A Duality

Proposition 1 (Strong duality)**.** *Let $\nu \in \mathcal{P}(\mathcal{X})$ and $f : \mathcal{P}(\mathcal{X}) \to [0, +\infty)$ a convex, lower semicontinuous and proper functional. Define $P_\nu^{\gamma,\tau}$ as in (7) and $D_\nu^{\gamma,\tau}$ as in (9). Assume $\gamma > 0$. Then*

$$\inf_{\mu\in\mathcal{P}(\mathcal{X})} P_\nu^{\gamma,\tau}(\mu) = \sup_{g\in L^2(\mathcal{X}),h\in L^2(\mathcal{X})} D_\nu^{\gamma,\tau}(g,h). \quad (16)$$

Suppose f is strictly convex and let g_*, h_* maximize $D_\nu^{\gamma,\tau}$. Then

$$\mu_* = \nabla f^*(-\frac{1}{\tau}g_*) \quad (17)$$

minimizes $P_\nu^{\gamma,\tau}$.

Proof. The domain of the primal problem is $\mathcal{P}(\mathcal{X})$, the space of probability measures, which is a closed subset of $\mathcal{M}(\mathcal{X})$, the Radon measures. $\mathcal{M}(\mathcal{X})$ is the topological dual of $\mathcal{C}_b(\mathcal{X})$, the space of continuous, bounded functions. This duality defines a product

$$\langle \mu, g\rangle \triangleq \int_{\mathcal{X}} g(\mathbf{x}) d\mu(\mathbf{x}), \quad (18)$$

for $\mu \in \mathcal{M}(\mathcal{X})$ and $g \in C_b(\mathcal{X})$.

For $\mathcal{W}_\gamma(\cdot,\nu)$ and f both convex, lower semicontinuous and proper, Fenchel's duality theorem (Rockafellar, 1970) has that

$$\begin{aligned}\inf_{\mu\in\mathcal{P}(\mathcal{X})} &\mathcal{W}_\gamma(\mu,\nu) + \tau f(\mu) \\ &= \sup_{g\in\mathcal{C}_b(\mathcal{X})} -\mathcal{W}_\gamma(\cdot,\nu)^*(g) - \tau f^*(-\frac{1}{\tau}g),\end{aligned} \quad (19)$$

with $\mathcal{W}_\gamma(\cdot,\nu)^*$ and f^* the convex conjugates,

$$\begin{aligned}\mathcal{W}_\gamma(\cdot,\nu)^*(g) &= \sup_{\mu\in\mathcal{P}(\mathcal{X})} \langle \mu, g\rangle - \mathcal{W}_\gamma(\mu,\nu), \\ (\tau f)^*(-g) = \tau f^*(-\frac{1}{\tau}g) &= \sup_{\mu\in\mathcal{P}(\mathcal{X})} -\langle \mu, g\rangle - \tau f(\mu).\end{aligned} \quad (20)$$

Let $\mathcal{U}(\nu)$ be the set of joint probability measures on $\mathcal{X}\times\mathcal{X}$ having second marginal equal to ν,

$$\mathcal{U}(\nu) \triangleq \{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) : \pi(\mathcal{X}, A) = \nu(A) \ \forall \text{ Borel } A \subseteq \mathcal{X}\}. \quad (21)$$

By expanding the term $\mathcal{W}_\gamma(\cdot,\nu)$, we can rewrite $\mathcal{W}_\gamma(\cdot,\nu)^*$ as an optimization over $\mathcal{U}(\nu)$,

$$\mathcal{W}_\gamma(\cdot,\nu)^*(g) = -\inf_{\mathcal{U}(\nu)} \langle \pi, c - g\rangle + \gamma R(\pi). \quad (22)$$

The Lagrangian dual for $\mathcal{W}_\gamma(\cdot,\nu)^*$ is

$$\begin{aligned}\mathcal{W}_\gamma(\cdot,\nu)^*(g) = \\ - \sup_{h\in\mathcal{C}_b(\mathcal{X}),\varepsilon\in C_b^+(\mathcal{X}\times\mathcal{X})} \inf_{\pi\in\mathcal{M}(\mathcal{X}\times\mathcal{X})} \\ \langle c - g, \pi\rangle + \gamma R(\pi) - \langle \varepsilon, \pi\rangle + \langle \nu - \pi, h\rangle.\end{aligned} \quad (23)$$

Here $\mathcal{C}_b^+(\mathcal{X}\times\mathcal{X})$ is the space of nonnegative, continuous, bounded functions.

