

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^a except for synthetic and delays10k datasets.

K = number of datapoints used to construct ν (approximately 10% of N_{train})

Name	N_{train}	N_{test}	d	K	Name	N_{train}	N_{test}	d	K
synthetic	1000	1000	1	100	abalone	3177	1000	8	300
delays10k ^b	8000	2000	8	800	airfoil	1103	400	5	100
CCPP	7568	2000	4	700	wine quality	3898	1000	11	300

^a <http://archive.ics.uci.edu/ml/index.php>

^b Hensman et al. [5]

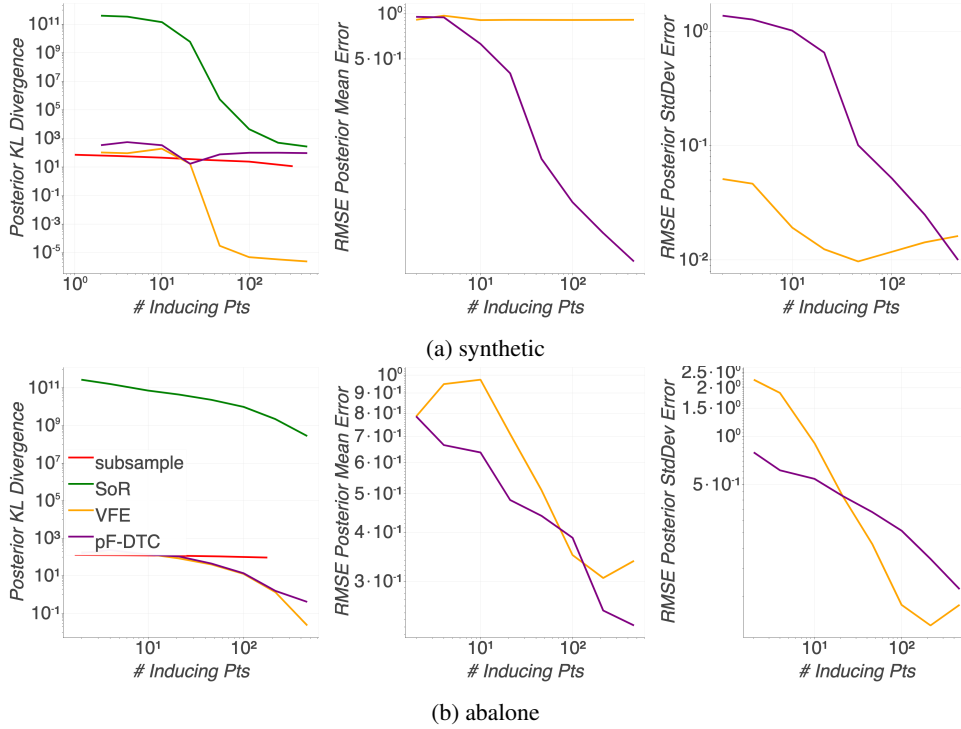


Figure A.1: KL divergences of the approximate posteriors and root mean squared error of the approximate posteriors for the VFE and pF-DTC trials with the smallest objective values.

