A. Alternative decompositions

As mentioned in the Section 3.2, the proposed representation of the GP posteriors—as the sum of a weight-space prior and a function-space update—is one of many possible choices. Here, we briefly reflect on two such alternatives.

To begin with, we may directly represent sparse GP posteriors in weight-space via a Bayesian linear model $f(\cdot) = \boldsymbol{\phi}(\cdot)^\top \boldsymbol{w}$. To this end, we may rewrite (12) for a given draw $\boldsymbol{u} \sim q(\boldsymbol{u})$ as

$$\boldsymbol{w} \mid \boldsymbol{u} \overset{\mathrm{d}}{=} \boldsymbol{w} + \boldsymbol{\Phi}^\top (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top)^{-1}(\boldsymbol{u} - \boldsymbol{\Phi}\boldsymbol{w}), \tag{18}$$

where $\boldsymbol{\Phi} = \boldsymbol{\phi}(\mathbf{Z})$ now denotes an $m \times \ell$ feature matrix. Prima facie, this appears to resolve many of the problems discussed earlier in the text: inducing distribution $q(\boldsymbol{u})$ relays information about $\boldsymbol{y}$ and the Bayesian linear model needs only explain for the function's behavior at $m \ll n$ locations. In practice, (18) does more harm than good however, since f must now exactly pass through $\boldsymbol{u}$ due to a lack of measurement noise σ^2.

Alternatively, we may think to employ an *orthogonal decomposition* $f(\cdot) = f_\parallel(\cdot) + f_\perp(\cdot)$ (Salimbeni et al., 2018; Shi et al., 2020). Here, we interpret "orthogonality" in the statistical sense of independent random variables (Rodgers et al., 1984). For Gaussian random variables, this distinction amounts to satisfying the definition $\mathrm{Cov}(f_\parallel, f_\perp) = 0$. In the case of sparse GPs, $f_\parallel$ is typically represented in terms of canonical basis functions $k(\cdot, \mathbf{Z})$ such that $(f_\parallel \mid \boldsymbol{u})(\cdot)$ denotes the posterior mean function given $q(\boldsymbol{u})$. Consequently, $f_\perp$ denotes the process residuals $(f_\perp \mid \boldsymbol{u})(\cdot) = (f \mid \boldsymbol{u})(\cdot) - (f_\parallel \mid \boldsymbol{u})(\cdot)$. By construction however, $f_\perp$ is independent of $f_\parallel$ and, hence, of particular values $\boldsymbol{u}$. Moreover, since $(f \mid \boldsymbol{u})(\mathbf{Z}) = (f_\parallel \mid \boldsymbol{u})(\mathbf{Z}) = \boldsymbol{u}$, it follows that $f_\perp(\mathbf{Z}) = (f_\perp \mid \boldsymbol{u})(\mathbf{Z}) = \mathbf{0}$.

Generating draws from this type of decomposition is made difficult by orthogonal component $f_\perp \mid \boldsymbol{u}$, whose covariance can readily be shown as

$$\mathrm{Cov}(f_\perp, f_\perp) = k(\cdot,\cdot) - k(\cdot, \mathbf{Z})\mathbf{K}_{m,m}^{-1} k(\mathbf{Z}, \cdot). \tag{19}$$

Sampling schemes based on random Fourier feature approximations of $f_\perp$ are nearly identical to (18): all that has changed is that the Bayesian linear model must now pass exactly through zero, rather than $\boldsymbol{u}$, at each of the m inducing locations. This approach to sampling therefore inherits the issues outlined above.

B. Error analysis

Definition 5 (Preliminaries)**.** *Consider a Gaussian process f defined on $\mathbb{R}^d$ and restricted to a compact subset $\mathcal{X} \subseteq \mathbb{R}^d$. Let $\boldsymbol{y} \in \mathbb{R}^n$. Assume a Gaussian likelihood $y_i \sim \mathcal{N}(f(x_i), \sigma^2)$, with $\sigma^2 \geq 0$. Let $f^{(w)}$ be a weight-space prior approximation. Let $f \mid \boldsymbol{y}$ be the true posterior, let $f^{(s)}$ be an inducing point approximate posterior, and let $f^{(d)}$ be the decoupled posterior approximation. Let $k, k^{(w)}, k^{(f|\boldsymbol{y})}, k^{(s)}, k^{(d)}$ be their respective kernels.*

