Supplementary Material for
Scalable Gaussian Process Inference with Finite-data
Mean and Variance Guarantees

A Experiments

Table 1: Datasets used for experiments. All datasets from the UCI Machine Learning Repository\textsuperscript{a} except for synthetic and delays10k datasets.

\begin{tabular}{|l|l|l|l|l|l|l|l|l|}
\hline
Name & $N_{\text{train}}$ & $N_{\text{test}}$ & $d$ & $K$ & Name & $N_{\text{train}}$ & $N_{\text{test}}$ & $d$ & $K$ \\
\hline
synthetic & 1000 & 1000 & 1 & 100 & abalone & 3177 & 1000 & 8 & 300 \\
delays10k\textsuperscript{b} & 8000 & 2000 & 8 & 800 & airfoil & 1103 & 400 & 5 & 100 \\
CCPP & 7568 & 2000 & 4 & 700 & wine quality & 3898 & 1000 & 11 & 300 \\
\hline
\end{tabular}

\textsuperscript{a} \url{http://archive.ics.uci.edu/ml/index.php}
\textsuperscript{b} Hensman et al. [5]

Figure A.1: KL divergences of the approximate posteriors and root mean squared error of the approximate posteriors for the VFE and pF-DTC trials with the smallest objective values.
Figure A.2: KL divergences of the approximate posteriors and root mean squared error of the approximate posteriors for the VFE and pF-DTC trials with the smallest objective values.
B Proof of Proposition 3.1

Choose the means and variances of \( \eta \) and \( \hat{\eta} \) such that \((\hat{\mu} - \mu)^2 = \hat{s}^2 \{ \exp(2\delta) - 1 \} \) and \( s^2 = \exp(2\delta) \hat{s}^2 \). We then have that

\[
\text{KL}(\hat{\eta}||\eta) = 0.5 \{ \hat{s}^2 / s^2 - 1 + \log(s^2 / \hat{s}^2) + (\hat{\mu} - \mu)^2 / s^2 \}
\]

Eqs. (C.2) and (C.3) together show that

\[
\text{C} = 0.5[k(0, \cdot)](1 + \sigma^2)^{-1} |t - \hat{t}|
\]

while their posterior means are, respectively, \( \mu(x) = (1 + \sigma^2)^{-1} e^{-x^2/2t} \) and \( \hat{\mu}(x) = (1 + \sigma^2)^{-1} e^{-x^2/2\hat{t}} \) Define the induced kernel \( k'(x, x') := \langle k_x, k_{x'} \rangle \). Since their covariance operators are equal, the 2-Wasserstein distance between the \( \eta \) and \( \hat{\eta} \) is [2, Thm. 3.5]

\[
\mathcal{W}_2(\eta, \hat{\eta}) = \| \mu - \hat{\mu} \| = \| k(0, \cdot) \| (1 + \sigma^2)^{-1} |t - \hat{t}|
\]

The log-likelihoods associated with \( \eta \) and \( \hat{\eta} \) are, respectively, \( \mathcal{L}(f) := -\frac{1}{2\sigma^2}(f(0) - t)^2 \) and \( \hat{\mathcal{L}}(f) := -\frac{1}{2\hat{\sigma}^2}(f(0) - \hat{t})^2 \). Using Lemma F.3, in the non-preconditioned case we have

\[
d_{F,\nu}(\eta||\hat{\eta}) = \mathbb{E}_{f,\nu}[\langle DL, DL \rangle + \langle D\hat{\mathcal{L}}, D\hat{\mathcal{L}} \rangle - 2\langle DL, D\hat{\mathcal{L}} \rangle] = \sigma^{-4}r(0, 0)(t - \hat{\mu}(0))^2 + (\hat{t} - \hat{\mu}(0))^2 - 2(t - \hat{\mu}(0))(\hat{t} - \hat{\mu}(0)) \]

We take \( \mathbb{H} \) to be the reproducing kernel Hilbert space with reproducing kernel \( r \). The posterior covariance functions for \( \eta \) and \( \hat{\eta} \) are equal to

\[
k_D(x, x') = e^{-|x-x'|^2/2} - (1 + \sigma^2)^{-1} e^{-|x-x'|^2/2 - (x')^2/2} \tag{C.1}
\]

C Details of Example 4.1

We take \( \mathbb{H} \) to be the reproducing kernel Hilbert space with reproducing kernel \( r \). The posterior covariance functions for \( \eta \) and \( \hat{\eta} \) are equal to

\[
k_D(x, x') = e^{-|x-x'|^2/2} - (1 + \sigma^2)^{-1} e^{-|x-x'|^2/2 - (x')^2/2} \tag{C.1}
\]

while their posterior means are, respectively, \( \mu(x) = (1 + \sigma^2)^{-1} e^{-x^2/2t} \) and \( \hat{\mu}(x) = (1 + \sigma^2)^{-1} e^{-x^2/2\hat{t}} \) Define the induced kernel \( k'(x, x') := \langle k_x, k_{x'} \rangle \). Since their covariance operators are equal, the 2-Wasserstein distance between the \( \eta \) and \( \hat{\eta} \) is [2, Thm. 3.5]

\[
\mathcal{W}_2(\eta, \hat{\eta}) = \| \mu - \hat{\mu} \| = \| k(0, \cdot) \| (1 + \sigma^2)^{-1} |t - \hat{t}|
\]

