Posterior and Computational Uncertainty in Gaussian Processes

Jonathan Wenger[1,2] **Geoff Pleiss**[2]

Marvin Pförtner[1] **Philipp Hennig**[1,3] **John P. Cunningham**[2]

[1] University of Tübingen
[2] Columbia University
[3] Max Planck Institute for Intelligent Systems, Tübingen

Abstract

Gaussian processes scale prohibitively with the size of the dataset. In response, many approximation methods have been developed, which inevitably introduce approximation error. This additional source of uncertainty, due to limited computation, is entirely ignored when using the approximate posterior. Therefore in practice, GP models are often as much about the approximation method as they are about the data. Here, we develop a new class of methods that provides consistent estimation of the combined uncertainty arising from *both* the finite number of data observed *and* the finite amount of computation expended. The most common GP approximations map to an instance in this class, such as methods based on the Cholesky factorization, conjugate gradients, and inducing points. For any method in this class, we prove (i) convergence of its posterior mean in the associated RKHS, (ii) decomposability of its combined posterior covariance into mathematical and computational covariances, and (iii) that the combined variance is a tight worst-case bound for the squared error between the method's posterior mean and the latent function. Finally, we empirically demonstrate the consequences of ignoring computational uncertainty and show how implicitly modeling it improves generalization performance on benchmark datasets.

1 Introduction

Gaussian processes (GPs) are an expressive probabilistic model class, but their prohibitive scaling necessitates approximation [1]. A range of approximations based on kernel [2–10] or precision matrix [11–14] estimates, inducing point methods [15–22], and iterative solvers [23–29] have been proposed. These methods all use an affordable amount of computation to obtain an approximation of the *mathematical* posterior, which exists theoretically but cannot be accessed given limited computational resources. The approximate posterior is then used as a direct replacement of the mathematical posterior in downstream applications. Doing so, however, completely ignores the fact that we only expended a limited amount of compute. By analogy to the typical GP operation, where *limited data* induces modeling error captured by *mathematical uncertainty*, our work is motivated by the fact that *limited computation* induces approximation error that must be captured by *computational uncertainty*.

Here, we introduce IterGP, a class of methods which return a *combined uncertainty* that is the sum of mathematical and computational uncertainty. Figure 1 illustrates the difference between ignoring computational uncertainty and explicitly modeling it. We perform GP regression using a Matérn($\frac{3}{2}$) kernel on a toy dataset and compare SVGP (🔴) [22] to its analog in our framework, IterGP-PI (🟢 + 🔵), for a fixed set of inducing points. The computational shortcuts of inducing point methods can lead to unavoidable biases in their posterior mean and covariance [30, 31]. As Figure 1 illustrates, SVGP

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

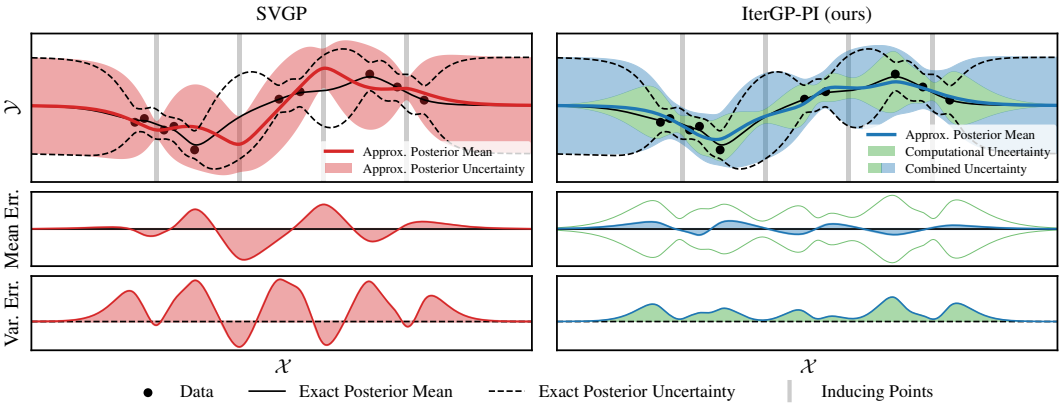


Figure 1: *Modeling computational uncertainty improves GP approximation.*

may underestimate the marginal variance where inducing points do not coincide with datapoints. In contrast, IterGP is guaranteed to overestimate the mathematical uncertainty – with the difference precisely given by the computational uncertainty (●). Additionally, the computational uncertainty is a worst-case bound (—) on the error of the approximate posterior mean.

To be clear, this overestimation is desirable: IterGP is not a typical approximation in the sense that its combined posterior attempts to approximate the mathematical posterior. Rather, IterGP recognizes that the mathematical posterior exists, but we do not have access to it, given computational constraints. Finite compute is as true a source of posterior uncertainty as finite data. Taking this view seriously, the true goal of GPs in the limited compute regime should in fact be to track combined uncertainty. This intuition motivates IterGP and is formally a feature of our results. We show that, if you update your GP via computation, specifically matrix-vector multiplication, then the combined uncertainty of the IterGP algorithm is precisely the correct object to capture your belief (Theorem 2) – in the same way the mathematical posterior is the correct object given finite data and unlimited computation.

Formally, IterGP is a probabilistic numerical method [32–35]. It treats the (unknown) representer weights as a latent variable with a prior belief that, when marginalized out, corresponds to a GP prior conditioned on no data. We then use a computational primitive (matrix-vector multiplication) that corresponds to tractable Bayesian updates on the representer weight distribution; that is, conditioning on computations performed on the data. The resulting belief can then be marginalized out to obtain a closed form, tractable expression for the combined – mathematical plus computational – uncertainty. This uncertainty quantification can be done *exactly* in quadratic time and linear space complexity.

Our framework admits three key theoretical properties. First, common GP approximations such as the partial Cholesky, the method of conjugate gradients and inducing point methods (e.g. SVGP) map to a corresponding IterGP instance. Therefore, these approaches can either be directly extended or modified to properly account for computational uncertainty. Second, the approximate posterior mean of any method in our proposed class converges to the mathematical posterior mean in RKHS norm in at most n steps, where the convergence rate is determined by the choice of method (Theorem 1). Third, the combined uncertainty is a tight worst case bound on the relative error between the approximate posterior mean and the latent function (Theorem 2). To the best of our knowledge no existing GP approximation has this last property; an analoguous guarantee only holds for exact GPs [36, Sec. 3.4].

Contribution This work introduces IterGP, which defines a new class of GP approximations that accounts for computational uncertainty arising from limited computation. Some IterGP instances extend classic methods with improved uncertainty quantification (Table 1). For any method in this class, we prove that the approximate posterior mean converges to the mathematical posterior mean (Theorem 1) and that the combined uncertainty is a tight worst-case bound on the relative distance to the latent function one is trying to learn (Theorem 2, Corollary 1). We demonstrate empirically that modeling computational uncertainty can either save computation or improve generalization on a set of regression benchmark datasets. In conclusion, we show that it is possible to exactly quantify the inevitable error of GP approximations at quadratic cost by propagating said error to the posterior in the form of computational uncertainty.

2 Computation-Aware Gaussian Process Inference

We aim to learn a latent function $h : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{X} \subseteq \mathbb{R}^d$ to $\mathcal{Y} \subseteq \mathbb{R}$ given a training dataset $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{n \times d}$ of n inputs $\boldsymbol{x}_j \in \mathbb{R}^d$ and corresponding outputs $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal \in \mathbb{R}^n$.

Gaussian Processes A stochastic process $f \sim \mathcal{GP}(\mu, k)$ with mean function $\mu : \mathbb{R}^d \to \mathbb{R}$ and kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called a *Gaussian process* (GP) if any collection of function values $\mathbf{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))^\intercal \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K})$ is jointly Gaussian with $\boldsymbol{\mu}_j = \mu(\boldsymbol{x}_j)$ and $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Assuming observation noise $\boldsymbol{y} \mid \mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2 \boldsymbol{I})$, the posterior distribution at test inputs $\boldsymbol{X}_\diamond$ is given by $\mathbf{f}_\diamond \sim \mathcal{N}(\mu_*(\boldsymbol{X}_\diamond), k_*(\boldsymbol{X}_\diamond, \boldsymbol{X}_\diamond))$ where the posterior mean and covariance functions are given by

$$\mu_*(\cdot) = \mu(\cdot) + k(\cdot, \boldsymbol{X}) \overbrace{\hat{\boldsymbol{K}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})}^{\boldsymbol{v}_*}, \quad \text{and} \quad k_*(\cdot, \cdot) = k(\cdot, \cdot) - k(\cdot, \boldsymbol{X}) \hat{\boldsymbol{K}}^{-1} k(\boldsymbol{X}, \cdot) \tag{1}$$

where $\hat{\boldsymbol{K}} \coloneqq \boldsymbol{K} + \sigma^2 \boldsymbol{I} \in \mathbb{R}^{n \times n}$. Computing the *representer weights* $\boldsymbol{v}_* = \hat{\boldsymbol{K}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$ exactly (as well as the posterior variance) is prohibitive given our limited computational budget.