Proposition 6. *We have that*

$$W_{2,L^2(\mathcal{X})}\left(f^{(d)}, f \mid \boldsymbol{y}\right) \leq W_{2,L^2(\mathcal{X})}\left(f^{(s)}, f \mid \boldsymbol{y}\right) + C_1\, W_{2,L^\infty(\mathcal{X})}\left(f^{(w)}, f\right) \tag{20}$$

where $C_1 = \sqrt{2\,\mathrm{diam}(\mathcal{X})^d \left(1 + \|k\|^2_{C(\mathcal{X}^2)} \|\mathbf{K}_{mm}^{-1}\|^2_{L(\ell^\infty;\ell^1)}\right)}$, $W_{2,L^2(\mathcal{X})}$ and $W_{2,C(\mathcal{X})}$ are the 2-Wasserstein distances over $L^2(\mathcal{X})$ and the space of continuous functions $C(\mathcal{X})$ equipped with the supremum norm, respectively, and $\|\cdot\|_{L(\ell^\infty;\ell^1)}$ is the corresponding operator norm of a matrix.

Proof. By the triangle inequality, we have

$$W_{2,L^2(\mathcal{X})}\left(f^{(d)}, f \mid \boldsymbol{y}\right) \leq W_{2,L^2(\mathcal{X})}\left(f^{(d)}, f^{(s)}\right) + W_{2,L^2(\mathcal{X})}\left(f^{(s)}, f \mid \boldsymbol{y}\right). \tag{21}$$

We proceed bound the first term pathwise. For arbitrary $x \in M$, write

$$\left|f^{(d)}(x) - f^{(s)}(x)\right|^2 \leq 2\left(\left|f^{(w)}(x) - f(x)\right|^2 + \left|\mathbf{K}_{xm}\mathbf{K}_{mm}^{-1}(f^{(w)}(\boldsymbol{z}) - f(\boldsymbol{z}))\right|^2\right) \tag{22}$$

$$\leq 2\left(\left\|f^{(w)} - f\right\|_{L^\infty(\mathcal{X})}^2 + \left\|\mathbf{K}_{xm}\mathbf{K}_{mm}^{-1}\right\|_{\ell^1}^2 \left\|f^{(w)}(\boldsymbol{z}) - f(\boldsymbol{z})\right\|_{\ell^\infty}^2\right) \tag{23}$$

$$\leq 2\left(\left\|f^{(w)} - f\right\|_{L^\infty(\mathcal{X})}^2 + \left\|\mathbf{K}_{xm}\right\|_{\ell^\infty}^2 \left\|\mathbf{K}_{mm}^{-1}\right\|_{L(\ell^\infty;\ell^1)}^2 \left\|f^{(w)} - f\right\|_{L^\infty(\mathcal{X})}^2\right) \tag{24}$$

$$\leq 2\left(1 + \left\|k\right\|_{C(\mathcal{X}^2)}^2 \left\|\mathbf{K}_{mm}^{-1}\right\|_{L(\ell^\infty;\ell^1)}^2\right) \left\|f^{(w)} - f\right\|_{L^\infty(\mathcal{X})}^2 \tag{25}$$

$$= 2\left(1 + \left\|k\right\|_{C(\mathcal{X}^2)}^2 \left\|\mathbf{K}_{mm}^{-1}\right\|_{L(\ell^\infty;\ell^1)}^2\right) \left\|f^{(w)} - f\right\|_{C(\mathcal{X})}^2 \tag{26}$$

where in (22) we have used Matheron's rule, in (23) we have used Hölder's inequality with $p = 1, q = \infty$, in (24) we have used the definition of an operator norm, and in (26) we have used that given sample paths are continuous so $\|\cdot\|_{L^\infty(\mathcal{X})}$ can be replaced with $\|\cdot\|_{C(\mathcal{X})}$. We now lift this to a bound on the Wasserstein distance by integrating both sides. With $\gamma \in \mathcal{C}$ denoting couplings between $\mathcal{GP}(0, k)$ and $\mathcal{GP}(0, k^{(w)})$, write

$$W_{2,L^2(\mathcal{X})}^2(f^{(d)}, f^{(s)}) \leq \inf_{\gamma \in \mathcal{C}} \int \left\|f^{(d)} - f^{(s)}\right\|_{L^2(\mathcal{X})}^2 \mathrm{d}\gamma \tag{27}$$

$$\leq C|\mathcal{X}| \inf_{\gamma \in \mathcal{C}} \int \left\|f^{(w)} - f\right\|_{C(\mathcal{X})}^2 \mathrm{d}\gamma \tag{28}$$

$$= C \operatorname{diam}(\mathcal{X})^d \, W_{2,C(\mathcal{X})}^2(f^{(w)}, f) \tag{29}$$

where C is the constant above. Finally, note that f is sample-continuous, and $C(\mathcal{X})$ is a separable metric space, so $W_{2,C(\mathcal{X})}$ is a proper metric. The claim follows. □

