Posterior Consistency for Gaussian Process Surrogate Models with Generalized Observations

Rujian Chen
CSAIL
MIT
Cambridge, MA 02139
rjchen@csail.mit.edu

John W. Fisher
CSAIL
MIT
Cambridge, MA 02139
fisher@csail.mit.edu

Abstract

Gaussian processes (GPs) are widely used as approximations to complex computational models. However, properties and implications of GP approximations on data analysis are not yet fully understood. In this work we study parameter inference in GP surrogate models that utilize generalized observations, and prove conditions and guarantees for the approximate parameter posterior to be consistent in terms of posterior expectations and KL-divergence.

1 Introduction

Bayesian inference is increasingly applied to automated systems in science and engineering. Complex models can incur significant computation and/or preclude closed-form analysis. We refer to such models as *black box* models and note that Gaussian processes (GPs) are commonly used as surrogates [1–7]. For example, the *GP emulator* model uses a GP approximation f of the relation of inputs θ to the response of some *black box* $f_*(\theta)$. Subsequently, one may utilize f for inference over *unknown* θ, or more generally the expectation of some $g(\theta)$, conditioned on response observations.

Let $\mathbb{E}_f\left\{g(\theta)\right\}$ be the expectation over the *approximate* posterior distribution $p(\theta\,|\,y, D_n)$ arising from f. Similarly, let $\mathbb{E}_{f_*}\left\{g(\theta)\right\}$ be the expectation over the *exact* posterior distribution $p_*(\theta \mid y)$ arising from f_*. We would like to know under what conditions does $\mathbb{E}_f\left\{g(\theta)\right\} \to \mathbb{E}_{f_*}\left\{g(\theta)\right\}$. While there are a variety results that establish the convergence of $f \to f_*$ [8–11], we consider a more general case where observations are *linear functionals* of f_* (which includes direct observations of f_* as a special case). Our main result is to establish detailed conditions and characterizations for the consistency of approximate GP model posteriors such that

$$\int_{\boldsymbol{\Theta}} g(\theta)p(\theta \mid y, D_n)\,\mathrm{d}\theta \xrightarrow{p} \int_{\boldsymbol{\Theta}} g(\theta)p_*(\theta \mid y)\,\mathrm{d}\theta.$$

Characterization of approximation and convergence properties in GP models have received increasing attention recently. New results have been obtained both for GP posterior distributions and surrogate model posteriors with embedded GPs [9–12]. Typically, one assumes that the GP training data D and real data y are direct observations of f_* values. Recent work shows that inclusion of generalized observations (*e.g.,* gradients, integrals, and higher-ordered derivatives) is efficient and can improve model performance [13, 14]. In general, GPs can efficiently model a class of observations represented by linear functionals on the reproducing kernel Hilbert space (RKHS) [15, 16]. Generalized observations may arise in simulation as cheap by-products [13, 17, 18], and in the real world from sensors of various types, e.g. measuring averages, values or rates.

Existing convergence results, while sound in their original context, are insufficient for generalized observations. For example, results establishing sup-norm type convergence [9, 10, 19] are insufficient

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

for studying derivative functionals; and analyses in stronger norms on noise-free versions [12, 15] do not carry over straightforwardly. Our main contribution is to prove asymptotic consistency for the case of generalized observations subject to additive noise. We show, by way of example, that the guarantees extend to systems with more complex model structure. To our knowledge, the result on this formulation is novel. Compared to existing work restricted to direct observations, our formulation admits both additive noise and flexible kernel choice aligning better to practice.

Notation In this paper, k denotes a positive definite kernel and H_k denotes its associated RKHS. For an RKHS H, H^* denotes the continuous dual space of H, consisting of linear functionals over H. $\mathcal{X}$ and $\mathbf{\Theta}$ are subsets of $\mathbb{R}^d$ for some d. $\mathrm{N}(\cdot;\, \mu, \Sigma)$ is the Gaussian density with mean μ and covariance matrix Σ. For a GP f, a linear functional L and data D, $m_{LD} = \mathbb{E}\left[L(f) \mid D\right]$ and $K_{LD} = \mathrm{Cov}\left[L(f) \mid D\right]$ are the conditional mean and variance of $L(f)$ given D. $\mathrm{KL}[\cdot \,\|\, \cdot]$ and $\mathrm{H}[\cdot]$ denote KL-divergence and entropy.

2 Essential problem formulation

We formulate a minimal problem for the purpose of establishing our main result. Subsequently, we expand to more complex problem structures with straightforward extension of the key analysis.

Exact model Consider a system with unknown parameters $\theta \in \mathbf{\Theta}$ and blackbox component f_*. Let $p(\theta)$ be the prior distribution over θ. Let f_* be a function over domain $\mathcal{X}$ that we may only query via an inefficient noisy simulator. We assume measurements of the system are given by a θ-dependent linear functional $L_\theta \in H_k^*$ operating on f_*, where k is some positive definite kernel. For example, L_θ can be the evaluation functional at the location θ: $L_\theta(f_*) = f_*(\theta)$, or a directional derivative functional along a vector v: $L_\theta(f_*) = \nabla_v f_*(\theta)$. More generally, θ can be an index into a set of linear functionals $\{L_\theta : \theta \in \mathbf{\Theta}\}$ where $\mathbf{\Theta}$ does not lie in $\mathcal{X}$. We assume the dependence of L_θ on θ is known and deterministic, and can be of arbitrary form. $u = L_\theta(f_*)$ denotes the black box's response to L_θ and $y = u + \xi$ is the measurement with noise $\xi \sim \mathcal{N}(0, \sigma_Y^2)$.

We assume each simulator query is a linear functional $\lambda \in H_k^*$ and the simulator outputs $\bar{y} = \lambda(f_*) + \epsilon$ with Gaussian noise ϵ. We use $D = \{(\lambda_j, \bar{y}_j)\}_{j=1}^{|D|}$ to denote a simulated dataset of $|D|$ observations.

Approximate model The surrogate model approximates f_* with a GP f with prior $\mathcal{GP}(0, k)$. It models the simulated data D as $\bar{y}_j = \lambda_j(f) + \epsilon_j$ and the system response as $u = L_\theta(f)$. We note that this model assumes D and u as generated by the GP f, as opposed to f_* in the exact model. This allows the approximate model to use GP formulae in inference rather than invoking the simulator. We analyze the resulting approximation error as part of establishing conditions for our main result.

Posterior consistency problem The posterior of θ under the exact and approximate models are:

$$\begin{aligned} &\text{Exact:} && p_*(\theta \mid y) \propto p(\theta) l_*(\theta;\, y) && \text{where } l_*(\theta;\, y) = \mathrm{N}\left(y;\, L_\theta(f_*), \sigma_Y^2\right) \\ &\text{Approx:} && p(\theta \mid y, D) \propto p(\theta) l(\theta;\, y, D) && \text{where } l(\theta;\, y, D) = \mathrm{N}\left(y;\, m_{L_\theta D}, K_{L_\theta D} + \sigma_Y^2\right) \end{aligned}$$

Due to the random noise in D, the approximate posterior distribution is a random function over θ. From an inference perspective, a desirable property is for the approximate posterior to converge to the exact one in some sense, as the GP approximation gets refined:

$$p(\cdot \mid y, D_n) \to p_*(\cdot \mid y) \quad \text{as } n \to \infty$$

where D_n is a sequence of increasingly refined datasets. In practice, one typically uses the posterior to compute summary statistics for downstream analysis. For a summary statistic functional F, we would like to have in some sense:

$$F\left[p(\cdot \mid y, D_n)\right] \to F\left[p_*(\cdot \mid y)\right] \quad \text{as } n \to \infty.$$

Viewing the approximate posterior as a (random) estimator of the exact posterior, we deem it consistent if such properties hold. While one might expect consistency intuitively, careful analysis is typically required to characterize the notions of consistency and their conditions for nonparametric models.

3 Results

In this section we present detailed conditions and characterizations for the consistency of approximate GP model posteriors. Our key intermediate result is a set of conditions for the consistency of the conditional distribution $L(f)|D_n$, given in Proposition 3.1 after stating needed assumptions. All proofs are given in the appendix.

Assumption 3.1. *For every $a \in \mathcal{X}$ and coordinate i , $\frac{\partial}{\partial x_i}|_{x=a} \in H_k^*$.*

Let $B_r(x)$ be the r-radius ball centered at x. Let $N_{x,r,n}$ be the number of evaluation functionals in D_n evaluating at points in $B_r(x)$. Let $R_{x,r,n} = N_{x,r,n}/|D_n|$.

Assumption 3.2. *$|D_n| \to \infty$ as $n \to \infty$ and for all $x \in \mathcal{X}$ and $r > 0$, $\liminf_{n\to\infty} R_{x,r,n} > 0$.*

Assumption 3.3. *Each D_n is a collection of pairs $\{(\lambda_{nj}, \bar{y}_{nj})\}_{j=1}^{|D_n|}$ where $\bar{y}_{nj} = \lambda_{nj}(f_*) + \epsilon_{nj}$ and λ_{nj} are linear operators. $\epsilon_{nj} \sim \mathcal{N}(0, \sigma^2_{\epsilon nj})$ are independent with $\sigma_{\epsilon nj} \leq C_1$ for some constant C_1.*

Assumption 3.1 is a mild assumption on the smoothness of the kernel which is satisfied by many commonly used kernels. Assumption 3.2 roughly requires each neighborhood in the GP domain be covered by a non-diminishing fraction of training samples.

Proposition 3.1. *Let k be a positive definite kernel over $\mathcal{X}$. Let $f \sim \mathcal{GP}(0, k)$. Let $L \in H_k^*$ be a linear functional. Let $\{D_n\}_{n=1}^\infty$ be a sequence of datasets generated from a fixed function $f_* \in H_k$. If k, f_* and $\{D_n\}_{n=1}^\infty$ satisfy Assumptions 3.1, 3.2 and 3.3, then $\mathbb{E}\left[m_{LD_n} - L(f_*)\right]^2 \to 0$ and $K_{LD_n} \to 0$ as $n \to \infty$.*

Proposition 3.1 provides a sense of pointwise consistency for the unnormalized approximate posterior $p(\theta)l(\theta;\, y, D_n)$. Our main result strengthens it to establish a sense of consistency for the (normalized) posterior $p(\theta \mid y, D_n)$ and posterior expectations:

Theorem 3.1. *Let k, $\mathcal{X}$, f_* and $\{D_n\}_{n=1}^\infty$ be defined as in Proposition 3.1 and satisfy Assumptions 3.1, 3.2 and 3.3. If $L_\theta \in H_k^*$ for every $\theta \in \mathbf{\Theta}$, then the approximate and exact posteriors satisfy for any measurable $g(\theta)$:*

$$\int_{\mathbf{\Theta}} g(\theta) p(\theta \mid y, D_n)\, d\theta \xrightarrow{p} \int_{\mathbf{\Theta}} g(\theta) p_*(\theta \mid y)\, d\theta.$$

If in addition $\|L_\theta\|_{H_k^*} \leq C$ for all $\theta \in \mathbf{\Theta}$ for some C,

$$\mathrm{KL}\left[p(\cdot \mid y, D_n) \,\|\, p_*(\cdot \mid y)\right] \xrightarrow{p} 0 \quad \textit{and} \quad \mathrm{H}\left[p(\cdot \mid y, D_n)\right] \xrightarrow{p} \mathrm{H}\left[p_*(\cdot \mid y)\right].$$

Theorem 3.1 guarantees that posterior expectations, such as moments and event probabilities which are widely used in practice, are consistent. Posterior entropy often appears in information-theoretic decision making. An intermediate proof result shows that the partition function $Z_n = \int p(\theta) l(\theta;\, y, D_n)\, \mathrm{d}\theta$ is consistent.

