m \Phi}(w) = \|w\|_2^{2m} \hat{\Phi}(w)$, i.e., $\hat{\Phi}(w) = \|w\|_2^{-2m} \widehat{\Delta^m \Phi}(w)$ for all $w \neq 0$. □

We now need to demonstrate the second property to complete the proof of Theorem 3, but in order to do so, we first will establish the lemma below. By Lemma 7, we know $\hat{\Phi}$ is radial and thus can write $\hat{\Phi}(w) = \phi_\wedge(\|w\|_2)$ for some continuous function $\phi_\wedge : (0,\infty) \to (0,\infty)$. Our first priority will be to lower bound $\phi_\wedge$ near the origin.

Lemma 8 (Log Inverse GFT Lower Bound)**.** *If Φ is the log inverse function on $\mathbb{R}^d$ from Lemma 7, then $\liminf_{r \to 0^+} r^d (\alpha + \log(1 + 1/r^2))^{-\beta+1} \phi_\wedge(r) > 0$ where $\hat{\Phi}(w) = \phi_\wedge(\|w\|_2)$ for all $w \neq 0$.*

Proof. First we will show that $\phi_\wedge$ is strictly decreasing. Since $r \mapsto (\alpha + \log(1+r))^\beta$ was shown to be completely monotone in Lemma 6, by (Wendland, 2004, Theorem 7.14) we must have $\Phi(z) = \int_0^\infty e^{-t\|z\|_2^2} \partial v(t)$ for some finite, non-negative

Borel measure v on $[0,\infty)$ that is not concentrated at zero. Let $(\varphi_m)_{m=1}^{\infty}$ be a sequence of Schwartz functions (Wendland, 2004, Definition 5.17) defined on $\mathbb{R}^d$. Then, for each m, both $\hat{\varphi}_m$ and $\Phi\hat{\varphi}_m$ are also Schwartz functions, and thus

$$\int_{\mathbb{R}^d}\left[\int_0^\infty |e^{-t\|x\|_2^2}\hat{\varphi}_m(x)|\partial v(t)\right]dx = \int_{\mathbb{R}^d}\left[\int_0^\infty e^{-t\|x\|_2^2}|\hat{\varphi}_m(x)|\partial v(t)\right]dx = \int_{\mathbb{R}^d}\Phi(x)|\hat{\varphi}_m(x)|\,dx < \infty,$$

as all Schwartz functions are integrable. This allows us to use Fubini's theorem in conjunction with Plancherel's Theorem to argue

$$\begin{aligned}\int_{\mathbb{R}^d}\hat{\Phi}(w)\varphi_m(w)\,dw = \int_{\mathbb{R}^d}\Phi(x)\hat{\varphi}_m(x)\,dx &= \int_{\mathbb{R}^d}\left[\int_0^\infty e^{-t\|x\|_2^2}\partial v(t)\right]\hat{\varphi}_m(x)\,dx\\ &= \int_{\mathbb{R}^d}\int_0^\infty e^{-t\|x\|_2^2}\hat{\varphi}_m(x)\partial v(t)\,dx\\ &= \int_0^\infty\int_{\mathbb{R}^d} e^{-t\|x\|_2^2}\hat{\varphi}_m(x)\,dx\partial v(t)\\ &= \int_0^\infty\int_{\mathbb{R}^d}(2t)^{d/2}e^{-\frac{1}{4t}\|w\|_2^2}\varphi_m(w)\,dw\partial v_+(t) + v_0\varphi_m(0),\end{aligned}$$

where we have used the decomposition $v \triangleq v_+ + v_0\delta_0$ for $v_0 \geq 0$ and v_+ non-zero and absolutely continuous with respect to Lebesgue measure on $[0,\infty)$. Let $\mathcal{B}:\mathbb{R}^d\to\mathbb{R}$ be a bump function, e.g., $\mathcal{B}(x) \triangleq \mathcal{Z}^{-1}\exp\{-1/(1-\|x\|_2^2)\}\mathbb{I}[\|x\|_2 < 1]$ where $\mathcal{Z}$ is the normalization constant chosen such that $\int_{\mathbb{R}^d}\mathcal{B}(x)\,dx = 1$. Then let us define $\varphi_m:\mathbb{R}^d\to\mathbb{R}$ via the mapping $\varphi_m(w) \triangleq m^d\mathcal{B}(m(w-w_0e_1)) - m^d\mathcal{B}(m(w-w_1e_1))$, where $0 < w_0 < w_1$ and $e_1 \in \mathbb{R}^d$ is the first standard basis vector. Then $\int_{\mathbb{R}^d}\hat{\Phi}(w)\varphi_m(w)\,dw \to \hat{\Phi}(w_0e_1) - \hat{\Phi}(w_1e_1) = \phi_\wedge(w_0) - \phi_\wedge(w_1)$ since $\hat{\Phi}$ is a continuous in neighborhoods of w_0e_1 and w_1e_1 (Wendland, 2004, Theorem 5.22).

Because v_+ cannot be the zero measure, there must be some finite interval $[a_0,b_0] \subset (0,\infty)$ such that $v_+([a_0,b_0]) > 0$. For each $t > 0$ and $m > \max(\frac{1}{w_0}, \frac{2}{w_1-w_0})$, we have

$$A_m(t) \triangleq \int_{\mathbb{R}^d}(2t)^{d/2}e^{-\frac{1}{4t}\|w\|_2^2}\varphi_m(w)\,dw = \int_{\mathbb{R}^d}(2t)^{d/2}(e^{-\frac{1}{4t}\|w-w_0e_1\|_2^2} - e^{-\frac{1}{4t}\|w-w_1e_1\|_2^2})m^d\mathcal{B}(mw)\,dw > 0,$$

since $\|w-w_0e_1\|_2 < \|w-w_1e_1\|_2$ when $\|w\|_2 < \min(\frac{w_0-w_1}{2}, w_0)$. Using (Wendland, 2004, Theorem 5.22) again, we have $A_m(t) \to (2t)^{d/2}(e^{-\frac{1}{4t}w_0^2} - e^{-\frac{1}{4t}w_1^2})$ as $m\to\infty$ for any $t > 0$. Moreover, for all $t \in [a_0,b_0]$ and $m \geq 1$, we have

$$|A_m(t)| \leq \int_{\mathbb{R}^d}(2t)^{d/2}e^{-\frac{1}{4t}\|w-w_0e_1\|_2^2}m^d\mathcal{B}(mw)\,dw \leq (2b_0)^{d/2}\sup_{\|w\|_2<1} e^{-\frac{1}{4b_0}\|w-w_0e_1\|_2^2} < \infty. \tag{13}$$

Hence, the dominated convergence theorem allows us to exchange the limit over m and integral over t below to conclude

$$\begin{aligned}\phi_\wedge(w_0) - \phi_\wedge(w_1) &= \lim_{m\to\infty}\int_0^\infty A_m(t)\partial v_+(t) + v_0\varphi_m(0) \geq \lim_{m\to\infty}\int_{a_0}^{b_0}A_m(t)\partial v_+(t) = \int_{a_0}^{b_0}\lim_{m\to\infty}A_m(t)\partial v_+(t)\\ &= \int_{a_0}^{b_0}(2t)^{d/2}(e^{-\frac{1}{4t}w_0^2} - e^{-\frac{1}{4t}w_1^2})\partial v_+(t) \geq v_+([a_0,b_0])\min_{t\in[a_0,b_0]}\left\{(2t)^{d/2}(e^{-\frac{1}{4t}w_0^2} - e^{-\frac{1}{4t}w_1^2})\right\} > 0,\end{aligned}$$

showing $\phi_\wedge$ is strictly decreasing as claimed.