Proposition 7. *Assume k is stationary continuous covariance defined on $\mathbb{R}^d \times \mathbb{R}^d$, $\mathcal{X} \subseteq \mathbb{R}^d$ is compact. We have that*

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{\omega} \sim \rho \\ v \sim U}} \left\|k^{(d)} - k^{(f|\boldsymbol{y})}\right\|_{C(\mathcal{X}^2)} \leq \left\|k^{(s)} - k^{(f|\boldsymbol{y})}\right\|_{C(\mathcal{X}^2)} + \frac{C_2 C_3}{\sqrt{\ell}} \tag{30}$$

where $\|\cdot\|_{C(\mathcal{X}^2)}$ is the supremum norm over continuous functions, C_2 is the constant given by Sutherland and Schneider (2015), which depends only on the Lipschitz constant of k, the rate of decay of the spectral density ρ, the dimension d, and the diameter of the domain $\mathcal{X}$, and $C_3 = m\left[1 + \left\|\mathbf{K}_{m,m}^{-1}\right\|_{C(\mathcal{X}^2)} \|k\|_{C(\mathcal{X}^2)}\right]^2$.

Proof. By the triangle inequality, we have

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{\omega} \sim \rho \\ v \sim U}} \left\|k^{(d)} - k^{f|\boldsymbol{y}}\right\|_{C(\mathcal{X}^2)} \leq \mathop{\mathbb{E}}_{\substack{\boldsymbol{\omega} \sim \rho \\ v \sim U}} \left\|k^{(d)} - k^{(s)}\right\|_{C(\mathcal{X}^2)} + \left\|k^{(s)} - k^{f|\boldsymbol{y}}\right\|_{C(\mathcal{X}^2)} \tag{31}$$

where we have used that the latter term does not depend on $\boldsymbol{\omega}$. We proceed to bound the inner portion of the first term. Define the bounded linear operator $M_k : C(\mathcal{X} \times \mathcal{X}) \to C(\mathcal{X} \times \mathcal{X})$ by the expression

$$(M_k c)(x, x') = c(x, x') - \mathbf{C}_{x,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,x'} - \mathbf{K}_{x,m}\mathbf{K}_{m,m}^{-1}\mathbf{C}_{m,x'} + \mathbf{K}_{x,m}\mathbf{K}_{m,m}^{-1}\mathbf{C}_{m,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,x'}. \tag{32}$$

Let $\boldsymbol{\Sigma} = \operatorname{Cov}(\boldsymbol{u})$. By explicit calculation, we have

$$k^{(d)}(x, x') = (M_k k^{(w)})(x, x') + \mathbf{K}_{x,m}\mathbf{K}_{m,m}^{-1}\boldsymbol{\Sigma}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,x'} \tag{33}$$

and we also have

$$k^{(s)}(x, x') = k^{(f|\boldsymbol{y})}(x, x') + \mathbf{K}_{x,m}\mathbf{K}_{m,m}^{-1}\boldsymbol{\Sigma}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,x'} \tag{34}$$

hence

$$\left\|k^{(d)} - k^{(s)}\right\|_{C(\mathcal{X}^2)} = \left\|M_k k^{(w)} - k^{(f|\boldsymbol{y})}\right\|_{C(\mathcal{X}^2)} = \left\|M_k k^{(w)} - M_k k\right\|_{C(\mathcal{X}^2)} \leq \left\|M_k\right\|_{L(C;C)} \left\|k^{(w)} - k\right\|_{C(\mathcal{X}^2)}. \tag{35}$$

We proceed to bound the operator norm $\left\|M_k\right\|_{L(C;C)}$. Write

$$\|M_k c\|_{C(\mathcal{X}^2)} \leq \|c\|_{C(\mathcal{X}^2)} + \left\|\mathbf{C}_{\cdot,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,\cdot}\right\|_{C(\mathcal{X}^2)} + \left\|\mathbf{K}_{\cdot,m}\mathbf{K}_{m,m}^{-1}\mathbf{C}_{m,\cdot}\right\|_{C(\mathcal{X}^2)} \tag{36}$$

$$+ \left\|\mathbf{K}_{\cdot,m}\mathbf{K}_{m,m}^{-1}\mathbf{C}_{m,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,\cdot}\right\|_{C(\mathcal{X}^2)}. \tag{37}$$