The theorem's conditions align with many common modeling choices. For example, we allow simulator noise to be assumed if needed; there is no restriction on the form of the kernel; and $\mathbf{\Theta}$ does not have to be compact, which removes the need for domain truncation.

Interestingly, consistency holds even if simulated and real observations come from different linear functionals (as long as theorem assumptions hold). For example, one may use a function value simulator and take real measurements with a gradient sensor. The following extended model shows that simulated and real data may both contain different sensor types. This feature offers flexibility to simulation and experiment design.

4 Extended model

Our guarantees extend to a larger class of models that can represent multiple-black-box systems with more complex structure. Figure 1 illustrates its graphical model. We note that the parameterization of linear functionals by θ can be very flexible, for example expressing shared and idiosyncratic structure across components, in which case $\mathbf{\Theta}$ may not be a simple Cartesian product of all the GP domains. Sensor measurements may combine different component responses and/or some coordinates of θ.

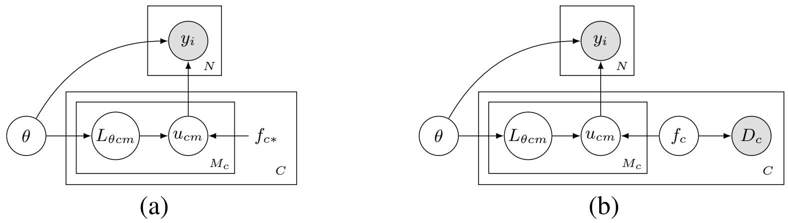

Figure 1: (a) Exact model with intractable black box models f_{c*}. (b) Approximate model with GP approximations f_c with training data D_c generated from f_{c*}.

Specifically, the exact model (Figure 1a) contains C black box components. Each component c has exact function f_{c*} over domain $\mathcal{X}_c$, with component responses $u_{cm} = L_{\theta cm}(f_{c*})$ generated by M_c linear functionals $\{L_{\theta cm}\}_{m=1}^{M_c}$. Measurements are $y = V^T\theta + W^T u + \xi$ with known weights V, W.

The approximate model (Figure 1b) approximates each black box component by a GP $f_c \sim \mathcal{GP}(0, k_c)$. Simulation data and response variables are modeled as generated from the GP: $D_c = \{\lambda_{cj}, \bar{y}_{cj}\}_{j=1}^{|D_c|}$ where $\bar{y}_{cj} = \lambda_{cj}(f_c) + \epsilon_{cj}$, and $u_{cm} = L_{\theta cm}(f_c)$.

The exact and approximate posteriors are:

$$p_*(\theta \mid y) \propto p(\theta) l_*(\theta;\, y) \qquad \text{where } l_*(\theta;\, y) = \mathrm{N}\left(y;\, V^T\theta + W^T L_\theta(f_*), \sigma_Y^2 I\right)$$

$$p(\theta \mid y, D) \propto p(\theta) l(\theta;\, y, D) \qquad \text{where } l(\theta;\, y, D) = \mathrm{N}\left(y;\, V^T\theta + W^T m_{L_\theta D}, W^T K_{L_\theta D} W + \sigma_Y^2 I\right)$$

where $D = \{D_1, \ldots, D_C\}$; $L_\theta(f_*)$ is the vector of exact functional values; $m_{L_\theta D}$ and $K_{L_\theta D}$ are the conditional means and covariances of all $L_{\theta cm}(f_c)$'s given D (details are given in appendix B).

Theorem 4.1. *Let $k_c, \mathcal{X}_c, f_{c*}$, $c = 1 \ldots C$, be defined as above. For each c, let $\{D_{cn}\}_{n=1}^{\infty}$ be a sequence of datasets generated by $f_{c*} \in H_{k_c}$. If for each θ and c, for each $1 \leq m \leq M_c$, $L_{\theta cm} \in H_{k_c}^*$ and $k_c, f_{c*}, \{D_{cn}\}_{n=1}^{\infty}$ satisfy Assumptions 3.1, 3.2 and 3.3, then the approximate and exact posteriors satisfy for any measurable $g(\theta)$:*

$$\int_{\mathbf{\Theta}} g(\theta) p(\theta \mid y, D_n)\, d\theta \xrightarrow{p} \int_{\mathbf{\Theta}} g(\theta) p_*(\theta \mid y)\, d\theta$$

Further suppose $p(\theta)$ has up to 4th finite moments if $V \neq 0$; then if $\|L_{\theta cm}\|_{H_k^*} \leq C$ for all θ, c, m for some C,

$$\mathrm{KL}\left[p(\cdot \mid y, D_n) \,\|\, p_*(\cdot \mid y)\right] \xrightarrow{p} 0 \quad \textit{and} \quad \mathrm{H}\left[p(\cdot \mid y, D_n)\right] \xrightarrow{p} \mathrm{H}\left[p_*(\cdot \mid y)\right].$$

5 Some comments on the result

We have proven consistency guarantees for posteriors inference in structured GP surrogate models with generalized measurements. The restriction to *linear* functionals poses a mild limitation. However, Sec. 4 provides an extension (by no means exhaustive) to a set of expressive models, while Sec. 1 discusses a variety of relevant applications fitting this assumption. Practically, the restriction to *linear* functionals precludes *nonlinear* sensor models. Local linearization methods might yield a means to extend the guarantees to some classes of nonlinear models, but that is speculative at present.

The primary distinction between our result and related work is the incorporation of generalized measurements. For example, Stuart and Teckentrup [12] are also motivated by parameter inference via GP approximations. They show posterior convergence in Hellinger distance, but with a more restrictive assumption of direct, noise-free GP observations. Wendland [15] provides an error analysis framework for the GP mean function, but similarly restricted.

Lederer [9, 10] derives probabilistic error bounds on the posterior GP mean for the noisy data case, but assumes direct observations. They also show sup-norm convergence for all continuous functions – not just those in the RKHS. Our consistency result, due to the assumption of linear functionals, does not hold for this enlarged set. Finally, Wynne *et al* [11] prove posterior mean consistency with respect to expected Sobolev norm, but restricted to Matérn kernels, whereas we do not restrict kernel forms. To our knowledge, our results provide new guarantees for a wide array of complex model structures combined with generalized measurements as compared to guarantees limited to direct observations.

References

[1] T. J. Santner, B. J. Williams, W. I. Notz, and B. J. Williams, *The design and analysis of computer experiments*, vol. 1. Springer, 2003.

[2] J. Oakley and A. O'Hagan, "Bayesian inference for the uncertainty distribution of computer model outputs," *Biometrika*, vol. 89, pp. 769–784, dec 2002.

[3] P. Di Achille, A. Harouni, S. Khamzin, O. Solovyova, J. J. Rice, and V. Gurev, "Gaussian process regressions for inverse problems and parameter searches in models of ventricular mechanics," *Frontiers in Physiology*, vol. 9, p. 1002, aug 2018.

[4] L. M. Paun and D. Husmeier, "Markov chain Monte Carlo with Gaussian processes for fast parameter estimation and uncertainty quantification in a 1D fluid-dynamics model of the pulmonary circulation," *International Journal for Numerical Methods in Biomedical Engineering*, vol. 37, no. 2, 2021.

[5] S. Conti and A. O'Hagan, "Bayesian emulation of complex multi-output and dynamic computer models," *Journal of Statistical Planning and Inference*, vol. 140, pp. 640–651, mar 2010.

[6] S. Conti, J. P. Gosling, J. E. Oakley, and A. O'Hagan, "Gaussian process emulation of dynamic computer codes," *Biometrika*, vol. 96, pp. 663–676, sep 2009.

[7] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 63, pp. 425–464, jan 2001.

[8] T. Choi and M. J. Schervish, "Posterior Consistency in Nonparametric Regression Problmes Under Gaussia Process Priors," *Journal of Multivariate Analysis*, vol. 98, no. 10, pp. 1–35, 2006.

[9] A. Lederer, J. Umlauft, and S. Hirche, "Uniform Error and Posterior Variance Bounds for Gaussian Process Regression with Application to Safe Control," jan 2021.

[10] A. Lederer, J. Umlauft, and S. Hirche, "Uniform error bounds for Gaussian process regression with application to safe control," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] G. Wynne, F. X. Briol, and M. Girolami, "Convergence guarantees for gaussian process means with misspecified likelihoods and smoothness," *Journal of Machine Learning Research*, vol. 22, pp. 1–40, 2021.

[12] A. M. Stuart and A. L. Teckentrup, "Posterior consistency for Gaussian process approximations of Bayesian posterior distributions," *Mathematics of Computation*, vol. 87, no. 310, pp. 721–753, 2017.

[13] O.-P. Koistinen, E. Maras, A. Vehtari, and H. Jónsson, "Minimum energy path calculations with Gaussian process regression," *Nanosystems: Physics, Chemistry, Mathematics*, pp. 925–935, mar 2016.

[14] J. Wu, M. Poloczek, A. G. Wilson, and P. I. Frazier, "Bayesian optimization with gradients," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 5268–5279, 2017.

[15] H. Wendland, "Scattered Data Approximation," *Scattered Data Approximation*, dec 2004.

[16] G. Wahba, "Spline Models for Observational Data," *Spline Models for Observational Data*, jan 1990.

[17] R.-E. Plessix, "A review of the adjoint-state method for computing the gradient of a functional with geophysical applications," *Geophysical Journal International*, vol. 167, no. 2, pp. 495–503, 2006.

[18] A. Jameson, "Re-engineering the design process through computation," *Journal of Aircraft*, vol. 36, no. 1, pp. 36–50, 1999.

[19] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, vol. 2. MIT press Cambridge, MA, 2006.

[20] A. Gut and A. Gut, *Probability: a graduate course*, vol. 200. Springer, 2005.

[21] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

A Background on GPs and linear functionals on GPs

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite function. A Gaussian process (GP), $\mathcal{GP}(0, k)$, over a subset $\mathcal{X}$ of $\mathbb{R}^n$ is a collection of random variables $f(x)$ indexed by $x \in \mathcal{X}$, with jointly Gaussian finite dimensional distributions: $\mathbb{E}f(x) = 0$, $\mathbb{E}f(x_1)f(x_2) = k(x_1, x_2)$.

The positive definite function k, also called the kernel, defines a useful subspace H_k of $C(\mathcal{X})$, called the reproducing kernel Hilbert space (RKHS) of the GP (also called the native space or the Cameron-Martin space). Given a kernel k, H_k can be constructed by (1) defining a pre-Hilbert space $H'_k \subset C(\mathcal{X})$ consisting of all functions of the form $\sum_i^n a_i k(\cdot, x_i)$, where $n \in \mathbb{Z}_+$ and $a_i \in \mathbb{R}$, equipped with the inner product $\langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{H'_k} = k(x_i, x_j)$, and (2) taking H_k as the completion of H'_k. An RKHS H_k has the reproducing property: $\langle f, k(\cdot, x)\rangle_{H_k} = f(x)$ for $f \in H_k$.

Since H_k is a Hilbert space, there is an isometric map ϕ from its continuous dual space H_k^* to H_k. For any $L_0 \in H_k^*$, $h_0 = \phi(L_0)$ is a representer of L_0 in H_k; $L_0(h) = \langle h, h_0\rangle_{H_k}$ for all $h \in H_k$; and L_0 is a continuous function over H_k.