Suppose $\psi:[0,\infty)\to\mathbb{R}$ is a C^∞ function with support $[a,b]$ for $0 < a < b$ such that $\psi(r) > 0$ for all $r \in (a,b)$ and $\int_0^\infty\psi(r)\,dr = 1$. Then because $\phi_\wedge$ is strictly decreasing, by the mean value theorem we have

$$\phi_\wedge(b/\lambda) \leq \int_0^\infty \lambda\phi_\wedge(r)\psi(\lambda r)\,dr \leq \phi_\wedge(a/\lambda) \tag{14}$$

for all $\lambda > 0$. If we assign $\Psi(w) \triangleq \psi(\|w\|_2)$ to be the radial continuation of ψ, by (Baker, 1999) the quantity sandwiched above becomes

$$\int_0^\infty \lambda\phi_\wedge(r)\psi(\lambda r)\,dr = \int_0^\infty \phi_\wedge(s/\lambda)\psi(s)\,ds = \frac{1}{dV_d}\int_{\mathbb{R}^d}\hat{\Phi}(w/\lambda)\frac{\Psi(w)}{\|w\|_2^{d-1}}\,dw.$$

Next suppose that $\xi:[0,\infty)\to\mathbb{R}$ is a Schwartz function satisfying $\xi^{(k)}(0) = 0$ for all integral $k \geq 0$, and let $\Xi:\mathbb{R}^d\to\mathbb{R}$ given by $\Xi(x) \triangleq \xi(\|x\|_2)$ be the radial continuation of ξ. Then by Plancherel's Theorem, scaling the input of a Fourier

transform as in (Wendland, 2004, Theorem 5.16), and the change to spherical coordinates in (Baker, 1999), for any $\lambda > 0$, we have

$$\int_{\mathbb{R}^d} \hat{\Phi}(w/\lambda)\Xi(w)\,dw = \int_{\mathbb{R}^d} \Phi(w)\hat{\Xi}(w/\lambda)\,dw = d\,V_d \int_0^\infty r^{d-1}\phi(r)\xi_\wedge(r/\lambda)\,dr = d\,V_d\,\lambda^d \int_0^\infty s^{d-1}\phi(\lambda s)\xi_\wedge(s)\,ds, \tag{15}$$

where $s = r/\lambda$ and $\xi_\wedge$ is the radial function associated with $\hat{\Xi}$, i.e., $\hat{\Xi}(w) = \xi_\wedge(\|w\|_2)$ for all w.

Let us define $\omega : [0,\infty) \to \mathbb{R}$ by the mapping $\omega(t) \triangleq (\alpha + t)^\beta$. Then by the mean value theorem and the fact that ω' is increasing, we have for all $s > 1$

$$-\omega'(\log(1+\lambda^2 s^2)) \leq -\frac{\omega(\log(1+\lambda^2 s^2)) - \omega(\log(1+\lambda^2))}{\log(1+\lambda^2 s^2) - \log(1+\lambda^2)} \leq -\omega'(\log(1+\lambda^2)).$$

By rearranging terms, this implies for all $\lambda > 0$

$$(-\beta)\left(\frac{\alpha + \log(1+\lambda^2 s^2)}{\alpha + \log(1+\lambda^2)}\right)^{\beta-1} \log\left(\frac{1+\lambda^2 s^2}{1+\lambda^2}\right) \leq -\frac{\omega(\log(1+\lambda^2 s^2)) - \omega(\log(1+\lambda^2))}{\omega(\log(1+\lambda^2))(\alpha + \log(1+\lambda^2))^{-1}} \leq (-\beta)\log\left(\frac{1+\lambda^2 s^2}{1+\lambda^2}\right).$$

Since $\log(\frac{1+\lambda^2 s^2}{1+\lambda^2}) \to 2\log s$ as $\lambda \to \infty$, and the sandwiched term above is $-(\alpha + \log(1+\lambda^2))(\phi(\lambda s)/\phi(\lambda) - 1)$, we have $(\alpha + \log(1+\lambda^2))(\phi(\lambda s)/\phi(\lambda) - 1) \to 2\beta \log s$ as $\lambda \to \infty$ for all $s > 1$. The case for $s \in (0,1]$ is analogous and yields the same asymptotic limit.

With this new asymptotic expansion in hand, we will revisit (15). We have

$$\begin{aligned}
\lambda^{-d}\phi(\lambda)^{-1}(\alpha + \log(1+\lambda^2)) \int_{\mathbb{R}^d} \hat{\Phi}(w/\lambda)\Xi(w)\,dw &= d\,V_d \phi(\lambda)^{-1}(\alpha + \log(1+\lambda^2)) \int_0^\infty \phi(\lambda s) s^{d-1}\xi_\wedge(s)\,ds \\
&= d\,V_d\,(\alpha + \log(1+\lambda^2)) \int_0^\infty \frac{\phi(\lambda s)}{\phi(\lambda)} s^{d-1}\xi_\wedge(s)\,ds \\
&= d\,V_d \int_0^\infty (\alpha + \log(1+\lambda^2))\left[\frac{\phi(\lambda s)}{\phi(\lambda)} - 1\right] s^{d-1}\xi_\wedge(s)\,ds.
\end{aligned}$$

Notice that final integrand converges to $2\beta s^{d-1}(\log s)\xi_\wedge(s)$ pointwise for all $s \geq 0$ as $\lambda \to \infty$. Since $\xi_\wedge$ is a Schwartz function on $[0,\infty)$, we can utilize the fact that $s \mapsto \log s$ is integrable near the origin to reason that $s^{d-1}(\log s)\xi_\wedge(s)$ is a Schwartz function as well, and thus integrable. Hence by the dominated convergence theorem, we have the integral above converges to $2\,\beta\,d\,V_d \int_0^\infty s^{d-1}(\log s)\xi_\wedge(s)\,ds$ as $\lambda \to \infty$.

Now suppose we choose $\Xi(x) \triangleq \|x\|_2^{1-d}\,\Psi(x)$. By (14) we have

$$\lim_{\lambda\to\infty} \lambda^{-d}\phi(\lambda)^{-1}(\alpha + \log(1+\lambda^2))\phi_\wedge(b/\lambda) \leq 2\beta \int_0^\infty s^{d-1}(\log s)\xi_\wedge(s)\,ds \leq \lim_{\lambda\to\infty} \lambda^{-d}\phi(\lambda)^{-1}(\alpha + \log(1+\lambda^2))\phi_\wedge(a/\lambda).$$

By Lemma 7, we know $\phi_\wedge(r) > 0$ for all $r > 0$, and thus the left-hand side above must be non-negative. Hence if we can show for some choice of ψ that the sandwiched term is non-zero, then the proof of the lemma will follow from choosing $r = a/\lambda$.

Let us define $L(x) = \log\|x\|_2$ with generalized Fourier transform $\hat{L}$. As usual, let $l : [0,\infty) \to \mathbb{R}$ and $l_\wedge : [0,\infty) \to \mathbb{R}$ be the radial functions associated with L and $\hat{L}$. Notice that again by Plancherel's Theorem

$$\begin{aligned}
\int_0^\infty s^{d-1}(\log s)\xi_\wedge(s)\,ds = \frac{1}{dV_d}\int_{\mathbb{R}^d} \hat{\Xi}(w)L(w)\,dw = \frac{1}{dV_d}\int_{\mathbb{R}^d} \Xi(x)\hat{L}(x)\,dx = \frac{1}{dV_d}\int_{\mathbb{R}^d} \frac{\Psi(x)}{\|x\|_2^{d-1}}\hat{L}(x)\,dx \\
= \int_0^\infty \psi(r) l_\wedge(r)\,dr.
\end{aligned} \tag{16}$$

Since we are free to choose ψ to be any Schwartz function with support $[a,b]$, if we could not find a function ψ such that the quantity in (16) is non-zero, this would imply the support of $l_\wedge$ is a subset of $\{0\}$. But this would mean $l_\wedge$ is some multiple of a point mass at zero, which would imply l is a constant function, a contradiction. Thus we must be able to find some ψ such that the integral above is non-zero, completing the lemma. □

Fix any $a_0 > 0$ and $\alpha_0 \in (0, \frac{1}{2})$. Our strategy for showing the KSD controls tightness will mimic (Gorham & Mackey, 2017, Lem. 16): we will show that a bandlimited approximation of the function $g_j(x) = 2\alpha_0 x_j (a_0^2 + \|x\|_2^2)^{\alpha_0 - 1}$ belongs to the inverse log RKHS and thus enforces tightness.

First note that in the proof of (Gorham & Mackey, 2017, Lem. 16), it was shown $h = \mathcal{T}_P g$ was a coercive, Lipschitz, and bounded-below function for $P \in \mathcal{P}$. Moreover, in the proof of (Gorham & Mackey, 2017, Lem. 12), a random vector Y with density $\rho(y)$ is constructed such that the support of $\hat{\rho}$ belongs to $[-4, 4]^d$ and also $\|Y\|_2$ is integrable. Consider the new function $g^\circ(x) \triangleq \mathbb{E}[g(x + Y)]$ for all $x \in \mathbb{R}^d$. By the convolution theorem, $\hat{g}_j^\circ = \hat{g}_j \hat{\rho}$ and so g_j° is bandlimited for all j. In the proof of (Gorham & Mackey, 2017, Lem. 16), $\hat{g}_j$ was shown to grow asymptotically at the rate $(iw_j)\|w\|_2^{-d-2\alpha_0}$ as $\|w\|_2 \to 0$. Thus

$$\sum_{j=1}^d \int_{\mathbb{R}^d} \frac{\hat{g}_j^\circ(w)\overline{\hat{g}_j^\circ(w)}}{\hat{\Phi}(w)}\, dw = \sum_{j=1}^d \int_{[-4,4]^d} \frac{\hat{g}_j(w)\overline{\hat{g}_j(w)}\hat{\rho}(w)^2}{\hat{\Phi}(w)}\, dw \le \kappa_0 \int_{[-4,4]^d} \frac{\|w\|_2^{-2d-4\alpha_0+2}}{\hat{\Phi}(w)}\, dw$$
$$\le \kappa_1 \int_0^{4\sqrt{d}} r^{-4\alpha_0+1} \log^{-\beta+1}(1 + r^{-2})\, dr,$$

for some constants $\kappa_0, \kappa_1 > 0$ where we used Lemma 8 in the final inequality. This integral is finite for all $\alpha_0 \in (0, \frac{1}{2})$ and any $\beta < 0$, which implies g° is in the log inverse RKHS by (Wendland, 2004, Theorem 10.21).