Now, note that

$$\left\|\mathbf{C}_{\cdot,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,\cdot}\right\|_{C(\mathcal{X}^2)} = \sup_{x,x'\in\mathcal{X}}\left[\mathbf{C}_{x,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,x'}\right] \tag{38}$$

$$\leq \sup_{x,x'\in\mathcal{X}}\left[\|\mathbf{C}_{x,m}\|_{\ell^\infty}\left\|\mathbf{K}_{m,m}^{-1}\right\|_{L(\ell^\infty;\ell^1)}\|\mathbf{K}_{m,x'}\|_{\ell^\infty}\right] \tag{39}$$

$$\leq \|c\|_{C(\mathcal{X}^2)}\left\|\mathbf{K}_{m,m}^{-1}\right\|_{L(\ell^\infty;\ell^1)}\|k\|_{C(\mathcal{X}^2)} \tag{40}$$

by Hölder's inequality with $p = 1$ and $q = \infty$, and then by the definition of the operator norm $\|\cdot\|_{L(\ell^\infty;\ell^1)}$. Similarly

$$\left\|\mathbf{K}_{\cdot,m}\mathbf{K}_{m,m}^{-1}\mathbf{C}_{m,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,\cdot}\right\|_{C(\mathcal{X}^2)} \leq m\|c\|_{C(\mathcal{X}^2)}\left\|\mathbf{K}_{m,m}^{-1}\right\|^2_{L(\ell^\infty;\ell^1)}\|k\|^2_{C(\mathcal{X}^2)} \tag{41}$$

hence

$$\|M_k c\|_{C(\mathcal{X}^2)} \leq \|c\|_{C(\mathcal{X}^2)} + 2\|c\|_{C(\mathcal{X}^2)}\left\|\mathbf{K}_{m,m}^{-1}\right\|_{L(\ell^\infty;\ell^1)}\|k\|_{C(\mathcal{X}^2)} + m\|c\|_{C(\mathcal{X}^2)}\left\|\mathbf{K}_{m,m}^{-1}\right\|^2_{L(\ell^\infty;\ell^1)}\|k\|^2_{C(\mathcal{X}^2)} \tag{42}$$

$$\leq \|c\|_{C(\mathcal{X}^2)}\left(m\left[1 + \left\|\mathbf{K}_{m,m}^{-1}\right\|_{L(\ell^\infty;\ell^1)}\|k\|_{C(\mathcal{X}^2)}\right]^2\right) \tag{43}$$

and therefore

$$\|M_k\|_{L(C;C)} = \sup_{c\neq 0}\frac{\|M_k c\|_{C(\mathcal{X}^2)}}{\|c\|_{C(\mathcal{X}^2)}} \leq m\left[1 + \left\|\mathbf{K}_{m,m}^{-1}\right\|_{L(\ell^\infty;\ell^1)}\|k\|_{C(\mathcal{X}^2)}\right]^2. \tag{44}$$

Note that this term is independent of $\boldsymbol{\omega}$, and hence constant with respect to the expectation. Finally, Sutherland and Schneider (2015) have shown that there exists a constant C_2 such that.

$$\mathop{\mathbb{E}}_{\substack{\boldsymbol{\omega}\sim\rho\\ v\sim U}}\left\|k^{(w)} - k\right\|_{C(\mathcal{X}^2)} \leq \frac{C_2}{\sqrt{\ell}}. \tag{45}$$

Putting together all the inequalities gives the result. □

C. Additional experiments

This appendix provides additional details regarding experiments discussed in Section 4. All experiments (and figures) were run using zero-mean GP priors with Matérn-$5/2$ kernels. For dynamical systems experiments, hyperparameters were learned (MLE type-2). In all other cases, hyperparameters were assumed to be known and specified as: lengthscales $l = \sqrt{d/100}$, measurement noise variance $\sigma^2 = 10^{-3}$, and kernel amplitude $\alpha = 1$.

C.1. 2-Wasserstein sample tests

In each trial, a set of training locations $\mathbf{X} \sim U[0,1]^{n\times d}$ was randomly generated and corresponding observations $\boldsymbol{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{n,n} + \sigma^2\mathbf{I})$ were subsequently drawn from the prior. Similarly, test sets $\mathbf{X}_* \sim U[0,1]^{*\times d}$ were sampled uniformly at random. For each sampling schemes, $100,000$ draws $\boldsymbol{f}_* \mid \boldsymbol{y}$ were then used to form an unbiased estimate $(\tilde{\boldsymbol{m}}_{*|n}, \widetilde{\mathbf{K}}_{*,*|n})$ to the true posterior moments $(\boldsymbol{m}_{*|n}, \mathbf{K}_{*,*|n})$. Given both sets of moments, 2-Wasserstein distances were computed as