Learning Representer Weights Consider the conditional distribution of the latent GP given its representer weights:

$$p(\mathbf{f}_\diamond \mid \boldsymbol{v}_*) = \mathcal{N}(\mu(\boldsymbol{X}_\diamond) + k(\boldsymbol{X}_\diamond, \boldsymbol{X}) \boldsymbol{v}_*, \;\; k_*(\boldsymbol{X}_\diamond, \boldsymbol{X}_\diamond)). \tag{2}$$

When $\boldsymbol{v}_*$ is known exactly, we recover eq. (1). However, if we instead treat $\boldsymbol{v}_*$ as a random variable with the prior $p(\boldsymbol{v}_*) = \mathcal{N}(\boldsymbol{v}_*; \boldsymbol{0}, \hat{\boldsymbol{K}}^{-1})$, then the resulting marginal $\int p(\mathbf{f}_\diamond \mid \boldsymbol{v}_*) p(\boldsymbol{v}_*) \, d\boldsymbol{v}_*$ recovers the GP prior $\mathcal{N}(\mu(\boldsymbol{X}_\diamond), k(\boldsymbol{X}_\diamond, \boldsymbol{X}_\diamond))$. Our goal is to update this prior by iteratively applying the tractable computational primitive (i.e. matrix-vector multiplies). More specifically, each iteration conditions the current belief distribution $p(\boldsymbol{v}_*) = \mathcal{N}(\boldsymbol{v}_*; \boldsymbol{v}_{i-1}, \boldsymbol{\Sigma}_{i-1})$ on a one-dimensional projection of the current *residual* $\boldsymbol{r}_{i-1} = (\boldsymbol{y} - \boldsymbol{\mu}) - \hat{\boldsymbol{K}} \boldsymbol{v}_{i-1}$, where the projection is defined by an arbitrary vector $\boldsymbol{s}_i$:

$$\alpha_i \coloneqq \boldsymbol{s}_i^\intercal \boldsymbol{r}_{i-1} = \boldsymbol{s}_i^\intercal ((\boldsymbol{y} - \boldsymbol{\mu}) - \hat{\boldsymbol{K}} \boldsymbol{v}_{i-1}) = \boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}} (\boldsymbol{v}_* - \boldsymbol{v}_{i-1}). \tag{3}$$

The choice of *actions* $\boldsymbol{s}_i$, which intuitively weight the approximation error of selected datapoints, defines different instances of our IterGP framework. Computing eq. (3) requires a single matrix-vector multiplication. After computing α_i, we can perform an exact Bayesian update of $p(\boldsymbol{v}_*)$ via linear Gaussian identities. The updated $p(\boldsymbol{v}_*)$ (conditioned on α_i) is $\mathcal{N}(\boldsymbol{v}_* \mid \boldsymbol{v}_i, \boldsymbol{\Sigma}_i)$, with

$$\boldsymbol{v}_i = \boldsymbol{v}_{i-1} + \underbrace{\boldsymbol{\Sigma}_{i-1} \hat{\boldsymbol{K}} \boldsymbol{s}_i}_{=: \boldsymbol{d}_i} \underbrace{(\boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{\Sigma}_{i-1} \hat{\boldsymbol{K}} \boldsymbol{s}_i)^{-1}}_{=: \eta_i} \underbrace{\boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}} (\boldsymbol{v}_* - \boldsymbol{v}_{i-1})}_{= \alpha_i} = \boldsymbol{C}_i (\boldsymbol{y} - \boldsymbol{\mu}) \tag{4}$$

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_{i-1} - \underbrace{\boldsymbol{\Sigma}_{i-1} \hat{\boldsymbol{K}} \boldsymbol{s}_i}_{= \boldsymbol{d}_i} \underbrace{(\boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{\Sigma}_{i-1} \hat{\boldsymbol{K}} \boldsymbol{s}_i)^{-1}}_{= \eta_i} \underbrace{\boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{\Sigma}_{i-1}}_{= \boldsymbol{d}_i^\intercal} = \hat{\boldsymbol{K}}^{-1} - \boldsymbol{C}_i. \tag{5}$$

where $\boldsymbol{C}_i \coloneqq \sum_{j=1}^{i} \frac{1}{\eta_j} \boldsymbol{d}_j \boldsymbol{d}_j^\intercal = \boldsymbol{S}_i (\boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{S}_i)^{-1} \boldsymbol{S}_i^\intercal$ is a rank-i matrix (see Proposition S3 for details). With each computation, the uncertainty about the representer weights contracts as $\boldsymbol{C}_i \to \hat{\boldsymbol{K}}^{-1} = \boldsymbol{\Sigma}_0$ as $i \to n$. After n iterations, $\boldsymbol{C}_n = \hat{\boldsymbol{K}}^{-1}$, meaning we fully recovered the representer weights with zero uncertainty. The consistent estimate for the representer weights is consequently $\boldsymbol{v}_i = \boldsymbol{C}_i (\boldsymbol{y} - \boldsymbol{\mu})$.

Combining Mathematical and Computational Uncertainty We now have a belief $p(\boldsymbol{v}_*) = \mathcal{N}(\boldsymbol{v}_*; \boldsymbol{v}_i, \boldsymbol{\Sigma}_i)$ about the representer weights reflecting the expended computation. To account for this computational uncertainty, we treat the representer weights as a latent variable of the mathematical posterior by reparameterizing $p(\mathbf{f}_\diamond \mid \boldsymbol{y}) = p(\mathbf{f}_\diamond \mid \boldsymbol{v}_*)$ and then marginalizing. The resulting marginal considers all possible representer weights which would have resulted in the same computational observations and therefore *implicitly* adds the uncertainty coming from *the computation itself*. Since the posterior mean of a GP is a linear function of the representer weights, the marginal distribution is given by $p(\mathbf{f}_\diamond) = \int p(\mathbf{f}_\diamond \mid \boldsymbol{v}_*) p(\boldsymbol{v}_*) \, d\boldsymbol{v}_* = \mathcal{N}(\mathbf{f}_\diamond; \mu_i(\boldsymbol{X}_\diamond), k_i(\boldsymbol{X}_\diamond, \boldsymbol{X}_\diamond))$, where

$$\begin{aligned} \mu_i(\cdot) &= \mu(\cdot) + k(\cdot, \boldsymbol{X}) \boldsymbol{v}_i \\ k_i(\cdot, \cdot) &= \underbrace{k(\cdot, \cdot) - k(\cdot, \boldsymbol{X}) \hat{\boldsymbol{K}}^{-1} k(\boldsymbol{X}, \cdot)}_{\text{mathematical uncertainty}} + \underbrace{k(\cdot, \boldsymbol{X}) \boldsymbol{\Sigma}_i k(\boldsymbol{X}, \cdot)}_{\text{computational uncertainty}} = \underbrace{k(\cdot, \cdot) - k(\cdot, \boldsymbol{X}) \boldsymbol{C}_i k(\boldsymbol{X}, \cdot)}_{\text{combined uncertainty}} \end{aligned} \tag{6}$$

since $\boldsymbol{\Sigma}_i = \hat{\boldsymbol{K}}^{-1} - \boldsymbol{C}_i$.[1] As we perform more computation, the computational uncertainty reduces and we approach the mathematical uncertainty. We note that, while the *individual* terms are computationally prohibitive, the *combined* uncertainty can be evaluated cheaply since the approximate

[1] While we derive the combined posterior from a probabilistic numerics perspective, we can alternatively interpret eq. (6) as the GP prior f conditioned on linearly transformed data $\boldsymbol{S}_i^\intercal \boldsymbol{y} \mid \mathbf{f} \sim \mathcal{N}(\boldsymbol{S}_i^\intercal \mathbf{f}, \sigma^2 \boldsymbol{S}_i^\intercal \boldsymbol{S}_i)$.

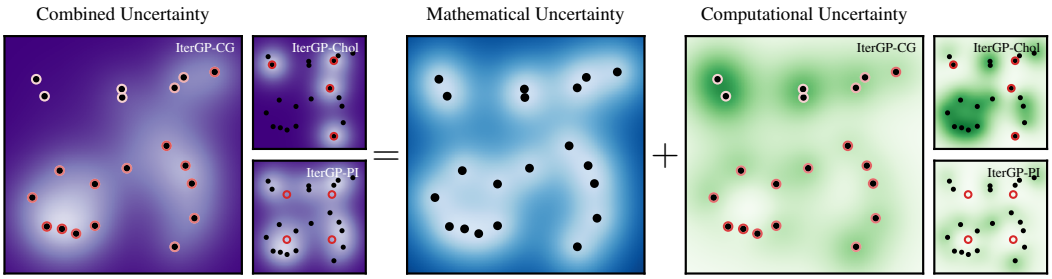


Figure 2: *Decomposition of the combined uncertainty.* The combined uncertainty (●) output by IterGP decomposes into the mathematical uncertainty (●) and computational uncertainty (●). After $i = 4$ iterations of Algorithm 1 computational uncertainty is small in parts of the input space where there either is no data (●) or computation was "targeted" (○). Which datapoints are targeted in each iteration and to what degree is defined by the magnitude of the action vector elements $(\boldsymbol{s}_i)_j$. Different instances of IterGP either reduce computational uncertainty locally (e.g. IterGP-Chol, IterGP-PI) or globally (e.g. IterGP-CG). After n iterations the mathematical uncertainty is recovered.

precision matrix $\boldsymbol{C}_i$ is of low rank. Figure 2 illustrates that computational uncertainty is large where there are data and we have not targeted computation yet. Different methods from our proposed class target computation in different parts of the input space. Where there is no data the prior is a good approximation of the posterior and therefore computational uncertainty is low.

Algorithm 1 computes an estimate of the representer weights $\boldsymbol{v}_i$ and the rank-i precision matrix approximation $\boldsymbol{C}_i$. A specific instance of IterGP is defined by a sequence of actions $\boldsymbol{s}_i$. To gain an intuition for how Algorithm 1 operates, it helps to interpret it as targeting a given computational budget towards certain datapoints defined by $\boldsymbol{s}_i$. Near datapoints $\boldsymbol{x}_j$ that are not targeted, i.e. $(\boldsymbol{s}_i)_j = 0$, computational uncertainty remains unchanged. In fact, datapoints $(\boldsymbol{x}_j, y_j)$ that are never targeted up to iteration i are not needed to compute $\mathcal{GP}(\mu_i, k_i)$, meaning that Algorithm 1 is *inherently online* and we can *observe data sequentially* without having to restart the algorithm (see Theorem S7).