Suppose $f \sim \mathcal{GP}(0, k)$, H_k is the RKHS of k and $L \in H_k^*$. It is known that a sample f is not in H_k almost surely, so $u = L(f)$ is not well-defined for every f. However, u is well defined in an L_2 sense. There is a sequence of L_n's with $\phi(L_n) \in H'_k$ s.t. $u = \lim_{n\to\infty} L_n(f)$ with respect to the L_2 norm on the probability space. Each L_n is a finite linear combination of GP function values.

Let k be a positive definite kernel; $f \in \mathcal{GP}(0, k)$ and let $D = \{(\lambda_j, \bar{y}_j)\}_{j=1}^{|D|}$ be a dataset generated from f, where $\lambda_j \in H_k^*$ and $\bar{y}_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_j^2)$. Let $S = \text{diag}(\sigma_1^2, \ldots, \sigma_{|D|}^2)$. The mean and covariance functions of f conditioned on D are:

$$m_D(x) = k_\lambda(x)^T \left(K_{\lambda\lambda} + S\right)^{-1} \bar{y},$$
$$k_D(x_1, x_2) = k(x_1, x_2) - k_\lambda(x_1)^T \left(K_{\lambda\lambda} + S\right)^{-1} k_\lambda(x_2).$$

where $k_\lambda(x)$ is the column vector whose j-th entry is $\text{Cov}\left[\lambda_j(f), f(x)\right]$. $K_{\lambda\lambda}$ is the matrix whose i, j-th entry is $\text{Cov}\left[\lambda_i(f), \lambda_j(f)\right]$.

For a linear functional L, let $L^{(x)}$ denote the operation of applying L to its argument as a collection, indexed by other variables besides x, of functions of x. (Thus $L^{(x')}k(x, x')$ results in a function of only x.) Then $\text{Cov}\left[\lambda_j(f), f(x)\right] = \lambda_j^{(x')}k(x', x)$ and $\text{Cov}\left[\lambda_i(f), \lambda_j(f)\right] = \lambda_j^{(x'')}\lambda_i^{(x')}k(x', x'')$.

Given a linear functional $L \in H_k^*$, the conditional mean m_{LD}, and the conditional variance K_{LD}, of the random variable $L(f)$ given D may be computed as $m_{LD} = L(m_D)$, $K_{LD} = L^{(x_2)}L^{(x_1)}k_D(x_1, x_2)$.

B Extended model details

Here we give detailed definitions of the exact functional value vector $L_\theta(f_*)$, the conditional mean vector $m_{L_\theta D}$ and the conditional covariance matrix $K_{L_\theta D}$ for the extended model defined in section 4. In this model, there are C components and each component c contains M_c linear functionals $\{L_{\theta cm}\}_{m=1}^{M_c}$. Each component c has a simulation dataset D_c generated from the exact f_{c*}.

We order vector and matrix elements by first grouping by components and then concatenating different groups. Namely, for a component c, define $L_{\theta c}(f_c) = [L_{\theta c1}(f_c), \ldots, L_{\theta cM_c}(f_c)]^T$, and similarly define $L_{\theta c}(f_{c*}) = [L_{\theta c1}(f_{c*}), ..., L_{\theta cM_c}(f_{c*})]^T$. Then $m_{L_\theta D} = \text{Cat}\big(\mathbb{E}\left[L_{\theta 1}(f_1) \mid D\right], \ldots, \mathbb{E}\left[L_{\theta C}(f_C) \mid D\right]\big)$, and $L_\theta(f_*) = \text{Cat}\big(L_{\theta 1}(f_{1*}), ..., L_{\theta M_c}(f_{M_c *})\big)$, where Cat means concatenation of column vectors. Letting $K_{L_{\theta c}D}$ be the matrix whose i, j-th entry is $\text{Cov}\left[L_{\theta ci}(f), L_{\theta cj}(f) \mid D\right]$, then $K_{L_\theta D}$ is a block diagonal matrix $\text{diag}\left(K_{L_{\theta 1}D}, \ldots, K_{L_{\theta C}D}\right)$.

C Proofs

In the proofs below, we write H for H_k, and similarly H^* for H_k^*, when the associated kernel k is unambiguous for notational simplicity.

In the following we collect some observations about linear functionals of GPs which will be used later.

Observation C.1. *Let H be the RKHS of a positive definite kernel k. Let $L \in H^*$ be a linear function on H. Let $f \sim \mathcal{GP}(0, k)$. Let $g(x) = \mathrm{Cov}\left[L(f), f(x)\right]$, then $g \in H$ and $\|g\|_H^2 = \mathrm{Cov}\left[L(f)\right]$.*

Proof. Using RKHS theory (see Appendix A and references there), we may choose a sequence of $L_n = \sum_{i=1}^{N_i} c_{ni}\delta_{x_{ni}}(\cdot)$ s.t. $\mathrm{Cov}\left[L_n(f) - L(f)\right] \to 0$ as $n \to \infty$, where $\delta_x(\cdot)$ denotes the evaluation functional at x. By Cauchy-Schwartz inequality we also have $\mathrm{Cov}\left[(L_n - L)(f), L(f)\right] \to 0$ and $\mathrm{Cov}\left[L_n(f)\right] = \mathrm{Cov}\left[(L_n - L)(f) + L(f)\right] = \mathrm{Cov}\left[(L_n - L)(f)\right] + \mathrm{Cov}\left[L(f)\right] + 2\,\mathrm{Cov}\left[(L_n - L)(f), L(f)\right] \to \mathrm{Cov}(L(f))$ as $n \to \infty$.

Let $g_n(x) = \mathrm{Cov}(L_n(f), f(x)) = \sum_{i=1}^{N_i} c_{ni} k(x_{ni}, x)$. The g_n sequence is Cauchy in H since L_n is Cauchy in H^* and

$$
\begin{aligned}
&\|g_n(\cdot) - g_m(\cdot)\|_H^2 = \|\sum_{i=1}^{N_n} c_{ni} k(x_{ni}, \cdot) - \sum_{i=1}^{N_m} c_{mi} k(x_{mi}, \cdot)\|_H^2 \\
&= \sum_{i=1}^{N_n}\sum_{j=1}^{N_n} c_{ni}c_{nj}k(x_{ni}, x_{nj}) + \sum_{i=1}^{N_m}\sum_{j=1}^{N_m} c_{mi}c_{mj}k(x_{mi}, x_{mj}) - 2\sum_{i=1}^{N_n}\sum_{j=1}^{N_m} c_{ni}c_{mj}k(x_{ni}, x_{mj}) \\
&= \mathrm{Cov}\left(\sum_{i=1}^{N_n} c_{ni} f(x_{ni})\right) + \mathrm{Cov}\left(\sum_{i=1}^{N_m} c_{mi} f(x_{mi})\right) - 2\,\mathrm{Cov}\left(\sum_{i=1}^{N_n} c_{ni} f(x_{ni})\right) \\
&= \mathrm{Cov}\left(\sum_{i=1}^{N_n} c_{ni} f(x_{ni}) - \sum_{i=1}^{N_m} c_{mi} f(x_{mi})\right) = \mathrm{Cov}\left((L_n(f) - L_m(f)\right). \qquad (1)
\end{aligned}
$$

Since $g_n \to g$ pointwise, g must be the H-limit of g_n, so $g \in H$.

$$
\|g\|_H^2 = \lim_{n\to\infty} \|g_n\|_H^2 = \lim_{n\to\infty} \mathrm{Cov}\left[L_n(f)\right] = \mathrm{Cov}\left[L(f)\right]
$$

where the second equality uses (1) with $g_m = 0$. ■

C.1 Proof of Proposition 3.1

Notation For an RKHS with kernel k over domain $\mathcal{X}$. For $a \in X$, define $k_a(x) = k(a, x)$. H' denotes the pre-Hilbert space of H, which contains functions of the form $f(x) = \sum_{i=1}^{N} c_i k_{a_i}(x)$ for finite N. For a dataset $D_n = \{(\lambda_{nj}, \bar{y}_{nj})\}_{j=1}^{|D_n|}$, let $k_{\lambda_{nj}}(\cdot) = \lambda_{nj}^{(x)}(k(x, \cdot))$ be the representer in H of λ_{nj}. The superscript (x) means applying the funtional along the dummy variable x. Let $k_{\lambda_n}(x)$ be the column vector of all $k_{\lambda_{nj}}(x)$'s. Let S_n be the diagonal matrix $\mathrm{diag}([\sigma_{n1}^2, \ldots, \sigma_{n|D_n|}^2])$. Let $K_{\lambda_n\lambda_n S_n} = K_{\lambda_n\lambda_n} + S_n$.

Proof. Overview: We will prove $\mathbb{E}\left[m_{LD_n} - L(f_*)\right]^2 \to 0$ as $n \to \infty$ Steps 1 and 2 below. We will show $K_{LD_n} \to 0$ in Step 3.

Step 1

We separate the conditional GP mean m_{D_n} into a deterministic part m_{fD_n} and a random part $m_{\epsilon D_n}$:

$$
\begin{aligned}
m_{D_n}(x) &= k_{\lambda_n}^T(x) K_{\lambda_n\lambda_n S_n}^{-1} \bar{y}_n \\
&= k_{\lambda_n}^T(x) K_{\lambda_n\lambda_n S_n}^{-1} \lambda_n(f_*) + k_{\lambda_n}^T(x) K_{\lambda_n\lambda_n S_n}^{-1} \epsilon \\
&:= m_{fD_n}(x) + m_{\epsilon D_n}(x)
\end{aligned}
$$

where we defined $m_{fD_n} = k_{\lambda_n}^T(x) K_{\lambda_n\lambda_n S_n}^{-1} \lambda_n(f_*)$ and $m_{\epsilon D_n} = k_{\lambda_n}^T(x) K_{\lambda_n\lambda_n S_n}^{-1} \epsilon$. For the deterministic part, we will show

$$
\|m_{fD_n} - f_*\|_H \to 0 \quad \text{as } n \to \infty
$$

which implies $L(m_{fD_n} - f_*) \to 0$ by the fact that L is continuous on H. We will show this for $f_* \in H'$ in Step 1a and for $f_* \in H$ in Step 1b. For the random part, we will show in Step 2:

$$\mathbb{E}\left[L(m_{\epsilon D_n})\right]^2 \to 0 \quad \text{as } n \to \infty.$$

Since $m_{LD_n} = L(m_{D_n})$, combining the two parts will give $\mathbb{E}\left[m_{LD_n} - L(f_*)\right]^2 \to 0$.