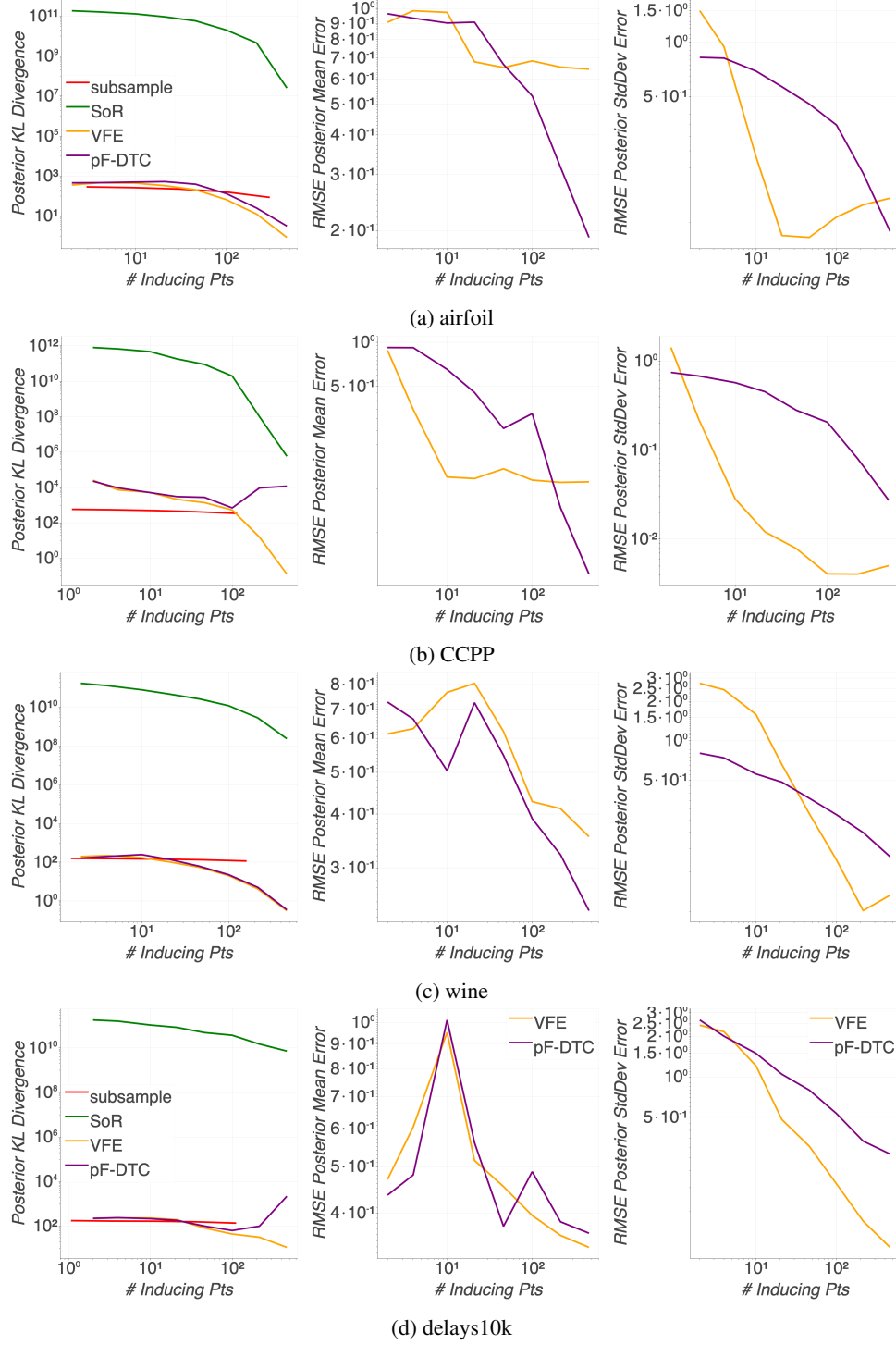


Figure A.2: KL divergences of the approximate posteriors and root mean squared error of the approximate posteriors for the VFE and pF-DTC trials with the smallest objective values.

B Proof of Proposition 3.1

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

$$\begin{aligned} \text{KL}(\tilde{\eta}||\eta) &= 0.5\{\tilde{s}^2/s^2 - 1 + \log(s^2/\tilde{s}^2) + (\tilde{\mu} - \mu)^2/s^2\} \\ &= 0.5[\tilde{s}^2/\{\exp(2\delta)\tilde{s}^2\} - 1 + \log\{\exp(2\delta)\tilde{s}^2/\tilde{s}^2\} + \tilde{s}^2\{\exp(2\delta) - 1\}/\{\exp(2\delta)\tilde{s}^2\}] \\ &= 0.5[\exp(-2\delta) - 1 + \log\{\exp(2\delta)\} + \{\exp(2\delta) - 1\}\exp(-2\delta)] \\ &= \delta. \end{aligned}$$

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 η and $\tilde{\eta}$ are equal to

$$k_{\mathcal{D}}(x, x') = e^{-(x-x')^2/2} - (1 + \sigma^2)^{-1} e^{-x^2/2 - (x')^2/2} \quad (\text{C.1})$$

while their posterior means are, respectively, $\mu(x) = (1 + \sigma^2)^{-1} e^{-x^2/2} t$ and $\tilde{\mu}(x) = (1 + \sigma^2)^{-1} e^{-x^2/2} \tilde{t}$. 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 η and $\tilde{\eta}$ is [2, Thm. 3.5]

$$\begin{aligned} \mathcal{W}_2(\eta, \tilde{\eta}) &= \|\mu - \tilde{\mu}\| = \|k(0, \cdot)\| (1 + \sigma^2)^{-1} |t - \tilde{t}| \\ &= \sqrt{k'(0, 0)} (1 + \sigma^2)^{-1} |t - \tilde{t}|. \end{aligned} \quad (\text{C.2})$$

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

$$\begin{aligned} d_{\text{F},\nu}(\eta, \tilde{\eta})^2 &= \mathbb{E}_{f \sim \nu}[\langle \mathcal{D}\mathcal{L}, \mathcal{D}\mathcal{L} \rangle + \langle \mathcal{D}\tilde{\mathcal{L}}, \mathcal{D}\tilde{\mathcal{L}} \rangle - 2\langle \mathcal{D}\mathcal{L}, \mathcal{D}\tilde{\mathcal{L}} \rangle] \\ &= \sigma^{-4} r(0, 0) [(t - \hat{\mu}(0))^2 + (\tilde{t} - \hat{\mu}(0))^2 - 2(t - \hat{\mu}(0))(\tilde{t} - \hat{\mu}(0))] \\ &= \sigma^{-4} r(0, 0) (t - \tilde{t})^2. \end{aligned} \quad (\text{C.3})$$

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 $f \in \mathbb{H}$,

$$\mathcal{C}_{\tilde{\eta}} \mathcal{D}\mathcal{L}(f) = -(1 + \sigma^2)^{-1} (f(0) - t) k(0, \cdot)$$

and similarly for $\mathcal{C}_{\tilde{\eta}} \mathcal{D}\tilde{\mathcal{L}}(f)$. Hence,

$$\begin{aligned} d_{\text{pF},\nu}(\eta||\tilde{\eta}) &= \mathbb{E}_{f \sim \nu}[\langle \mathcal{C}_{\tilde{\eta}} \mathcal{D}\mathcal{L}, \mathcal{C}_{\tilde{\eta}} \mathcal{D}\mathcal{L} \rangle + \langle \mathcal{C}_{\tilde{\eta}} \mathcal{D}\tilde{\mathcal{L}}, \mathcal{C}_{\tilde{\eta}} \mathcal{D}\tilde{\mathcal{L}} \rangle - 2\langle \mathcal{C}_{\tilde{\eta}} \mathcal{D}\mathcal{L}, \mathcal{C}_{\tilde{\eta}} \mathcal{D}\tilde{\mathcal{L}} \rangle] \\ &= (1 + \sigma^2)^{-2} k'(0, 0) [(t - \hat{\mu}(0))^2 + (\tilde{t} - \hat{\mu}(0))^2 - 2(t - \hat{\mu}(0))(\tilde{t} - \hat{\mu}(0))] \\ &= (1 + \sigma^2)^{-2} k'(0, 0) (t - \tilde{t})^2. \end{aligned} \quad (\text{C.4})$$

Eqs. (C.2) and (C.4) together show that $d_{\text{pF},\nu}(\eta||\tilde{\eta}) = \mathcal{W}_2(\eta, \tilde{\eta})$.