Algorithm 1: A Class of Computation-Aware Iterative Methods for GP Approximation

Input: prior mean function μ, prior covariance function / kernel k, training inputs $\boldsymbol{X}$, labels $\boldsymbol{y}$
Output: (combined) GP posterior $\mathcal{GP}(\mu_i, k_i)$

procedure ITERGP($\mu, k, \boldsymbol{X}, \boldsymbol{y}$)
  $(\mu_0, k_0) \leftarrow (\mu, k)$ ▷ Initialize mean and covariance function with prior.
  $\boldsymbol{\mu} \leftarrow \mu(\boldsymbol{X})$ ▷ Prior predictive mean.
  $\hat{\boldsymbol{K}} \leftarrow k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}$ ▷ Prior predictive kernel matrix.
  while not STOPPINGCRITERION() **do** ▷ Stopping criterion.
    $\boldsymbol{s}_i \leftarrow$ POLICY() ▷ Select action via policy (see Table 1 for examples).
    $\boldsymbol{r}_{i-1} \leftarrow (\boldsymbol{y} - \boldsymbol{\mu}) - \hat{\boldsymbol{K}} \boldsymbol{v}_{i-1}$ ▷ Predictive residual.
    $\alpha_i \leftarrow \boldsymbol{s}_i^\intercal \boldsymbol{r}_{i-1}$ ▷ Observation via information operator.
    $\boldsymbol{d}_i \leftarrow \boldsymbol{\Sigma}_{i-1} \hat{\boldsymbol{K}} \boldsymbol{s}_i = (\boldsymbol{I} - \boldsymbol{C}_{i-1} \hat{\boldsymbol{K}}) \boldsymbol{s}_i$ ▷ Search direction.
    $\eta_i \leftarrow \boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{\Sigma}_{i-1} \hat{\boldsymbol{K}} \boldsymbol{s}_i = \boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{d}_i$ ▷ Normalization constant.
    $\boldsymbol{C}_i \leftarrow \boldsymbol{C}_{i-1} + \frac{1}{\eta_i} \boldsymbol{d}_i \boldsymbol{d}_i^\intercal$ ▷ Precision matrix approximation $\boldsymbol{C}_i \approx \hat{\boldsymbol{K}}^{-1}$.
    $\boldsymbol{Q}_i \leftarrow \boldsymbol{Q}_{i-1} + \frac{1}{\eta_i} \hat{\boldsymbol{K}} \boldsymbol{d}_i \boldsymbol{d}_i^\intercal \hat{\boldsymbol{K}}$ ▷ Kernel matrix approximation $\boldsymbol{Q}_i \approx \hat{\boldsymbol{K}}$.
    $\boldsymbol{v}_i \leftarrow \boldsymbol{v}_{i-1} + \frac{\alpha_i}{\eta_i} \boldsymbol{d}_i$ ▷ Representer weights estimate.
    $\boldsymbol{\Sigma}_i \leftarrow \boldsymbol{\Sigma}_0 - \boldsymbol{C}_i$ ▷ Computational representer weights uncertainty.
  $p(\boldsymbol{v}_*) \leftarrow \mathcal{N}(\boldsymbol{v}_*; \boldsymbol{v}_i, \boldsymbol{\Sigma}_i)$ ▷ Belief about representer weights.
  $\mu_i(\cdot) \leftarrow \mu(\cdot) + k(\cdot, \boldsymbol{X}) \boldsymbol{v}_i$ ▷ Approximate posterior mean function.
  $k_i(\cdot, \cdot) \leftarrow k(\cdot, \cdot) - k(\cdot, \boldsymbol{X}) \boldsymbol{C}_i k(\boldsymbol{X}, \cdot)$ ▷ Combined uncertainty.
  return $\mathcal{GP}(\mu_i, k_i)$

Greyed out quantities are *not* needed to compute the combined posterior and are only included for clarity of exposition.

Table 1: *Instances of Algorithm 1, which map to commonly used GP approximations.*

Method	Actions $\boldsymbol{s}_i$	Classic Analog	Reference
IterGP-Chol	$\boldsymbol{e}_i$	(partial) Cholesky	Theorem S3
IterGP-PBR	$\mathrm{ev}_i(\hat{\boldsymbol{K}})$	(partial) EVD / SVD	Theorem S4
IterGP-CG	$\boldsymbol{s}_i^{\mathrm{PCG}}$ or $\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i$	(preconditioned) CG	Theorem S5 and Corollary S2
IterGP-PI	$k(\boldsymbol{X}, \boldsymbol{z}_i)$	$\approx$ Nyström (SoR, DTC), SVGP	Section 2.1 and Theorem S6

2.1 Connection to Other GP Approximation Methods

IterGP extends the most commonly used GP approximations to include computational uncertainty, with at most quadratic cost (see Table 1 for a summary and Figure 2, Figure S3 for illustration).

Cholesky Decomposition The (partial) Cholesky decomposition iteratively chooses datapoints or pivots $\boldsymbol{x}_i$ based on a given ordering. The resulting Cholesky factor is lower triangular and increases in rank each iteration, and a well-chosen ordering achieves fast convergence (cf. [37, Thm. 2.3]). If one chooses standard unit vectors $\boldsymbol{e}_i$ as actions corresponding to the selected datapoint per iteration, then Algorithm 1 recovers the partial Cholesky factorization exactly (Theorem S3).

Conjugate Gradients CG [38] with preconditioning for GP inference has become increasingly popular [24–29, 39, 40]. Algorithm 1 recovers preconditioned CG exactly, if we choose either preconditioned conjugate gradients or residuals as actions (see Theorem S5 and Corollary S2). In fact, Algorithm 1 can even construct its own diagonal-plus-low-rank preconditioner by first running a few iterations with an arbitrary policy and then using the byproducts of these iterations for the preconditioner. For example, if we run IterGP-Chol initially, we can construct an incomplete Cholesky preconditioner for subsequent CG iterations.

Inducing Point Methods Inducing point methods, such as variants of the Nyström approximation [16], i.e. subset of regressors (SoR) [15, 41] and deterministic training conditional (DTC) [18, 42], as well as SVGP [22] share a posterior mean, which by Theorem S6 takes the form

$$\mu_{\mathrm{SVGP}}(\cdot) = q(\cdot, \boldsymbol{X})\boldsymbol{K}_{\boldsymbol{XZ}}(\boldsymbol{K}_{\boldsymbol{ZX}}(q(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2\boldsymbol{I})\boldsymbol{K}_{\boldsymbol{XZ}})^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu}) \tag{7}$$

where $\boldsymbol{Z} \in \mathbb{R}^{n \times i}$ is a set of inducing points and $q(\cdot, \cdot) = k(\cdot, \boldsymbol{Z})\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}k(\boldsymbol{Z}, \cdot)$. These approximations also have very closely related posterior covariance functions [20, 43]. If we choose actions $\boldsymbol{s}_i = k(\boldsymbol{X}, \boldsymbol{z}_i)$, by Proposition S3, Algorithm 1 returns a posterior mean given by

$$\mu_i(\cdot) = k(\cdot, \boldsymbol{X})\boldsymbol{K}_{\boldsymbol{XZ}}(\underbrace{\boldsymbol{K}_{\boldsymbol{ZX}}(k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2\boldsymbol{I})\boldsymbol{K}_{\boldsymbol{XZ}}}_{\text{Gram matrix } \boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}}\boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i})^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu}). \tag{8}$$

Choosing such actions, given by kernel functions $k(\cdot, \boldsymbol{z}_i)$ centered at inducing points $\boldsymbol{z}_i$, reduces computational uncertainty in regions close to inducing points (see IterGP-PI in Figure 2), where closeness is determined by the kernel. Comparing SVGP's and IterGP-PI's posterior mean provides a probabilistic numerical perspective on why even for small KL-divergence between the approximating distribution of SVGP and the true posterior, the mean estimate can be far from the true mean [31, Prop. 3.1]. As outlined in Section 2, eq. (8) is a Bayesian update on the initially unknown representer weights $\boldsymbol{v}_* = \hat{\boldsymbol{K}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$. The Gram matrix in eq. (8) describes how surprising the computational observations $\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{S}_i^\intercal(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{v}_*$ of the representer weights should be, given the prior uncertainty $\boldsymbol{\Sigma}_0$ about them. SVGP uses a similar form for the posterior mean (c.f. (7) and (8)), but the Gram matrix is "smaller" since $q(\boldsymbol{X}, \boldsymbol{X}) \preceq k(\boldsymbol{X}, \boldsymbol{X})$. This can be interpreted as inducing point methods being overconfident in their update of the representer weight estimates to achieve linear time complexity. As the inducing points approach the data points the two posterior mean functions μ_{SVGP} and μ_i become closer and are equivalent if the inducing points equal the training data.

2.2 The Cost of Computational Uncertainty

Quantifying combined uncertainty has greater cost than linear time GP approximations such as inducing point methods, due to its use of matrix-vector multiplication as the computational operation to condition on the data. Algorithm 1 in its most general form performs three matrix-vector products per iteration resulting in a quadratic time complexity $\mathcal{O}(n^2 i)$ overall for i iterations. In this sense,

Algorithm 1 represents a middle ground between the mathematical posterior—which incurs a cubic time complexity—and $\mathcal{O}(ni^2)$ approximations—which can only estimate their computational error through potentially loose theoretical bounds which may [e.g. 21, 22, 44] or may not be computable in less than $\mathcal{O}(n^3)$ [4, 37]. At any point during a run of Algorithm 1, computing the predictive mean on $n_\diamond$ new data points has cost $\mathcal{O}(n_\diamond n)$, while the marginal predictive (co-)variance can be evaluated in $\mathcal{O}(n_\diamond ni)$ since $\boldsymbol{C}_i$ is of rank i. Additionally, using Matheron's rule [45–47], sampling from the approximate posterior at $n_\diamond$ evaluation points also only requires $\mathcal{O}(n_\diamond ni)$ computation (assuming we can sample from the prior—see Section S3.3). The objects required to make predictions and draw samples are the vector $\boldsymbol{v}_i$ and low rank matrix $\boldsymbol{C}_i$ which both require $\mathcal{O}(ni)$ memory. Finally, the memory cost of Algorithm 1 is only linear in n, since matrix multiplication $\boldsymbol{v} \mapsto \hat{\boldsymbol{K}}\boldsymbol{v}$ can be computed without explicitly forming $\hat{\boldsymbol{K}}$ [48].