Step 1a

Suppose $f_*(x) = \sum_{i=1}^N c_i k_{a_i}(x)$. $m_{fD_n}(x)$ has the form $m_{fD_n}(x) = \sum_{j=1}^{|D_n|} d_j k_{\lambda_{nj}}(x)$, where $d = K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n a} c$. Then the RKHS norm of the error is:

$$\begin{aligned}\|f_* - m_{fD_n}\|_H^2 &= \|\sum_i c_i k_{a_i} - \sum_j d_j k_{\lambda_{nj}}\|_H^2 \\ &= c^T K_{aa} c + d^T K_{\lambda_n\lambda_n} d - 2c^T K_{a\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n a} c \\ &= A + B_n - 2E_n \end{aligned} \tag{2}$$

where we defined $A = c^T K_{aa} c$, $B_n = d^T K_{\lambda_n\lambda_n} d$, $E_n = c^T K_{a\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n a} c$. Compute

$$B_n - E_n = c^T K_{a\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n}(-S) K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n a} c \le 0, \tag{3}$$

so the error norm is bounded by:

$$0 \le \|f_* - m_{fD_n}\|_H^2 \le A - E_n. \tag{4}$$

Combining (3) and (4),

$$B_n \le E_n \le A. \tag{5}$$

Since $f_*(x) = \sum_{i=1}^N c_i k_{a_i}(x)$, we can check that $f_*(a) = K_{aa} c$ and $\lambda_n(f_*) = K_{\lambda_n a} c$. Then the upper bound $A - E_n$ can be expressed as a weighted sum of residues:

$$A - E_n = c^T f_*(a) - c^T K_{a\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} \lambda_n(f_*) = c^T (f_*(a) - m_{fD_n}(a)).$$

Since c is a fixed vector, it will suffice to show $f_*(a_i) - m_{fD_n}(a_i) \to 0$ for each a_i, $i = 1, \ldots, N$.

Suppose this is not the case for some a_i. The variational definition of m_{fD_n} is:

$$m_{fD_n} = \underset{g_n \in H}{\arg\min}\, R(g_n), \text{ where } R(g_n) = \frac{1}{|D_n|}\|g_n\|_H^2 + \frac{1}{|D_n|}\sum_{j=1}^{|D_n|}\left(\frac{\lambda_{nj}(f_*) - \lambda_{nj}(g_n)}{\sigma_{nj}}\right)^2.$$

We note if any $\sigma_{nj} = 0$, the above holds with the convention $(x/0)^2 = 0$ if $x = 0$ and ∞ if $x \neq 0$. So $R(m_{fD_n}) \le R(f_*)$, which gives the inequality:

$$\frac{1}{|D_n|}\sum_{j=1}^{|D_n|}\left(\frac{\lambda_{nj}(f_*) - \lambda_{nj}(m_{fD_n})}{\sigma_{nj}}\right)^2 \le R(m_{fD_n}) \le R(f_*) = \frac{1}{|D_n|}\|f_*\|_H^2.$$

By Assumption 3.3, $\sigma_{nj} \le C_1$, so

$$\frac{1}{|D_n|}\sum_{j=1}^{|D_n|}\left(\lambda_{nj}(f_*) - \lambda_{nj}(m_{fD_n})\right)^2 \le \frac{C_1^2}{|D_n|}\|f_*\|_H^2, \tag{6}$$

so the average squared residues over λ_n goes to 0, as $|D_n| \to \infty$. It is also easy to check that the above holds if any $\sigma_{nj} = 0$. Next we will show that at each $a \in \mathcal{X}$, $f_* - m_{fD_n}$ have uniformly bounded gradient $\nabla(f_* - m_{fD_n})(a)$ over n, and deduce that the inequality above is false, reaching a contradiction.

Let $D_{p,a} = \frac{\partial}{\partial x_p}|_{x=a}$. By assumption, $D_{p,a} \in H^*$ for all $a \in \mathcal{X}$. Then $|\frac{\partial f_*}{\partial x_p}(a)| = |D_{p,a}(f_*)| \le \|D_{p,a}\|_{H^*}\|f_*\|_H$. Since $B_n = \|m_{fD_n}\|$, $A = \|f_*\|$, and $B_n \le A$ by inequality (5), $|\frac{\partial m_{fD_n}}{\partial x_p}(a)| = |D_{p,a}(m_{fD_n})| \le \|D_{p,a}\|_{H^*}\|m_{fD_n}\|_H \le \|D_{p,a}\|_{H^*}\|f_*\|_H$. Thus

$|\frac{\partial(f_* - m_{fD_n})}{\partial x_p}(a)| \leq 2\|D_{p,a}\|_{H^*}\|f_*\|_H$. Therefore, at each $a \in \mathcal{X}$ there exists a gradient bound $D(a)$ independent of n s.t.

$$\|\nabla(f_* - m_{fD_n})(a)\| \leq D(a).$$

By hypothesis, there exists a $\delta > 0$ s.t. $|f_*(a_i) - m_{fD_{n_k}}(a_i)| > \delta$ for $k = 1, 2, \ldots$. Choosing $r = \frac{\delta}{2D(a_i)}$, by the gradient bound,

$$|(f_* - m_{fD_{n_k}})(x)| \geq \delta/2, \text{ for } x \in B_r(a_i).$$

By Assumption 3.2, there exists a $\rho > 0$ and $N \in \mathbb{N}$ s.t. $R_{a_i,r,n} > \rho$ for all $n \geq N$. Let $\{\bar{x}_{nj}, f_*(\bar{x}_{nj})\}_{j=1}^{N_{a_i,r,n}}$ denote the evaluation functional data in D_n in the r-ball of a_i, then

$$N_{a_i,r,n} \geq \rho|D_n|, \text{ for } n \geq N.$$

The above two inequalities, and the fact that $\bar{x}_{nj} \in B_r(a_i)$, give

$$\begin{aligned}\frac{1}{|D_{n_k}|}\sum_{j=1}^{|D_{n_k}|}\Big(\lambda_{n_kj}(f_*) - \lambda_{n_kj}(m_{fD_{n_k}})\Big)^2 &\geq \frac{1}{|D_{n_k}|}\sum_{j=1}^{N_{a_i,r,n_k}}[(f_* - m_{fD_{n_k}})(\bar{x}_{n_kj})]^2\\ &\geq \frac{1}{|D_{n_k}|}\rho|D_{n_k}|\delta/2 = \rho\delta/2, \text{ for } n_k \geq N, k \geq 1.\end{aligned}$$

which contradicts (6), thus Step 1a is proven.

Step 1b

In this step we show $\mathbb{E}(m_{LD_n} - L(f_*))^2 \to 0$ for a general $f_* \in H$.

Let $f_* \in H$. Then $f_* = \lim_{i\to\infty} f_i$ in H, where $f_i(x) = c_i^T k_{a_i}(x) = \sum_{j=1}^{N_i} c_{ij}k_{a_{ij}}(x)$. Let m_{ifD_n} be the conditional GP mean function data $\{\lambda_{nj}, \lambda_{nj}(f_i)\}_{j=1}^{|D_n|}$ generate from the f_i. m_{fD_n} and m_{ifD_n} can be expressed by:

$$m_{fD_n}(x) = k_{\lambda_n}(x)^T d, \quad m_{ifD_n}(x) = k_{\lambda_n}(x)^T d_i.$$

where $d = K^{-1}_{\lambda_n\lambda_n S_n}\lambda_n(f_*)$ and $d_i = K^{-1}_{\lambda_n\lambda_n S_n}\lambda_n(f_i)$.

Since each $\lambda \in H^*$ is a continuous functional on H, $\lim_{i\to\infty} d_i = d$, so for fixed n, m_{fD_n} is the H-limit of m_{ifD_n}:

$$\lim_{i\to\infty}\|m_{fD_n} - m_{ifD_n}\|_H^2 = \lim_{i\to\infty}\|k_{\lambda_n}(x)^T(d - d_i)\|_H^2 = \lim_{i\to\infty}(d - d_i)^T K_{\lambda_n\lambda_n}(d - d_i) = 0. \quad (7)$$

Define

$$\begin{aligned}A_i &= \|f_i\|_H^2 = c_i^T K_{aa} c_i,\\ E_{in} &= \langle f_i, m_{ifD_n}\rangle_H = c_i^T K_{a_i\lambda_n}K^{-1}_{\lambda_n\lambda_n S_n}K_{\lambda_n a_i}c_i,\\ A &= \|f_*\|,\\ E_n &= \langle f_*, m_{fD_n}\rangle_H.\end{aligned}$$

$f_i \to f_*$ in H and equation (7) imply

$$\begin{aligned}A = \|f_*\| &= \lim_{i\to\infty} A_i,\\ E_n = \langle f_*, m_{fD_n}\rangle_H &= \lim_{i\to\infty} E_{in}.\end{aligned}$$

Using bound 4 for each i, the target error norm can be bounded by:

$$\begin{aligned}\|f - m_{fD_n}\|_H^2 &= \lim_{i\to\infty}\|f_i - m_{ifD_n}\|_H^2 \leq \lim_{i\to\infty} A_i - E_{in} = A - E_n\\ &\leq |A - A_i| + |A_i - E_{in}| + |E_{in} - E_n|\end{aligned}$$

Suppose $\|f_i - f_*\|_H \le \epsilon$ for some i, the first term can be bounded:

$$
\begin{aligned}
|A_i - A| &= \left|\|f_i\|_H^2 - \|f_*\|_H^2\right| = \left|\|(f_i - f_*) + f_*\|_H^2 - \|f_*\|_H^2\right| \\
&\le \left|\|f_i - f_*\|_H^2 + 2\langle f_i - f_*, f_*\rangle_H\right| \le \|f_i - f_*\|_H^2 + 2\|f_i - f_*\|_H\|f_*\|_H \\
&\le \epsilon^2 + 2\epsilon\|f_*\|_H.
\end{aligned}
$$

Next we show that for fixed i, $|E_{in} - E_n|$ can be uniformly bounded over n. Define a semi-inner product Q_{E_n} on H' by:

$$
\langle f, g\rangle_{Q_{E_n}} = \lambda_n(f)^T K^{-1}_{\lambda_n\lambda_n S_n}\lambda_n(g) \quad \text{and} \quad \|f\|^2_{Q_{E_n}} = \langle f, f\rangle_{Q_{E_n}},
$$

for $f, g \in H'$. We can check that Q_{E_n} is binear and positive semidefinite, and is a valid semi-inner product on H'. Q_{E_n} is also continuous on $H' \times H'$, so Q_{E_n} can be extended to a semi-inner product on H. We use Q_{E_n} in the following to denote this extended definition.

From inequality (4) in Step 1a, it follows that $\|f\|^2_{Q_{E_n}} \le \|f\|_H^2$ for all $f \in H'$. By continuity, $\|f\|^2_{Q_{E_n}} \le \|f\|_H^2$ for actually all $f \in H$. Expressing E_{in} as $\|f_i\|^2_{Q_{E_n}}$ and E_n as $\|f_*\|^2_{Q_{E_n}}$, we have the bound:

$$
\begin{aligned}
|E_{in} - E_n| &= \left|\|f_i\|^2_{Q_{E_n}} - \|f_*\|^2_{Q_{E_n}}\right| \le \|f_i - f_*\|^2_{Q_{E_n}} + 2\|f_i - f_*\|_{Q_{E_n}}\|f_*\|_{Q_{E_n}} \\
&\le \|f_i - f_*\|_H^2 + 2\|f_i - f_*\|_H\|f_*\|_H \le \epsilon^2 + 2\epsilon\|f_*\|_H.
\end{aligned}
$$

Thus, given an $\epsilon > 0$, we can choose i s.t. $|A - A_i| \le \epsilon^2 + 2\epsilon\|f_*\|_H$ and for all n, $|E_{in} - E_n| \le \epsilon^2 + 2\epsilon\|f_*\|_H$. For this value of i, choose N s.t. $|A_i - E_{in}| \le \epsilon$ for all $n \ge N$, which is possible from Step 1a. Then $A - E_n \le |A - A_i| + |A_i - E_{in}| + |E_{in} - E_n| \le 2\epsilon^2 + 4\epsilon\|f_*\|_H + \epsilon$ for all $n \ge N$, which means $\|f_* - m_{fD_n}\|_H^2 \le A - E_n \to 0$ as $n \to \infty$. Thus Step 1b is proven.