Finally, notice that via the argument proving (Gorham & Mackey, 2017, Lemma 12),

$$\sup_{x \in \mathbb{R}^d} |\mathcal{T}_P g^\circ(x) - h(x)| \le \frac{3d \log 2}{\pi} \left(\sup_{x \in \mathbb{R}^d} \|\nabla h(x)\|_2 + \sup_{x \in \mathbb{R}^d} \left\|\nabla^2 \log p(x)\right\|_{op} \cdot \sup_{x \in \mathbb{R}^d} \|g(x)\|_2 \right) < \infty.$$

Since h is bounded below and coercive, these properties are inherited by $\mathcal{T}_P g^\circ$. This allows us to apply (Gorham & Mackey, 2017, Lemma 17) to argue $D_{\mathcal{K}_0, P}(\mu_m) \to 0$ implies the measures μ_m are uniformly tight. Combining this with Lemma 7 allows us to utilize (Gorham & Mackey, 2017, Theorem 7) for the log inverse kernel, thereby concluding the proof.

A.3. Proof of Theorem 4: IMQ Score KSD Convergence Control

For $b = \nabla \log p$, introduce the alias $k_b = k_3$, let $\mathcal{K}_b$ denote the RKHS of k_b, and let C_c represent the set of continuous compactly supported functions on X. Since $P \in \mathcal{P}$, the proof of Thm. 13 in (Gorham & Mackey, 2017) shows that if, for each $h \in C^1 \cap C_c$ and $\epsilon > 0$, there exists $h_\epsilon \in \mathcal{K}_b$ such that $\sup_{x \in X} |(\mathcal{T}_P h)(x) - (\mathcal{T}_P h_\epsilon)(x)| \le \epsilon$, then $\mu_m \Rightarrow P$ whenever $D_{\mathcal{K}_0, P}(\mu_m) \to 0$ and $(\mu_m)_{m=1}^\infty$ is uniformly tight. Hence, to establish our result, it suffices to show (1) that, for each $h \in C^1 \cap C_c$ and $\epsilon > 0$, there exists $h_\epsilon \in \mathcal{K}_b$ such that $\sup_{x \in X} \max(\|\nabla(h - h_\epsilon)(x)\|_2, \|b(x)(h - h_\epsilon)(x)\|_2) \le \epsilon$ and (2) that $D_{\mathcal{K}_0, P}(\mu_m) \to 0$ implies $(\mu_m)_{m=1}^\infty$ is uniformly tight.

A.3.1. Approximating $C^1 \cap C_c$ with $\mathcal{K}_b$

Fix any $f \in C^1 \cap C_c$ and $\epsilon > 0$, and let $\mathcal{K}$ denote the RKHS of $k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$. Since p is strictly log-concave, b is invertible with $\det(\nabla b(x))$ never zero. Since $P \in \mathcal{P}$, b is Lipschitz. By the following theorem, proved in Section A.4, it therefore suffices to show that there exists $f_\epsilon \in \mathcal{K}$ such that $\sup_{x \in X} \max(\|\nabla(f - f_\epsilon)(x)\|_2, \|x(f - f_\epsilon)(x)\|_2) \le \epsilon$.

Theorem 9 (Composition Kernel Approximation). *For $b : X \to X$ invertible and k a reproducing kernel on X with induced RKHS $\mathcal{K}$, define the composition kernel $k_b(x, y) = k(b(x), b(y))$ with induced RKHS $\mathcal{K}_b$. Suppose that, for each $f \in C^1 \cap C_c$ and $\epsilon > 0$, there exists $f_\epsilon \in \mathcal{K}$ such that*

$$\sup_{x \in X} \max(\|\nabla(f - f_\epsilon)(x)\|_2, \|x(f - f_\epsilon)(x)\|_2) \le \epsilon.$$

If b is Lipschitz and $\det(\nabla b(x))$ is never zero, then, for each $h \in C^1 \cap C_c$ and $\epsilon > 0$, there exists $h_\epsilon \in \mathcal{K}_b$ such that

$$\sup_{x \in X} \max(\|\nabla(h - h_\epsilon)(x)\|_2, \|b(x)(h - h_\epsilon)(x)\|_2) \le \epsilon.$$

Since the identity map $x \mapsto x$ is Lipschitz and $f \in L^2$ because it is continuous and compactly supported, (Gorham & Mackey, 2017, Lem. 12) provides an explicit construction of $f_\epsilon \in \mathcal{K}$ satisfying our desired property whenever $k(x, y) = \Phi(x - y)$ for $\Phi \in C^2$ with non-vanishing Fourier transform. Our choice of IMQ k satisfies these properties by (Wendland, 2004, Thm. 8.15).

A.3.2. Controlling Tightness

Since P is distantly dissipative,

$$-\|b(x)\|_2 \|x\|_2 \leq \langle b(x), x\rangle \leq -\kappa \|x\|_2^2 + C + \langle b(0), x\rangle \leq -\kappa \|x\|_2^2 + C + \|b(0)\|_2 \|x\|_2$$

by Cauchy-Schwarz. Hence, b is *norm-coercive*, i.e., $\|b(x)\|_2 \to \infty$ whenever $\|x\|_2 \to \infty$. Since ∇b is bounded, our desired result follows from the following lemma which guarantees tightness control on b under weaker conditions.

Lemma 10 (Coercive Score Kernel KSDs Control Tightness)**.** *If $b : X \to X$ is norm coercive and differentiable, and $\nabla_j b_j(x) = o(\|b(x)\|_2^2)$ as $\|x\|_2 \to \infty$, then $\limsup_m D_{\mathcal{K}_0,P}(\mu_m) < \infty$ implies $(\mu_m)_{m=1}^\infty$ is tight.*

Proof. Fix any $a > c/2$ and $\alpha \in (0, \frac{1}{2}(\beta+1))$. The proof of (Gorham & Mackey, 2017, Lem. 16) showed that the function $g_j(x) = 2\alpha x_j(a^2 + \|x\|_2^2)^{\alpha-1} \in \mathcal{K}$ for each $j \in \{1,\dots,d\}$. Hence $g_{b,j}(x) \triangleq g_j(b(x)) \in \mathcal{K}_b$ for each $j \in \{1,\dots,d\}$ by Lemma 12. By our assumptions on ∇b, we have

$$\begin{aligned}(\mathcal{T}_P g_b)(x) &= 2\alpha(\|b(x)\|_2^2(a^2 + \|b(x)\|_2^2)^{\alpha-1} + \sum_{j=1}^d \nabla_j b_j(x)(a^2 + \|b(x)\|_2^2)^{\alpha-1} + b_j(x)^2 2(\alpha-1)(a^2 + \|b(x)\|_2^2)^{\alpha-2}\nabla_j b_j(x)) \\ &= 2\alpha\|b(x)\|_2^2(a^2 + \|b(x)\|_2^2)^{\alpha-1} + o(\|b(x)\|_2^{2\alpha}),\end{aligned}$$

so $\mathcal{T}_P g_b$ is coercive, and the proof of (Gorham & Mackey, 2017, Lem. 17) therefore gives the result $(\mu_m)_{m=1}^\infty$ is uniformly tight whenever $\limsup_m D_{\mathcal{K}_0,P}(\mu_m)$ finite. □

A.4. Proof of Theorem 9: Composition Kernel Approximation

Let $c = b^{-1}$ represent the inverse of b, and for any function f on X, let $f_c(y) = f(c(y))$ denote the composition of f and c so that $f_c(b(x)) = f(x)$. The following lemma shows that f_c inherits many of the properties of f under suitable restrictions on b.

Lemma 11 (Composition Properties)**.** *For any function f on X and invertible function b on X, define $f_c(y) = f(c(y))$ for $c = b^{-1}$. The following properties hold.*

1. *If f has compact support and b is continuous, then f_c has compact support.*
2. *If $f \in C^1$, $b \in C^1$, and $\det(\nabla b(x))$ is never zero, then $f_c \in C^1$.*

Proof. We prove each claim in turn.