The regularization functional R is differentiable in the sense of measures (Santambrogio, 2015), meaning that for any $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$ there exists a function $\nabla R(\pi) \in \mathcal{C}_b(\mathcal{X}\times\mathcal{X})$ such that, for all $\xi_1, \xi_2 \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$ and $t > 0$,

$$R(\pi + t(\xi_2 - \xi_1)) = R(\pi) + \langle \xi_2 - \xi_1, \nabla R(\pi)\rangle + \mathcal{O}(t). \quad (24)$$

Because R is Legendre, ∇R is a bijection between $\operatorname{int}\operatorname{dom} R$ and $\operatorname{int}\operatorname{dom} R^*$ whose inverse is ∇R^*, defined by

$$\nabla R^*(z) = \operatorname*{argmax}_{\pi\in\mathcal{P}(\mathcal{X})} \langle \pi, z\rangle - R(\pi), \quad (25)$$

for any $z \in C_b(\mathcal{X}\times\mathcal{X})$ (Bernhard and Rapaport, 1995, Theorem C.1).

Necessary conditions for optimality of (23) (Luenberger, 1997, Section 9) are

$$\begin{aligned}&c - g + \gamma\nabla R(\pi) - \varepsilon - h = 0 \\ &\varepsilon, \pi \geq 0 \\ &\varepsilon(\mathbf{x},\mathbf{y}) = 0 \ \forall \ (\mathbf{x},\mathbf{y}) \in \operatorname{supp}\pi.\end{aligned}$$

The first condition implies

$$\begin{aligned}\nabla R(\pi) &= \frac{1}{\gamma}(g + h - c + \varepsilon) \\ \Rightarrow \pi &= \nabla R^*\left(\frac{1}{\gamma}(g + h - c + \varepsilon)\right).\end{aligned}$$

The third condition (complementary slackness) then guarantees, for any optimal (g, h, ε),

$$\varepsilon(\mathbf{x},\mathbf{y}) > 0 \ \Rightarrow \ (\mathbf{x},\mathbf{y}) \notin \operatorname{supp}\nabla R^*\left(\frac{1}{\gamma}(g + h - c + \varepsilon)\right). \quad (26)$$

And the second condition (nonnegativity) then implies

$$\begin{aligned}&\nabla R^*\left(\frac{1}{\gamma}(g + h - c + \varepsilon)\right)(A) \\ &\quad = \nabla R^*\left(\max\left\{\frac{1}{\gamma}(g + h - c), \nabla R(\mathbf{0})\right\}\right)(A), \\ &\quad \forall \text{ Borel } A \subseteq \mathcal{X}\times\mathcal{X},\end{aligned} \quad (27)$$

with $\mathbf{0}$ the zero measure. Note that we've abused notation slightly, with $\nabla R(\mathbf{0})$ here referring to the natural extension of ∇R to the nonnegative measures, whose range lies in $\mathcal{C}(\mathcal{X} \times \mathcal{X})$, the continuous (but possibly unbounded) functions.

By definition of the convex conjugate,

$$R\left(\nabla R^*(\xi)\right) = \langle \xi, \nabla R^*(\xi) \rangle - R^*(\xi),$$

so plugging optimal π into the Lagrangian dual for $\mathcal{W}_\gamma(\cdot, \nu)^*$, we get

$$\begin{aligned} \mathcal{W}_\gamma(\cdot, \nu)^*(g) = - \sup_{h \in C_b(\mathcal{X})} \\ \langle h, \nu \rangle - \gamma R^* \left(\max \left\{ \frac{1}{\gamma}(g + h - c), \nabla R(0) \right\} \right). \end{aligned} \tag{28}$$

From (19), then we get the Fenchel dual

$$\begin{aligned} D_\nu^{\gamma,\tau}(g, h) = \\ - \tau f^* \left(-\frac{1}{\tau} g \right) + \langle h, \nu \rangle \\ - \gamma R^* \left(\max \left\{ \frac{1}{\gamma}(g + h - c), \nabla R(0) \right\} \right). \end{aligned} \tag{29}$$

Suppose $g_*, h_* \in C_b(\mathcal{X})$ optimize the dual objective $D_\nu^{\gamma,\tau}$. Then μ_* optimal for $P_\nu^{\gamma,\tau}$ satisfies

$$\mu_* \in \partial(\tau f)^*(-g_*).$$

When f is strictly convex, this is $\mu_* = \nabla(\tau f)^*(-g_*) = \nabla f^*(-\frac{1}{\tau} g_*)$. □