$$W_{2,\ell_*^2}\left(\mathcal{N}(\boldsymbol{m}_{*|n}, \mathbf{K}_{*,*|n}), \mathcal{N}(\tilde{\boldsymbol{m}}_{*|n}, \widetilde{\mathbf{K}}_{*,*|n})\right)^2 = \left\|\boldsymbol{m}_{*|n} - \widetilde{\boldsymbol{m}}_{*|n}\right\|^2 + \operatorname{tr}\left(\mathbf{K}_{*,*|n} + \widetilde{\mathbf{K}}_{*,*|n} - 2\left(\mathbf{K}_{*,*|n}^{1/2}\widetilde{\mathbf{K}}_{*,*|n}\mathbf{K}_{*,*|n}^{1/2}\right)^{1/2}\right), \tag{46}$$

where $\mathbf{K}_{*,*|n}^{1/2}$ denotes the symmetric matrix square root, and W_{2,ℓ_*^2} denotes the 2-Wasserstein distance between probability measures over $*$-dimensional vectors equipped with Euclidean distance.

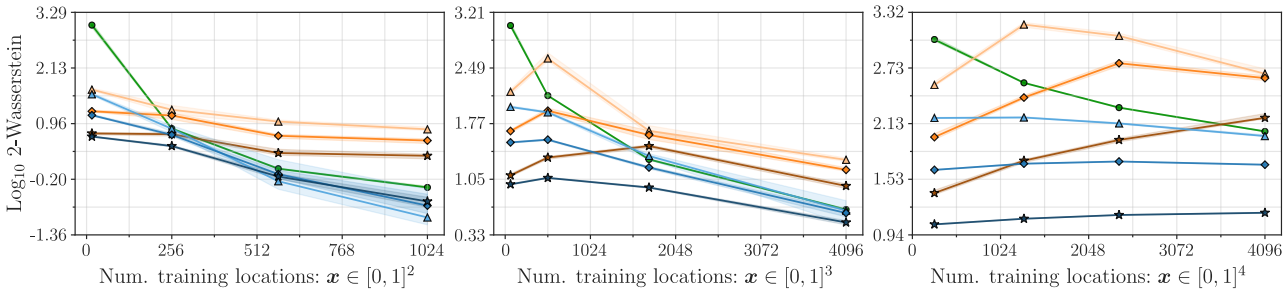

Figure 6: Medians and interquartile ranges of empirically estimated 2-Wasserstein distances measured over 32 independent trials consisting of 100,000 samples. LOVE (green) improves as the regularly spaced grids of training locations fill the space. Weight-space (orange) and decoupled (blue) sampling utilized a total of $b = m + \ell$ basis functions. Results using $\ell \in \{1024, 4096, 16384\}$ initial bases correspond with $\{\text{light}, \text{medium}, \text{dark}\}$ tones and $\{\triangle, \diamond, \star\}$ markers.

As an additional baseline, we compared decoupled sampling with a LanczOs Variance Estimates (LOVE) based alternative (Pleiss et al., 2018). The LOVE approach to sampling from GP posteriors exploits structured covariance matrices in conjunction with fast (approximate) solvers to achieve linear time complexity with respect to number of test locations. For example, when inducing locations $\mathbf{Z}$ are defined to be a regularly spaced grid, the prior covariance $\mathbf{K}_{m,m} = k(\mathbf{Z}, \mathbf{Z})$ can be expressed as the Kronecker product of Toeplitz matrices—a property that can be used to dramatically expedite much of the related linear algebra (Zimmerman, 1989; Saatçi, 2012; Wilson and Nickisch, 2015).

Here, we are interested in comparing the performance of sampling schemes themselves and not that of approximate GPs. As before, we will therefore sample from exact GPs with known hyperparameters. As an additional caveat however, we now define training locations as regularly spaced grids, such that LOVE may represent the data exactly. Similarly, we allow LOVE to utilize n conjugate gradient iterations during precomputation.