Intermediate results

From the arguments above, we have an intermediate result useful for later:

Lemma C.1. *Let $f \in H$. Define the semi-inner products Q_{E_n} and Q_{B_n} as:*

$$
\begin{aligned}
\langle f, g\rangle_{Q_{E_n}} &= \lambda_n(f)^T K^{-1}_{\lambda_n\lambda_n S_n}\lambda_n(g) \\
\langle f, g\rangle_{Q_{B_n}} &= \lambda_n(f)^T K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n}\lambda_n(g).
\end{aligned}
$$

Then $\|f\|_{Q_{E_n}} \le \|f\|_H$ and $\|f\|_{Q_{E_n}} \to \|f\|_H$ as $n \to \infty$. Similarly $\|f\|_{Q_{B_n}} \le \|f\|_H$ and $\|f\|_{Q_{B_n}} \to \|f\|_H$ as $n \to \infty$.

Proof. By noting that in Step 1b, $E_n = \|f_*\|^2_{Q_{E_n}}$, $A = \|f_*\|_H^2$, and using the intermediate result that $0 \le A - E_n \to 0$ as $n \to \infty$, we conclude that $\|f_*\|_{Q_{E_n}} \le \|f_*\|_H$ and $\|f_*\|_{Q_{E_n}} \to \|f_*\|_H$ as $n \to \infty$. Since $f_* \in H$ is arbitrary, the claim for Q_{E_n} follows.

B_n of equation (2) in Step 1a can be expressed as $B_n = \|f_*\|^2_{Q_{B_n}}$. By the positivity of (2) and inequality (5), $0 \le E_n - B_n \le A - E_n$. Since $A - E_n \to 0$, $E_n - B_n \to 0$. Thus $B_n \to E_n$ and $E_n \to A$, so $B_n \to A$. This shows the claims for Q_{B_n} for $f_* \in H'$, which can be extended to $f \in H$ by continuity. ■

Step 2

In this step we show $\mathbb{E}\left[L(m_{\epsilon D_n})\right]^2 \to 0$, which is equivalent to $\mathrm{Cov}\left[L(m_{\epsilon D_n})\right] \to 0$ since $\mathbb{E}\left[m_{\epsilon D_n}\right] = 0$, as $n \to \infty$. Our approach is bound the target quantity by analyzing the posterior distribution of the related Bayesian GP model. Specifically, let $f \sim \mathrm{GP}(0, k)$, $u = L(f)$ and $v_n = \lambda_n(f)$. Define $K_{u\lambda_n}$ as the row vector whose j-th entry is $\mathrm{Cov}\left[u, \lambda_{nj}(f)\right]$ and define $K_{\lambda_n u} = K^T_{u\lambda_n}$. Define random variables $\epsilon \sim \mathcal{N}(0, S_n)$ and $y = v_n + \epsilon$. Using the conditional distribution $p(u\,|\,y) = \mathrm{N}(u;\, K_{u\lambda_n}K^{-1}_{\lambda_n\lambda_n S_n}y, K_{uu|y})$, where $K_{uu|y} = K_{uu} - K_{u\lambda_n}K^{-1}_{\lambda_n\lambda_n S_n}K_{\lambda_n u}$, we have

$$p(u) = \mathbb{E}\, p(u \mid y) = \mathbb{E}\, p(u \mid v_n + \epsilon) = \int p(u \mid v_n + \epsilon) p(v_n, \epsilon)\, \mathrm{d}v_n\, \mathrm{d}\epsilon$$
$$= \int \mathrm{N}\left(u \mid K_{u\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n}(v_n + \epsilon), K_{uu|y}\right) p(v_n, \epsilon)\, \mathrm{d}v_n\, \mathrm{d}\epsilon.$$

Let z be a random variable defined as $z = K_{u\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n}(v_n + \epsilon) + \gamma$, where $\gamma \sim \mathcal{N}(0, K_{uu|y})$ and $\gamma \perp\!\!\!\perp (v_n, \epsilon)$. Then z clearly has the same density as u, so u is equal in distribution to z, i.e.:

$$u \overset{d}{=} K_{u\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} v_n + K_{u\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} \epsilon + \gamma$$

Since v_n, ϵ, γ are independent,

$$\mathrm{Cov}(u) = \mathrm{Cov}\left(K_{u\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} v_n\right) + \mathrm{Cov}\left(K_{u\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} \epsilon\right) + \mathrm{Cov}(\gamma)$$
$$= K_{u\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n u} + \mathrm{Cov}\left[L(m_{\epsilon D_n})\right] + \mathrm{Cov}(\gamma)$$

This gives an upper bound on the target quantity:

$$\mathrm{Cov}\left[L(m_{\epsilon D_n})\right] \leq \mathrm{Cov}(u) - K_{u\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n u} \tag{8}$$

By noting that $K_{u\lambda_n} = \lambda_n\left[\mathrm{Cov}(u, f(\cdot))\right]$, and using the semi-inner product Q_{B_n} defined in Lemma C.1, the target bound can be written as:

$$\mathrm{Cov}\left[L(m_{\epsilon D_n})\right] \leq \mathrm{Cov}(u) - \|\mathrm{Cov}\left[u, f(\cdot)\right]\|^2_{Q_{B_n}}.$$

Taking $\limsup$ and by Lemma C.1 and Observation C.1,

$$\limsup_{n\to\infty} \mathrm{Cov}\left[L(m_{\epsilon D_n})\right]$$
$$\leq \limsup_{n\to\infty} \mathrm{Cov}(u) - \|\mathrm{Cov}\left[u, f(\cdot)\right]\|^2_{Q_{B_n}}$$
$$= \mathrm{Cov}(u) - \|\mathrm{Cov}\left[u, f(\cdot)\right]\|^2_{H} = \mathrm{Cov}(u) - \mathrm{Cov}(u) = 0.$$

Therefore $\lim_{n\to\infty} \mathrm{Cov}\left[L(m_{\epsilon D_n})\right] = 0$ and Step 2 is proven.

Step 3

In this step we show $K_{LD_n} \to 0$ as $n \to \infty$. Let $u = L(f)$. The expression for the target quantity is

$$K_{LD_n} = \mathrm{Cov}(u \mid D_n) = \mathrm{Cov}(u) - K_{u\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n u}.$$

Observing that $K^{-1}_{\lambda_n\lambda_n S_n} - K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} = K^{-1}_{\lambda_n\lambda_n S_n} S_n K^{-1}_{\lambda_n\lambda_n S_n} \succcurlyeq 0$, the target quantity can be bounded by

$$K_{LD_n} \leq \mathrm{Cov}(u \mid D_n) \leq \mathrm{Cov}(u) - K_{u\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n\lambda_n} K^{-1}_{\lambda_n\lambda_n S_n} K_{\lambda_n u},$$

which is the same as (8), which goes to 0 as $n \to \infty$. Thus Step 3 is proven.

This finishes the proof of Proposition 3.1. ■

C.2 Proof of Theorems 3.1 and 4.1

We will prove Theorem 4.1, which contains Theorem 3.1 as a special case.

Proof. Overview: We will prove expectation consistency, $\int g(\theta) p(\theta \mid y, D_n)\, \mathrm{d}\theta \xrightarrow{p} \int g(\theta) p_*(\theta \mid y)\, \mathrm{d}\theta$, in Step 1, which will consist of two substeps, Step 1a and Step 1b.

We will prove KL-divergence and entropy consistency, $\mathrm{KL}\left[p(\cdot \mid y, D_n) \,\|\, p_*(\cdot \mid y)\right] \xrightarrow{p} 0$ and $\mathrm{H}\left[p(\cdot \mid y, D_n)\right] \xrightarrow{p} \mathrm{H}\left[p_*(\cdot \mid y)\right]$, in Step 2, which will also consists of two substeps, Step 2a and Step 2b.

Consistency of expectations

Step 1

Define shorthands $l_n(\theta) = l(\theta;\, y, D_n)$ and $l_*(\theta) = l_*(\theta;\, y)$. We want to show

$$\int \frac{1}{Z_n} g(\theta)p(\theta)l_n(\theta)\,\mathrm{d}\theta \xrightarrow{p} \int \frac{1}{Z} g(\theta)p(\theta)l_*(\theta)\,\mathrm{d}\theta \tag{9}$$

where $Z_n = \int g(\theta)p(\theta)l_n(\theta)\,\mathrm{d}\theta$ and $Z = \int g(\theta)p(\theta)l_*(\theta)\,\mathrm{d}\theta$. Define

$$I_n(g) = \int g(\theta)p(\theta)l_n(\theta)\,\mathrm{d}\theta \quad \text{and} \quad I_*(g) = \int g(\theta)p(\theta)l_*(\theta)\,\mathrm{d}\theta.$$

Our approach is to show $I_n(g) \xrightarrow{p} I_*(g)$ and $Z_n \xrightarrow{p} Z$, and Slutsky's theorem implies the desired result (9). Since $Z_n = I_n(g)$ and $Z = I_*(g)$ for $g(\theta) = 1$, we just have to show

$$I_n(g) \xrightarrow{p} I_*(g). \tag{10}$$

In Step 1a below we show

$$\mathbb{E}I_n(g) \to I_*(g) \quad \text{as } n \to \infty.$$

In Step 1b we show

$$\mathbb{E}\left[I_n(g)\right]^2 \to \left[I_*(g)\right]^2 \quad \text{as } n \to \infty.$$

Then we have

$$\mathbb{E}\left[I_n(g) - I_*(g)\right]^2 = \mathbb{E}\left[I_n(g)\right]^2 + \left[I_*(g)\right]^2 - 2I_*(g)\mathbb{E}I_n(g) \to 0,$$

which implies (10) (e.g. [20]).

Step 1a

We can avoid dealing with specific forms of $p(\theta)g(\theta)$ by writing $I_n(g) = \int l_n(\theta)\,\mathrm{d}G$, where G is the signed measure $\mathrm{d}G = p(\theta)g(\theta)\,\mathrm{d}\theta$. Then we compute:

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}I_n(g) &= \lim_{n\to\infty} \mathbb{E} \int \mathrm{N}\left(y;\, V^T\theta + W^T m_{L_\theta D_n}, W^T K_{L_\theta D_n} W + \sigma_Y^2 I\right)\,\mathrm{d}G \\
&= \lim_{n\to\infty} \int \mathbb{E}\mathrm{N}\left(y;\, V^T\theta + W^T m_{L_\theta D_n}, W^T K_{L_\theta D_n} W + \sigma_Y^2 I\right)\,\mathrm{d}G && (11) \\
&= \lim_{n\to\infty} \int \mathbb{E}\mathrm{N}\left(y;\, V^T\theta + W^T L_\theta(m_{fD_n} + m_{\epsilon D_n}), W^T K_{L_\theta D_n} W + \sigma_Y^2 I\right)\,\mathrm{d}G && (12) \\
&= \lim_{n\to\infty} \int \mathbb{E}\mathrm{N}\left(y;\, V^T\theta + W^T L_\theta(m_{fD_n}) + W^T L_\theta(m_{\epsilon D_n}), W^T K_{L_\theta D_n} W + \sigma_Y^2 I\right)\,\mathrm{d}G \\
&= \lim_{n\to\infty} \int \mathrm{N}\left(y;\, V^T\theta + W^T L_\theta(m_{fD_n}), W^T K_{L_\theta D_n} W + \sigma_Y^2 I + \mathrm{Cov}\left[W^T L_\theta(m_{\epsilon D_n})\right]\right)\,\mathrm{d}G && (13) \\
&= \int \lim_{n\to\infty} \mathrm{N}\left(y;\, V^T\theta + W^T L_\theta(m_{fD_n}), W^T K_{L_\theta D_n} W + \sigma_Y^2 I + \mathrm{Cov}\left[W^T L_\theta(m_{\epsilon D_n})\right]\right)\,\mathrm{d}G && (14) \\
&= \int \mathrm{N}\left(y;\, V^T\theta + W^T L_\theta(f_*), \sigma_Y^2 I\right)\,\mathrm{d}G && (15) \\
&= \mathbb{E}I_*(g)
\end{aligned}$$