1. If f is compactly supported and b is continuous, then $\mathrm{supp}(f_c) = b(\mathrm{supp}(f))$ is also compact, since continuous functions are compact-preserving (Joshi, 1983, Prop. 1.8).
2. If $f \in C^1$, $b \in C^1$, and $\det(\nabla b(x))$ is never zero, then c is continuous by the inverse function theorem (Spivak, 1965, Thm. 2-11), $x \mapsto (\nabla b(x))^{-1}$ is continuous, and hence $\nabla f_c(y) = (\nabla c(y))(\nabla f)(c(y)) = ((\nabla b)(c(y)))^{-1}(\nabla f)(c(y))$ is continuous.

□

Our next lemma exposes an important relationship between the RKHSes $\mathcal{K}$ and $\mathcal{K}_b$.

Lemma 12. *Suppose f is in the RKHS $\mathcal{K}$ of a reproducing kernel k on X and $b : X \to X$ is invertible. Then f_b is in the RKHS $\mathcal{K}_b$ of k_b for $f_b(x) = f(b(x))$ and $k_b(x,y) = k(b(x), b(y))$.*

Proof. Since $f \in \mathcal{K}$, there exist $f_m = \sum_{j=1}^{J_m} a_{m,j} k(x_{m,j}, \cdot)$ for $m \in \mathbb{N}, a_{m,j} \in \mathbb{R}$, and $x_{m,j} \in X$ such that $\lim_{m\to\infty} \|f_m - f\|_{\mathcal{K}} = 0$ and $\lim_{m\to\infty} f_m(x) = f(x)$ for all $x \in X$. Now let $c = b^{-1}$, and define

$$f_{m,b}(x) = f_m(b(x)) = \sum_{j=1}^{J_m} a_{m,j} k(x_{m,j}, b(x)) = \sum_{j=1}^{J_m} a_{m,j} k_b(c(x_{m,j}), x).$$

Since $\mathcal{K}_b = \overline{\{\sum_{j=1}^{J} a_j k_b(y_j, \cdot) : J \in \mathbb{N}, a_j \in \mathbb{R}, y_j \in X\}}$, each $f_{m,b} \in \mathcal{K}_b$. Since $(f_m)_{m=1}^{\infty}$ is a Cauchy sequence, and $\langle f_{m,b}, f_{m',b} \rangle_{\mathcal{K}_b}^2 = \sum_{j=1}^{J_m} a_{m,j} \sum_{j'=1}^{J_{m'}} a_{m',j'} k_b(c(x_{m,j}), c(x_{m',j'})) = \langle f_m, f_{m'} \rangle_{\mathcal{K}}$ so that $\|f_{m,b} - f_{m',b}\|_{\mathcal{K}_b} = \|f_m - f_{m'}\|_{\mathcal{K}}$ for all m, m', the sequence $(f_{m,b})_{m=1}^{\infty}$ is also Cauchy and converges in $\|\cdot\|_{\mathcal{K}_b}$ to its pointwise limit f_b. Since an RKHS is complete, $f_b \in \mathcal{K}_b$. □

With our lemmata in hand, we now prove the advertised claim. Suppose b is Lipschitz, $\det(\nabla b(x))$ is never zero, and for each $f \in C^1 \cap C_c$ and $\epsilon > 0$ there exists $f_\epsilon \in \mathcal{K}$ such that $\sup_{x \in X} \max(\|\nabla(f - f_\epsilon)(x)\|_2, \|x(f - f_\epsilon)(x)\|_2) \leq \epsilon$. Select any $h \in C^1 \cap C_c$ and any $\epsilon > 0$. By Lemma 11, $h_c \in C^1 \cap C_c$, and hence there exists $h_{c,\epsilon} \in \mathcal{K}$ such that $\sup_{y \in X} \max(\|\nabla(h_c - h_{c,\epsilon})(y)\|_2, \|y(h_c - h_{c,\epsilon})(y)\|_2) \leq \epsilon / \max(1, M_1(b))$. Now define $h_\epsilon(x) = h_{c,\epsilon}(b(x))$ so that $h_\epsilon \in \mathcal{K}_b$ by Lemma 12. We have $\sup_{x \in X} \|b(x)(h_\epsilon - h)(x)\|_2 \leq \sup_{y \in X} \|y(h_{c,\epsilon} - h_c)(y)\|_2 \leq \epsilon$, and

$$\sup_{x \in X} \|\nabla h_\epsilon(x) - \nabla h(x)\|_2 = \sup_{x \in X} \|(\nabla b(x))((\nabla h_{c,\epsilon})(b(x)) - (\nabla h_c)(b(x)))\|_2 \leq M_1(b)\epsilon / \max(1, M_1(b)) \leq \epsilon.$$

B. Implementational Detail

B.1. Benchmark Methods

In this section we briefly describe the MED and SVGD methods used as our empirical benchmark, as well as the (block) coordinate descent method that was used in conjunction with Stein Points.

B.1.1. Minimum Energy Designs

The first class of method that we consider is due to (Joseph et al., 2015). That work restricted attention to $X = [0,1]^d$ and constructed an energy functional:

$$\mathcal{E}_{\delta,P}(\{x_i\}_{i=1}^n) \quad := \quad \sum_{i \neq j} \left[\frac{p(x_i)^{-\frac{1}{2d}} p(x_j)^{-\frac{1}{2d}}}{\|x_i - x_j\|_2} \right]^{\delta}$$

for some tuning parameter $\delta \in [1, \infty)$ to be specified. In (Joseph et al., 2017) the rule-of-thumb $\delta = 4d$ was recommended. A heuristic argument in (Joseph et al., 2015) suggests that the points $\{x_i\}_{i=1}^n$ that minimise $\mathcal{E}_{\delta,P}(\{x_i\}_{i=1}^n)$ form an empirical approximation that converges weakly to P. The argument was recently made rigorous in (Joseph et al., 2017).

Minimisation of $\mathcal{E}_{\delta,P}$ does not require knowledge of how p is normalised. However, the actual minimisation of $\mathcal{E}_{\delta,P}$ can be difficult. In (Joseph et al., 2015) an extensible (greedy) method was considered, wherein the first point is selected as

$$x_1 \in \arg\max_{x \in X} \quad p(x)$$

and subsequent points are selected as

$$x_n \quad \in \quad \arg\min_{x \in X} \quad p(x)^{-\frac{\delta}{2d}} \sum_{i=1}^{n-1} \frac{p(x_i)^{-\frac{\delta}{2d}}}{\|x_i - x\|_2^{\delta}}.$$

However, alternative approaches could easily be envisioned. For instance, if n were fixed then one could consider e.g. applying the Newton method for optimisation over the points $\{x_i\}_{i=1}^n$.

Remark: There is a connection between certain minimum energy methods and discrepancy measures in RKHS; see (Sejdinovic et al., 2013).

Remark: Several potential modifications to $\mathcal{E}_{\delta,P}$ were suggested in (Joseph et al., 2017), but that report appeared after this work was completed. These could be explored in future work.

Remark: The MED objective function is typically numerically unstable due to the fact that the values of the density $p(\cdot)$ can be very small. In contrast, our proposed methods operate on $\log p(\cdot)$ and its gradient, which is more numerically robust.

B.1.2. Stein Variational Gradient Descent

The second method that we considered was due to (Liu & Wang, 2016; Liu, 2017) and recently generalised in (Liu & Zhu, 2017). The idea starts by formulating a continuous version of gradient descent on $\mathcal{P}(X)$ with the Kullback-Leibler

divergence $\mathrm{KL}(\cdot||P)$ as a target. To this end, restrict attention to $X = \mathbb{R}^d$ and consider the dynamics

$$S_f(x) \quad = \quad x + \epsilon f(x)$$

parametrised by a function $f \in \mathcal{K}^d$. For infinitesimal values of ϵ we can lift S_f to a pushforward map on $\mathcal{P}(X)$; i.e. $Q \mapsto S_f Q$. It was then shown in (Liu & Wang, 2016) that

$$-\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}(S_f Q||P)\Big|_{\epsilon=0} \quad = \quad \int \mathcal{T}_P f \, \mathrm{d}Q \tag{17}$$

where $\mathcal{T}_P$ is the Langevin Stein operator in Eqn. 5. Recall that this operator can be decomposed as $\mathcal{T}_P f = \sum_{j=1}^d \mathcal{T}_{P,j} f_j$ with $\mathcal{T}_{P,j} = \nabla_j + \nabla_j \log p$, where ∇_j denotes differentiation with respect to the jth coordinate in X. Then the direction of fastest descent

$$f^*(\cdot) \quad := \quad \underset{f \in B(\mathcal{K}^d)}{\arg\max} \; -\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}(S_f Q||P)\Big|_{\epsilon=0}$$

has a closed-form, with jth coordinate

$$f_j^*(\cdot; Q) \quad = \quad \int \mathcal{T}_{P,j} k(x, \cdot) \, \mathrm{d}Q(x).$$

The algorithm proposed in (Liu & Wang, 2016) discretises this dynamics in both space X, through the use of n points, and in time, through the use of a positive step size $\epsilon > 0$, leading to a sequence of empirical measures based on point sets $\{x_i^m\}_{i=1}^n$ for $m \in \mathbb{N}$. Thus, given an initialisation $\{x_i^0\}_{i=1}^n$ of the points, at iteration $m \geq 1$ of the algorithm we update

$$x_i^m \quad = \quad x_i^{m-1} + \epsilon f^*(x_i^{m-1}; Q_n^m)$$

in parallel, where

$$Q_n^m \quad = \quad \frac{1}{n}\sum_{i=1}^n \delta_{x_i^{m-1}}$$

is the empirical measure, at a computational cost of $O(n)$. The output is the empirical measure Q_n^m.