Results of these experiments are show in Figure 6. LOVE's performance improves significantly as $m = n$ increases but still lags behind that of decoupled sampling for matching m. Several points are immediately worth addressing here. First, kernel interpolation methods such as LOVE offer improved scaling w.r.t. m when compared to naïve inducing point methods (even when additional structure is imposed on $\mathbf{Z}$). LOVE can therefore utilize many more inducing locations than traditional sparse GPs in exchange for the imposed structural constraints. Assessing the relative merits of these inducing paradigms is beyond the scope of this work. Second, during sample generation, LOVE exhibits $\mathcal{O}(m + *)$ time complexity, compared to decoupled sampling's $\mathcal{O}(m \times *)$. Third, LOVE samples function values $\boldsymbol{f}_*$ at locations $\mathbf{X}_*$ whereas decoupled sampling generates function draws $(f \mid \boldsymbol{u})(\cdot)$, the implications of which were previously explored in Section 4. Fourth and finally, the techniques and ideas espoused by these frameworks are complementary: just as we may approximate the prior via a collection of Fourier features, we may approximate the update via, e.g., kernel interpolation.

C.2. Thompson sampling

As baselines, we compared against Random Search (Bergstra and Bengio, 2012) and Dividing Rectangles (Jones et al., 1993), the latter of which was run in strictly sequential fashion (i.e., $\kappa = 1$). Minimization tasks were drawn from a known GP prior (see above) and their global minimums were estimated by running gradient descent from a large number of starting locations (for purposes of measuring regret). Here, we discuss algorithmic differences between variants of TS.

For function-space TS, batches were constructed as follows.

1. Construct a mesh $\mathbf{X}_*$ consisting of $|\mathbf{X}_*| = 10^6$ random points.
2. Draw a vector of independent values $\boldsymbol{f}_* \mid \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}_{*|n}, \mathbf{K}_{*,*|n} \odot \mathbf{I})$, where $\odot$ is the element-wise product.
3. Define an active set $\mathbf{X}_s \subseteq \mathbf{X}_*$ corresponding to the $s = 2048$ smallest elements of $\boldsymbol{f}_* \mid \boldsymbol{y}$.
4. Jointly sample $\boldsymbol{f}_s \mid \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}_{s|n}, \mathbf{K}_{s,s|n})$.
5. Select $\boldsymbol{x}_i \in \arg\min_{1 \le i \le s} \boldsymbol{f}_s \mid \boldsymbol{y}$ as the i-th batch element.

For simplicity, a new mesh $\mathbf{X}_*$ was generated at each TS iteration and shared between batch elements, but steps (2-5) we