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 $\mathcal{C}$ -Wiener process [1, Definition 4.2], where $\mathcal{C} : \mathbb{H} \rightarrow \mathbb{H}$ is the linear, self-adjoint, positive semi-definite, trace-class operator. Let $\mu \in \mathbb{H}$ and let

$b, \tilde{b} : \mathbb{H} \rightarrow \mathbb{R}$ and consider the following infinite-dimensional stochastic differential equations (SDEs) in $\mathbb{H}$:

$$dX_t = (\mu - X_t)dt + b(X_t)dt + \sqrt{2}dW_t \quad (\text{D.1})$$

$$dY_t = (\mu - Y_t)dt + \tilde{b}(Y_t)dt + \sqrt{2}dW_t. \quad (\text{D.2})$$

We will need the constructions from the following lemma, the proof of which is deferred to Appendix D.1.

Lemma D.1. *Let $\tilde{\mathbb{H}} := \mathbb{H} \oplus \mathbb{H}$, the direct sum of $\mathbb{H}$ with itself, for which the inner product is given by*

$$\langle (x_1, x_2), (y_1, y_2) \rangle_{\tilde{\mathbb{H}}} = \langle x_1, y_1 \rangle + \langle x_2, y_2 \rangle.$$

Define the self-adjoint operator $\tilde{\mathcal{C}} : \tilde{\mathbb{H}} \rightarrow \tilde{\mathbb{H}}$ 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

$$d(X_t, Y_t) = (\mu, \mu)dt - (X_t, Y_t)dt + (b(X_t), \tilde{b}(Y_t))dt + \sqrt{2}d(W_t, W_t) \quad (\text{D.3})$$

or

$$(X_t, Y_t) = \int_0^t (\mu, \mu)ds - \int_0^t (X_s, Y_s)ds + \int_0^t (b(X_s), \tilde{b}(Y_s))ds + \sqrt{2}(W_t, W_t),$$

where $t \mapsto (X_t, Y_t)$ is a process on $\tilde{\mathbb{H}}$ and $t \mapsto (W_t, W_t)$ is a $\tilde{\mathcal{C}}$ -Wiener process on $\tilde{\mathbb{H}}$.

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 $A : \mathbb{H} \rightarrow \mathbb{H}$ 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 η 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*

$$\langle b(x) - b(y), x - y \rangle \leq (-\alpha + 1) \|x - y\|^2 \text{ for all } x, y \in \mathbb{H}.$$

Then

$$\mathcal{W}_2(\eta, \tilde{\eta}) \leq \alpha^{-1} \|b - \tilde{b}\|_\eta. \quad (\text{D.4})$$

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 $\nu \ll \eta$,*

$$\mathcal{W}_2(\eta, \tilde{\eta}) \leq \alpha^{-1} \left\| \frac{d\eta}{d\nu} \right\|_\infty^{1/2} \|b - \tilde{b}\|_\nu \quad (\text{D.5})$$

Proof. Using Hölder's inequality, we have

$$\begin{aligned} \|b - \tilde{b}\|_\eta^2 &= \int \|b(x) - \tilde{b}(x)\|^2 \eta(dx) \\ &= \int \frac{d\eta}{d\nu}(x) \|b(x) - \tilde{b}(x)\|^2 \nu(dx) \\ &\leq \left\| \frac{d\eta}{d\nu} \right\|_\infty \int \|b(x) - \tilde{b}(x)\|^2 \nu(dx). \end{aligned}$$

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 $\mathbb{H} = \mathbb{H}_r$, then for all $\mathbf{x} \in \mathcal{X}$,*

$$\begin{aligned} |\mu_\eta(\mathbf{x}) - \mu_\nu(\mathbf{x})| &\leq r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon \\ |k_\eta(\mathbf{x}, \mathbf{x})^{1/2} - k_\nu(\mathbf{x}, \mathbf{x})^{1/2}| &\leq \sqrt{6} r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon \\ |k_\eta(\mathbf{x}, \mathbf{x}) - k_\nu(\mathbf{x}, \mathbf{x})| &\leq 3 r(\mathbf{x}, \mathbf{x})^{1/2} \min(k_\eta(\mathbf{x}, \mathbf{x}), k_\nu(\mathbf{x}, \mathbf{x}))^{1/2} \varepsilon + 6 r(\mathbf{x}, \mathbf{x}) \varepsilon^2. \end{aligned}$$

We defer the proof to Appendix D.3.