The log-likelihoods associated with \( \eta \) and \( \hat{\eta} \) are, respectively, \( \mathcal{L}(f) := -\frac{1}{2\sigma^2}(f(0) - t)^2 \) and \( \hat{\mathcal{L}}(f) := -\frac{1}{2\hat{\sigma}^2}(f(0) - \hat{t})^2 \). Using Lemma F.3, in the non-preconditioned case we have

\[
d_{F,\nu}(\eta||\hat{\eta}) = \mathbb{E}_{f,\nu}[\langle DL, DL \rangle + \langle D\hat{\mathcal{L}}, D\hat{\mathcal{L}} \rangle - 2\langle DL, D\hat{\mathcal{L}} \rangle] = \sigma^{-4}r(0, 0)(t - \hat{\mu}(0))^2 + (\hat{t} - \hat{\mu}(0))^2 - 2(t - \hat{\mu}(0))(\hat{t} - \hat{\mu}(0)) \]

Eqs. (C.2) and (C.3) together show that \( c = \sqrt{r(0, 0)/k'(0, 0)} \).

The preconditioned case is almost identical to Eq. (C.3). Using Lemmas F.1 and F.4 and Eq. (C.1), for any \( f \in \mathbb{H} \),

\[
C_{\eta}DL(f) = -(1 + \sigma^2)^{-1}(f(0) - t)k(0, \cdot)
\]

and similarly for \( C_{\hat{\eta}}DL(f) \). Hence,

\[
d_{F,\nu}(\eta||\hat{\eta}) = \mathbb{E}_{f,\nu}[(C_{\eta}DL, C_{\eta}DL) + (C_{\hat{\eta}}DL, C_{\hat{\eta}}DL) - 2(C_{\eta}DL, C_{\hat{\eta}}DL)] = (1 + \sigma^2)^{-2}k'(0, 0)(t - \hat{\mu}(0))^2 + (\hat{t} - \hat{\mu}(0))^2 - 2(t - \hat{\mu}(0))(\hat{t} - \hat{\mu}(0)) \]

Eqs. (C.2) and (C.4) together show that \( d_{F,\nu}(\eta||\hat{\eta}) = \mathcal{W}_2(\eta, \hat{\eta}) \).

D Proof of Theorem 4.3

Theorem 4.3 will follow almost immediately after we develop a number of preliminary results. For more details on infinite-dimensional SDEs and related ideas, we recommend Hairer et al. [3, 4] and Da Prato and Zabczyk [1].

The notation in this section differs slightly from the rest of the paper in order to follow the conventions of the stochastic processes literature. Let \( W \) denote a \( C \)-Wiener process [1, Definition 4.2], where \( C : \mathbb{H} \rightarrow \mathbb{H} \) is the linear, self-adjoint, positive semi-definite, trace-class operator. Let \( \mu \in \mathbb{H} \) and let
Then we will need the constructions from the following lemma, the proof of which is deferred to Appendix D.1.

**Lemma D.1.** Let \( \mathbb{H} := \mathbb{H} \oplus \mathbb{H} \), the direct sum of \( \mathbb{H} \) with itself, for which the inner product is given by
\[
\langle (x_1, x_2), (y_1, y_2) \rangle_{\mathbb{H}} = \langle x_1, y_1 \rangle + \langle x_2, y_2 \rangle.
\]
Define the self-adjoint operator \( \tilde{\mathcal{H}} : \mathbb{H} \to \mathbb{H} \) given by \( (x, y) \mapsto (\mathcal{C}(x+y), \mathcal{C}(x+y)) \). Then Eqs. (D.1) and (D.2) can be written on a common probability space as
\[
d(X_t, Y_t) = (\mu, \nu) dt - (X_t, Y_t) dt + \langle b(X_t), \tilde{b}(Y_t) \rangle dt + \sqrt{2} d(W_t, W_t)
\]
or
\[
(X_t, Y_t) = \int_0^t (\mu, \nu) ds - \int_0^t (X_s, Y_s) ds + \int_0^t \langle b(X_s), \tilde{b}(Y_s) \rangle ds + \sqrt{2} \int_0^t d(W_t, W_t),
\]
where \( t \mapsto (X_t, Y_t) \) is a process on \( \mathbb{H} \) and \( t \mapsto (W_t, W_t) \) is a \( \tilde{\mathcal{C}} \)-Wiener process on \( \mathbb{H} \).