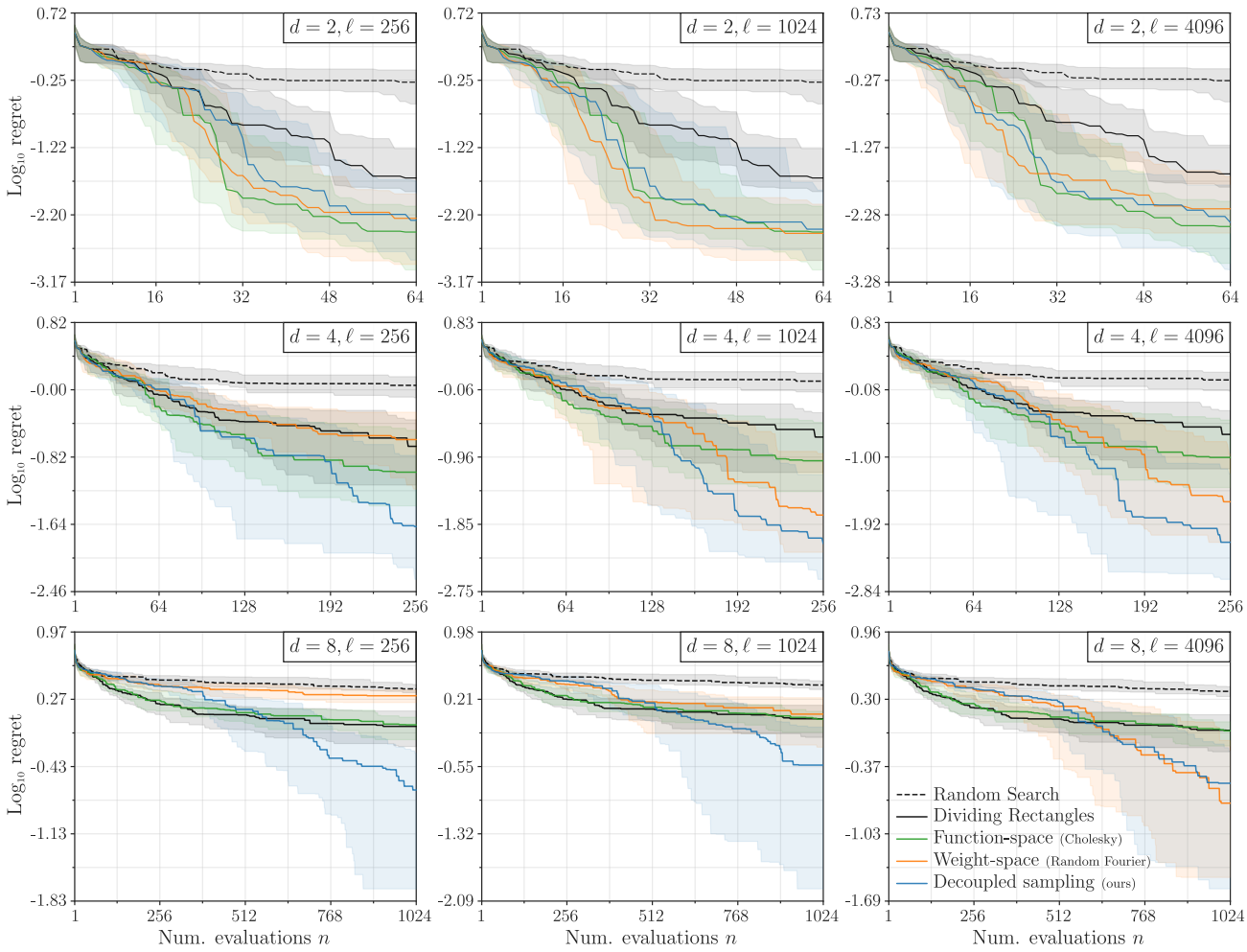

Figure 7: Results for parallel Thompson sampling, shown as quartiles over 32 independent runs with matched seeds.

run independently. Weight-space and decoupled TS employed a similar procedure, with minor differences stemming from use of function draws.

1. Construct a mesh $\mathbf{X}_*$ consisting of $|\mathbf{X}_*| = 250,000$ random points.
2. Generate a function draw $(f \mid \boldsymbol{y})(\cdot)$.
3. Define starting locations $\mathbf{X}_s \subseteq \mathbf{X}_*$ corresponding to the $s = 32$ smallest elements of $(f \mid \boldsymbol{y})(\mathbf{X}_*)$.
4. Run multi-start gradient-based optimization: we employed an off-the-shelf version of L-BFGS-B.
5. Select $\boldsymbol{x}_i \in \arg\min_{1 \leq i \leq s} (f \mid \boldsymbol{y})(\mathbf{X}'_*)$ as the i-th batch element, where $\mathbf{X}'_s$ denotes the optimized locations.

As before, steps (2-5) we run independently. Optimization performance and runtimes are shown below.

C.3. Dynamical systems

We investigated decoupled sampling's impact on (sequential) Monte Carlo methods' runtimes by using a sparse GP to simulate a simple dynamical system, the FitzHugh-Nagumo model neuron (FitzHugh, 1961; Nagumo et al., 1962) with diffusion coefficient $\boldsymbol{\Sigma} = 0.01 \cdot \mathbf{I}$. Training and simulation were both performed using a step size $\Delta t = 0.25$.

During training, independent sparse GPs with $m = 32$ shared inducing locations were fit to 3-dimensional inputs $\boldsymbol{x}_t = [\boldsymbol{s}_t, \boldsymbol{c}_t]$, where $\boldsymbol{s} \in [0, 1]^2$ denotes the (normalized) state vector at time t and $\boldsymbol{c} \in [0, 1]$ the coinciding (normalized) control input, with targets defined as the i-th element of the Euler-Maruyama transition vectors specified by (17). Owing to the