Remark: The step size ϵ is a tuning parameter of the method.

Remark: At present there are not theoretical guarantees for this method. Initial steps toward this goal are presented in (Liu, 2017).

B.1.3. Block Coordinate Descent

The Stein Point methods developed in the main text can be adapted to return a fixed number n of points for a given finite computational budget by first iteratively generating a size n point set, as described in the main text, and then performing (block) coordinate descent on this point set. The (block) coordinate descent procedure is now described:

Fix an initial configuration $\{x_i^0\}_{i=1}^n$. Then at iteration $m \geq 1$ of the algorithm, perform the following sequence of operations:

$$\begin{aligned} \forall i \qquad & x_i^m \leftarrow x_i^{m-1} \quad \text{then:} \\ \text{for } i = 1, \ldots, n \qquad & x_i^m \leftarrow \underset{x \in X}{\arg\min} \, D_{\mathcal{K}_0, P}(\{x_j^m\}_{j \neq i} \cup \{x\}) \end{aligned}$$

The output is the point set $\{x_i^m\}_{i=1}^n$.

Remark: The block coordinate descent method can equally be applied to MED; this was not considered in our empirical work.

Remark: Any numerical optimisation method can be used to solve the global optimisation problem in the inner loop. In this work we considered the same three candidates in the main text; Monte Carlo, Nelder-Mead and grid search. These are described next.

B.2. Numerical Optimisation Methods

Computation of the nth term in the proposed Stein Point sequences, given the previous $n-1$ terms, requires that a global optimisation is performed over $x_n \in X$. The same is true for both MED and KSD in the coordinate descent context. For all experiments reported in the main text, three different numerical methods were considered for this task, denoted `NM`, `MC`, `GS` in the main text. In this section we provide full details for how these methods were implemented.

B.2.1. Nelder-Mead

The Nelder-Mead (`NM`) method (Nelder & Mead, 1965) proceeds as in Algorithm 1. The function NM takes the following inputs: f is the objective function; t is the iteration count; n_{init} is the number of initial points to be drawn from a proposal distribution; n_{delay} is the number of iterations after which the proposal distribution becomes adaptive; μ_0 and Σ_0 are the mean vector and the covariance matrix of the initial proposal distribution; $\{x_j^{\text{curr}}\}_{j=1}^{n_{\text{curr}}}$ is the set of existing points; λ is the variance of each mixture component of the adaptive proposal distribution; l and u are the lower- and upper-bounds of the search space. The non-adaptive initial proposal distribution is a truncated multivariate Gaussian $\mathcal{N}(\mu_0, \Sigma_0)$ whose support is bounded by the hypercube $[l, u]$. The adaptive proposal distribution is a truncated Gaussian mixture $\Pi(\{x_j^{\text{curr}}\}_{j=1}^{n_{\text{curr}}}, \lambda) := \frac{1}{n_{\text{curr}}-1} \sum_{j=1}^{n_{\text{curr}}-1} \mathcal{N}(x_j^{\text{curr}}, \lambda I)$ with $\lambda > 0$ and support $[l, u]$. The expression $\text{NelderMead}_x \left[f(x), x_i^{\text{init}}, l, u \right]$ denotes the standard Nelder-Mead procedure for objective function f, initial point x_i^{init}, and bound constraint $x \in [l, u]$. We use the symbol $\leftsquigarrow$ to denote the assignment of a realised independent draw. The operator $\text{trunc}_l^u[\cdot]$ bounds the support of a distribution by the hypercube $[l, u]$.

Algorithm 1 Nelder-Mead

input $f, t, n_{\text{init}}, n_{\text{delay}}, \mu_0, \Sigma_0, \{x_j^{\text{curr}}\}_{j=1}^{n_{\text{curr}}}, \lambda, l, u$
output x^*

function NM
  for $i \leftarrow 1 : n_{\text{init}}$ **do**
    if $t \leq n_{\text{delay}}$ **then**
      $x_i^{\text{init}} \leftsquigarrow \text{trunc}_l^u \left[\mathcal{N}(\mu_0, \Sigma_0) \right]$
    else
      $x_i^{\text{init}} \leftsquigarrow \text{trunc}_l^u \left[\Pi(\{x_j^{\text{curr}}\}_{j=1}^{n_{\text{curr}}}, \lambda) \right]$
    end if
    $x_i^{\text{local}} \leftarrow \text{NelderMead}_x \left[f(x), x_i^{\text{init}}, l, u \right]$
  end for
  $i^* \leftarrow \arg\min_{i \in \{1 \ldots n_{\text{init}}\}} f(x_i^{\text{local}})$
  $x^* \leftarrow x_{i^*}^{\text{local}}$
end function

B.2.2. Monte Carlo

The Monte Carlo (`MC`) optimisation method proceeds as in Algorithm 2. The function MC takes the following inputs: f is the objective function; t is the iteration count; n_{test} is the number of test points to be drawn from a proposal distribution; n_{delay} is the number of iterations after which the proposal distribution becomes adaptive; μ_0 and Σ_0 are the mean vector and the covariance matrix of the initial proposal distribution; $\{x_j^{\text{curr}}\}_{j=1}^{n_{\text{curr}}}$ is the set of existing points; λ is the variance of each mixture component of the adaptive proposal distribution; l and u are the lower- and upper-bounds of the search space. The non-adaptive initial proposal distribution is a truncated multivariate Gaussian $\mathcal{N}(\mu_0, \Sigma_0)$ whose support is bounded by the hypercube $[l, u]$. The adaptive proposal distribution is a truncated Gaussian mixture $\Pi(\{x_j^{\text{curr}}\}_{j=1}^{n_{\text{curr}}}, \lambda) := \frac{1}{n_{\text{curr}}-1} \sum_{j=1}^{n_{\text{curr}}-1} \mathcal{N}(x_j^{\text{curr}}, \lambda I)$ with $\lambda > 0$ and support $[l, u]$.

Algorithm 2 Monte Carlo

input $f, t, n_{\text{test}}, n_{\text{delay}}, \mu_0, \Sigma_0, \{x_j^{\text{curr}}\}_{j=1}^{n_{\text{curr}}}, \lambda, l, u$
output x^*
function MC
  if $t \leq n_{\text{delay}}$ **then**
    $\{x_i^{\text{test}}\}_{i=1}^{n_{\text{test}}} \rightsquigarrow \text{trunc}_l^u\left[\mathcal{N}(\mu_0, \Sigma_0)\right]$
  else
    $\{x_i^{\text{test}}\}_{i=1}^{n_{\text{test}}} \rightsquigarrow \text{trunc}_l^u\left[\Pi(\{x_j^{\text{curr}}\}_{j=1}^{n_{\text{curr}}}, \lambda)\right]$
  end if
  $i^* \leftarrow \arg\min_{i \in \{1 \ldots n_{\text{test}}\}} f(x_i^{\text{test}})$
  $x^* \leftarrow x_{i^*}^{\text{test}}$
end function

B.2.3. GRID SEARCH

The grid search (GS) optimisation method proceeds as in Algorithm 3. The function GS takes the following inputs: f is the objective function; t is the iteration count; l and u are the lower- and upper-bounds of the grid; n_0 is the initial grid size.

Algorithm 3 Grid Search

input f, t, l, u, n_0
output x^*
function GS
  $n_{\text{grid}} \leftarrow n_0 + \text{Round}(\sqrt{t})$
  $\delta_{\text{grid}} \leftarrow (u - l)/(n_{\text{grid}} - 1)$
  $X_{\text{grid}} \leftarrow \{l, l + \delta_{\text{grid}}, \ldots, u\}^d$
  $x^* \leftarrow \arg\min_{x \in X_{\text{grid}}} f(x)$
end function

B.3. Remark on Application to a Reference Point Set

It is interesting to comment on the behaviour of our proposed methods in the case where X is a finite set or the global optimisation over X is replaced by a discrete optimisation over a pre-determined fixed set $Y = \{y_i\}_{i=1}^N \subseteq X$. In this case it can be shown that:

- The algorithm after n iterations will have selected n points $\{y_{\pi(i)}\}_{i=1}^n$ with replacement from Y. (Here $\pi(i)$ indexes the point that was selected at iteration i of the algorithm.)
- The empirical measure $\frac{1}{n}\sum_{i=1}^n \delta_{y_{\pi(i)}}$ can be expressed as $\sum_{i=1}^N w_i y_i$ for some weights w_i.
- The weights w_i converge to

$$(*) \;=\; \underset{\substack{w \geq 0 \\ w_1 + \cdots + w_N = 1}}{\arg\min} \sqrt{\frac{1}{N^2}\sum_{i,j=1}^N w_i w_j k_0(y_i, y_j)}.$$

- At iteration n, it holds that $D_{\mathcal{K}_0, P}(\{y_{\pi(i)}\}_{i=1}^n) = (*) + O(\sqrt{\log(n)/n})$.