B Representer theorem

Proposition 2 (Representation for general RKHS)**.** *Let $\nu \in \mathcal{P}(\mathcal{X})$ and $\gamma, \tau, N > 0$. Let $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^N \subset \mathcal{X} \times \mathcal{X}$. Then there exist $g_*, h_* \in \mathcal{H}$ maximizing* (14) *such that*

$$(g_*, h_*) = \sum_{i=1}^N \left(\alpha_g^{(i)} \kappa(\mathbf{x}^{(i)}, \cdot), \alpha_h^{(i)} \kappa(\mathbf{y}^{(i)}, \cdot) \right),$$

for some sequences of scalar coefficients $\{\alpha_g^{(i)}\}_{i=1}^N$ and $\{\alpha_h^{(i)}\}_{i=1}^N$, with $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ the reproducing kernel for $\mathcal{H}$.

Proof. Let $\mathcal{H}$ be the RKHS having kernel κ, and let $\langle \cdot, \cdot \rangle_\mathcal{H} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ be the associated inner product. Let $g \in \mathcal{H}$. From the reproducing property of $\mathcal{H}$, we have that pointwise evaluation is a linear functional such that $g(\mathbf{x}) = \langle g, \kappa(\mathbf{x}, \cdot) \rangle_\mathcal{H}$, for all $\mathbf{x} \in \mathcal{X}$.

Let $\mathcal{H}_N \subset \mathcal{H}$ be the linear span of the functions $\kappa(\mathbf{x}^{(i)}, \cdot)$, and $\mathcal{H}_N^\perp$ its orthogonal complement. For any $g \in \mathcal{H}$, we can decompose it as $g = g^\parallel + g^\perp$, with $g^\parallel \in \mathcal{H}_N$ and $g^\perp \in \mathcal{H}_N^\perp$. Moreover, $D_{\nu,N}^{\gamma,\tau}(g, h) = D_{\nu,N}^{\gamma,\tau}(g^\parallel, h)$, as $D_{\nu,N}^{\gamma,\tau}$ depends on its first argument only via the evaluation functional at each point,

$$g(\mathbf{x}^{(i)}) = \langle \kappa(\mathbf{x}^{(i)}, \cdot), g \rangle_\mathcal{H} = \langle \kappa(\mathbf{x}^{(i)}, \cdot), g^\parallel \rangle_\mathcal{H}.$$

Hence if $D_{\nu,N}^{\gamma,\tau}$ is maximized by g_*, it is also maximized by $g_*^\parallel \in \mathcal{H}_N$. The same argument holds for h_*. □

C Consistency

We make the following assumptions.

A1 $\mathcal{X} \times \mathcal{X}$ is compact.

A2 μ_0 and ν_0 are bounded away from zero: $\mu_0(\mathbf{x}) \geq U_0^{\min} > 0$, $\nu_0(\mathbf{y}) \geq V_0^{\min} > 0$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

A3 $\mathcal{G}$ is compact and convex, with $\|g\|_\mathcal{H} \leq H$ for all $g \in \mathcal{G}$.

A4 $\mathcal{H}$ has reproducing kernel κ that is bounded: $\max_{\mathbf{x} \in \mathcal{X}} \sqrt{\kappa(\mathbf{x}, \mathbf{x})} = K < \infty$.

A5 $\bar{f}^*$ is convex and L_{f^*}-Lipschitz.

A6 $\operatorname{dom} \bar{R}^* = \mathbb{R}$.

The assumptions guarantee that the Monte Carlo dual objective (14) is L-Lipschitz.

Proposition 4 (Lipschitz property for $d_\nu^{\gamma,\tau}$)**.** *Let $d_\nu^{\gamma,\tau}$ be defined as in* (13) *and suppose Assumptions* **A1**-**A6** *hold. Let $U^{\max} = \max_{\mathbf{x} \in \mathcal{X}, g \in \mathcal{H}} \frac{\nabla f^*(-\frac{1}{\tau} g(\mathbf{x}))}{\mu_0(\mathbf{x})}$ and $V^{\max} = \max_{\mathbf{y} \in \mathcal{X}} \frac{\nu(\mathbf{y})}{\nu_0(\mathbf{y})}$. Then for all $g, g', h, h' \in \mathcal{H}$, $d_\nu^{\gamma,\tau}$ satisfies*