Let \( \mathcal{P} \) denote the space of Borel measures on \( \mathbb{H} \). Recall that for any \( \eta \in \mathcal{P} \), the \( \|\cdot\|_{\eta} \)-norm acting on functions \( A : \mathbb{H} \to \mathbb{H} \) is defined by
\[
\|A\|_{\eta} := \left( \int \|A(x)\|^2 \eta(dx) \right)^{1/2}.
\]

**Theorem D.2.** Assume that Eq. (D.3) has a unique stationary law with the marginal stationary laws of Eqs. (D.1) and (D.2) given by \( \tilde{\eta} \) and \( \eta \) respectively. Suppose that for \( X \sim \tilde{\eta} \) and \( Y \sim \eta \), \( \mathbb{E} \|X\|^2 < \infty \) and \( \mathbb{E} \|Y\|^2 < \infty \). Suppose that for some \( \alpha > 0 \), \( b \) satisfies the one-sided Lipschitz condition
\[
(b(x) - b(y), x - y) \leq (-\alpha + 1) \|x - y\|^2 \text{ for all } x, y \in \mathbb{H}.
\]
Then
\[
\mathcal{W}_2(\eta, \tilde{\eta}) \leq \alpha^{-1} \|b - \tilde{b}\|_{\eta}.
\]
We defer the proof to Appendix D.2.

**Proposition D.3.** If the hypotheses of Theorem D.2 hold, then for any distribution \( \nu \in \mathcal{P} \) such that \( \nu \ll \eta \),
\[
\mathcal{W}_2(\eta, \tilde{\eta}) \leq \alpha^{-1} \left\| \frac{d\eta}{d\nu} \right\|_\infty^{1/2} \|b - \tilde{b}\|_{\nu}
\]
**Proof.** Using Hölder’s inequality, we have
\[
\|b - \tilde{b}\|^2_{\eta} = \int \|b(x) - \tilde{b}(x)\|^2 \eta(dx)
\]
\[
= \left\| \frac{d\eta}{d\nu} \right\|_\infty \int \|b(x) - \tilde{b}(x)\|^2 \nu(dx)
\]
\[
\leq \left\| \frac{d\eta}{d\nu} \right\|_\infty \int \|b(x) - \tilde{b}(x)\|^2 \nu(dx).
\]
Eq. (D.5) follows by plugging the previous display into Eq. (D.4). \( \square \)

**Proposition D.4.** If \( \eta, \nu \in \mathcal{P} \), \( \mathcal{W}_2(\eta, \nu) \leq \varepsilon \) and \( \mathbb{H} = \mathbb{H}_{\nu} \), then for all \( x \in X \),
\[
|\mu_\eta(x) - \mu_\nu(x)| \leq \varepsilon r(x, x)^{1/2}
\]
\[
|k_\eta(x, x)^{1/2} - k_\nu(x, x)^{1/2}| \leq \sqrt{6} \varepsilon r(x, x)^{1/2}
\]
\[
|k_\eta(x, x) - k_\nu(x, x)| \leq 3 r(x, x)^{1/2} \min(k_\eta(x, x), k_\nu(x, x))^{1/2} \varepsilon + 6 r(x, x) \varepsilon^2.
\]
We defer the proof to Appendix D.3.

The result will follow by taking $C = C_0$. With this choice of $C$, $b = 0$ and $\mu = \mu_0$, so $b$ satisfies the one-sided Lipschitz condition with $\alpha = 1$. The remaining hypotheses of Theorem D.2 hold by construction, so Theorem 4.3 follows by applying Propositions D.3 and D.4.

D.1 Proof of Lemma D.1

We first check that the process $t \mapsto (W_t, W_t)$ satisfies the definition of a Wiener process. It starts from 0, has continuous trajectories and independent increments. Furthermore, 

$$\mathcal{L} ((W_t, W_t) - (W_s, W_s)) = N(0, (t-s)\tilde{C}).$$

To see that for $t \geq s$ the variance of $(W_t, W_t) - (W_s, W_s)$ in $\mathbb{H}$ is indeed equal to $(t-s)\tilde{C}$, note that, for any $(x_1, x_2), (y_1, y_2) \in \mathbb{H}$,

$$\mathbb{E} [\langle(x_1, x_2), (W_t, W_t) - (W_s, W_s)\rangle] \mathbb{E} [\langle(y_1, y_2), (W_t, W_t) - (W_s, W_s)\rangle]$$

$$= \mathbb{E} [\langle(x_1, W_t - W_s) + (x_2, W_t - W_s)\rangle] \mathbb{E} [\langle(y_1, W_t - W_s) + (y_2, W_t - W_s)\rangle]$$