2.3 Related Work

GP inference based on matrix-vector multiplies, particularly CG [38], has become popular recently [5, 24–29, 39]. Advances in specialized hardware has boosted their scalability without excessive memory footprint [27, 48]. Such iterative methods typically rely on preconditioning, which has been shown to significantly improve their performance [25, 26, 29]. Our method generalizes CG in this setting and thus retains the same benefits. At its core Algorithm 1 employs a (Bayesian) probabilistic numerical method [32–35], more specifically a probabilistic linear solver (PLS) [49–54] applied to the linear system $\hat{\boldsymbol{K}}\boldsymbol{v}_* = \boldsymbol{y}$. The fact that a PLS using CG actions can recover CG in its posterior mean was observed previously [49, 51, 53]. Here, we extend this result to residual actions and preconditioning. Further, we also demonstrate the connection to the Cholesky and singular value decompositions. For randomized actions, the PLS as part of Algorithm 1 also recovers the randomized Kaczmarz method in its posterior mean [55–58]. Employing a PLS for GP approximation by updating beliefs over the kernel and precision matrix was suggested previously [53, 59]. Our work differs in that it updates a belief over the representer weights, as opposed to the kernel function or matrix, considers more general projections than just conjugate residuals, and, most importantly, provides a theoretically motivated combined posterior which can be computed exactly.

3 Theoretical Analysis

The main goals of our theoretical analysis will be to prove

(a) *convergence of IterGP's posterior mean* in norm (Theorem 1) and pointwise (Corollary 1)

and to provide rigorous justification for the combined and computational uncertainty. Importantly, the

(b) *combined uncertainty is a tight worst-case bound on the relative distance to all potential latent functions* consistent with our (computational) observations (Theorem 2).

We will demonstrate a similar interpretation of the computational uncertainty as a bound on the relative error to the mathematical posterior mean (see eqs. (14) and (16)).

3.1 Estimation of Representer Weights

At the heart of Algorithm 1 is a probabilistic linear solver [49–51, 53] iteratively updating a belief about the representer weights. It constructs an expanding subspace $\operatorname{span}\{\boldsymbol{s}_1, \dots, \boldsymbol{s}_i\} = \operatorname{span}\{\boldsymbol{d}_1, \dots, \boldsymbol{d}_i\}$ spanned by the actions in which the inverse $\hat{\boldsymbol{K}}^{-1}$ is perfectly identified. Each step $\boldsymbol{d}_i$ expanding this explored subspace is $\hat{\boldsymbol{K}}$-orthogonal to the previous ones.

Proposition 1 (Conjugate Direction Method)
Let the actions $\boldsymbol{s}_i$ of Algorithm 1 be linearly independent. Then Algorithm 1 is a conjugate direction method, i.e. it holds that $\boldsymbol{d}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{d}_j = 0$ for all $i \neq j$.

Proof. Without loss of generality assume $i > j$. Then the result follows directly from Lemma S1. □

Geometrically, Algorithm 1 iteratively projects the representer weights onto the expanding subspace $\operatorname{span}\{\boldsymbol{S}_i\}$ with respect to $\langle \cdot, \cdot \rangle_{\hat{\boldsymbol{K}}}$. We can use this intuition to understand the convergence of the representer weights estimate. The relative error $\rho(i)$ at iteration i is given by how small the "angle" between this subspace and the representer weights vector is.

Proposition 2 (Relative Error Bound for the Representer Weights)
For any choice of actions a relative error bound $\rho(i)$, s.t. $\|\boldsymbol{v}_* - \boldsymbol{v}_i\|_{\hat{\boldsymbol{K}}} \leq \rho(i)\|\boldsymbol{v}_*\|_{\hat{\boldsymbol{K}}}$ is given by

$$\rho(i) = (\bar{\boldsymbol{v}}_*^\intercal \underbrace{(\boldsymbol{I} - \boldsymbol{C}_i\hat{\boldsymbol{K}})}_{\textit{projection onto } \operatorname{span}\{\boldsymbol{S}_i\}^{\perp_{\hat{\boldsymbol{K}}}}} \bar{\boldsymbol{v}}_*)^{\frac{1}{2}} \leq \lambda_{\max}(\boldsymbol{I} - \boldsymbol{C}_i\hat{\boldsymbol{K}}) \leq 1 \tag{9}$$

where $\bar{\boldsymbol{v}}_* = \boldsymbol{v}_*/\|\boldsymbol{v}_*\|_{\hat{\boldsymbol{K}}}$. If the actions $\{\boldsymbol{s}_i\}_{i=1}^n$ are linearly independent, then $\rho(i) \leq \delta_{n=i}$.

Proof. See Section S2.2. □

Proposition 2 guarantees convergence in at most n iterations, if the actions are chosen to be linearly independent, since $\boldsymbol{C}_i\hat{\boldsymbol{K}}$ is a $\hat{\boldsymbol{K}}$-orthogonal projection onto $\operatorname{span}\{\boldsymbol{S}_i\}$ (see Lemma S1). Therefore, if our finite computational budget is large enough, we eventually recover the mathematical posterior. This is reflected by the contraction of the posterior over the representer weights (see Proposition S4). The bound in Proposition 2 is tight without further assumptions on the actions, since there exists an adverserial sequence of actions such that the first $(n-1)$ are in $\operatorname{span}\{\boldsymbol{v}_*\}^{\perp_{\hat{\boldsymbol{K}}}}$. Then the inverse is perfectly identified in that subspace, but $\boldsymbol{v}_i = \boldsymbol{C}_i\boldsymbol{y} = \boldsymbol{C}_i\hat{\boldsymbol{K}}\boldsymbol{v}_* = \boldsymbol{0}$. In practice, one can derive tighter convergence bounds for specific sequences of actions. For example, for randomized actions the bound depends on their distribution [56, 57]. If residuals $\boldsymbol{r}_i$ are chosen as actions, we obtain

$$\rho(i) = 2\left(\tfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i \text{ or } \rho(i) = \left(\tfrac{\lambda_{n-i}-\lambda_1}{\lambda_{n-i}+\lambda_1}\right) \tag{10}$$

since then Algorithm 1's estimate of the representer weights equals that of CG (Corollary S2). Here κ is the condition number and λ_j the eigenvalues of either (i) the kernel matrix $\hat{\boldsymbol{K}}$ if $\boldsymbol{s}_i = \boldsymbol{r}_i$, or (ii) the preconditioned kernel matrix $\hat{\boldsymbol{P}}^{-\frac{1}{2}}\hat{\boldsymbol{K}}\hat{\boldsymbol{P}}^{-\frac{\intercal}{2}}$ if $\boldsymbol{s}_i = \hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i$.

3.2 Convergence in RKHS Norm of the Posterior Mean

Having established convergence of the representer weights estimate, we can use this result to prove convergence in norm of IterGP's posterior mean to the mathematical posterior at the same rate.

Theorem 1 (Convergence in RKHS Norm of the Posterior Mean Approximation)
Let $\mathcal{H}_k$ be the RKHS associated with kernel $k(\cdot,\cdot)$, $\sigma^2 > 0$ and let $\mu_* - \mu \in \mathcal{H}_k$ be the unique solution to the regularized empirical risk minimization problem

$$\arg\min_{f\in\mathcal{H}_k} \tfrac{1}{n}\big(\textstyle\sum_{j=1}^n (f(\boldsymbol{x}_j) - y_j + \mu(\boldsymbol{x}_j))^2 + \sigma^2\|f\|_{\mathcal{H}_k}^2\big) \tag{11}$$

which is equivalent to the mathematical posterior mean up to shift by the prior μ [e.g. 1, Sec. 6.2]. Then for $i \in \{0,\dots,n\}$ the posterior mean $\mu_i(\cdot)$ computed by Algorithm 1 satisfies

$$\boxed{\|\mu_* - \mu_i\|_{\mathcal{H}_k} \leq \rho(i)c(\sigma^2)\|\mu_* - \mu_0\|_{\mathcal{H}_k}} \tag{12}$$

where $\mu_0 = \mu$ is the prior mean and the constant $c(\sigma^2) = \sqrt{1 + \frac{\sigma^2}{\lambda_{\min}(\boldsymbol{K})}} \to 1$ as $\sigma^2 \to 0$.

Proof. See Section S2.3. □

Theorem 1 gives a bound on the RKHS-norm error between the posterior mean μ_i of IterGP and the mathematical posterior mean μ_*. If for the given prior kernel a bound on the RKHS-norm error $\|h - \mu_*\|_{\mathcal{H}_k}$ between the latent function h and the mathematical posterior mean μ_* is known, Theorem 1 can be directly used to bound the RKHS-norm error between IterGP's posterior mean and the latent function h via the triangle inequality: $\|h - \mu_i\|_{\mathcal{H}_k} \leq \underbrace{\|h - \mu_*\|_{\mathcal{H}_k}}_{\to 0 \text{ as } n\to\infty} + \underbrace{\|\mu_* - \mu_i\|_{\mathcal{H}_k}}_{\to 0 \text{ as } i\to n}$.

3.3 Combined and Computational Uncertainty as Worst Case Errors

While Theorem 1 shows convergence in norm for IterGP's posterior mean, the convergence rate $\rho(i)$ may contain expressions which cannot be evaluated at runtime with the limited computation at our disposal. For example, for residual actions evaluating eq. (10) requires computation of the kernel matrix spectrum. However, the combined uncertainty of IterGP is a tight bound on the *pointwise* relative error to all possible latent functions which would have resulted in the same computations.