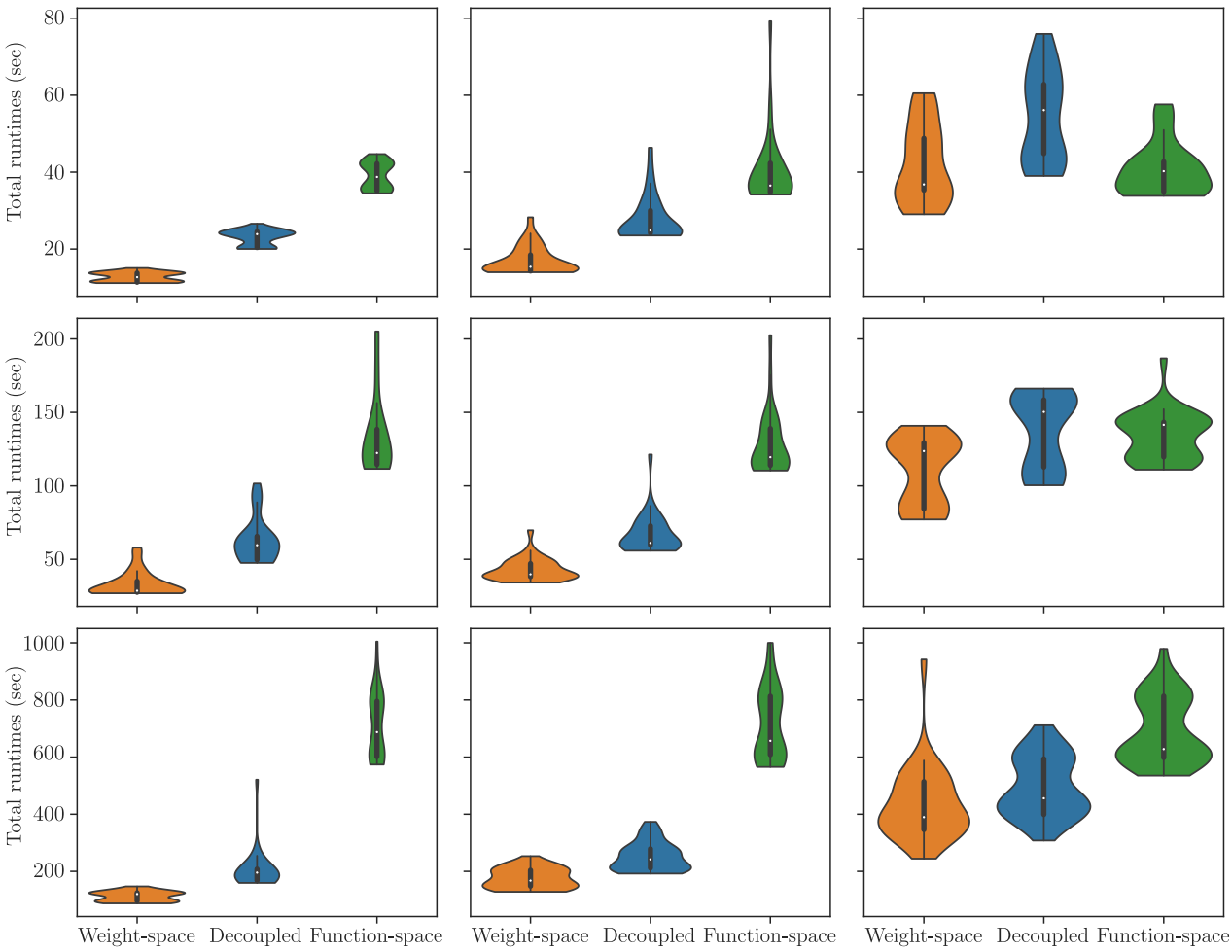

Figure 8: Empirical distributions of per trial runtimes for parallel TS with different sampling strategies; subplots are 1-to-1 with those in Figure 7.

need to separate out signal from noise, the training set consisted of $10,000$ uniform random training points and training was performed using stochastic gradient descent.

At test time, a baseline was constructed by iteratively drawing drift vectors $f_{t+1} \mid \boldsymbol{f}_{1:t}$. At each iteration, the current input $\boldsymbol{x}_t$ is added to the set of inducing locations $\mathbf{Z}_{t+1} = \mathbf{Z}_t \cup \{\boldsymbol{x}_t\}$ and the i-th inducing distribution is augmented to incorporate the sampled drift as

$$q_{t+1}^{(i)}(\boldsymbol{u}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_t^{(i)} \\ f_t^{(i)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_t - \boldsymbol{v}\boldsymbol{v}^\top & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix}\right) \tag{47}$$

where $\boldsymbol{v} = k_t(\boldsymbol{x}_t, \mathbf{Z}_t) k_t(\boldsymbol{x}_t, \boldsymbol{x}_t)^{-1/2}$ is defined in terms of the posterior covariance given the $m + t$ preceding inducing locations. When the inducing covariance is parameterized by its Cholesky factor, $\boldsymbol{\Sigma}_{t+1}^{1/2}$ can be directly computed via a rank-1 downdate (Gill et al., 1974; Seeger, 2004). Since only the m-th leading principal submatrix of $\boldsymbol{\Sigma}_{t+1}^{1/2}$ needs to be modified (the remaining terms are all zero because $\boldsymbol{f}_t$ is directly observed), this downdate incurs $\mathcal{O}(m^2)$ time complexity per iteration. In similar fashion, the prior covariance and its Cholesky factor may be maintained online. Here, however, as well as when computing posterior marginals, the matrices are no longer sparse, resulting in $\mathcal{O}((m+t)^2)$ cost per step. Overall, the iterative approach to unrolling scales cubically in the number of steps.