The result will follow by taking $\mathcal{C} = \mathcal{C}_{\tilde{\eta}}$. With this choice of $\mathcal{C}$, $b = 0$ and $\mu = \mu_{\tilde{\eta}}$, 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,

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

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

$$\begin{aligned} & \mathbb{E} [\langle (x_1, x_2), (W_t, W_t) - (W_s, W_s) \rangle_{\tilde{\mathbb{H}}} \langle (y_1, y_2), (W_t, W_t) - (W_s, W_s) \rangle_{\tilde{\mathbb{H}}}] \\ &= \mathbb{E} [\langle (x_1, W_t - W_s) + (x_2, W_t - W_s) \rangle_{\tilde{\mathbb{H}}} \langle (y_1, W_t - W_s) + (y_2, W_t - W_s) \rangle_{\tilde{\mathbb{H}}}] \\ &= \langle (t-s)\mathcal{C}x_1, y_1 \rangle + \langle (t-s)\mathcal{C}x_1, y_2 \rangle + \langle (t-s)\mathcal{C}x_2, y_1 \rangle + \langle (t-s)\mathcal{C}x_2, y_2 \rangle \\ &= \langle (t-s)\mathcal{C}(x_1 + x_2), y_1 \rangle + \langle (t-s)\mathcal{C}(x_1 + x_2), y_2 \rangle \\ &= \langle (t-s)\tilde{\mathcal{C}}(x_1, x_2), (y_1, y_2) \rangle_{\tilde{\mathbb{H}}}. \end{aligned}$$

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

$$\begin{aligned} \langle \tilde{\mathcal{C}}(x_1, x_2), (y_1, y_2) \rangle_{\tilde{\mathbb{H}}} &= \langle \mathcal{C}(x_1 + x_2), y_1 + y_2 \rangle \\ &= \langle x_1 + x_2, \mathcal{C}(y_1 + y_2) \rangle \\ &= \langle (x_1, x_2), \tilde{\mathcal{C}}(y_1, y_2) \rangle_{\tilde{\mathbb{H}}}. \end{aligned}$$

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 Φ is an $L_2(U_0, H)$ -valued process stochastically integrable in $[0, T]$, φ 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 $t \in [0, T]$:

$$\begin{aligned} F(t, X_t) &= F(0, X_0) + \int_0^t \langle F_x(s, X_s), \Phi(s) dW_t \rangle + \int_0^t F_t(s, X_s) ds \\ &\quad + \int_0^t \langle F_x(s, X_s), \varphi(s) \rangle ds \\ &\quad + \int_0^t \frac{1}{2} \text{Tr} \left[F_{xx}(s, X_s) (\Phi(s) Q^{1/2}) (\Phi(s) Q^{1/2})^* \right] ds. \end{aligned}$$

Let $F : [0, \infty) \times \tilde{\mathbb{H}} \rightarrow \mathbb{R}$ 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

$$F_{(x,y)}(t; x, y)[(h_1, h_2)] = 2e^{2\alpha t} \langle x - y, h_1 - h_2 \rangle. \quad (\text{D.6})$$

Eq. (D.6) holds because

$$\begin{aligned}
& \frac{\left| \|x + h_1 - y - h_2\|^2 - \|x - y\|^2 - 2\langle x - y, h_1 - h_2 \rangle \right|}{\sqrt{\|h_1\|^2 + \|h_2\|^2}} \\
&= \frac{\|h_1 - h_2\|^2}{\sqrt{\|h_1\|^2 + \|h_2\|^2}} \\
&\leq 2\sqrt{\|h_1\|^2 + \|h_2\|^2} \xrightarrow{\|h_1\|, \|h_2\| \rightarrow 0} 0.
\end{aligned}$$

Furthermore, the second Fréchet derivative with respect to the space parameters is

$$F_{(x,y),(x,y)}[(h_1, h_2), (h_3, h_4)] = 2e^{2\alpha t} \langle h_3 - h_4, h_1 - h_2 \rangle.$$

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

$$\begin{aligned}
& \langle b(X_t) - \tilde{b}(Y_t), X_t - Y_t \rangle \\
&= \langle b(X_t) - b(Y_t), X_t - Y_t \rangle + \langle b(Y_t) - \tilde{b}(Y_t), X_t - Y_t \rangle \\
&\leq (-\alpha + 1) \|X_t - Y_t\|^2 + \|b(Y_t) - \tilde{b}(Y_t)\| \|X_t - Y_t\|.
\end{aligned} \tag{D.7}$$

We will assume that we start the process $t \mapsto (X_t, Y_t)$ 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 (so that $\varphi(t) = (b(X_t), \tilde{b}(Y_t)) - (X_t, Y_t)$ in Theorem D.5):

$$\begin{aligned}
e^{2\alpha t} \|X_t - Y_t\|^2 &= \|X_0 - Y_0\|^2 + \int_0^t 2\sqrt{2} e^{2\alpha s} \langle X_s - Y_s, dW_s - dW_s \rangle \\
&\quad + \int_0^t 2\alpha e^{2\alpha s} \|X_s - Y_s\|^2 ds \\
&\quad + \int_0^t 2e^{2\alpha s} \langle X_s - Y_s, b(X_s) - X_s - \tilde{b}(Y_s) + Y_s \rangle ds \\
&\quad + \int_0^t e^{2\alpha s} \text{Tr}[(x, y) \mapsto (\mathcal{C}(x + y) - \mathcal{C}(x + y), \mathcal{C}(x + y) - \mathcal{C}(x + y))] ds \\
&= \|X_0 - Y_0\|^2 + \int_0^t 2\alpha e^{2\alpha s} \|X_s - Y_s\|^2 ds \\
&\quad + \int_0^t 2e^{2\alpha s} \langle X_s - Y_s, b(X_s) - X_s - \tilde{b}(Y_s) + Y_s \rangle ds.
\end{aligned}$$