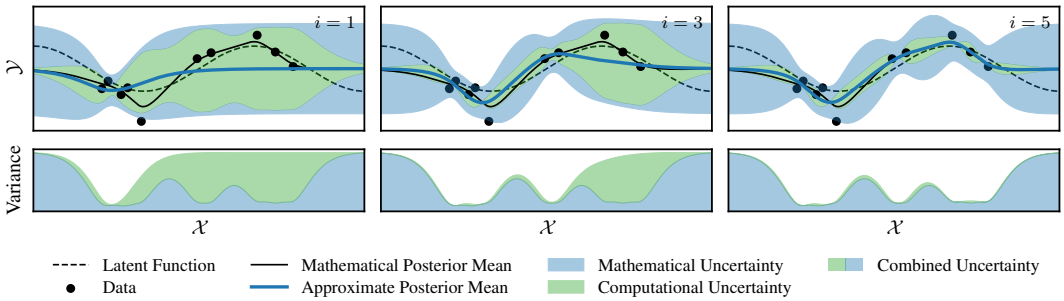


Figure 3: *Computational and combined uncertainty of IterGP as worst-case bounds.*[2]

Theorem 2 (Combined and Computational Uncertainty as Worst Case Errors)
Let $\sigma^2 \geq 0$ and let $k_i(\cdot,\cdot) = k_*(\cdot,\cdot) + k_i^{\text{comp}}(\cdot,\cdot)$ be the combined uncertainty computed by Algorithm 1. Then, for any $\boldsymbol{x} \in \mathcal{X}$ (assuming $\boldsymbol{x} \notin \boldsymbol{X}$ if $\sigma^2 > 0$) we have

$$\sup_{g \in \mathcal{H}_{k^\sigma} : \|g\|_{\mathcal{H}_{k^\sigma}} \leq 1} \overbrace{\underbrace{g(\boldsymbol{x}) - \mu_*^g(\boldsymbol{x})}_{\text{error of math. post. mean}} + \underbrace{\mu_*^g(\boldsymbol{x}) - \mu_i^g(\boldsymbol{x})}_{\text{computational error}}}^{\text{error of approximate posterior mean}} = \sqrt{k_i(\boldsymbol{x},\boldsymbol{x}) + \sigma^2}, \quad \textit{and} \tag{13}$$

$$\sup_{g \in \mathcal{H}_{k^\sigma} : \|g\|_{\mathcal{H}_{k^\sigma}} \leq 1} \underbrace{\mu_*^g(\boldsymbol{x}) - \mu_i^g(\boldsymbol{x})}_{\text{computational error}} = \sqrt{k_i^{\text{comp}}(\boldsymbol{x},\boldsymbol{x})} \tag{14}$$

where $\mu_*^g(\cdot) = k(\cdot,\boldsymbol{X})\hat{\boldsymbol{K}}^{-1} g(\boldsymbol{X})$ is the mathematical and $\mu_i^g(\cdot) = k(\cdot,\boldsymbol{X})\boldsymbol{C}_i g(\boldsymbol{X})$ IterGP's posterior mean for the latent function $g \in \mathcal{H}_{k^\sigma}$. If $\sigma^2 = 0$, then the above also holds for $\boldsymbol{x} \in \boldsymbol{X}$.

Proof. See Section S2.4. □

Theorem 2 rigorously explains why the combined (mathematical + computational) uncertainty k_i is the correct object characterizing our belief about the latent function h, given that we are in the limited compute regime. In the same way that the mathematical uncertainty is a tight bound on the distance to all functions g which could have produced the data (see [36, Prop. 3.8]), the combined uncertainty is a tight bound on all functions g which would have produced the same computations.

3.4 Pointwise Convergence of the Posterior Mean

In particular, as Corollary 1 shows and Figure 3 illustrates, the computational uncertainty (🟢) is a pointwise bound on the relative distance to the mathematical posterior mean (16) and *the combined uncertainty (🟢 + 🔵) is a pointwise bound on the relative distance to the true latent function* (15).

Corollary 1 (Pointwise Convergence of the Posterior Mean)
Assume the conditions of Theorem 2 hold and assume the latent function $h \in \mathcal{H}_{k^\sigma}$. Let μ_* be the corresponding mathematical posterior mean and μ_i the posterior mean computed by Algorithm 1. Then it holds that

$$\frac{|h(\boldsymbol{x}) - \mu_i(\boldsymbol{x})|}{\|h\|_{\mathcal{H}_{k^\sigma}}} \leq \sqrt{k_i(\boldsymbol{x},\boldsymbol{x}) + \sigma^2}, \quad \textit{and} \tag{15}$$

$$\frac{|\mu_*(\boldsymbol{x}) - \mu_i(\boldsymbol{x})|}{\|h\|_{\mathcal{H}_{k^\sigma}}} \leq \sqrt{k_i^{\text{comp}}(\boldsymbol{x},\boldsymbol{x})}. \tag{16}$$

Proof. This follows immediately from Theorem 2 by recognizing that $h/\|h\|_{\mathcal{H}_{k^\sigma}}$ has unit norm. □

[2]The combined (co-)variance decomposes into mathematical and computational covariances, as opposed to the combined standard deviation since $\sqrt{🟢 + 🔵} \neq \sqrt{🟢} + \sqrt{🔵}$. The bottom panel thus illustrates the variance decomposition. However, to better illustrate Theorem 2, in the upper panel we plot the combined standard deviation $\sqrt{🟢 + 🔵}$ and computational standard deviation $\sqrt{🟢}$ within it, in line with standard GP plotting practice.

It is worth noting that Theorem 2 and Corollary 1 generally *do not hold for other GP approximations*. They explicitly rely on $\boldsymbol{C}_i\hat{\boldsymbol{K}}$ being the $\hat{\boldsymbol{K}}$-orthogonal projection onto the space spanned by the actions (see Lemma S1). Since orthogonal projections are unique, if another GP approximation is such a projection and therefore satisfies Theorem 2, it is in fact an instance of IterGP.

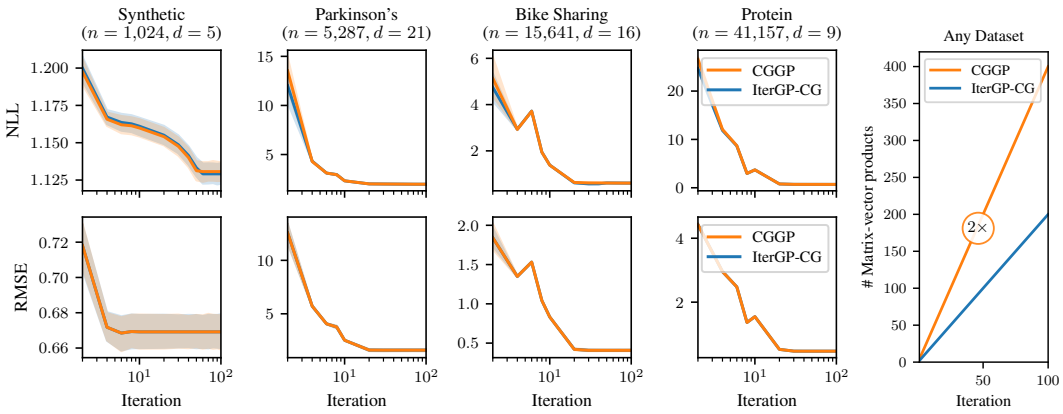
4 Experiments

To demonstrate the effects of quantifying computational uncertainty we perform GP regression on synthetic and benchmark datasets for the two most common GP approximations in the large-scale setting, SVGP [22] and CGGP [26], and their direct analogs from our class of methods. An implementation of Algorithm 1, based on KeOps [48] and ProbNum [60], is available at:

`https://github.com/JonathanWenger/itergp`

Experimental Setup We consider a synthetic dataset of iid uniformly sampled inputs $\boldsymbol{x}_j \in [-1, 1]^d$ with $y(\boldsymbol{x}) = \sin(\pi \boldsymbol{x}^\intercal \mathbf{1}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, as well as a range of UCI datasets [61] with training set sizes $n = 5,287$ to $57,247$, dimensions $d = 9$ to 26 and standardized features. All experiments were run on an NVIDIA GeForce RTX 2080 Ti graphics card. We perform GP regression using a zero mean prior and a Matérn($\frac{1}{2}$) kernel (for other kernels see Section S4). All experiments were run 10 times with randomly sampled training and test splits of $90/10$ and we report average metrics with 95% confidence intervals.

IterGP reduces the necessary computations for CG-based GP inference. We compare IterGP to the CG-based GP inference used in the GPyTorch library [26]. For all datasets, we select hyperparameters using the training procedure of Wenger et al. [29]. As we show in Theorem S5, the posterior mean of IterGP with (conjugate) residual actions is exactly equivalent to performing CG to compute the representer weights. Therefore, both methods produce the exact same posterior mean estimate and thus achieve the same RMSE as a function of CG iterations (Figure 4, bottom). The primary difference between the two methods is in the posterior variance. The combined variance estimate of IterGP is essentially "free" in the sense that it reuses terms from the posterior mean calculation. In contrast, computing the posterior variance with CG requires $n_\diamond$ additional linear solves $(\hat{\boldsymbol{K}}^{-1}\boldsymbol{x}_{\diamond 1}, \ldots, \hat{\boldsymbol{K}}^{-1}\boldsymbol{x}_{\diamond n_\diamond})$. GPyTorch relies on the Lanczos Variance Estimate technique [62] which essentially warm-starts each of these solves by reusing quantities from the linear solve $\hat{\boldsymbol{K}}^{-1}k(\boldsymbol{X}, \boldsymbol{x}_{\diamond 1})$. While this approach produces reliable variance estimates that converge to the true posterior variance, it requires additional computation: at least one set of additional CG iterations to compute $\hat{\boldsymbol{K}}^{-1}k(\boldsymbol{X}, \boldsymbol{x}_{\diamond 1})$. In Figure 4(a) (top), we see that IterGP and GPyTorch's CGGP achieve nearly identical NLL, suggesting that both methods produce variances that yield similar generalization.