$$= \langle (t-s)C_{x_1, y_1} + ((t-s)C_{x_2, y_1} + ((t-s)C_{x_1, y_2}, y_1) + ((t-s)C_{x_2, y_2}, y_1)$$

$$= \langle (t-s)\tilde{C}(x_1 + x_2, y_1) + ((t-s)\tilde{C}(x_1 + x_2, y_2)$$

Given that $\tilde{C}$ is self-adjoint, it follows that $\tilde{C}$ is self-adjoint as well:

$$\langle \tilde{C}(x_1, x_2), (y_1, y_2)\rangle = \langle C(x_1 + x_2, y_1 + y_2)$$

$$= \langle x_1 + x_2, C(y_1 + y_2)$$

$$= \langle (x_1, x_2), \tilde{C}(y_1, y_2)\rangle.$$

D.2 Proof of Theorem D.2

We begin by quoting the Itô formula we will be using (see Da Prato and Zabczyk [1] for complete details):

**Theorem D.5 (Itô formula, Da Prato and Zabczyk [1, Theorem 4.32]).** Let $H$ and $U$ be two Hilbert spaces and $W$ be a Q-Wiener process for a symmetric non-negative operator $Q \in L(U)$. Let $U_0 = Q^{1/2}(U)$ and let $L_2(U_0, H)$ be the space of all Hilbert-Schmidt operators from $U_0$ to $H$. Assume that $\Phi$ is an $L_2(U_0, H)$-valued process stochastically integrable in $[0, T], \phi$ is an $H$-valued predictable process Bochner integrable on $[0, T]$ almost surely, and $X(0)$ a $H$-valued random variable. Then the following process:

$$X_t = X_0 + \int_0^t \varphi(s) ds + \int_0^t \Phi(s) dW_s, \quad t \in [0, T]$$

is well defined. Assume that a function $F : [0, T] \times H \rightarrow \mathbb{R}$ and its partial derivatives $F_t, F_x, F_{xx}$ are uniformly continuous on bounded subsets of $[0, T] \times H$. Under these conditions, almost surely, for all $t \in [0, T]$:

$$F(t, X_t) = F(0, X_0) + \int_0^t \langle F_x(s, X_s), \Phi(s) dW_t \rangle + \int_0^t F_t(s, X_s) ds$$

$$+ \int_0^t \langle F_x(s, X_s), \varphi(s) \rangle ds$$

$$+ \int_0^t \frac{1}{2} \text{Tr} \left[ F_{xx}(s, X_s)(\Phi(s)Q^{1/2})(\Phi(s)Q^{1/2})^* \right] ds.$$

Let $F : [0, \infty) \times \mathbb{H} \rightarrow \mathbb{R}$ be given by $F(t; x, y) = e^{2\alpha t} \|x - y\|^2$. Then the Fréchet derivative of $F$ with respect to the space parameters is given by

$$F_{(x,y)}(t; x, y)(h_1, h_2) = 2e^{2\alpha t}(x - y, h_1 - h_2). \quad (D.6)$$
Eq. (D.6) holds because
\[
\frac{\|x + h_1 - y - h_2\|^2 - \|x - y\|^2 - 2\langle x - y, h_1 - h_2 \rangle}{\sqrt{\|h_1\|^2 + \|h_2\|^2}} = \frac{\|h_1 - h_2\|^2}{\sqrt{\|h_1\|^2 + \|h_2\|^2}}.
\]
Furthermore, the second Fréchet derivative with respect to the space parameters is
\[
\mathbb{E} \left[ \frac{\|x + h_1 - y - h_2\|^2 - \|x - y\|^2 - 2\langle x - y, h_1 - h_2 \rangle}{\sqrt{\|h_1\|^2 + \|h_2\|^2}} \right] \leq 2\sqrt{\|h_1\|^2 + \|h_2\|^2} \|h_1\| \|h_2\| \to 0.
\]
Note that \( \tilde{C}^{1/2}(x, y) = \frac{\sqrt{2}}{2} (C^{1/2}(x + y), C^{1/2}(x + y)) \). Using the one-sided Lipschitz condition and the Cauchy-Schwarz inequality, we obtain
\[
\langle b(X_t) - \tilde{b}(Y_t), X_t - Y_t \rangle = \langle b(X_t) - b(Y_t), X_t - Y_t \rangle + \langle b(Y_t) - \tilde{b}(Y_t), X_t - Y_t \rangle \leq (-\alpha + 1) \|X_t - Y_t\|^2 + \|b(Y_t) - \tilde{b}(Y_t)\| \|X_t - Y_t\|.
\]
(D.7)
We will assume that we start the process \( (X_t, Y_t) \) at joint stationarity (with \( X_0 \sim \eta \) and \( Y_0 \sim \nu \). By the Itô formula given by Theorem D.5, applied to the process described by Eq. (D.3) and function \( F(\cdot) \)

\[
|\mathbb{E}(x,y)\rangle \langle (1 + \beta(x+y), C(x+y) - C(x+y))| ds
\]