$$\begin{aligned} |d_\nu^{\gamma,\tau}(\mathbf{x}, \mathbf{y}, g, h) - d_\nu^{\gamma,\tau}(\mathbf{x}, \mathbf{y}, g', h')| \\ \leq L \|(g(\mathbf{x}), h(\mathbf{y})) - (g'(\mathbf{x}), h'(\mathbf{y}))\|_1 \end{aligned} \tag{30}$$

with constant L defined by $L = \max \left\{ U^{\max}, V^{\max}, \frac{\nabla \bar{R}^*\left(\frac{2}{\gamma} KH\right)}{U_0^{\min} V_0^{\min}} \right\}$.

Proof. Note that $U^{\max}$ and $V^{\max}$ are finite by assumptions **A2** and **A5**.

By **A3-A4**, we have that $K = \min_{\mathbf{x} \in \mathcal{X}} \sqrt{\kappa(\mathbf{x}, \mathbf{x})} < \infty$, and $\mathcal{G} \times \mathcal{G}$ is bounded, such that $\|g\|_\mathcal{H}, \|h\|_\mathcal{H} \leq H$. Therefore $|g(\mathbf{x})|, |h(\mathbf{y})| \leq KH$, because by the reproducing property

$$\begin{aligned} |g(\mathbf{x})| &= |\langle \kappa(\mathbf{x}, \cdot), g \rangle_\mathcal{H}| \\ &\leq \|\kappa(\mathbf{x}, \cdot)\|_\mathcal{H} \|g\|_\mathcal{H} \\ &\leq K \|g\|_\mathcal{H}, \\ &\leq KH, \end{aligned}$$

with the second step from Cauchy-Schwarz. The analogous result holds for $|h(\mathbf{y})|$.

Let $q(g(\mathbf{x}), h(\mathbf{y})) = \frac{1}{\gamma}(g(\mathbf{x}) + h(\mathbf{y}) - d^2(\mathbf{x}, \mathbf{y}))$. Then $d_\nu^{\gamma,\tau}$ has subderivatives

$$\frac{\partial d_\nu^{\gamma,\tau}}{\partial g(\mathbf{x})} = \frac{\nabla \bar{f}^*(-\frac{1}{\tau}g(\mathbf{x}))}{\mu_0(\mathbf{x})} - \frac{\gamma}{\mu_0(\mathbf{x})\nu_0(\mathbf{y})} \begin{cases} \frac{1}{\gamma}\nabla \bar{R}^* \left(q(g(\mathbf{x}), h(\mathbf{y}))\right) \\ \quad \text{if } \ q(g(\mathbf{x}), h(\mathbf{y})) > \nabla \bar{R}(0) \\ [0, \frac{1}{\gamma}\nabla \bar{R}^*(q(g(\mathbf{x}), h(\mathbf{y})))] \\ \quad \text{if } \ q(g(\mathbf{x}), h(\mathbf{y})) = \nabla \bar{R}(0) \\ 0 \ \text{ otherwise} \end{cases} \tag{31}$$

in $g(\mathbf{x})$ and

$$\frac{\partial d_\nu^{\gamma,\tau}}{\partial h(\mathbf{y})} = \frac{\nu(\mathbf{y})}{\nu_0(\mathbf{y})} - \frac{\gamma}{\mu_0(\mathbf{x})\nu_0(\mathbf{y})} \begin{cases} \frac{1}{\gamma}\nabla \bar{R}^* \left(q(g(\mathbf{x}), h(\mathbf{y}))\right) \\ \quad \text{if } \ q(g(\mathbf{x}), h(\mathbf{y})) > \nabla \bar{R}(0) \\ [0, \frac{1}{\gamma}\nabla \bar{R}^*(q(g(\mathbf{x}), h(\mathbf{y})))] \\ \quad \text{if } \ q(g(\mathbf{x}), h(\mathbf{y})) = \nabla \bar{R}(0) \\ 0 \ \text{ otherwise} \end{cases} \tag{32}$$

in $h(\mathbf{y})$. In both cases, the second term subtracts a nonnegative quantity while the first term is nonnegative. As $g(\mathbf{x})$ and $h(\mathbf{y})$ are bounded, q is bounded from above, with

$$q(g(\mathbf{x}), h(\mathbf{y})) \leq \frac{2}{\gamma}KH.$$

$\nabla \bar{R}^*$ is monotonic, so it is bounded above by $\nabla \bar{R}^* \left(\frac{2}{\gamma}KH\right)$. We therefore have