Thus in this scenario the algorithms that we have proposed act to ensure that these points are optimally weighted in the sense just described.

C. Experimental Protocol and Additional Numerical Results

This section contains additional numerical results that elaborate on the three experiments reported in the main text.

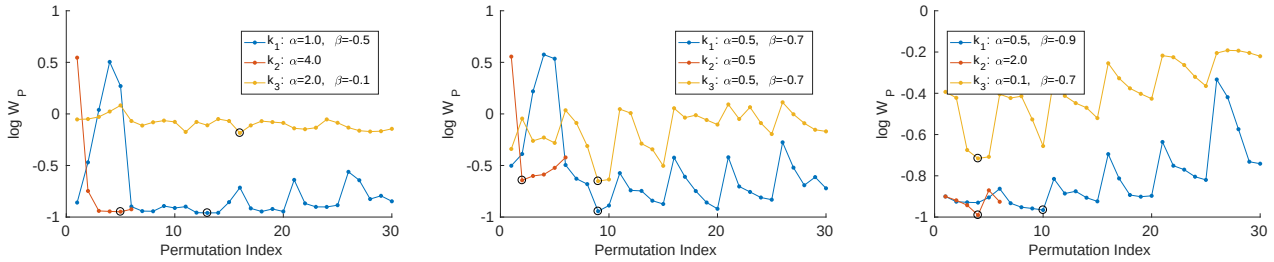
C.1. Gaussian Mixture Test

Recall from the main text that the kernels k_1, k_2 and k_3 contain either one or two hyper-parameters that must be selected. For each of the methods (a)-(f) reported in Figure 2 in the main text we optimised these parameters over a discrete set, with respect to an objective function of W_P based on a point set of size $n = 100$ and the Nelder-Mead optimisation method. The set of possible values for α was $\{0.1\eta, 0.5\eta, \eta, 2\eta, 4\eta, 8\eta\}$, where η is a problem dependent "base scale" and chosen to be 1 for the Gaussian mixture test. The set of possible values for β was $\{-0.1, -0.3, -0.5, -0.7, -0.9\}$. The sensitivity of the reported results to the variation in hyper-parameters is shown, for the Gaussian mixture test, in Figure 5. Point sets obtained under representatives of each method class are shown in Figure 6.

For all the global optimisation methods we imposed a bounding box $(-5, 5) \times (-5, 5)$; for the Nelder-Mead method, we set $n_{\text{init}} = 3$, $n_{\text{delay}} = 20$, $\mu_0 = (0, 0)$, $\Sigma_0 = 25I$, and $\lambda = 1$; for the Monte Carlo method, we set $n_{\text{test}} = 20$, $n_{\text{delay}} = 20$, $\mu_0 = (0, 0)$, $\Sigma_0 = 25I$, and $\lambda = 1$; for the grid search, we set $n_0 = 100$.

For MED the tuning parameter δ was considered for $\delta = 4$, $\delta = 8$ or $\delta = 16$, with $\delta = 4d = 8$ being the recommendation in (Joseph et al., 2017).

For SVGD we set the initial point-set to be an equally spaced rectangular grid over the bounding box. Following (Liu & Wang, 2016), the step-size ϵ for SVGD was determined by AdaGrad with a master step-size of 0.1 and a momentum factor of 0.9.



(a) Stein Points (Greedy) (b) Stein Points (Herding) (c) SVGD

Figure 5: Kernel parameter selection results for the Gaussian mixture test. Parameters α, β in the kernels k_1, k_2, k_3 were optimised over a discrete set with respect to the Wasserstein distance W_P for a point set of size $n = 100$. The values $\log W_P$ (y-axis) are shown for all different configurations of parameters (x-axis) considered. Optimal parameter configurations are circled and detailed in the legend.

C.2. Gaussian Process Test

For the Gaussian process test, the base scale η is also set to 1. The sensitivity of results to the selection of kernel parameters was reported in Figure 7. Point sets obtained under representatives of each method class are shown in Figures 8 and 9. Detailed results for each method considered are contained in Figure 10.

For all the global optimisation methods we imposed a bounding box of $(-5, 5) \times (-13, -7)$; for the Nelder-Mead method, we set $n_{\text{init}} = 3$, $n_{\text{delay}} = 20$, $\mu_0 = (0, -10)$, $\Sigma_0 = 25I$, and $\lambda = 1$; for the Monte Carlo method, we set $n_{\text{test}} = 20$, $n_{\text{delay}} = 20$, $\mu_0 = (0, -10)$, $\Sigma_0 = 25I$, and $\lambda = 1$; for the grid search, we set $n_0 = 100$.

For SVGD we set the initial point-set to be an equally spaced rectangular grid over the bounding box. Following (Liu & Wang, 2016), the step-size ϵ for SVGD was determined by AdaGrad with a master step-size of 0.1 and a momentum factor of 0.9.

C.3. IGARCH Test

For the IGARCH test, we choose the base scale η to be 1e-5. The sensitivity of results to the selection of kernel parameters was reported in Figure 11. Point sets obtained under representatives of each method class are shown in Figures 12 and 13. Detailed results for each method considered are contained in Figure 14.

For all the global optimisation methods we impose a bounding box of $(0.002, 0.04) \times (0.05, 0.2)$; for the Nelder-Mead

method, we set n_{init} = 3, n_{delay} = 20, μ_0 = $(0.021, 0.125)$, Σ_0 = $\text{diag}[(\text{1e-4}, \text{1e-3})]$, and λ = 1e-5; for the Monte Carlo method, we set n_{test} = 20, n_{delay} = 20, μ_0 = $(0.021, 0.125)$, Σ_0 = $\text{diag}[(\text{1e-4}, \text{1e-3})]$, and λ = 1e-5; for the grid search, we set n_0 = 100.

For SVGD we set the initial point-set to be an equally spaced rectangular grid over the bounding box. Following (Liu & Wang, 2016), the step-size ϵ for SVGD was determined by AdaGrad with a master step-size of 1e-3 and a momentum factor of 0.9.

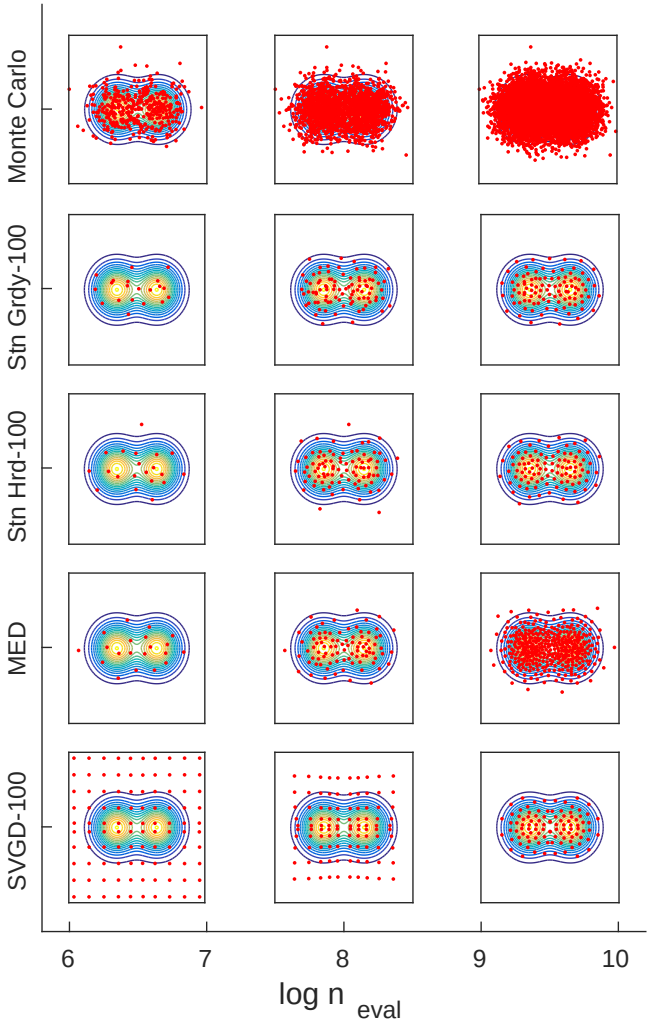
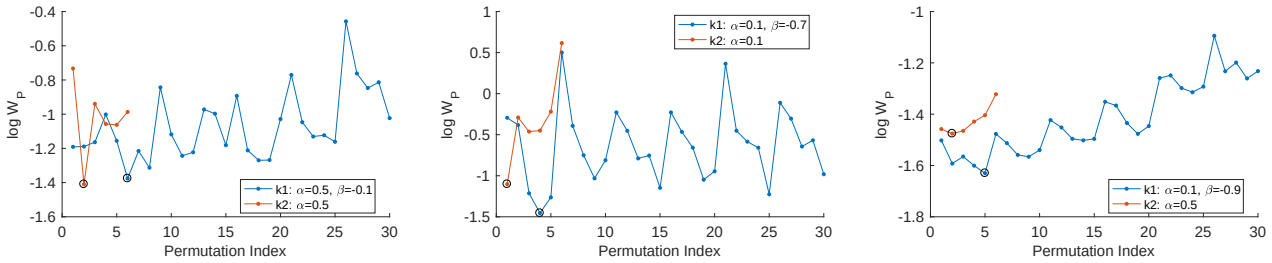