Taking expectations on both sides (with respect to everything that is random and at the fixed time \( t \), multiplying by \( e^{-2\alpha t} \) and applying Eq. (D.7)

\[
\mathbb{E} \|X_t - Y_t\|^2 \leq e^{-2\alpha t} \mathbb{E} \|X_0 - Y_0\|^2 + e^{-2\alpha t} \mathbb{E} \left[ \int_0^t 2e^{2\alpha(s-t)} \|b(Y_s) - \tilde{b}(Y_s)\| \|X_s - Y_s\| ds \right]
\]

\[
\leq e^{-2\alpha t} \mathbb{E} \|X_0 - Y_0\|^2 + \left( \int_0^t 2e^{2\alpha(s-t)} \mathbb{E} \|b(Y_s) - \tilde{b}(Y_s)\|^2 \right)^{1/2} \left( \int_0^t 2e^{2\alpha(s-t)} \mathbb{E} \|X_s - Y_s\|^2 \right)^{1/2}
\]

\[
= e^{-2\alpha t} \mathbb{E} \|X_0 - Y_0\|^2 + \left( \int_0^t 2e^{2\alpha(s-t)} \mathbb{E} \|b(Y_s) - \tilde{b}(Y_s)\|^2 \right)^{1/2} \left( \mathbb{E} \|X_t - Y_t\|^2 \right)^{1/2}
\]

\[
= e^{-2\alpha t} \mathbb{E} \|X_0 - Y_0\|^2 + \left( \alpha^{-1/2}(1 - e^{-2\alpha t})^{1/2} \|b - \tilde{b}\| \mathbb{E} \|X_t - Y_t\|^2 \right)^{1/2}.
\]

(D.8)

(D.9)
where Eq. (D.8) follows by the Cauchy-Schwarz inequality and Eq. (D.9) follows from the assumption that we start the process \( t \mapsto (X_t, Y_t) \) at joint stationarity.

Now, dividing by \( (E\|X_t - Y_t\|^2)^{1/2} \), taking \( t \to \infty \) and noting that the process \( t \mapsto (X_t, Y_t) \) remains at joint stationarity, we obtain the result.

### D.3 Proof of Proposition D.4

Let \( f \sim \eta \) and \( g \sim \nu \) and define \( \tilde{k}_v(x, x') := E[g(x)g(x')] \). By Cauchy-Schwarz and Jensen’s inequalities,

\[
|\mu_\eta(x) - \mu_\nu(x)| = |E[f(x) - g(x)]| = |E[(f - g, r_x)]|
\]

\[
\leq \|f - g\|_r \leq r(x, x)^{1/2}E[\|f - g\|^2]^{1/2}
\]

\[
\leq r(x, x)^{1/2}\varepsilon.
\]

Without loss of generality we can assume \( \mu_\eta = 0 \), since if not then we consider the random variables \( \bar{f} := f - \mu_\eta \) and \( \bar{g} := g - \mu_\eta \) instead. It follows from the Cauchy-Schwarz inequality that

\[
|k_\eta(x, x) - \bar{k}_v(x, x)| = E[f(x)^2 - g(x)^2]
\]

\[
= E[(f(x) - g(x))(f(x) + g(x))]
\]

\[
\leq \sqrt{E[(f(x) - g(x))^2]} \sqrt{E[(f(x) + g(x))^2]}
\]

\[
\leq r(x, x)^{1/2}\varepsilon \sqrt{2E[f(x)^2] + E[g(x)^2]}
\]

\[
\leq \sqrt{2}r(x, x)^{1/2}\varepsilon (k_\eta(x, x)^{1/2} + \bar{k}_v(x, x)^{1/2})
\]

\[
|k_\eta(x, x)^{1/2} - \bar{k}_v(x, x)^{1/2}| \leq \sqrt{2}r(x, x)^{1/2}\varepsilon.
\]

Also,

\[
\bar{k}_v(x, x)^{1/2} \leq \sqrt{k_v(x, x) + \mu_\nu(x)^2} \leq k_v(x, x)^{1/2} + r(x, x)^{1/2}\varepsilon.
\]

We now have that

\[
|k_\eta(x, x) - k_v(x, x)| = |k_\eta(x, x) - \bar{k}_v(x, x) + \mu_\nu(x)^2|
\]

\[
\leq |k_\eta(x, x) - \bar{k}_v(x, x)| + \mu_\nu(x)^2
\]

\[
\leq \sqrt{2}r(x, x)^{1/2}\varepsilon (k_\eta(x, x)^{1/2} + \bar{k}_v(x, x)^{1/2}) + r(x, x)\varepsilon^2
\]

\[
\leq \sqrt{2}r(x, x)^{1/2}\varepsilon (k_\eta(x, x)^{1/2} + k_v(x, x)^{1/2}) + (1 + \sqrt{2})r(x, x)\varepsilon^2
\]

\[
|k_\eta(x, x)^{1/2} - k_v(x, x)^{1/2}| \leq \sqrt{2}r(x, x)^{1/2}\varepsilon + \frac{(1 + \sqrt{2})r(x, x)\varepsilon^2}{k_\eta(x, x)^{1/2} + k_v(x, x)^{1/2}}.
\]