Equality (11) is Fubini's theorem. Equality (12) is the decomposition of the conditional GP mean into the deterministic and random parts. Equality (13) uses the fact that $L_\theta(m_{\epsilon D_n})$ is zero-mean Gaussian distributed: $L_\theta(m_{\epsilon D_n}) \sim \mathcal{N}\left(0, \mathrm{Cov}\left[L_\theta(m_{\epsilon D_n})\right]\right)$. Together with the Gaussian form of l_n, we may compute the expectation in closed-form, which is (13). In (14), we use the bounded convergence theorem (BCT) for signed measures to interchange limit and integration. To apply BCT, we must check the integrand of (13) is bounded. The integrand is a Gaussian density so has the form (omitting multiplicative constants) $\left[\det\left(M(\theta)\right)\right]^{-\frac{1}{2}} \exp\left(-a(\theta)\right)$, where $M(\theta) = W^T K_{L_\theta D_n} W + \sigma_Y^2 I + \mathrm{Cov}\left[W^T L_\theta(m_{\epsilon D_n})\right]$ and $a(\theta) \geq 0$. $\exp(-a(\theta)) \leq 1$. For the determinant term $M \succeq$

$\sigma_Y^2 I$, so $\det(M) \geq \sigma_Y^{2N}$ by determinant inequality for the Leowner partial order (e.g. [21]) and $[\det(M(\theta))]^{-\frac{1}{2}} \leq \sigma_Y^{-N}$. Thus BCT applies. Equality (15) uses Proposition 3.1, which implies $L_{\theta cm}(m_{f_c D_{cn}}) \to L_{\theta cm}(f_{c*})$ and diagonal entries of $K_{L_\theta D_n}$ and $\mathrm{Cov}\left[L_\theta(m_{\epsilon D_n})\right]$ go to 0 as $n \to \infty$. Cauchy-Schwartz ensures the off-diagonal entries also go to 0. Finally we use the continuity of $\mathrm{N}(y;\, m, V)$ at m and $V \succeq 0$ to conclude (15). Thus Step 1a is proven.

Step 1b

In this step we show $\lim_{n\to\infty} \mathbb{E}\left[I_n(g)\right]^2 = \left[I_*(g)\right]^2$. We compute:

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\left[I_n(g)\right]^2 = \lim_{n\to\infty} \mathbb{E} \iint l_n(\theta_1) l_n(\theta_2)\, \mathrm{d}G_1\, \mathrm{d}G_2 \\
&= \lim_{n\to\infty} \iint \mathbb{E} l_n(\theta_1) l_n(\theta_2)\, \mathrm{d}G_1\, \mathrm{d}G_2 \\
&= \lim_{n\to\infty} \iiint \mathrm{N}(y_1';\, J_1 x, V_{y_1'|x}) \mathrm{N}(y_2';\, J_2 x, V_{y_2'|x}) \mathrm{N}(x;\, 0, V_X)\, \mathrm{d}x\, \mathrm{d}G_1\, \mathrm{d}G_2 \qquad (16)
\end{aligned}$$

where in (16) we defined the following notation for clearer analysis:

$$\begin{aligned}
x &= \begin{bmatrix} W^T L_{\theta_1}(m_{\epsilon D_n}) \\ W^T L_{\theta_2}(m_{\epsilon D_n}) \end{bmatrix}, \quad V_X = \mathrm{Cov}(x) \\
y_1' &= y - V^T\theta_1 - W^T L_{\theta_1}(m_{f D_n}), \quad V_{y_1'|x} = W^T K_{L_{\theta_1} D_n} W + \sigma_Y^2 I \\
y_2' &= y - V^T\theta_2 - W^T L_{\theta_2}(m_{f D_n}), \quad V_{y_2'|x} = W^T K_{L_{\theta_2} D_n} W + \sigma_Y^2 I \\
J_1 &= \left[I_N \quad 0\right], \quad J_2 = \left[0 \quad I_N\right], \quad J = I_{2N}
\end{aligned}$$

Further define

$$\begin{aligned}
y' &= \begin{bmatrix} y_1' \\ y_2' \end{bmatrix}, \quad V_{y'|x} = \begin{bmatrix} V_{y_1'|x} & 0 \\ 0 & V_{y_2'|x} \end{bmatrix} \\
a &= V_X (V_X + V_{y'|x})^{-1} y' \\
B &= V_{y'|x} (V_X + V_{y'|x})^{-1} V_X
\end{aligned}$$

Continue from (16):

$$\begin{aligned}
(16) &= \lim_{n\to\infty} \iiint \mathrm{N}(y';\, Jx, V_{y'|x}) \mathrm{N}(x; 0, V_X)\, \mathrm{d}x\, \mathrm{d}G_1\, \mathrm{d}G_2 \\
&= \lim_{n\to\infty} \iiint \mathrm{N}(y';\, 0, J V_X J^T + V_{y'|x}) \mathrm{N}(a, B)\, \mathrm{d}x\, \mathrm{d}G_1\, \mathrm{d}G_2 \qquad (17) \\
&= \lim_{n\to\infty} \iint \mathrm{N}(y';\, 0, V_X + V_{y'|x})\, \mathrm{d}G_1\, \mathrm{d}G_2 \\
&= \iint \lim_{n\to\infty} \mathrm{N}(y';\, 0, V_X + V_{y'|x})\, \mathrm{d}G_1\, \mathrm{d}G_2 \qquad (18) \\
&= \iint \mathrm{N}(y; L_{\theta_1}(f_*), \sigma_Y^2 I) \mathrm{N}(y; L_{\theta_2}(f_*), \sigma_Y^2 I)\, \mathrm{d}G_1\, \mathrm{d}G_2 = \left[I_*(g)\right]^2 \qquad (19)
\end{aligned}$$

The arguments for the equalities are similar to Step 1a: in (17) we rewrote the joint Gaussian distribution $p(x, y')$ as $p(y')p(x \,|\, y')$; in (18) we interchanged limit and integration by BCT, using the Gaussian density form and the Loewner determinant inequality; in (19) we used Proposition 3.1 for the conditional GP mean and diagonal variance terms and Cauchy-Schwartz for off-diagonal terms, combined with the continuity of $\mathrm{N}(\cdot;\, m, V)$ w.r.t. m and V. Thus Step 1b is proven and we have shown claim (9).

KL-divergence and entropy consistency

Step 2

In this step we show $\mathrm{KL}\left[p(\cdot \mid y, D_n) \,\|\, p_*(\cdot \mid y)\right] \xrightarrow{p} 0$ and $\mathrm{H}\left[p(\cdot \mid y, D_n)\right] \xrightarrow{p} \mathrm{H}\left[p_*(\cdot \mid y)\right]$ as $n \to \infty$. These quantities can be written as:

$$\begin{aligned}
&\mathrm{KL}\left[p(\cdot \mid y, D_n) \,\|\, p_*(\cdot \mid y)\right] \\
&\quad= \int p(\theta \mid y, D_n) \log \frac{p(\theta \mid y, D_n)}{p_*(\theta \mid y)} \,\mathrm{d}\theta \\
&\quad= \frac{1}{Z_n} \int p(\theta) l_n(\theta) \log l_n(\theta) \,\mathrm{d}\theta + \int p(\theta \mid y, D_n) \log \frac{p(\theta)}{p_*(\theta \mid y)} \,\mathrm{d}\theta - \log Z_n
\end{aligned} \tag{20}$$

$$\begin{aligned}
&\mathrm{H}\left[p(\theta \mid y, D_n)\right] \\
&\quad= -\int p(\theta \mid y, D_n) \log p(\theta \mid y, D_n) \,\mathrm{d}\theta \\
&\quad= -\frac{1}{Z_n} \int p(\theta) l_n(\theta) \log l_n(\theta) \,\mathrm{d}\theta - \int p(\theta \mid y, D_n) \log p(\theta) \,\mathrm{d}\theta + \log Z_n
\end{aligned} \tag{21}$$

By Step 1, the second terms of (20) and (21) converge in probability to their exact values. $\log Z_n \xrightarrow{p} \log Z$ since $Z_n \xrightarrow{p} Z$ where Z is a number and $\log$ is continuous. We only need to show

$$T_n := -\int p(\theta) l_n(\theta) \log l_n(\theta) \,\mathrm{d}\theta \xrightarrow{p} -\int p(\theta) l_*(\theta) \log l_*(\theta) \,\mathrm{d}\theta =: T_*,$$

then the claims of the theorem will follow by Slutsky's theorem.

We will use a similar argument as that for expectation consistency to show KL and entropy consistency. Namely, with T_n and T_* defined as above, we will show in Steps 2a and 2b:

$$\text{Step 2a: } \lim_{n\to\infty} \mathbb{E} T_n \to T_*. \quad \text{Step 2b: } \lim_{n\to\infty} \mathbb{E} T_n^2 \to T_*^2.$$

Step 2a

First we define notations to facilitate analysis:

$$\begin{aligned}
x &= W^T L_\theta(m_{\epsilon D_n}), \quad V_X = \mathrm{Cov}(x) \\
y' &= y - V^T\theta - W^T L_\theta(m_{f D_n}), \quad V_{y'|x} = W^T K_{L_\theta D_n} W + \sigma_Y^2 I \\
a &= V_X (V_X + V_{y'|x})^{-1} y' \\
B &= V_{y'|x} (I + V_X^{-1} V_{y'|x})^{-1}.
\end{aligned}$$