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)

$$\begin{aligned}
& \mathbb{E} \|X_t - Y_t\|^2 \\
&\leq e^{-2\alpha t} \mathbb{E} \|X_0 - Y_0\|^2 + \mathbb{E} \left[\int_0^t 2e^{2\alpha(s-t)} \|b(Y_s) - \tilde{b}(Y_s)\| \|X_s - Y_s\| ds \right] \\
&\leq e^{-2\alpha t} \mathbb{E} \|X_0 - Y_0\|^2 \\
&\quad + \left(\int_0^t 2e^{2\alpha(s-t)} \mathbb{E} \|b(Y_s) - \tilde{b}(Y_s)\|^2 ds \right)^{1/2} \left(\int_0^t 2e^{2\alpha(s-t)} \mathbb{E} \|X_s - Y_s\|^2 ds \right)^{1/2} \tag{D.8} \\
&= e^{-2\alpha t} \mathbb{E} \|X_0 - Y_0\|^2 \\
&\quad + \left(\alpha^{-1/2} (1 - e^{-2\alpha t})^{1/2} \|b - \tilde{b}\|_\nu \right) \left(\alpha^{-1/2} (1 - e^{-2\alpha t})^{1/2} \left(\mathbb{E} \|X_t - Y_t\|^2 \right)^{1/2} \right) \tag{D.9} \\
&= e^{-2\alpha t} \mathbb{E} \|X_0 - Y_0\|^2 + \alpha^{-1} (1 - e^{-2\alpha t}) \|b - \tilde{b}\|_\nu \left(\mathbb{E} \|X_t - Y_t\|^2 \right)^{1/2},
\end{aligned}$$

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 $(\mathbb{E}\|X_t - Y_t\|^2)^{1/2}$, taking $t \rightarrow \infty$ 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 $\bar{k}_\nu(\mathbf{x}, \mathbf{x}') := \mathbb{E}[g(\mathbf{x})g(\mathbf{x}')]$. By Cauchy-Schwarz and Jensen's inequalities,

$$\begin{aligned} |\mu_\eta(\mathbf{x}) - \mu_\nu(\mathbf{x})| &= |\mathbb{E}[f(\mathbf{x}) - g(\mathbf{x})]| = |\mathbb{E}[\langle f - g, r_{\mathbf{x}} \rangle]| \\ &\leq \mathbb{E}[\|f - g\| \|r_{\mathbf{x}}\|] \leq r(\mathbf{x}, \mathbf{x})^{1/2} \mathbb{E}[\|f - g\|^2]^{1/2} \\ &\leq r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon. \end{aligned}$$

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

$$\begin{aligned} |k_\eta(\mathbf{x}, \mathbf{x}) - \bar{k}_\nu(\mathbf{x}, \mathbf{x})| &= |\mathbb{E}[f(\mathbf{x})^2 - g(\mathbf{x})^2]| \\ &= \mathbb{E}[(f(\mathbf{x}) - g(\mathbf{x}))(f(\mathbf{x}) + g(\mathbf{x}))] \\ &\leq \sqrt{\mathbb{E}[(f(\mathbf{x}) - g(\mathbf{x}))^2]} \sqrt{\mathbb{E}[(f(\mathbf{x}) + g(\mathbf{x}))^2]} \\ &\leq r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon \sqrt{2\mathbb{E}[f(\mathbf{x})^2 + g(\mathbf{x})^2]} \\ &\leq \sqrt{2} r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon (k_\eta(\mathbf{x}, \mathbf{x})^{1/2} + \bar{k}_\nu(\mathbf{x}, \mathbf{x})^{1/2}) \\ |k_\eta(\mathbf{x}, \mathbf{x})^{1/2} - \bar{k}_\nu(\mathbf{x}, \mathbf{x})^{1/2}| &\leq \sqrt{2} r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon. \end{aligned}$$

Also,

$$\bar{k}_\nu(\mathbf{x}, \mathbf{x})^{1/2} \leq \sqrt{k_\nu(\mathbf{x}, \mathbf{x}) + \mu_\nu(\mathbf{x})^2} \leq k_\nu(\mathbf{x}, \mathbf{x})^{1/2} + r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon.$$

We now have that

$$\begin{aligned} |k_\eta(\mathbf{x}, \mathbf{x}) - k_\nu(\mathbf{x}, \mathbf{x})| &= |k_\eta(\mathbf{x}, \mathbf{x}) - \bar{k}_\nu(\mathbf{x}, \mathbf{x}) + \mu_\nu(\mathbf{x})^2| \\ &\leq |k_\eta(\mathbf{x}, \mathbf{x}) - \bar{k}_\nu(\mathbf{x}, \mathbf{x})| + \mu_\nu(\mathbf{x})^2 \\ &\leq \sqrt{2} r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon (k_\eta(\mathbf{x}, \mathbf{x})^{1/2} + \bar{k}_\nu(\mathbf{x}, \mathbf{x})^{1/2}) + r(\mathbf{x}, \mathbf{x}) \varepsilon^2 \\ &\leq \sqrt{2} r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon (k_\eta(\mathbf{x}, \mathbf{x})^{1/2} + k_\nu(\mathbf{x}, \mathbf{x})^{1/2}) + (1 + \sqrt{2}) r(\mathbf{x}, \mathbf{x}) \varepsilon^2 \\ |k_\eta(\mathbf{x}, \mathbf{x})^{1/2} - k_\nu(\mathbf{x}, \mathbf{x})^{1/2}| &\leq \sqrt{2} r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon + \frac{(1 + \sqrt{2}) r(\mathbf{x}, \mathbf{x}) \varepsilon^2}{k_\eta(\mathbf{x}, \mathbf{x})^{1/2} + k_\nu(\mathbf{x}, \mathbf{x})^{1/2}}. \end{aligned}$$