$$\left|\frac{\partial d_\nu^{\gamma,\tau}}{\partial g(\mathbf{x})}\right| \leq \max\left\{U^{\max}, \frac{\nabla \bar{R}^* \left(\frac{2}{\gamma}KH\right)}{U_0^{\min}V_0^{\min}}\right\} \triangleq L_g$$

$$\left|\frac{\partial d_\nu^{\gamma,\tau}}{\partial h(\mathbf{y})}\right| \leq \max\left\{V^{\max}, \frac{\nabla \bar{R}^* \left(\frac{2}{\gamma}KH\right)}{U_0^{\min}V_0^{\min}}\right\} \triangleq L_h.$$

$\bar{R}^*$ is smooth on $\operatorname{int}\operatorname{dom}\bar{R}^*$, and $\frac{2}{\gamma}KH \in \operatorname{int}\operatorname{dom}\bar{R}^*$ by Assumption **A6**, so $\nabla \bar{R}^* \left(\frac{2}{\gamma}KH\right)$ is finite.

Letting $L = \max\{L_g, L_h\}$, this implies

$$\begin{aligned} &|d_\nu^{\gamma,\tau}(\mathbf{x}, \mathbf{y}, g, h) - d_\nu^{\gamma,\tau}(\mathbf{x}, \mathbf{y}, g', h')| \\ &\quad \leq L\|(g(\mathbf{x}), h(\mathbf{y})) - (g'(\mathbf{x}), h'(\mathbf{y}))\|_1, \end{aligned} \tag{33}$$

for all $(g, h), (g', h') \in \mathcal{G} \times \mathcal{G}$ and $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{X}$. □

Note that assumption **A5** is satisfied by an advection-diffusion, so long as we assume w is bounded below, as

$$\begin{aligned} &\max_{g\in\mathcal{G}, \mathbf{x}\in\mathcal{X}} \left|\nabla f^*(-\frac{1}{\tau}g(\mathbf{x}))\right| \\ &\quad = \max_{g\in\mathcal{G}, \mathbf{x}\in\mathcal{X}} \exp(-\frac{\beta}{\tau}g(\mathbf{x}) - w(\mathbf{x})) \\ &\quad \leq \exp(\frac{\beta}{\tau}KH - \beta W) \end{aligned} \tag{34}$$

with $W = \min_{\mathbf{x}\in\mathcal{X}} w(\mathbf{x})$.

Under the assumptions, then, we get uniform convergence of the stochastic dual objective (14) to its expectation (12), and this suffices to guarantee consistency.

Proposition 3 (Consistency of stochastic program)**.** *Let $D_\nu^{\gamma,\tau}$ and $D_{\nu,N}^{\gamma,\tau}$ be defined as in* (12) *and* (14)*, respectively, with $\gamma, \tau, N > 0$, and suppose Assumptions* **A1**-**A6** *hold. Let (g_N, h_N) optimize $D_{\nu,N}$ and (g_∞, h_∞) optimize $D_\nu^{\gamma,\tau}$. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the sample of size N,*

$$\begin{aligned} &D_\nu^{\gamma,\tau}(g_\infty, h_\infty) - D_\nu^{\gamma,\tau}(g_N, h_N) \\ &\quad \leq \mathcal{O}\left(\sqrt{\frac{(HKL)^2 \log(1/\delta)}{N}}\right). \end{aligned} \tag{35}$$

Proof. Note that $d_\nu^{\gamma,\tau}$ is jointly convex in $g(\mathbf{x})$ and $h(\mathbf{y})$, and these are in linear in g and h, respectively. They can be written $g(\mathbf{x}) = \langle g, \kappa(\mathbf{x}, \cdot)\rangle_{\mathcal{H}}$ with $\|\kappa(\mathbf{x}, \cdot)\|_{\mathcal{H}} \leq K$ and $\|g\|_{\mathcal{H}} \leq H$, and similarly for $h(\mathbf{y})$, with the same bounds.