Then compute the target quantity as:

$$\lim_{n\to\infty} \mathbb{E}T_n = \lim_{n\to\infty} \mathbb{E}\int \left[-\log l_n(\theta)\right] l_n(\theta)\,\mathrm{d}P(\theta)$$

$$= \lim_{n\to\infty} \mathbb{E}\int \left(\frac{1}{2}(y'-x)^T V_{y'|x}^{-1}(y'-x) + \frac{1}{2}\log\det V_{y'|x}\right) \mathrm{N}(y';\, x, V_{y'|x})\,\mathrm{d}P(\theta)$$

$$= \lim_{n\to\infty} \iint \left(\frac{1}{2}(y'-x)^T V_{y'|x}^{-1}(y'-x) + \frac{1}{2}\log\det V_{y'|x}\right) \mathrm{N}(y';\, x, V_{y'|x})\mathrm{N}(x;\, 0, V_X)\,\mathrm{d}x\,\mathrm{d}P(\theta)$$

$$= \lim_{n\to\infty} \iint \left(\frac{1}{2}(y'-x)^T V_{y'|x}^{-1}(y'-x) + \frac{1}{2}\log\det V_{y'|x}\right) \mathrm{N}(y';\, 0, V_X + V_{y'|x}) \\ \mathrm{N}(x;\, a, B)\,\mathrm{d}x\,\mathrm{d}P(\theta) \tag{22}$$

$$= \lim_{n\to\infty} \iint \frac{1}{2} z^T z \mathrm{N}\left(z;\, V_{y'|x}^{-\frac{1}{2}}(a-y'), V_{y'|x}^{-\frac{1}{2}} B V_{y'|x}^{-\frac{1}{2}}\right)\,\mathrm{d}z \mathrm{N}(y';\, 0, V_X + V_{y'|x})\,\mathrm{d}P(\theta) \\ + \int \frac{1}{2}\log\det V_{y'|x}\mathrm{N}(y';\, 0, V_X + V_{y'|x})\,\mathrm{d}P(\theta) \tag{23}$$

$$= \lim_{n\to\infty} \int \frac{1}{2}\left[(y'-a)^T V_{y'|x}^{-1}(y'-a) + \mathrm{tr}\left(V_{y'|x}^{-\frac{1}{2}} B V_{y'|x}^{-\frac{1}{2}}\right) + \log\det V_{y'|x}\right] \\ \mathrm{N}(y';\, 0, V_X + V_{y'|x})\,\mathrm{d}P(\theta) \tag{24}$$

$$= \lim_{n\to\infty} \int \frac{1}{2}\left[{y'}^T (V_X + V_{y'|x})^{-1} V_{y'|x} (V_X + V_{y'|x})^{-1} y' + \mathrm{tr}\left(I + V_{y'|x}^{\frac{1}{2}} V_X^{-1} V_{y'|x}^{\frac{1}{2}}\right)^{-1} + \log\det V_{y'|x}\right] \\ \mathrm{N}(y';\, 0, V_X + V_{y'|x})\,\mathrm{d}P(\theta) \tag{25}$$

where in (22) we rewrite $p(x)p(y' \mid x)$ as $p(y')p(x \mid y')$ using Gaussian density formulae, followed by a change of variables in (23) and integration of x in (24). Applying the definition of a and B and simplifying results in (25). We now show that the limit and integration can be interchanged, by considering the integrand in three parts from the terms in the bracket of (25).

For the first term, $(V_X + V_{y'|x})^{-1} V_{y'|x} (V_X + V_{y'|x})^{-1} = \left[V_X V_{y'|x}^{-1} V_X + V_{y'|x} + 2V_X\right]^{-1} \preceq V_{y'|x}^{-1} \preceq (\sigma_Y^2 I)^{-1}$ implies

$$\begin{aligned} &{y'}^T (V_X + V_{y'|x})^{-1} V_{y'|x} (V_X + V_{y'|x})^{-1} y' && (26)\\ \le &\frac{1}{\sigma_Y^2} {y'}^T y' = \frac{1}{\sigma_Y^2} \|y - V^T\theta - W^T L_\theta(m_{fD_n})\|^2 \\ = &\frac{1}{\sigma_Y^2} \|A(\theta, n) - V^T\theta\|^2 && (27)\end{aligned}$$

where $A(\theta, n) = y - W^T L_\theta(m_{fD_n})$. We first show $A(\theta, n)$ is absolutely bounded uniformly over all $\theta \in \mathbf{\Theta}$ and $n > 0$. This is true because y and W are fixed, $\|L_{\theta cm}\|_{H^*}$ is bounded by hypothesis, and arguments from Step 1 of C.1 show that $\|m_{fD_{cn}}\|_H \le \|f_{c*}\|_H$ for all n. With fixed y and W, we have the claimed uniform absolute boundedness of $A(\theta, n)$. Let $\bar{A}$ be the bound of $A(\theta, n)$.

Expanding the quadratic form in (27) results in a 2nd-order polynomial in θ's coordinates with coefficients given by monomials in $A(\theta, n)$ and entries of V. This polynomial is dominated by the 2nd-order polynomial in coordinates of $|\theta|$, obtained by replacing $A(\theta, n)$, V_{ij}, θ_i in the original polynomial with $\bar{A}$, $|V_{ij}|$, $|\theta_i|$. Let $P(|\theta|)$ denote this dominating polynomial. Using similar arguments as (14) we have $\mathrm{N}(y';\, 0, V_X + V_{y'|x})$ is bounded uniformly over θ and n, so the first part of the integrand, ${y'}^T (V_X + V_{y'|x})^{-1} V_{y'|x} (V_X + V_{y'|x})^{-1} y' \mathrm{N}(y';\, 0, V_X + V_{y'|x})$, is dominated by $CP(|\theta|)$ for all n, for some constant C. By the finite moments (need up to 2nd moments here) hypothesis, this dominating polynomial is integrable, so dominated convergence theorem (DCT) applies.

In the argument above, the finite moments condition is needed only if $V = 0$, otherwise it is not required for DCT to apply.

For the second term, all eigenvalues of $\left(I + V_{y'|x}^{\frac{1}{2}} V_X^{-1} V_{y'|x}^{\frac{1}{2}}\right)^{-1}$ are no greater than 1 for all n, so its trace is bounded for all n and BCT applies.

For the third term,

$$\sigma_Y^2 I \preceq V_{y'|x} = W^T K_{L_\theta D_n} W + \sigma_Y^2 I \preceq W^T K_{L_\theta} W + \sigma_Y^2 I \tag{28}$$

where K_{L_θ} is the prior covariance matrix whose entries are $L^{(x)}_{\theta c m_1} L^{(x')}_{\theta c m_2} k_c(x, x')$ (and 0's for linear functional values on different components). First we show entries of K_{L_θ} are uniformly and absolutely bounded over θ. To show this, each diagonal entry is $L^{(x)}_{\theta cm} L^{(x')}_{\theta cm} k_c(x, x') = \text{Cov}\left[L_{\theta cm}(f_c)\right]$, which by definition is $\|L_{\theta cm}\|^2_{H_c^*}$. By hypothesis, $L_{\theta cm}$ are uniformly H^*-norm bounded, thus diagonal entries of K_{L_θ} are absolutely and uniformly bounded over θ. Off-diagonal entries are bounded by diagonal entries by Cauchy-Schwartz, so the claim is shown. With fixed W, the entries of $W^T K_{L_\theta} W + \sigma_Y^2 I$ are also absolutely and uniformly bounded, which implies the same for its determinant. Using the Loewner determinant inequality again on (28), we have $\log\det V_{y'|x}$ is lower and upper bounded uniformly for all θ and n, then BCT applies after shifting.

Therefore we may interchange limit and integration in (25) and use results in Proposition 3.1:

$$
\begin{aligned}
&(25)\\
&= \int \lim_{n\to\infty} \frac{1}{2}\left[{y'}^T (V_X + V_{y'|x})^{-1} V_{y'|x} (V_X + V_{y'|x})^{-1} y' + \text{tr}\left(I + V_{y'|x}^{\frac{1}{2}} V_X^{-1} V_{y'|x}^{\frac{1}{2}}\right)^{-1} + \log\det V_{y'|x}\right] \\
&\qquad\qquad \text{N}(y';\, 0, V_X + V_{y'|x})\,\text{d}P(\theta) \\
&= \int \frac{1}{2}\left[\left(y - V^T\theta - W^T L_\theta(f_*)\right)^T (\sigma_Y^2 I)^{-1} \left(y - V^T\theta - W^T L_\theta(f_*)\right) + \log\det(\sigma_Y^2 I)\right] \\
&\qquad\qquad \text{N}(y - V^T\theta - W^T L_\theta(f_*);\, 0, \sigma_Y^2 I)\,\text{d}P(\theta) \\
&= \int \left[-\log l_*(\theta)\right] l_*(\theta)\,\text{d}P(\theta) = T_*
\end{aligned}
$$

Thus Step 2a is proven.

Step 2b

Again we define notations to facilitate analysis:

$$
\begin{aligned}
J_1 &= [I_N \quad 0], \quad J_2 = [0 \quad I_N], \quad J = I_{2N} \\
x &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} W^T L_{\theta_1}(m_{\epsilon D_n}) \\ W^T L_{\theta_2}(m_{\epsilon D_n}) \end{bmatrix} \\
V_X &= \text{Cov}(x), \quad V_{X_1} = J_1 V_X J_1^T, \quad V_{X_2} = J_2 V_X J_2^T \\
\theta &= \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \quad \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}, \quad \boldsymbol{\Theta}_2 = \boldsymbol{\Theta} \\
y_1' &= y - V^T\theta_1 - W^T L_{\theta_1}(m_{fD_n}), \quad V_{y_1'|x} = W^T K_{L_{\theta_1} D_n} W + \sigma_Y^2 I \\
y_2' &= y - V^T\theta_2 - W^T L_{\theta_2}(m_{fD_n}), \quad V_{y_2'|x} = W^T K_{L_{\theta_2} D_n} W + \sigma_Y^2 I \\
y' &= \begin{bmatrix} y_1' \\ y_2' \end{bmatrix}, \quad V_{y'|x} = \begin{bmatrix} V_{y_1'|x} & 0 \\ 0 & V_{y_2'|x} \end{bmatrix} \\
a &= V_X (V_X + V_{y'|x})^{-1} y' \\
B &= V_{y'|x} (V_X + V_{y'|x})^{-1} V_X
\end{aligned}
$$

Using these defintions, compute the target quantity

$$
\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}(T_n)^2 \\
&\quad= \lim_{n\to\infty} \mathbb{E} \iint \left[-\log l_n(\theta_1)\right]\left[-\log l_n(\theta_2)\right] l_n(\theta_1) l_n(\theta_2)\,\text{d}P(\theta_1)\,\text{d}P(\theta_2) \\
&\quad= \lim_{n\to\infty} \iiint \frac{1}{4}\left[(y_1' - J_1 x)^T V_{y_1'|x}^{-1} (y_1' - J_1 x) + \log\det V_{y_1'|x}\right] \\
&\qquad\qquad \left[(y_2' - J_2 x)^T V_{y_2'|x}^{-1} (y_2' - J_2 x) + \log\det V_{y_2'|x}\right] \text{N}(y_1';\, x, V_{y_1'|x}) \\
&\qquad\qquad \text{N}(y_2';\, x, V_{y_2'|x}) \text{N}(x;\, 0, V_X)\,\text{d}x\,\text{d}P(\theta_1)\,\text{d}P(\theta_2)
\end{aligned}
$$

Rewriting $\mathrm{N}(y_1';\, J_1x, V_{y_1'|x})\mathrm{N}(y_2';\, J_2x, V_{y_2'|x})\mathrm{N}(x;\, 0, V_X)$ as $\mathrm{N}(y';\, x, V_{y'|x})\mathrm{N}(x;\, 0, V_X)$, and changing the order of conditioning to $\mathrm{N}(x;\, a, B)\mathrm{N}(y';\, 0, V_X + V_{y'|x})$, we get

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}(T_n)^2 \\
&= \lim_{n\to\infty} \iiint \frac{1}{4}\Big[(y_1' - J_1x)^T V_{y_1'|x}^{-1}(y_1' - J_1x)(y_2' - J_2x)^T V_{y_2'|x}^{-1}(y_2' - J_2x) \\
&\qquad + (y_1' - J_1x)^T V_{y_1'|x}^{-1}(y_1' - J_1x)\log\det V_{y_2'|x} \\
&\qquad + (y_2' - J_2x)^T V_{y_2'|x}^{-1}(y_2' - J_2x)\log\det V_{y_1'|x} \\
&\qquad + \log\det V_{y_1'|x}\log\det V_{y_2'|x}\Big] \\
&\qquad \mathrm{N}(x;\, a, B)\mathrm{N}(y';\, 0, V_X + V_{y'|x})\,\mathrm{d}x\,\mathrm{d}P(\theta_1)\,\mathrm{d}P(\theta_2)
\end{aligned} \tag{29}$$

Now we show it is valid to interchange limit and integration with respect to θ_1 and θ_2. (We do not exchange limit and integration over x.) We consider the integrand in four separate parts, each coming from one of the four summands in the bracket in (29).