Let $a := \frac{1 + \sqrt{3 + 2\sqrt{2}}}{\sqrt{2}} r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon$. If $\max(k_\eta(\mathbf{x}, \mathbf{x})^{1/2}, k_\nu(\mathbf{x}, \mathbf{x})^{1/2}) \leq a$, then clearly $|k_\eta(\mathbf{x}, \mathbf{x})^{1/2} - k_\nu(\mathbf{x}, \mathbf{x})^{1/2}| \leq a$. Otherwise we have

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

Hence we conclude unconditionally that

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

Thus, we also have that

$$\begin{aligned} |k_\eta(\mathbf{x}, \mathbf{x}) - k_\nu(\mathbf{x}, \mathbf{x})| &\leq \sqrt{2} r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon (k_\eta(\mathbf{x}, \mathbf{x})^{1/2} + k_\nu(\mathbf{x}, \mathbf{x})^{1/2}) + (1 + \sqrt{2}) r(\mathbf{x}, \mathbf{x}) \varepsilon^2 \\ &< \sqrt{2} r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon (2k_\eta(\mathbf{x}, \mathbf{x})^{1/2} + \sqrt{6} r(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon) + (1 + \sqrt{2}) r(\mathbf{x}, \mathbf{x}) \varepsilon^2 \\ &= 2\sqrt{2} r(\mathbf{x}, \mathbf{x})^{1/2} k_\eta(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon + (1 + \sqrt{2} + \sqrt{12}) r(\mathbf{x}, \mathbf{x}) \varepsilon^2 \\ &< 3 r(\mathbf{x}, \mathbf{x})^{1/2} k_\eta(\mathbf{x}, \mathbf{x})^{1/2} \varepsilon + 6 r(\mathbf{x}, \mathbf{x}) \varepsilon^2. \end{aligned}$$

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 $\mathbb{H}_k$:

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

Define

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

If $\sum_{j \geq 1} \lambda_j^{-1} < \infty$ then r dominates k . So given inputs $\mathbf{X} = (\mathbf{x}_n)_{n=1}^N$, and defining $a_{nm,j} := e_j(\mathbf{x}_n) e_j(\mathbf{x}_m)$, 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 $j \in 1, \dots, J$ 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 $\sqrt{a_j} = o(j^{-1})$, $\lambda_j = \omega(j)$, and so $\lambda_j^{-1} = o(j^{-1})$ yielding $\sum_{j \geq 1} \lambda_j^{-1} < \infty$.

F Proof of Proposition 5.1

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

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

$$\mathcal{D}\mathcal{L}_n(f) = -\sigma^{-2} (f(\mathbf{x}_n) - y_n) r_{\mathbf{x}_n}.$$

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

$$\begin{aligned} & |\mathcal{L}_n(f + g) - \mathcal{L}_n(f) + \langle \sigma^{-2} (f(\mathbf{x}_n) - y_n) r(\mathbf{x}_n, \cdot), g \rangle| \\ &= \left| -\frac{1}{2\sigma^2} (f(\mathbf{x}_n) + g(\mathbf{x}_n) - y_n)^2 + \frac{1}{2\sigma^2} (f(\mathbf{x}_n) - y_n)^2 + \sigma^{-2} (f(\mathbf{x}_n) - y_n) g(\mathbf{x}_n) \right| \\ &\leq \frac{1}{2\sigma^2} g(\mathbf{x}_n)^2 = \frac{1}{2\sigma^2} \langle r(\mathbf{x}_n, \cdot), g \rangle^2 \leq \frac{r(\mathbf{x}_n, \mathbf{x}_n)}{2\sigma^2} \|g\|^2. \end{aligned}$$

□

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

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

and

$$\mathcal{D}\tilde{\mathcal{L}}(f) = -\sigma^{-2} (\bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}}) - \mathbf{y})^\top \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} r_{\tilde{\mathbf{X}}}.$$

Proof. Both results follow directly from Lemma F.1.

□

Lemma F.3. If $\nu = \text{GP}(\hat{\mu}, \hat{k})$, then

$$\mathbb{E}_{f \sim \nu}[\langle \mathcal{DL}_n(f), \mathcal{DL}_m(f) \rangle] = \sigma^{-4} r(\mathbf{x}_n, \mathbf{x}_m) [\hat{k}(\mathbf{x}_n, \mathbf{x}_m) + (y_n - \hat{\mu}(\mathbf{x}_n))(y_m - \hat{\mu}(\mathbf{x}_m))].$$

Proof. Using Lemma F.1, we have

$$\begin{aligned} \mathbb{E}_{f \sim \nu}[\langle \mathcal{DL}_n(f), \mathcal{DL}_m(f) \rangle] &= \sigma^{-4} \langle r_{\mathbf{x}_n}, r_{\mathbf{x}_m} \rangle \mathbb{E}_{f \sim \nu}[(f(\mathbf{x}_n) - y_n)(f(\mathbf{x}_m) - y_m)] \\ &= \sigma^{-4} r(\mathbf{x}_n, \mathbf{x}_m) [\hat{k}(\mathbf{x}_n, \mathbf{x}_m) + (y_n - \hat{\mu}(\mathbf{x}_n))(y_m - \hat{\mu}(\mathbf{x}_m))]. \end{aligned}$$

□

Lemma F.4. If $\eta = \text{GP}(0, \ell)$ then $(\mathcal{C}_\eta f)(\mathbf{x}) = \langle f, \ell_{\mathbf{x}} \rangle$.