From (Shalev-Shwartz et al., 2009) Thm. 1, then, we have uniform convergence of the empirical functional to its expectation, such that with probability $1 - \delta$

$$\begin{aligned} &\sup_{g,h\in\mathcal{H}} \left|D_\nu^{\gamma,\tau}(g, h) - D_{\nu,N}^{\gamma,\tau}(g, h)\right| \\ &\quad \leq \mathcal{O}\left(\sqrt{\frac{(HKL)^2 \log(1/\delta)}{N}}\right), \end{aligned} \tag{36}$$

for any $g, h \in \mathcal{G}$. This implies

$$\begin{aligned} &D_\nu^{\gamma,\tau}(g_\infty, h_\infty) - D_{\nu,N}^{\gamma,\tau}(g_\infty, h_\infty) \\ &\qquad + D_{\nu,N}^{\gamma,\tau}(g, h) - D_\nu^{\gamma,\tau}(g, h) \\ &\qquad \leq \mathcal{O}\left(\sqrt{\frac{(HKL)^2 \log(1/\delta)}{N}}\right) \\ \Rightarrow &D_\nu^{\gamma,\tau}(g_\infty, h_\infty) - D_\nu^{\gamma,\tau}(g, h) \\ &\qquad \leq \left(D_{\nu,N}^{\gamma,\tau}(g_\infty, h_\infty) - D_{\nu,N}^{\gamma,\tau}(g, h)\right) \\ &\qquad\quad + \mathcal{O}\left(\sqrt{\frac{(HKL)^2 \log(1/\delta)}{N}}\right) \\ &\qquad \leq \left(D_{\nu,N}^{\gamma,\tau}(g_N, h_N) - D_{\nu,N}^{\gamma,\tau}(g, h)\right) \\ &\qquad\quad + \mathcal{O}\left(\sqrt{\frac{(HKL)^2 \log(1/\delta)}{N}}\right) \end{aligned} \tag{37}$$

for any $g, h \in \mathcal{G}$. In particular, it's true for $g = g_N$ and $h = h_N$, which yields the statement. □

D Gradient flow approximates exact diffusion

In Figure 1, the diffusion is an Ornstein-Uhlenbeck process with potential $w(x) = x^2$ and dispersion $\beta = 1$. The exact solution for the probability density is computed by Chang and Cooper's method on a grid of 400 points on the interval $[-3, 3]$. The initial condition is a mixture of two Gaussians, centered at ± 1, and each having standard deviation 1. The Wasserstein gradient flow is computed using a Gaussian kernel supported at 40 points chosen uniformly at random from $[-3, 3]$, with bandwidth $5 \cdot 10^{-2}$. The objective is approximated with $3 \cdot 10^4$ Monte Carlo samples. We use an entropic regularizer for the Wasserstein distance, with $\gamma = 10^{-2}$, and set timestep $\tau = 1 \cdot 10^{-2}$. The figure shows the density at times $t = 0.05, 0.2, 0.5$.

E Accuracy in High Dimensions: Ornstein-Uhlenbeck Process

The process we are approximating in Figure 2a is an Ornstein-Uhlenbeck process, having potential $w(\mathbf{x}) = (\mathbf{x} - \mathbf{b})\mathbf{A}(\mathbf{x} - \mathbf{b})$, with $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ chosen randomly: $\mathbf{A}$ is a diagonal matrix with diagonal elements gamma distributed with shape 2 and scale 0.5, while $\mathbf{b}$ has independent normally distributed elements with standard deviation 0.5. The process has dispersion $\beta = 1$, and initial density a delta function at 0. Its density is computed exactly, in closed form, at time $\Delta t = 1$.

Our baseline is a particle simulation. For each particle, we forward simulate from time $t = 0$, using the Euler-Maruyama method with timestep 10^{-3}. We use $N = 1000, 10000$ particles.

For the Wasserstein gradient flow, we approximate the objective using $2 \cdot 10^4$ Monte Carlo samples. We use a polynomial kernel of degree three, and an L^2 regularizer for the Wasserstein distance, with $\gamma = 10^{-6}$. We set timestep $\tau = 0.2$.

To evaluate the accuracy, we estimate the symmetric KL divergence between the estimated and exact densities by Monte Carlo, sampling $4 \cdot 10^4$ points randomly from the exact solution distribution at $t = 1$. For both estimation methods, we care about the accuracy up to normalization of the estimated distribution. Before computing the divergence, we choose the normalization constant that minimizes the sum of squared errors between the estimated and exact distribution.

We repeat the experiment 20 times, for 20 different random potentials, with Figure 2a showing the median and 95% interval for each method.