(a) Generalization on synthetic and UCI benchmark datasets. (b) Comp. Cost

Figure 4: *Generalization of CGGP and its closest IterGP analog.* (a) GP regression using a Matérn($\frac{1}{2}$) kernel on UCI datasets. The plot shows the average generalization error in terms of NLL and RMSE for an increasing number of solver iterations. The posterior mean of IterGP-CG and CGGP is identical, which explains the identical RMSE. However, CGGP performs additional computation for the posterior covariance as (b) illustrates, which is not needed since IterGP-CG has identical NLL.

The key difference between the methods is that 1) unlike CGGP, IterGP's variances exactly capture both mathematical and computational uncertainty, and 2) IterGP's variances require no additional solves, resulting in half as much computation as GPyTorch's CGGP implementation (see Figure 4(b)).

Quantifying computational uncertainty improves generalization of inducing point methods. To understand the benefits of quantifying computational uncertainty, we compare the linear-time SVGP method (which does not quantify computational uncertainty) with the closest (quadratic-time) inducing point analog from our proposed IterGP framework (see Section 2.1). While the IterGP method is inherently more expensive than SVGP, our goal is simply to demonstrate that inducing points can yield far more accuracy if one has the budget to account for computational uncertainty. To that end, we compare SVGP against IterGP using the same set of randomly-placed inducing points. We identify a set of kernel hyperparameters by optimizing the ELBO of SVGP on the training data, using these for both SVGP and IterGP. As Figure 5 shows, we find that across all datasets that IterGP offers better RMSE and NLL than SVGP, despite the fact that the hyperparameters are chosen to favor SVGP. This suggests that the extra computation needed to quantify computational uncertainty can more "effectively" utilize a set of inducing points for predictive models.

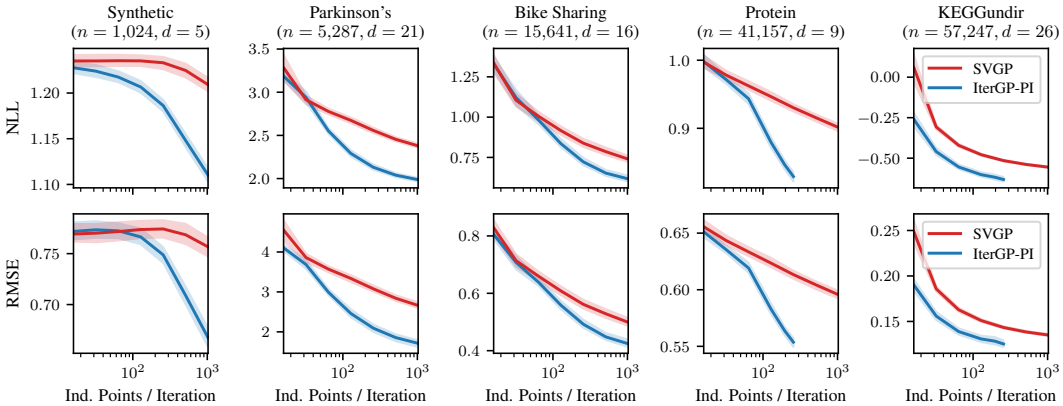


Figure 5: *Generalization of SVGP and its closest IterGP analog.* GP regression using a Matérn($\frac{1}{2}$) kernel on UCI datasets. The plot shows the average generalization error in terms of NLL and RMSE for an increasing number of identical inducing points. After a small number of inducing points relative to the size of the training data, IterGP has significantly lower generalization error than SVGP.

5 Conclusion

Scalable GP approximations inevitably introduce error, leading to a worse model for the latent function in question. This work demonstrates that it is possible to account for *both* uncertainty arising from limited data *and* uncertainty arising from limited computation *exactly* – which as we show improves model performance. IterGP methods return this combined uncertainty which crucially represents a dataset-specific, pointwise worst-case bound on the error to the true latent function. At its core, IterGP performs repeated matrix-vector multiplication resulting in quadratic complexity. Since modern computing architectures (i.e. GPUs) have been specifically designed for this operation at scale, iterative approaches for GP approximation are becoming competitive with theoretically cheaper approximations, like inducing point methods [26, 27]. Finally, in addition to the general utility of IterGP, we expect this class of methods to be particularly useful in applications where accurate uncertainty quantification is important or, due to its inherently online nature, where data is acquired sequentially such as in active learning and Bayesian optimization.

Acknowledgments and Disclosure of Funding

JW, MP and PH gratefully acknowledge financial support by the European Research Council through ERC StG Action 757275 / PANAMA; the DFG Cluster of Excellence "Machine Learning - New Perspectives for Science", EXC 2064/1, project number 390727645; the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A); and funds from the Ministry of Science, Research and Arts of the State of Baden-Württemberg. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting JW and MP. JPC and GP are supported by the Simons Foundation, the McKnight Foundation, the Grossman Center, and the Gatsby Charitable Trust. The authors would like to thank Hanna Dettki, Frank Schneider, Lukas Tatzel and all anonymous reviewers for helpful feedback on an earlier version of this manuscript.

References

[1] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[2] Huaiyu Zhu, Christopher KI Williams, Richard Rohwer, and Michal Morciniec. Gaussian regression and optimal finite dimensional linear models. In *Neural Networks and Machine Learning*, 1997.

[3] Giancarlo Ferrari Trecate, Christopher KI Williams, and Manfred Opper. Finite-dimensional approximation of Gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 218–224, 1999.

[4] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

[5] Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[6] Andrew Gordon Wilson, Christoph Dann, and Hannes Nickisch. Thoughts on massively scalable Gaussian processes. *arXiv pre-print*, 2015. URL http://arxiv.org/abs/1511.01870.

[7] Zichao Yang, Andrew Wilson, Alex Smola, and Le Song. A la carte–learning fast kernels. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1098–1106, 2015.

[8] Pavel Izmailov, Alexander Novikov, and Dmitry Kropotov. Scalable Gaussian processes with billions of inducing inputs via tensor train decomposition. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[9] Trefor Evans and Prasanth Nair. Scalable Gaussian processes with grid-structured eigenfunctions (GP-GRIEF). In *International Conference on Machine Learning (ICML)*, 2018.

[10] Amir Zandieh, Navid Nouri, Ameya Velingker, Michael Kapralov, and Ilya Razenshteyn. Scaling up kernel ridge regression via locality sensitive hashing. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[11] Aldo V Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312, 1988.

[12] Abhirup Datta, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812, 2016.

[13] Matthias Katzfuss and Joseph Guinness. A General Framework for Vecchia Approximations of Gaussian Processes. *Statistical Science*, 36(1):124 – 141, 2021.

[14] Florian Schäfer, Matthias Katzfuss, and Houman Owhadi. Sparse Cholesky factorization by Kullback-Leibler minimization. *SIAM Journal on Scientific Computing*, 43(3):A2019–A2046, 2021.

[15] Alex J Smola and Peter Bartlett. Sparse greedy Gaussian process regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 598–604, 2000.

[16] Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2001.

[17] Petros Drineas and Michael W Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

[18] Matthias W. Seeger, C. K. Williams, and N. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2003.

[19] Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems (NeurIPS)*, 18:1257–1264, 2005.

[20] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[21] Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 567–574. PMLR, 2009.

[22] James Hensman, Nicolò Fusi, and Neil D Lawrence. Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 282–290, 2013.

[23] Mark Gibbs. Bayesian Gaussian processes for classification and regression. *University of Cambridge, Cambridge*, 1997.

[24] Iain Murray. Gaussian processes and fast matrix-vector multiplies. In *Numerical Mathematics in Machine Learning Workshop (ICML)*, 2009.

[25] Kurt Cutajar, Michael Osborne, John Cunningham, and Maurizio Filippone. Preconditioning kernel matrices. In *International Conference on Machine Learning (ICML)*, 2016.

[26] Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018:7576–7586, 2018.

[27] Ke Alexander Wang, Geoff Pleiss, Jacob R Gardner, Stephen Tyree, Kilian Q Weinberger, and Andrew Gordon Wilson. Exact Gaussian processes on a million data points. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[28] Artem Artemev, David R Burt, and Mark van der Wilk. Tighter bounds on the log marginal likelihood of Gaussian process regression using conjugate gradients. In *International Conference on Machine Learning (ICML)*, 2021.

[29] Jonathan Wenger, Geoff Pleiss, Philipp Hennig, John P. Cunningham, and Jacob R. Gardner. Preconditioning for scalable Gaussian process hyperparameter optimization. In *International Conference on Machine Learning (ICML)*, 2022.

[30] Matthias Bauer, Mark van der Wilk, and Carl Edward Rasmussen. Understanding probabilistic sparse Gaussian process approximations. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016.

[31] Jonathan H. Huggins, Trevor Campbell, Mikolaj Kasprzak, and Tamara Broderick. Scalable Gaussian process inference with finite-data mean and variance guarantees. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[32] Philipp Hennig, Mike A. Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471(2179), 2015.

[33] Jon Cockayne, Chris Oates, TJ Sullivan, and Mark Girolami. Bayesian probabilistic numerical methods. *SIAM Review*, 61(4):756–789, 2019.

[34] Chris Oates and TJ Sullivan. A modern retrospective on probabilistic numerics. *Statistics and Computing*, 10 2019.

[35] Philipp Hennig, Michael A. Osborne, and Hans P. Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, June 2022. ISBN 9781316681411. doi: 10.1017/9781316681411.

[36] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint*, 2018. URL http://arxiv.org/abs/1807.02582.

[37] Florian Schäfer, TJ Sullivan, and Houman Owhadi. Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. *Multiscale Modeling and Simulation*, 19 (2):688–730, 2021.