Proof. Since $(\mathcal{C}_\eta r_{\mathbf{x}'}) = \langle r_{\mathbf{x}'}, \ell_{\cdot} \rangle = \ell_{\mathbf{x}'}$, for $f \sim \eta$,

$$\langle r_{\mathbf{x}}, \mathcal{C}_\eta r_{\mathbf{x}'} \rangle = \langle r_{\mathbf{x}}, \ell_{\mathbf{x}'} \rangle = \ell(\mathbf{x}, \mathbf{x}') = \text{Cov}(f(\mathbf{x}), f(\mathbf{x}')).$$

□

Lemma F.5. For the DTC log-likelihood approximation $\tilde{\pi}$,

$$(\mathcal{C}_{\tilde{\pi}} f)(\mathbf{x}) = (\mathcal{C}_{\pi_0} f)(\mathbf{x}) - \langle f, k_{\tilde{\mathbf{X}}} \rangle (k_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}^{-1} - \tilde{\Sigma}) k_{\tilde{\mathbf{X}}\mathbf{x}},$$

where $\tilde{\Sigma} := (k_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} + \sigma^{-2} k_{\tilde{\mathbf{X}}\mathbf{X}} k_{\mathbf{X}\tilde{\mathbf{X}}})^{-1}$.

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

It follows from Lemmas F.2 and F.5 that

$$\begin{aligned} \mathcal{C}_{\tilde{\pi}} \mathcal{DL}(f) &= -\sigma^{-2} (\bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}}) - \mathbf{y})^\top K_{\mathbf{X}\tilde{\mathbf{X}}} \tilde{\Sigma} k_{\tilde{\mathbf{X}}} \\ \mathcal{C}_{\tilde{\pi}} \mathcal{DL}(f) &= -\sigma^{-2} (f(\mathbf{X}) - \mathbf{y})^\top (k_{\mathbf{X}} - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} k_{\tilde{\mathbf{X}}} + K_{\mathbf{X}\tilde{\mathbf{X}}} \tilde{\Sigma} k_{\tilde{\mathbf{X}}}). \end{aligned}$$

We therefore have that

$$\begin{aligned} &-\sigma^2 \mathcal{C}_{\tilde{\pi}} \mathcal{D}(\mathcal{L} - \tilde{\mathcal{L}})(f) \\ &= (f(\mathbf{X}) - \mathbf{y})^\top (k_{\mathbf{X}} - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} k_{\tilde{\mathbf{X}}}) + (f(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}}))^\top K_{\mathbf{X}\tilde{\mathbf{X}}} \tilde{\Sigma} k_{\tilde{\mathbf{X}}} \end{aligned}$$

Consider the limit $r \rightarrow k$, so $k' \rightarrow k$. Then

$$\begin{aligned} &\sigma^4 \|\mathcal{C}_{\tilde{\pi}} \mathcal{D}(\mathcal{L} - \tilde{\mathcal{L}})(f)\|^2 \\ &= (f(\mathbf{X}) - \mathbf{y})^\top (K_{\mathbf{X}\mathbf{X}} + \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} K_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}}^\top - 2K_{\mathbf{X}\tilde{\mathbf{X}}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}}^\top) (f(\mathbf{X}) - \mathbf{y}) \\ &\quad + (f(\mathbf{X}) - \mathbf{y})^\top (K_{\mathbf{X}\tilde{\mathbf{X}}} \tilde{\Sigma} K_{\tilde{\mathbf{X}}\mathbf{X}} - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} K_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \tilde{\Sigma} K_{\tilde{\mathbf{X}}\mathbf{X}}) (f(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}})) \\ &\quad + (f(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}}))^\top K_{\mathbf{X}\tilde{\mathbf{X}}} \tilde{\Sigma} K_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \tilde{\Sigma} K_{\tilde{\mathbf{X}}\mathbf{X}} (f(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}})) \\ &= (f(\mathbf{X}) - \mathbf{y})^\top (K_{\mathbf{X}\mathbf{X}} - Q_{\mathbf{X}\mathbf{X}}) (f(\mathbf{X}) - \mathbf{y}) \\ &\quad + (f(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}}))^\top S_{\mathbf{X}\mathbf{X}} (f(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}})), \end{aligned}$$

where $S_{\mathbf{X}\mathbf{X}} := K_{\mathbf{X}\tilde{\mathbf{X}}} \tilde{\Sigma} K_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \tilde{\Sigma} K_{\tilde{\mathbf{X}}\mathbf{X}}$. Let $E_{\mathbf{X}\mathbf{X}} := K_{\mathbf{X}\mathbf{X}} - Q_{\mathbf{X}\mathbf{X}}$. Taking expectations we get