F Nonlinear filtering

F.1 Problem Setup and Data Generation

Latent state trajectories in $\mathbb{R}$ are generated from the SDE model

$$d\mathbf{x}_t = -\left(2\cos(2\pi\mathbf{x}_t) + \frac{1}{2}\mathbf{x}_t\right)dt + d\mathbf{W}_t$$

which is an advection-diffusion with potential $w(\mathbf{x}) = \frac{1}{\pi}\sin(2\pi\mathbf{x}) + \frac{1}{4}\mathbf{x}_t^2$ and inverse dispersion coefficient $\beta = 1$. The latent system is observed at a time interval of $\Delta t = 1$, with additive Gaussian noise having standard deviation $\sigma = 1$. State trajectories are generated by simulating the SDE using an Euler-Maruyama method with timestep 10^{-3}, starting from $\mathbf{x}_0 = 0$.

F.2 Baselines

Discretized numerical integration. We construct a regularly-spaced grid of 1000 points on the interval $[-4, 4]$, and use Chang and Cooper's method (Chang and Cooper, 1970) to integrate the Fokker-Planck equation for the dynamics. We use a timestep of 10^{-3} for the integration.

When filtering, we obtain the posterior state distribution by first propagating forward the posterior at the previous observation time, via integrating the Fokker-Planck equation, then multiplying the resulting distribution pointwise by the observation likelihood and normalizing to sum to one.

Extended Kalman filter. The extended Kalman filter is implemented as described in (Brown and Hwang, 1997). We use Scipy's `odeint` to integrate the ODE for the mean and covariance. The EKF is initialized with a Gaussian of whose mean is drawn from a normal distribution having mean 0 and standard deviation 0.1, and whose variance is 10^{-4}.

Unscented Kalman filter. The unscented Kalman filter is implemented as described in (Sarkka, 2007). We use Scipy's `odeint` to integrate the ODE for the mean and covariance. The UKF is initialized with a Gaussian of mean 0 and variance 10^{-4}. We use parameters $\alpha = \frac{1}{2}$, $\beta = 2$, $\kappa = 1$. (β here refers to the parameter in (Sarkka, 2007), rather than the inverse dispersion coefficient in the main text.)

Gaussian sum filter. We implement a Gaussian sum filter as described in (Alspach and Sorenson, 1972). The filter is initialized with a mixture of eight Gaussians,

having means drawn independently from a normal distribution with mean 0 and standard deviation 1, and each having variance 10^{-4}.

Bootstrap particle filter. The bootstrap particle filter is implemented as described in (Gordon et al., 1993). For propagating particles forward in time, we simulate the system dynamics using an Euler-Maruyama method with timestep 10^{-3}. We resample trajectories after each observation. To extrapolate the posterior to new points, we use Gaussian kernel density estimation on sampled support points with bandwidth chosen by Scott's rule.

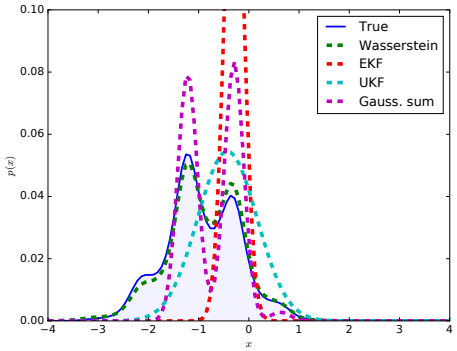
F.3 Example Posterior Evolution

Figure 3 shows an example of the evolution of the posterior distribution for consecutive timesteps. We simulate system trajectories and observations as described above and use the stochastic program for the Wasserstein gradient flow (Section 4) to propagate the posterior at one observation time to the next. The resulting distribution is multiplied pointwise by the likelihood to obtain an unnormalized posterior. The sampling distribution for the stochastic program is uniform on the interval $[-4, 4]$. We use an L2 regularizer with $\gamma = 10^{-6}$, a Gaussian kernel with bandwidth 0.1, and 10^4 samples for approximating the stochastic program objective. We solve the stochastic program using L-BFGS (from `scipy.optimize`), stopping when the norm of the gradient is less than 10^{-8}.

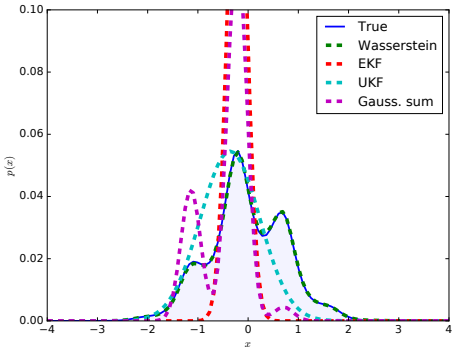
We additionally overlay posterior distributions for the baseline algorithms. The distribution obtained from discretized numerical integration is shaded in blue. For visualization, all distributions are sampled on a grid and normalized to sum to one.