Figure 6: Typical point sets obtained in the Gaussian mixture test, where the budget-constrained methods Stein Greedy-100 (Stn Grdy-100) and Stein Herding-100 (Stn Hrd-100) are considered. [Here each row corresponds to an algorithm, and each column corresponds to a chosen level of computational cost. The left border of each sub-plot is aligned to the exact value of $\log n_{\text{eval}}$ spent to obtain each point-set.]



(a) Stein Points (Greedy) (b) Stein Points (Herding) (c) SVGD

Figure 7: Kernel parameter selection results for the Gaussian process test. Parameters α, β in the kernels k_1, k_2, k_3 were optimised over a discrete set with respect to the Wasserstein distance W_P for a point set of size $n = 100$. The values $\log W_P$ (y-axis) are shown for all different configurations of parameters (x-axis) considered. Optimal parameter configurations are circled and detailed in the legend.

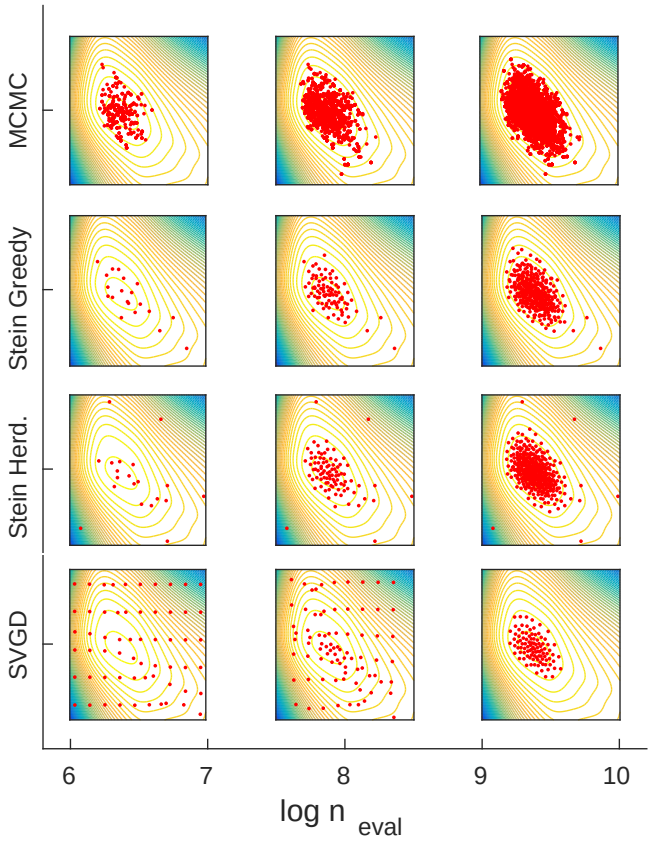


Figure 8: Typical point sets obtained in the Gaussian process test. [Here each row corresponds to an algorithm, and each column corresponds to a chosen level of computational cost. The left border of each sub-plot is aligned to the exact value of $\log n_{\text{eval}}$ spent to obtain each point-set. MCMC represents a random-walk Metropolis algorithm with a proposal distribution optimised according to acceptance rate.]

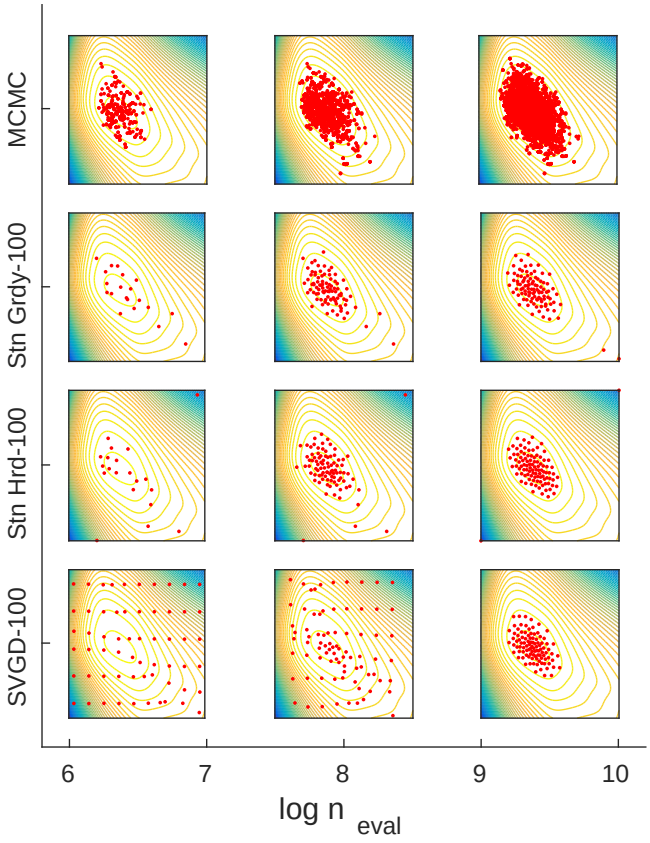
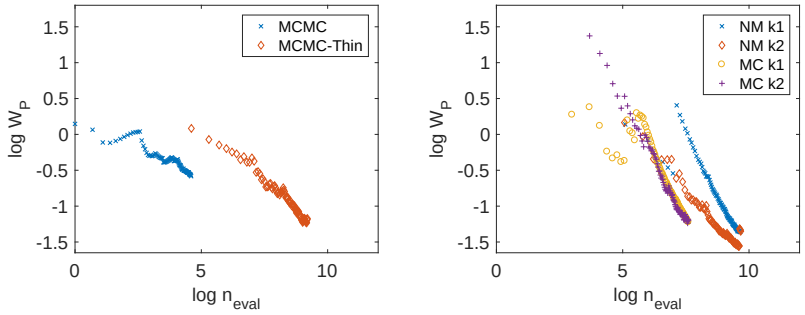
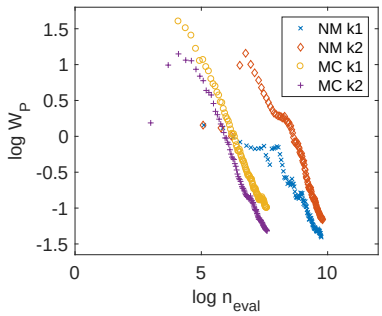


Figure 9: Typical point sets obtained in the Gaussian process test, where the budget-constrained methods Stein Greedy-100 (Stn Grdy-100) and Stein Herding-100 (Stn Hrd-100) are considered. [Here each row corresponds to an algorithm, and each column corresponds to a chosen level of computational cost. The left border of each sub-plot is aligned to the exact value of $\log n_{\text{eval}}$ spent to obtain each point-set. MCMC represents a random-walk Metropolis algorithm with a proposal distribution optimised according to acceptance rate.]