$$\begin{aligned} &\mathbb{E}_\nu[(f(\mathbf{X}) - \mathbf{y})^\top E_{\mathbf{X}\mathbf{X}} (f(\mathbf{X}) - \mathbf{y})] \\ &= \mathbb{E}_\nu[(f(\mathbf{X}) - \hat{\mu}(\mathbf{X}) + \hat{\mu}(\mathbf{X}) - \mathbf{y})^\top E_{\mathbf{X}\mathbf{X}} (f(\mathbf{X}) - \hat{\mu}(\mathbf{X}) + \hat{\mu}(\mathbf{X}) - \mathbf{y})] \\ &= \text{Tr}(\hat{K}_{\mathbf{X}\mathbf{X}} E_{\mathbf{X}\mathbf{X}}) + (\hat{\mu}(\mathbf{X}) - \mathbf{y})^\top E_{\mathbf{X}\mathbf{X}} (\hat{\mu}(\mathbf{X}) - \mathbf{y}) \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}_\nu[(f(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}}))^\top S_{\mathbf{X}\mathbf{X}} (f(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}}))] \\ &= \mathbb{E}_\nu[\|(f(\mathbf{X}) - \hat{\mu}(\mathbf{X}) + \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} \hat{\mu}(\tilde{\mathbf{X}}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} f(\tilde{\mathbf{X}}) + \hat{\mu}(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} \hat{\mu}(\tilde{\mathbf{X}}))^\top S_{\mathbf{X}\mathbf{X}}^{1/2}\|_2^2] \\ &= \text{Tr}(\hat{K}_{\mathbf{X}\mathbf{X}} S_{\mathbf{X}\mathbf{X}}) + \text{Tr}(\hat{K}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}}^\top S_{\mathbf{X}\mathbf{X}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}}) - 2 \text{Tr}(\hat{K}_{\tilde{\mathbf{X}}\mathbf{X}} S_{\mathbf{X}\mathbf{X}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}}) \\ &\quad + (\hat{\mu}(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} \hat{\mu}(\tilde{\mathbf{X}}))^\top S_{\mathbf{X}\mathbf{X}} (\hat{\mu}(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} \hat{\mu}(\tilde{\mathbf{X}})). \end{aligned}$$

Let $S'_{\mathbf{X}\tilde{\mathbf{X}}} := K_{\mathbf{X}\tilde{\mathbf{X}}} \tilde{\Sigma} K_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \tilde{\Sigma}$. Putting everything together, conclude that

$$\begin{aligned}
& \sigma^4 \|\mathcal{C}_{\tilde{\pi}} \mathcal{D}(\mathcal{L} - \tilde{\mathcal{L}})\|_{\nu}^2 \\
&= \text{Tr}((\hat{K}_{\mathbf{X}\mathbf{X}} + (\hat{\mu}(\mathbf{X}) - \mathbf{y})(\hat{\mu}(\mathbf{X}) - \mathbf{y})^\top)(K_{\mathbf{X}\mathbf{X}} - Q_{\mathbf{X}\mathbf{X}})) \\
&\quad + \text{Tr}(\hat{K}_{\mathbf{X}\mathbf{X}} S_{\mathbf{X}\mathbf{X}}) + \text{Tr}(\hat{K}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}}^\top S_{\mathbf{X}\mathbf{X}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}}) - 2 \text{Tr}(\hat{K}_{\tilde{\mathbf{X}}\mathbf{X}} S_{\mathbf{X}\mathbf{X}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}}) \\
&\quad + (\hat{\mu}(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} \hat{\mu}(\tilde{\mathbf{X}}))^\top S_{\mathbf{X}\mathbf{X}} (\hat{\mu}(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} \hat{\mu}(\tilde{\mathbf{X}})). \\
&= -\text{Tr}(K_{\tilde{\mathbf{X}}\mathbf{X}}(\hat{K}_{\mathbf{X}\mathbf{X}} + (\hat{\mu}(\mathbf{X}) - \mathbf{y})(\hat{\mu}(\mathbf{X}) - \mathbf{y})^\top) \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}}) \\
&\quad + \text{Tr}((K_{\tilde{\mathbf{X}}\mathbf{X}} \hat{K}_{\mathbf{X}\mathbf{X}} + K_{\tilde{\mathbf{X}}\mathbf{X}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} \hat{K}_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}}^\top - 2K_{\tilde{\mathbf{X}}\mathbf{X}} \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} \hat{K}_{\tilde{\mathbf{X}}\mathbf{X}}) S'_{\mathbf{X}\tilde{\mathbf{X}}}) \quad (\text{F.1}) \\
&\quad + (\hat{\mu}(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} \hat{\mu}(\tilde{\mathbf{X}}))^\top S'_{\mathbf{X}\tilde{\mathbf{X}}} K_{\tilde{\mathbf{X}}\mathbf{X}} (\hat{\mu}(\mathbf{X}) - \bar{Q}_{\mathbf{X}\tilde{\mathbf{X}}} \hat{\mu}(\tilde{\mathbf{X}})) + C(\mathbf{X}).
\end{aligned}$$

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 $C(\mathbf{X})$, the pF divergence can be computed in $O(NM^2)$ time and $O(NM)$ space.

References

- [1] G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*. Cambridge University Press, New York, NY, 2nd edition, 2014.
- [2] M. Gelbrich. On a Formula for the L2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.
- [3] M. Hairer, A. M. Stuart, J. Voss, and P. Wiberg. Analysis of SPDEs arising in path sampling part I: The Gaussian case. *Communications in Mathematical Sciences*, 3(4):587–603, Dec. 2005.
- [4] M. Hairer, A. M. Stuart, and J. Voss. Analysis of SPDEs arising in path sampling part II: The nonlinear case. *The Annals of Applied Probability*, 17(5/6):1657–1706, Oct. 2007.
- [5] J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian Processes for Big Data. In *Uncertainty in Artificial Intelligence*, 2013.
- [6] J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.
- [7] C. Villani. *Optimal transport: old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.