Let \( a := \frac{1 + \sqrt{3 + 2\sqrt{2}}}{\sqrt{2}}r(x, x)^{1/2} \). If \( \max(k_\eta(x, x)^{1/2}, k_v(x, x)^{1/2}) \leq a \), then clearly

\[
|k_\eta(x, x)^{1/2} - k_v(x, x)^{1/2}| \leq \sqrt{2}r(x, x)^{1/2}\varepsilon + \frac{(1 + \sqrt{2})r(x, x)\varepsilon^2}{a} = a.
\]

Hence we conclude unconditionally that

\[
|k_\eta(x, x)^{1/2} - k_v(x, x)^{1/2}| \leq \frac{1 + \sqrt{3 + 2\sqrt{2}}}{\sqrt{2}}r(x, x)^{1/2}\varepsilon < \sqrt{6}r(x, x)^{1/2}\varepsilon.
\]

Thus, we also have that

\[
|k_\eta(x, x) - k_v(x, x)| \leq \sqrt{2}r(x, x)^{1/2}\varepsilon (k_\eta(x, x)^{1/2} + k_v(x, x)^{1/2}) + (1 + \sqrt{2})r(x, x)\varepsilon^2
\]

\[
< \sqrt{2}r(x, x)^{1/2}\varepsilon (2k_\eta(x, x)^{1/2} + \sqrt{6}r(x, x)^{1/2}\varepsilon) + (1 + \sqrt{2})r(x, x)\varepsilon^2
\]

\[
= 2\sqrt{2}r(x, x)^{1/2}\varepsilon k_\eta(x, x)^{1/2} + (1 + \sqrt{2} + \sqrt{12})r(x, x)\varepsilon^2
\]

\[
< 3r(x, x)^{1/2}k_\eta(x, x)^{1/2} + 6r(x, x)\varepsilon^2.
\]

The final inequality follows from Jensen’s inequality (which implies that the 1-Wasserstein distance lower bound the 2-Wasserstein distance) and [7, Rmk. 6.5].
E  Proof of Proposition 4.5

We first write $k$ in terms of the orthonormal basis of $\mathbb{H}_k$:

$$k(x, x') = \sum_{j \geq 1} e_j(x)e_j(x').$$

Define

$$r(x, x') := \sum_{j \geq 1} \lambda_j e_j(x)e_j(x').$$

If $\sum_{j \geq 1} \lambda_j^{-1} < \infty$ then $r$ dominates $k$. So given inputs $X = (x_n)_{n=1}^N$, and defining $a_{nm,j} := e_j(x_n)e_j(x_m)$, to show the existence of the required kernel $r$ we need to show there exists a solution to

$$\forall (n, m) \in [N]^2, \quad \left| \sum_{j \geq 1} \lambda_j a_{nm,j} - \sum_{j \geq 1} a_{nm,j} \right| \leq \epsilon, \quad \sum_{j \geq 1} \lambda_j^{-1} < \infty, \quad \text{and} \quad \forall j \in \mathbb{N}, \lambda_j \geq 0.$$

By assumption on the pointwise decay of orthonormal basis elements, for all $(n, m) \in [N]^2$, $|a_{nm,j}| = o(j^{-2})$. Define $a_j := \max_{(n, m) \in [N]^2} |a_{nm,j}|$. Therefore $\sqrt{a_j} = o(j^{-1})$, $\sum_{j \geq 1} \sqrt{a_j} < \infty$, and there exists a $J > 0$ such that

$$\forall j > J, \sqrt{a_j} < 1 \quad \text{and} \quad \sum_{j \geq J} \sqrt{a_j} < \epsilon.$$

Setting $\lambda_j = 1$ for each $j \in 1, \ldots, J$ and $\lambda_j = 1 + \sqrt{a_j}^{-1}$ for $j > J$, we have that for any $(n, m) \in [N]^2$,

$$\left| \sum_{j \geq 1} \lambda_j a_{nm,j} - \sum_{j \geq 1} a_{nm,j} \right| = \left| \sum_{j \geq J} \frac{a_{nm,j}}{\sqrt{a_j}} \right| \leq \sum_{j \geq J} \sqrt{a_j} < \epsilon.$$

Finally since $\sqrt{a_j} = o(j^{-1})$, $\lambda_j = o(j)$, and so $\lambda_j^{-1} = o(j^{-1})$ yielding $\sum_{j \geq 1} \lambda_j^{-1} < \infty$.