For the first term, define $z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = V_{y'|x}^{-\frac{1}{2}}(y' - x)$, then

$$\begin{aligned}
&\int (y_1' - J_1x)^T V_{y_1'|x}^{-1}(y_1' - J_1x)(y_2' - J_2x)^T V_{y_2'|x}^{-1}(y_2' - J_2x)\mathrm{N}(x;\, a, B)\,\mathrm{d}x \\
&= \int (z_1^T z_1)(z_2^T z_2)\mathrm{N}(z;\, m_Z, V_Z)\,\mathrm{d}z = \int \sum_{i=1}^{N}\sum_{j=1}^{N} z_{1i}^2 z_{2j}^2 \mathrm{N}(z; m_Z, V_Z)\,\mathrm{d}z
\end{aligned} \tag{30}$$

where using the definition of a and B,

$$\begin{aligned}
m_Z &= V_{y'|x}^{-\frac{1}{2}}(y' - a) = (V_X V_{y'|x}^{-\frac{1}{2}} + V_{y'|x}^{\frac{1}{2}})^{-1} y' \\
V_Z &= V_{y'|x}^{-\frac{1}{2}} B V_{y'|x}^{-\frac{1}{2}} = \left(I + V_{y'|x}^{\frac{1}{2}} V_X^{-1} V_{y'|x}^{\frac{1}{2}}\right)^{-1}
\end{aligned}$$

By the definition of $V_{y'|x}$, its eigenvalues $\lambda_i(V_{y'|x}) \geq \sigma_Y^2$, where $\lambda_i(\cdot)$ denotes the i-th largest eigenvalue. Therefore we have

$$\begin{aligned}
&\lambda_i\left(V_X V_{y'|x}^{-\frac{1}{2}} + V_{y'|x}^{\frac{1}{2}}\right) \geq \lambda_i\left(V_{y'|x}^{\frac{1}{2}}\right) \geq \sigma_Y \implies \lambda_i\left[\left(V_X V_{y'|x}^{-\frac{1}{2}} + V_{y'|x}^{\frac{1}{2}}\right)^{-1}\right] \leq 1/\sigma_Y \\
&\implies \mathrm{tr}\left[\left(V_X V_{y'|x}^{-\frac{1}{2}} + V_{y'|x}^{\frac{1}{2}}\right)^{-1}\right] \leq 2N/\sigma_Y
\end{aligned}$$

By positive definiteness, its absolute diagonal entries no greater than $2N/\sigma_Y$, which implies the same for off-diagonal entries by Cauchy-Schwartz. As in Step 2a, entries of y' have the form $A_0(\theta, n) + A_1(\theta, n)^T\theta$ with absolutely bounded A_0 and A_1 entries. Taken together m_Z entries have the form $B_0(\theta, n) + B_1(\theta, n)^T\theta$ with $B_0(\theta, n)$ and $B_1(\theta, n)$ absolutely and uniformly bounded over θ and n.

Similarly, since eigenvalues of the positive definite matrix V_Z are no greater than 1, absolute V_Z entries are no greater than $2N$.

To evaluate (30), we use the following formula for non-central Gaussian moments:

$$\int u_1^2 u_2^2 \mathrm{N}(u;\, m, V)\,\mathrm{d}u = V_{11}V_{22} + 2V_{12}^2 + V_{11}m_2^2 + V_{22}m_1^2 + m_1^2 m_2^2 + 4V_{12}m_1 m_2 \tag{31}$$

Therefore (30) is a polynomial in m_Z's components, with coefficients given by monomials of V_Z's components. More specifically, since V_Z entries are bounded and m_Z entries have the form $B_0(\theta, n) + B_1(\theta, n)^T\theta$ with bounded B_0 and B_1, (31) implies that (30) can be written as a polynomial of degree 4 in coordinates of θ, with θ, n-dependent coefficients which are bounded absolutely and uniformly over θ and n. This implies that (30) is dominated a polynomial of degree 4 in absolute coordinates of θ, with common coefficients for all n, taken as the absolute and uniform coefficient bounds in the previous statement. As in Step 2a, $\mathrm{N}(y'; 0, V_X + V_{y'|x})$ is bounded, together with the existence of finite 4th moments by hypothesis, DCT applies for this part. Since $B_1(\theta, n)$ is non-zero only if $V \neq 0$, the moment condition is not required if $V = 0$.

Using (31), definition of y', and noting that $\lim_{n\to\infty} V_Z \to 0$ entry-wise by Proposition 3.1, we can compute the limit of (30):

$$\lim_{n\to\infty} \int \sum_{i=1}^{N} \sum_{j=1}^{N} z_{1i}^2 z_{2j}^2 \mathrm{N}(z; m_Z, V_Z)\, \mathrm{d}z = \lim_{n\to\infty} (m_{Z_1}^T m_{Z_1})(m_{Z_2}^T m_{Z_2})$$
$$= \frac{1}{\sigma_Y^4} \|y - V^T\theta_1 - W^T L_{\theta_1}(f_*)\|^2 \|y - V^T\theta_2 - W^T L_{\theta_2}(f_*)\|^2$$

For the second term, define change of variable $z = V_{y_1'|x}^{-\frac{1}{2}}(J_1 x - y_1')$ and compute:

$$\log\det V_{y_2'|x} \int (y_1' - J_1 x)^T V_{y_1'|x}^{-1} (y_1' - J_1 x) \mathrm{N}(x;\, a, B)\, \mathrm{d}x \tag{32}$$
$$= \log\det V_{y_2'|x} \int z^T z \mathrm{N}\left(z;\, V_{y_1'|x}^{-\frac{1}{2}}(J_1 a - y_1'), V_{y_1'|x}^{-\frac{1}{2}} J_1 B J_1^T V_{y_1'|x}^{-\frac{1}{2}}\right)$$
$$= \log\det V_{y_2'|x} \left((J_1 a - y_1') V_{y_1'|x}^{-1} (J_1 a - y_1') + \mathrm{tr}\left(V_{y_1'|x}^{-\frac{1}{2}} J_1 B J_1^T V_{y_1'|x}^{-\frac{1}{2}} \right) \right)$$

With definitions of a, B and J_1:

$$\begin{aligned}
&(J_1 a - y_1') V_{y_1'|x}^{-1} (J_1 a - y_1') \\
&\quad = (a - y')^T J_1^T V_{y_1'|x}^{-1} J_1 (a - y') \\
&\quad = {y'}^T \left(V_X + V_{y'|x}\right)^{-1} V_{y'|x} J_1^T V_{y_1'|x}^{-1} J_1 V_{y'|x} \left(V_X + V_{y'|x}\right)^{-1} y' \\
&\quad = {y'}^T \left(V_X + V_{y'|x}\right)^{-1} \begin{bmatrix} V_{y_1'|x} & 0 \\ 0 & 0 \end{bmatrix} \left(V_X + V_{y'|x}\right)^{-1} y' \\
&\quad \leq {y'}^T \left(V_X + V_{y'|x}\right)^{-1} V_{y'|x} \left(V_X + V_{y'|x}\right)^{-1} y'.
\end{aligned} \tag{33}$$

$$\begin{aligned}
&\mathrm{tr}\left(V_{y_1'|x}^{-\frac{1}{2}} J_1 B J_1^T V_{y_1'|x}^{-\frac{1}{2}} \right) \\
&\quad = \mathrm{tr}\left(V_{y_1'|x}^{-\frac{1}{2}} J_1 V_{y'|x}^{\frac{1}{2}} \left(I + V_{y'|x}^{\frac{1}{2}} V_X^{-1} V_{y'|x}^{\frac{1}{2}} \right)^{-1} V_{y'|x}^{\frac{1}{2}} J_1^T V_{y'|x}^{-\frac{1}{2}} \right) \\
&\quad = \mathrm{tr}\left(J_1 \left(I + V_{y'|x}^{\frac{1}{2}} V_X^{-1} V_{y'|x}^{\frac{1}{2}} \right)^{-1} J_1 \right) \\
&\quad \leq \mathrm{tr}\left(\left(I + V_{y'|x}^{\frac{1}{2}} V_X^{-1} V_{y'|x}^{\frac{1}{2}} \right)^{-1} \right).
\end{aligned} \tag{34}$$

In (33) we used $\begin{bmatrix} V_{y_1'|x} & 0 \\ 0 & 0 \end{bmatrix} \preceq V_{y'|x}$. The trace term (34) is bounded by $2N$. $\log\det V_{y_2'|x}$ is bounded from the reasoning for (28). Then from (33) we can follow the same argument as (26) to validate BCT. Similarly the boundedness of $\boldsymbol{\Theta}$ is not required if $V = 0$.

The value of the limit of (32) is:

$$\lim_{n\to\infty} \log\det V_{y_2'|x} \int (y_1' - J_1 x)^T V_{y_1'|x}^{-1} (y_1' - J_1 x) \mathrm{N}(x;\, a, B)\, \mathrm{d}x$$
$$= \log\det(\sigma_Y^2 I_N) \frac{1}{\sigma_Y^2} \|y - V^T\theta_1 - W^T L_{\theta_1}(f_*)\|^2$$

The third part of the integrand of (29) interchanges θ_1 and θ_2 of the second part. Criteria for BCT is not affected by this, so BCT holds.

The fourth part of the integrand of (29) is $\log\det V_{y_1'|x} \log\det V_{y_2'|x} \mathrm{N}\left(y';\, 0, V_X + V_{y'|x}\right)$. By the same argument as for (28) in Step 2a, both $\log\det$ terms are absolutely and uniformly bounded over

θ and n, as is $\mathrm{N}\left(y';\, 0, V_X + V_{y'|x}\right)$. So BCT applies. The value of its limit is:

$$\lim_{n\to\infty} \log\det V_{y_1'|x} \log\det V_{y_2'|x} = \left(\log\det(\sigma_Y^2 I_N)\right)^2$$

Thus we may interchange limit and integration in (29), and using the computed limit of each part of its integrand, we get:

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}(T_n)^2 \\
&\quad = \iint \frac{1}{4}\Big[\frac{1}{\sigma_Y^4}\|y - V^T\theta_1 - W^T L_{\theta_1}(f_*)\|^2 \|y - V^T\theta_2 - W^T L_{\theta_2}(f_*)\|^2 \\
&\qquad\qquad + \log\det(\sigma_Y^2 I_N)\frac{1}{\sigma_Y^2}\|y - V^T\theta_1 - W^T L_{\theta_1}(f_*)\|^2 \\
&\qquad\qquad + \log\det(\sigma_Y^2 I_N)\frac{1}{\sigma_Y^2}\|y - V^T\theta_2 - W^T L_{\theta_2}(f_*)\|^2 \\
&\qquad\qquad + \left(\log\det(\sigma_Y^2 I_N)\right)^2\Big]\mathrm{N}(y;\, 0, \sigma_Y^2 I_{2N})\,\mathrm{d}P(\theta_1)\,\mathrm{d}P(\theta_2) \\
&\quad = \iint \left[-\log l_*(\theta_1)\right]\left[-\log l_*(\theta_2)\right] l_*(\theta_1) l_*(\theta_2)\,\mathrm{d}P(\theta_1)\,\mathrm{d}P(\theta_2) \\
&\quad = T_*^2
\end{aligned}$$

Thus Step 2 is proven. This completes the proof of Theorem 4.1.

■