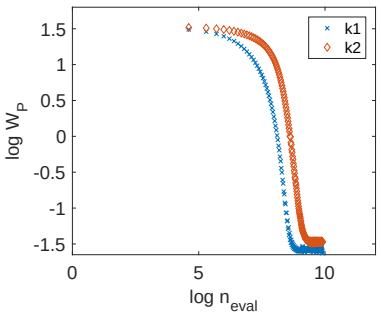


(a) Monte Carlo

(b) Stein Points (Greedy)

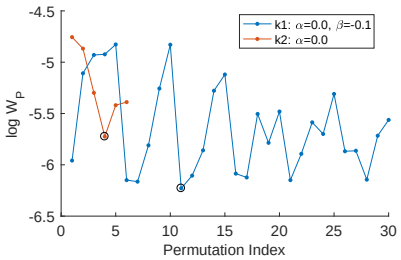


(c) Stein Points (Herding)

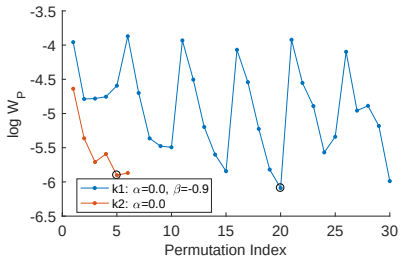


(d) SVGD

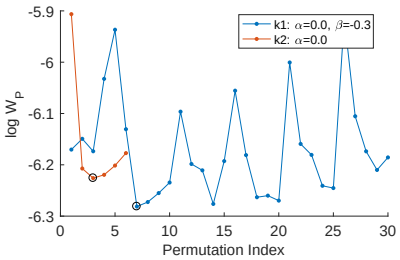
Figure 10: Results for the Gaussian process test. [Here $n = 100$. x-axis: log of the number n_{eval} of model evaluations that were used. y-axis: log of the Wasserstein distance $W_P(\{x_i\}_{i=1}^n)$ obtained. Kernel parameters α, β were optimised according to W_P. In sub-figure 10a, MCMC represents a random-walk Metropolis algorithm with a proposal distribution optimised according to acceptance rate. MCMC-Thin represents a thinned chain by taking every 100th observation.]



(a) Stein Points (Greedy)



(b) Stein Points (Herding)



(c) SVGD

Figure 11: Kernel parameter selection results for the IGARCH test. Parameters α, β in the kernels k_1, k_2, k_3 were optimised over a discrete set with respect to the Wasserstein distance W_P for a point set of size $n = 100$. The values $\log W_P$ (y-axis) are shown for all different configurations of parameters (x-axis) considered. Optimal parameter configurations are circled and detailed in the legend.

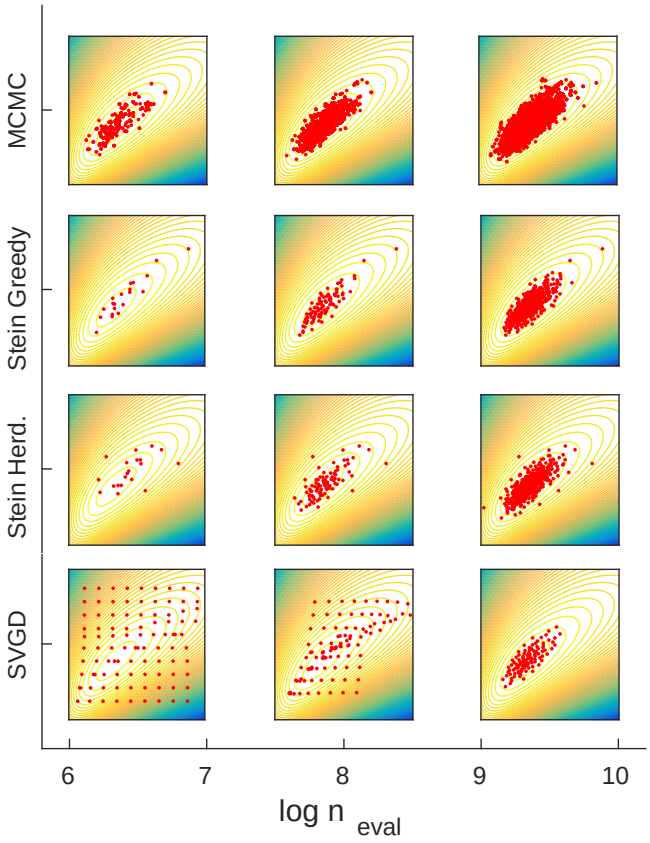


Figure 12: Typical point sets obtained in the IGARCH test. [Here each row corresponds to an algorithm, and each column corresponds to a chosen level of computational cost. The left border of each sub-plot is aligned to the exact value of $\log n_{\text{eval}}$ spent to obtain each point-set. MCMC represents a random-walk Metropolis algorithm with a proposal distribution optimised according to acceptance rate.]

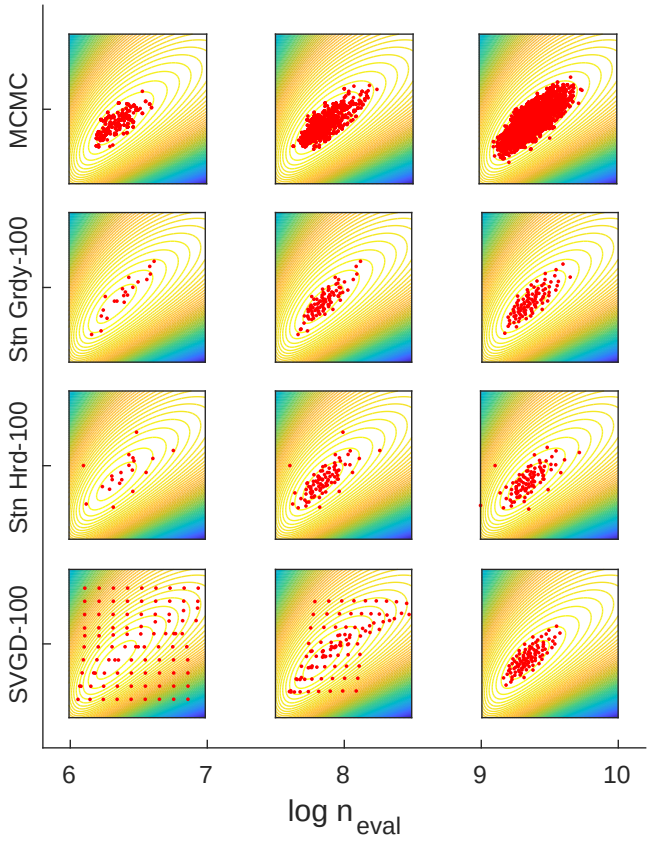
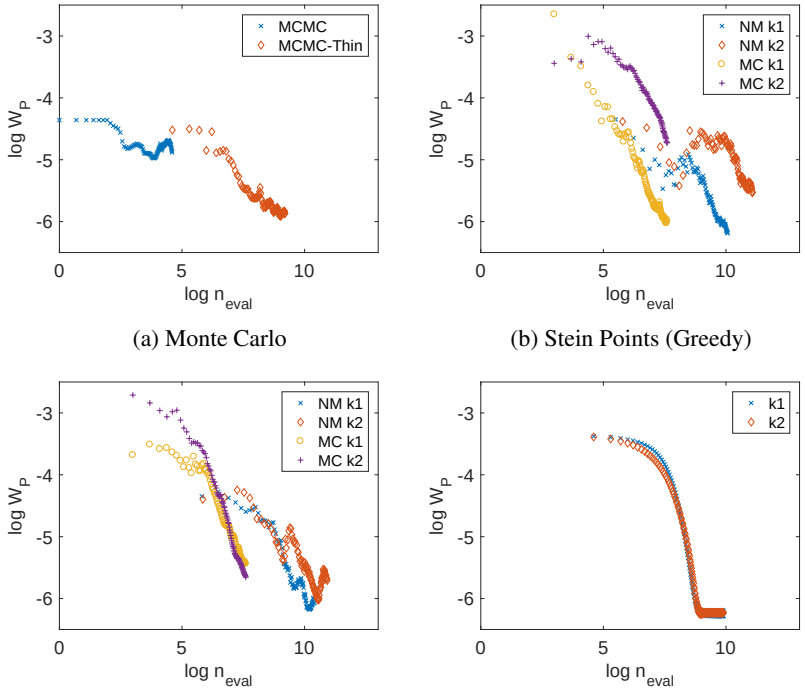


Figure 13: Typical point sets obtained in the IGARCH test, where the budget-constrained methods Stein Greedy-100 (Stn Grdy-100) and Stein Herding-100 (Stn Hrd-100) are considered. [Here each row corresponds to an algorithm, and each column corresponds to a chosen level of computational cost. The left border of each sub-plot is aligned to the exact value of $\log n_{\text{eval}}$ spent to obtain each point-set. MCMC represents a random-walk Metropolis algorithm with a proposal distribution optimised according to acceptance rate.]



(a) Monte Carlo

(b) Stein Points (Greedy)

(c) Stein Points (Herding)

(d) SVGD

Figure 14: Results for the IGARCH test. [Here $n = 100$. x-axis: log of the number n_{eval} of model evaluations that were used. y-axis: log of the Wasserstein distance $W_P(\{x_i\}_{i=1}^n)$ obtained. Kernel parameters α, β were optimised according to W_P. In sub-figure 14a, MCMC represents a random-walk Metropolis algorithm with a proposal distribution optimised according to acceptance rate. MCMC-Thin represents a thinned chain by taking every 100th observation.]