F  Proof of Proposition 5.1

Let $L_n(f) := -\frac{1}{2\sigma^2}(f(x_n) - y_n)^2$ denote the log-likelihood of the $n$th observation and recall that $\mathbb{H} = \mathbb{H}_r$.

**Lemma F.1.** For any $f \in \mathbb{H}$,

$$D L_n(f) = -\sigma^{-2}(f(x_n) - y_n) r x_n.$$

**Proof.** For $g \in \mathbb{H}$,

$$|L_n(f + g) - L_n(f) + (\sigma^{-2}(f(x_n) - y_n)r(x_n, \cdot), g)|$$

$$= \left| -\frac{1}{2\sigma^2}(f(x_n) + g(x_n) - y_n)^2 + \frac{1}{2\sigma^2}(f(x_n) - y_n)^2 + \sigma^{-2}(f(x_n) - y_n)g(x_n) \right|$$

$$\leq \frac{1}{2\sigma^2}g(x_n)^2 = \frac{1}{2\sigma^2}(r(x_n, \cdot), g)^2 \leq \frac{r(x_n, x_n) ||g||^2}{2\sigma^2}.$$

$\Box$

**Lemma F.2.** For any $f \in \mathbb{H}$,

$$D L(f) = -\sigma^{-2}(f(X) - y) X$$

and

$$D L(f) = -\sigma^{-2}(\bar{Q} X f(X) - y) \bar{Q} X$$

**Proof.** Both results follow directly from Lemma F.1. $\Box$
Lemma F.3. If $\nu = \text{GP}(\hat{\mu}, \hat{k})$, then
\[
\mathbb{E}_{f \sim \nu}[(\mathcal{D}_n(f), \mathcal{D}_m(f))] = \sigma^{-4}r(x_n, x_m)[\hat{k}(x_n, x_m) + (y_n - \hat{\mu}(x_n))(y_m - \hat{\mu}(x_m))].
\]

Proof. Using Lemma F.1, we have
\[
\mathbb{E}_{f \sim \nu}[(\mathcal{D}_n(f), \mathcal{D}_m(f))] = \sigma^{-4}(r(x_n, x_m)\mathbb{E}_{f \sim \nu}[(f(x_n) - y_n)(f(x_m) - y_m)] = \sigma^{-4}r(x_n, x_m)[\hat{k}(x_n, x_m) + (y_n - \hat{\mu}(x_n))(y_m - \hat{\mu}(x_m))].
\]

Lemma F.4. If $\eta = \text{GP}(0, \ell)$ then $(C_\eta f)(x) = \langle f, \ell_x \rangle$.

Proof. Since $(C_\eta r_{x'}) = (r_{x'}, \ell_x) = \ell_{x'}$, for $f \sim \eta$,
\[
\langle r_{x}, C_\eta r_{x'} \rangle = \langle r_{x}, \ell_{x'} \rangle = \ell(x, x') = \text{Cov}(f(x), f(x')).
\]

Lemma F.5. For the DTC log-likelihood approximation $\hat{\pi}$,
\[
(C_\pi f)(x) = (C_\eta f)(x) - \langle f, k_\hat{X}\rangle(k^{-1}_\hat{X} - \hat{\Sigma})k_{\hat{X}x},
\]
where $\hat{\Sigma} := (k_{\hat{X}\hat{X}} + \sigma^{-2}k_{\hat{X}x}k_{x\hat{X}})^{-1}$.

Proof. Since $\hat{\pi}$ has covariance function $k(x, x') - Q_{xx'} + k_{x\hat{X}}\hat{\Sigma}k_{\hat{X}x}$ [6], the result follows from Lemma F.4.