[38] Magnus Rudolph Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49, 1952.

[39] John P Cunningham, Krishna V Shenoy, and Maneesh Sahani. Fast Gaussian process methods for point process intensity estimation. In *International Conference on Machine Learning (ICML)*, 2008.

[40] Andres Potapczynski, Luhuan Wu, Dan Biderman, Geoff Pleiss, and John P Cunningham. Bias-free scalable Gaussian processes via randomized truncations. In *International Conference on Machine Learning (ICML)*, 2021.

[41] Bernhard W Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):1–21, 1985.

[42] Lehel Csató and Manfred Opper. Sparse on-line Gaussian processes. *Neural computation*, 14(3):641–668, 2002.

[43] Veit Wild, Motonobu Kanagawa, and Dino Sejdinovic. Connections and equivalences between the Nyström method and sparse variational Gaussian processes. *arXiv pre-print*, 2021. URL http://arxiv.org/abs/2106.01121.

[44] David R. Burt, Carl Edward Rasmussen, and Mark Van Der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *International Conference on Machine Learning (ICML)*, 2019.

[45] Andre G Journel and Charles J Huijbregts. *Mining geostatistics*. Academic Press London, 1976.

[46] James T Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Deisenroth. Efficiently sampling functions from Gaussian process posteriors. In *International Conference on Machine Learning (ICML)*, pages 10292–10302. PMLR, 2020.

[47] James T Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Peter Deisenroth. Pathwise conditioning of Gaussian processes. *arXiv pre-print*, 2020. URL http://arxiv.org/abs/2011.04026.

[48] Benjamin Charlier, Jean Feydy, Joan Alexis Glaunès, François-David Collin, and Ghislain Durif. Kernel operations on the GPU, with autodiff, without memory overflows. *Journal of Machine Learning Research (JMLR)*, 22(74):1–6, 2021. URL http://jmlr.org/papers/v22/20-275.html.

[49] Philipp Hennig. Probabilistic interpretation of linear solvers. *SIAM Journal on Optimization*, 25(1): 234–260, 2015.

[50] Simon Bartels, Jon Cockayne, Ilse CF Ipsen, and Philipp Hennig. Probabilistic linear solvers: A unifying view. *Statistics and Computing*, 29(6):1249–1263, 2019.

[51] Jon Cockayne, Chris Oates, Ilse C. Ipsen, and Mark Girolami. A Bayesian conjugate gradient method. *Bayesian Analysis*, 14(3):937–1012, 2019.

[52] Jon Cockayne, Ilse C. F. Ipsen, Chris J. Oates, and Tim W. Reid. Probabilistic iterative methods for linear systems. *arXiv pre-print*, 2020. URL http://arxiv.org/abs/2012.12615.

[53] Jonathan Wenger and Philipp Hennig. Probabilistic linear solvers for machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[54] Tim W. Reid, Ilse C. F. Ipsen, Jon Cockayne, and Chris J. Oates. BayesCG as an uncertainty aware version of CG. *arXiv pre-print*, 2022. URL http://arxiv.org/abs/2008.03225.

[55] Stefan Kaczmarz. Angenäherte Auflösung von Systemen linearer Gleichungen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques*, pages 355–357, 1937.

[56] Thomas Strohmer and Roman Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[57] Robert M Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

[58] Robert M Gower. Sketch and project: Randomized iterative methods for linear systems and inverting matrices. *arXiv preprint*, 2016. URL http://arxiv.org/abs/1612.06013.

[59] Simon Bartels and Philipp Hennig. Conjugate gradients for kernel machines. *Journal of Machine Learning Research*, 21(55):1–42, 2020.

[60] Jonathan Wenger, Nicholas Krämer, Nathanael Bosch, Marvin Pförtner, Jonathan Schmidt, Nina Effenberger, Maren Mahsereci, Johannes Zenn, Thomas Gläßle, Alexandra Gessner, Toni Karvonen, Philipp Hennig, and ProbNum Contributors. ProbNum: Probabilistic Numerics in Python, 2021. URL https://github.com/probabilistic-numerics/probnum.

[61] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[62] Geoff Pleiss, Jacob Gardner, Kilian Weinberger, and Andrew Gordon Wilson. Constant-time predictive distributions for Gaussian processes. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4114–4123, 2018.

[63] Yousef Saad, Manshung Yeung, Jocelyne Erhel, and Frédéric Guyomarc'h. A deflated version of the conjugate gradient algorithm. *SIAM Journal on Scientific Computing*, 21(5):1909–1926, 2000.

[64] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[65] Jason Frank and Cornelis Vuik. On the construction of deflation-based preconditioners. *SIAM Journal on Scientific Computing*, 23(2):442–462, 2001.

[66] Jean-Paul Chiles and Pierre Delfiner. *Geostatistics: modeling spatial uncertainty*, volume 497. John Wiley & Sons, 2009.

[67] Arnaud Doucet. A note on efficient conditional simulation of Gaussian distributions. Technical report, Departments of Computer Science and Statistics, University of British Columbia, 2010.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs are included in the supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Material: Posterior and Computational Uncertainty in Gaussian Processes

Jonathan Wenger[1,2] **Geoff Pleiss**[2]

Marvin Pförtner[1] **Philipp Hennig**[1,3] **John P. Cunningham**[2]

[1] University of Tübingen
[2] Columbia University
[3] Max Planck Institute for Intelligent Systems, Tübingen

This supplementary material contains additional results and in particular proofs for all theoretical statements. References referring to sections, equations or theorem-type environments within this document are prefixed with 'S', while references to, or results from, the main paper are stated as is.

S1 Connections to Other GP Approximations **16**
S1.1 Pivoted Cholesky Decomposition . . . 16
S1.2 Singular / Eigenvalue Decomposition . . . 17
S1.3 Conjugate Gradient Method . . . 18
S1.4 Inducing Point Methods . . . 21

S2 Theoretical Results and Proofs **21**
S2.1 Properties of Algorithm 1 . . . 21
S2.2 Approximation of Representer Weights . . . 25
S2.3 Convergence Analysis of the Posterior Mean Approximation . . . 26
S2.4 Combined Uncertainty as Worst Case Error . . . 27

S3 Implementation of Algorithm 1 **28**
S3.1 Policy Choice . . . 28
S3.2 Stopping Criterion . . . 29
S3.3 Efficient Sampling from the Combined Posterior . . . 29

S4 Additional Experimental Results **29**

S1 Connections to Other GP Approximations

S1.1 Pivoted Cholesky Decomposition

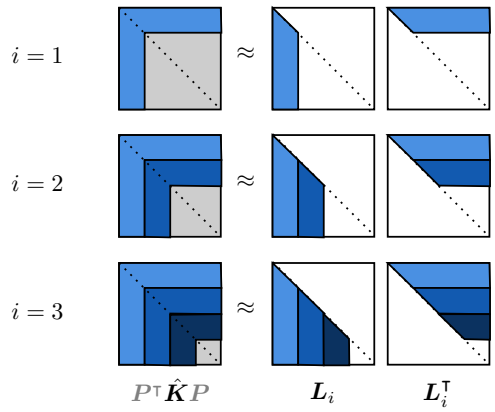
Theorem S3 (Pivoted Cholesky Decomposition)
Let $(j_i)_{i=1}^n$ be a set of indices defining the pivot elements of the pivoted Cholesky decomposition and $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ the corresponding permutation matrix. Assume the actions of Algorithm 1 are given by the standard unit vectors $\boldsymbol{s}_i = \boldsymbol{P}\boldsymbol{e}_i = \boldsymbol{e}_{j_i}$, i.e.

$$(\boldsymbol{s}_i)_j = (\boldsymbol{e}_{j_i})_j = \begin{cases} 1 & \text{if } j = j_i \\ 0 & \text{otherwise} \end{cases}. \tag{S17}$$

Then Algorithm 1 recovers the pivoted Cholesky decomposition, i.e. it holds for all $i \in \{0, \dots, n\}$ that

$$\boldsymbol{P}^\intercal \boldsymbol{Q}_i \boldsymbol{P} = \boldsymbol{L}_i \boldsymbol{L}_i^\intercal, \tag{S18}$$

where $\boldsymbol{L}_i \in \mathbb{R}^{n \times i}$ is the (partial) Cholesky factor of $\boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P}$ as computed by Algorithm S2.



Algorithm S2: (Pivoted) Cholesky Decomposition

Input: kernel matrix $\hat{\boldsymbol{K}}$, permutation matrix $\boldsymbol{P}$
Output: lower triangular $\boldsymbol{L}_i$, s.t. $\boldsymbol{L}_i \boldsymbol{L}_i^\intercal \approx \boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P}$

procedure CHOLESKY($\hat{\boldsymbol{K}}$, $\boldsymbol{P}$)
  $\boldsymbol{A} \leftarrow \boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P}$
  for $i \in \{1, \dots, n\}$ **do**
    $\boldsymbol{l}_i \leftarrow \boldsymbol{A}_{:i} / \sqrt{\boldsymbol{A}_{ii}}$
    $\boldsymbol{A} \leftarrow \boldsymbol{A} - \boldsymbol{l}_i \boldsymbol{l}_i^\intercal = \boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P} - \boldsymbol{L}_i \boldsymbol{L}_i^\intercal$
    $\boldsymbol{L}_i = (\boldsymbol{L}_{i-1} \quad \boldsymbol{l}_i)$
  return $\boldsymbol{L}_i$

Figure S1: *Cholesky decomposition.* Every column added to the lower triangular Cholesky factor $\boldsymbol{L}$ defines the ith "right angle ruler"-pattern in $\boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P}$. The bottom right matrix in gray given by $\boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P} - \boldsymbol{L}_i \boldsymbol{L}_i^\intercal = \boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P} - \sum_{j=1}^{i} \boldsymbol{l}_j \boldsymbol{l}_j^\intercal$ changes every iteration.