F.4 Quantitative Comparison of Methods

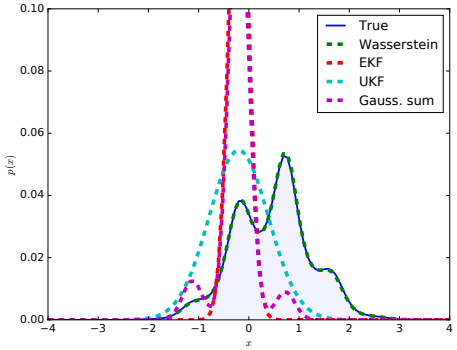
We simulate 100 independent latent state trajectories and their observations. For each we obtain posterior distributions for the proposed Wasserstein gradient flow approximation and the baseline methods, as described above. We sample the resulting distributions on the same grid as was used for discretized numerical integration and normalize to sum to one. We compute the symmetric KL-divergence between the exact distribution from discretized numerical integration and the approximate distribution from the given method.



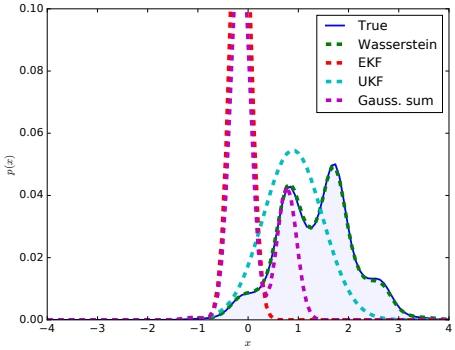
(a) $t = 1$.



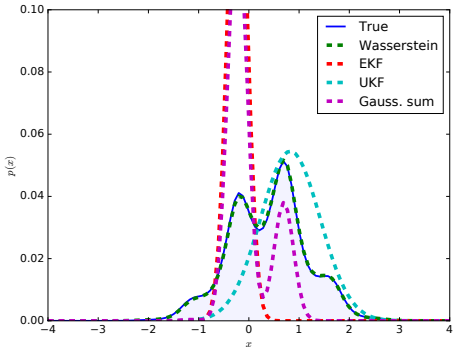
(b) $t = 2$.



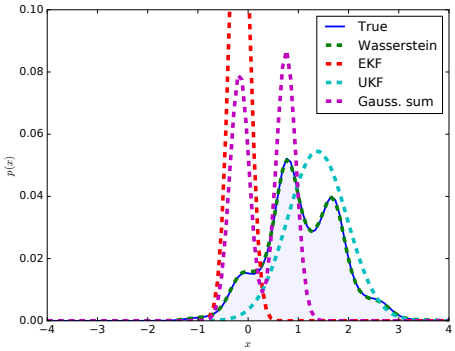
(c) $t = 3$.



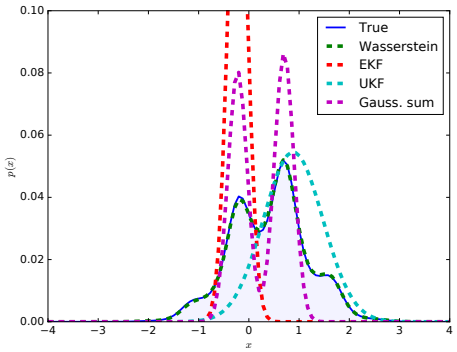
(d) $t = 4$.



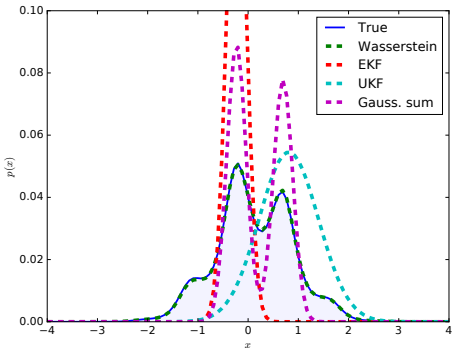
(e) $t = 5$.



(f) $t = 6$.



(g) $t = 7$.



(h) $t = 8$.

Figure 3: Filtering in a sine potential with noisy observations ($\sigma = 1$). Evolution of the posterior density, with estimates from the various methods overlaid. Shaded region is the exact solution.