It follows from Lemmas F.2 and F.5 that
\[
\begin{align*}
C_\pi D\hat{\mathcal{L}}(f) &= -\sigma^{-2}(\hat{Q}_{\hat{X}X}f(\hat{X}) - y)\top K_{XX}\hat{\Sigma}k_{\hat{X}} \\
C_\pi D\mathcal{L}(f) &= -\sigma^{-2}(f(X) - y)\top (K_{XX} - \hat{Q}_{XX}k_{\hat{X}} + K_{XX}\hat{\Sigma}k_{\hat{X}}).
\end{align*}
\]
We therefore have that
\[
\begin{align*}
-\sigma^2C_\pi D(\mathcal{L} - \hat{\mathcal{L}})(f) &= (f(X) - y)\top (k_{XX} - \hat{Q}_{XX}k_{\hat{X}}) + (f(X) - \hat{Q}_{XX}f(\hat{X}))\top K_{XX}\hat{\Sigma}k_{\hat{X}}.
\end{align*}
\]
Consider the limit $r \to k$, so $k' \to k$. Then
\[
\sigma^4|C_\pi D(\mathcal{L} - \hat{\mathcal{L}})(f)|^2 = (f(X) - y)\top (K_{XX} + \hat{Q}_{XX}K_{XX}Q_{XX} - 2K_{XX}\hat{Q}_{XX})(f(X) - y) + (f(X) - y)\top (K_{XX}\hat{\Sigma}K_{XX} - \hat{Q}_{XX}K_{XX}\hat{\Sigma}K_{XX})(f(X) - \hat{Q}_{XX}f(\hat{X})) + (f(X) - \hat{Q}_{XX}f(\hat{X}))\top K_{XX}\hat{\Sigma}K_{XX}f(\hat{X}) - \hat{Q}_{XX}f(\hat{X}))
\]
\[
= (f(X) - y)\top (K_{XX} - \hat{Q}_{XX})(f(X) - y) + (f(X) - \hat{Q}_{XX}f(\hat{X}))\top S_{XX}(f(X) - \hat{Q}_{XX}f(\hat{X})),
\]
where $S_{XX} := K_{XX}\hat{\Sigma}K_{XX}\hat{\Sigma}K_{XX}$. Let $E_{XX} := K_{XX} - \hat{Q}_{XX}$. Taking expectations we get
\[
\begin{align*}
\mathbb{E}_\nu[(f(X) - y)\top E_{XX}(f(X) - y)] &= \mathbb{E}_\nu[(f(X) - \hat{\mu}(X) + \hat{\mu}(X) - y)\top E_{XX}(f(X) - \hat{\mu}(X) + \hat{\mu}(X) - y)] \\
&= \text{Tr}(K_{XX}E_{XX}) + \hat{\mu}(X) - y)\top E_{XX}(\hat{\mu}(X) - y)
\end{align*}
\]
and
\[
\begin{align*}
\mathbb{E}_\nu[(f(X) - \hat{Q}_{XX}f(\hat{X}))\top S_{XX}(f(X) - \hat{Q}_{XX}f(\hat{X}))] &= \mathbb{E}_\nu[(f(X) - \hat{\mu}(X) + \hat{Q}_{XX}\hat{\mu}(X) - \hat{Q}_{XX}f(\hat{X}) + \hat{\mu}(X) - \hat{Q}_{XX}\hat{\mu}(\hat{X}))\top S_{XX}^{1/2}\\
&= \text{Tr}(K_{XX}S_{XX}) + \text{Tr}(K_{XX}\hat{Q}_{XX}S_{XX}\hat{Q}_{XX}f(\hat{X})) - 2\text{Tr}(K_{XX}S_{XX}\hat{Q}_{XX}) + (\hat{\mu}(X) - \hat{Q}_{XX}\hat{\mu}(\hat{X}))\top S_{XX}(\hat{\mu}(X) - \hat{Q}_{XX}\hat{\mu}(\hat{X})).
\end{align*}
\]
Let $S'_{XX} := K_{XX} \hat{\Sigma} K_{XX} \hat{\Sigma}$. Putting everything together, conclude that

$$
\sigma^4 \|C_{\hat{\pi}} D(L - \hat{L})\|_2^2
= \text{Tr}( (\hat{K}_{XX} + (\hat{\mu}(X) - y)(\hat{\mu}(X) - y)^\top) (K_{XX} - Q_{XX}) )
+ \text{Tr}(\hat{K}_{XX} S_{XX}) + \text{Tr}(\hat{K}_{XX} \hat{\Sigma} S_{XX} \hat{Q}_{XX} \hat{\Sigma}) - 2 \text{Tr}(\hat{K}_{XX} S_{XX} \hat{Q}_{XX} \hat{\Sigma})
+ (\hat{\mu}(X) - \hat{Q}_{XX} \hat{\mu}(\hat{X}))^\top S_{XX} (\hat{\mu}(X) - \hat{Q}_{XX} \hat{\mu}(\hat{X})).
$$

$$
= - \text{Tr}(\hat{K}_{XX} (\hat{K}_{XX} + (\hat{\mu}(X) - y)(\hat{\mu}(X) - y)^\top) \hat{Q}_{XX})
+ \text{Tr}(\hat{K}_{XX} \hat{K}_{XX} \hat{Q}_{XX} \hat{K}_{XX} \hat{Q}_{XX} \hat{Q}_{XX} \hat{K}_{XX} \hat{K}_{XX} S'_{XX})
+ (\hat{\mu}(X) - \hat{Q}_{XX} \hat{\mu}(\hat{X}))^\top S'_{XX} K_{XX} (\hat{\mu}(X) - \hat{Q}_{XX} \hat{\mu}(\hat{X})).
$$

(F.1)

It is clear from Eq. (F.1) that all quantities can be computed while never instantiating a matrix larger than $N \times M$, hence, up to the constant $C(X)$, the pF divergence can be computed in $O(N^2 M^2)$ time and $O(NM)$ space.

References