Proof. We proceed by induction. Assume (S18) holds after i iterations of Algorithm 1. For the base case $i = 0$, it holds by assumption that $\boldsymbol{P}^\intercal \boldsymbol{Q}_0 \boldsymbol{P} = \boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{C}_0 \hat{\boldsymbol{K}} \boldsymbol{P} = \boldsymbol{0}$. Now for the induction step $i \to i + 1$, we have

$$
\begin{aligned}
\frac{1}{\eta_{i+1}} \hat{\boldsymbol{K}} \boldsymbol{d}_i \boldsymbol{d}_i^\intercal \hat{\boldsymbol{K}} &= \frac{1}{\eta_{i+1}} \hat{\boldsymbol{K}} \boldsymbol{\Sigma}_i \hat{\boldsymbol{K}} \boldsymbol{s}_{i+1} \boldsymbol{s}_{i+1}^\intercal \hat{\boldsymbol{K}} \boldsymbol{\Sigma}_i \hat{\boldsymbol{K}} \\
&= \frac{1}{\eta_{i+1}} \hat{\boldsymbol{K}} (\boldsymbol{\Sigma}_0 - \boldsymbol{C}_i) \hat{\boldsymbol{K}} \boldsymbol{s}_{i+1} \boldsymbol{s}_{i+1}^\intercal \hat{\boldsymbol{K}} (\boldsymbol{\Sigma}_0 - \boldsymbol{C}_i) \hat{\boldsymbol{K}} \\
&= \frac{1}{\eta_{i+1}} (\hat{\boldsymbol{K}} - \boldsymbol{Q}_i) \boldsymbol{s}_{i+1} \boldsymbol{s}_{i+1}^\intercal (\hat{\boldsymbol{K}} - \boldsymbol{Q}_i) \\
&\overset{\text{IH}}{=} \frac{1}{\eta_{i+1}} (\hat{\boldsymbol{K}} - \boldsymbol{P} \boldsymbol{L}_i \boldsymbol{L}_i^\intercal \boldsymbol{P}^\intercal) \boldsymbol{s}_{i+1} \boldsymbol{s}_{i+1}^\intercal (\hat{\boldsymbol{K}} - \boldsymbol{P} \boldsymbol{L}_i \boldsymbol{L}_i^\intercal \boldsymbol{P}^\intercal) \\
&= \frac{(\hat{\boldsymbol{K}} - \boldsymbol{P} \boldsymbol{L}_i \boldsymbol{L}_i^\intercal \boldsymbol{P}^\intercal) \boldsymbol{P} \boldsymbol{e}_{i+1}}{\sqrt{\boldsymbol{e}_{i+1}^\intercal \boldsymbol{P}^\intercal (\hat{\boldsymbol{K}} - \boldsymbol{P} \boldsymbol{L}_i \boldsymbol{L}_i^\intercal \boldsymbol{P}^\intercal) \boldsymbol{P} \boldsymbol{e}_{i+1}}} \frac{\boldsymbol{e}_{i+1}^\intercal \boldsymbol{P}^\intercal (\hat{\boldsymbol{K}} - \boldsymbol{P} \boldsymbol{L}_i \boldsymbol{L}_i^\intercal \boldsymbol{P}^\intercal)}{\sqrt{\boldsymbol{e}_{i+1}^\intercal \boldsymbol{P}^\intercal (\hat{\boldsymbol{K}} - \boldsymbol{P} \boldsymbol{L}_i \boldsymbol{L}_i^\intercal \boldsymbol{P}^\intercal) \boldsymbol{P} \boldsymbol{e}_{i+1}}} \\
&= \frac{\boldsymbol{P} (\boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P} - \boldsymbol{L}_i \boldsymbol{L}_i^\intercal) \boldsymbol{e}_{i+1}}{\sqrt{\boldsymbol{e}_{i+1}^\intercal (\boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P} - \boldsymbol{L}_i \boldsymbol{L}_i^\intercal) \boldsymbol{e}_{i+1}}} \frac{\boldsymbol{e}_{i+1}^\intercal (\boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P} - \boldsymbol{L}_i \boldsymbol{L}_i^\intercal) \boldsymbol{P}^\intercal}{\sqrt{\boldsymbol{e}_{i+1}^\intercal (\boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P} - \boldsymbol{L}_i \boldsymbol{L}_i^\intercal) \boldsymbol{e}_{i+1}}} \\
&= \boldsymbol{P} \boldsymbol{l}_{i+1} \boldsymbol{l}_{i+1}^\intercal \boldsymbol{P}^\intercal.
\end{aligned}
$$

where $\boldsymbol{l}_{i+1}$ is given by Algorithm S2. Combining this with one more use of the induction hypothesis we obtain

$$
\begin{aligned}
\boldsymbol{P}^\intercal \boldsymbol{Q}_{i+1} \boldsymbol{P} &= \boldsymbol{P}^\intercal \boldsymbol{Q}_i \boldsymbol{P} + \frac{1}{\eta_{i+1}} \boldsymbol{P}^\intercal \hat{\boldsymbol{K}} \boldsymbol{d}_{i+1} \boldsymbol{d}_{i+1}^\intercal \hat{\boldsymbol{K}} \boldsymbol{P} \\
&= \boldsymbol{L}_i \boldsymbol{L}_i^\intercal + \boldsymbol{l}_{i+1} \boldsymbol{l}_{i+1}^\intercal = (\boldsymbol{L}_i \quad \boldsymbol{l}_{i+1}) \begin{pmatrix} \boldsymbol{L}_i^\intercal \\ \boldsymbol{l}_{i+1}^\intercal \end{pmatrix} = \boldsymbol{L}_{i+1} \boldsymbol{L}_{i+1}^\intercal
\end{aligned}
$$

This proves the claim.

□

S1.2 Singular / Eigenvalue Decomposition

Theorem S4 (Singular / Eigenvalue Decomposition)
Let the actions $\boldsymbol{s}_i = \mathbf{u}_i$ of Algorithm 1 be given by the eigenvectors $\mathbf{u}_i$ of $\hat{\boldsymbol{K}}$ in arbitrary order. Then

for $i \in \{1, \dots, n\}$ it holds that

$$\boldsymbol{C}_i = \boldsymbol{U}_i \boldsymbol{\Lambda}_i^{-1} \boldsymbol{U}_i^\intercal = \mathrm{SVD}_i(\hat{\boldsymbol{K}}^{-1})$$
$$\boldsymbol{Q}_i = \boldsymbol{U}_i \boldsymbol{\Lambda}_i \boldsymbol{U}_i^\intercal = \mathrm{SVD}_i(\hat{\boldsymbol{K}}),$$

where $\boldsymbol{U} = (\mathbf{u}_1, \dots, \mathbf{u}_i) \in \mathbb{R}^{n \times i}$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_i) \in \mathbb{R}^{i \times i}$ is the diagonal matrix of eigenvalues of $\hat{\boldsymbol{K}}$ with the order given by the order of the actions.

Proof. It holds by assumption and eq. (S37), that

$$\boldsymbol{C}_i = \boldsymbol{S}_i (\boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{S}_i)^{-1} \boldsymbol{S}_i^\intercal = \boldsymbol{U}_i (\boldsymbol{U}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{U}_i)^{-1} \boldsymbol{U}_i^\intercal = \boldsymbol{U}_i \boldsymbol{\Lambda}_i^{-1} \boldsymbol{U}_i^\intercal,$$

as well as

$$\boldsymbol{Q}_i = \hat{\boldsymbol{K}} \boldsymbol{C}_i \hat{\boldsymbol{K}} = \hat{\boldsymbol{K}} \boldsymbol{U}_i \boldsymbol{\Lambda}_i^{-1} \boldsymbol{U}_i^\intercal \hat{\boldsymbol{K}} = \boldsymbol{U}_i \boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^{-1} \boldsymbol{\Lambda}_i \boldsymbol{U}_i^\intercal = \boldsymbol{U}_i \boldsymbol{\Lambda}_i \boldsymbol{U}_i^\intercal$$

This proves the claim. □

S1.3 Conjugate Gradient Method

Algorithm S3: Preconditioned Conjugate Gradient Method [38]

Input: kernel matrix $\hat{\boldsymbol{K}}$, labels $\boldsymbol{y}$, prior mean $\boldsymbol{\mu}$, preconditioner $\hat{\boldsymbol{P}}$
Output: representer weights $\boldsymbol{v}_i \approx \hat{\boldsymbol{K}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$

procedure CG($\hat{\boldsymbol{K}}, \boldsymbol{y} - \boldsymbol{\mu}, \hat{\boldsymbol{P}}$)
  $\boldsymbol{v}_0 \leftarrow \mathbf{0}$
  $\boldsymbol{s}_0 \leftarrow \mathbf{0}$
  while $\|\boldsymbol{r}_i\|_2 > \max(\delta_{\text{rtol}} \|\boldsymbol{y}\|_2, \delta_{\text{atol}})$ **and** $i < i_{\max}$ **do**
    $\boldsymbol{r}_{i-1} \leftarrow (\boldsymbol{y} - \boldsymbol{\mu}) - \hat{\boldsymbol{K}} \boldsymbol{v}_{i-1}$
    $\boldsymbol{s}_i \leftarrow \hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_{i-1} - \frac{(\hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_{i-1})^\intercal \hat{\boldsymbol{K}} \boldsymbol{s}_{i-1}}{\boldsymbol{s}_{i-1}^\intercal \hat{\boldsymbol{K}} \boldsymbol{s}_{i-1}} \boldsymbol{s}_{i-1}$
    $\boldsymbol{v}_i \leftarrow \boldsymbol{v}_{i-1} + \frac{(\hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_{i-1})^\intercal \boldsymbol{r}_{i-1}}{\boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{s}_i} \boldsymbol{s}_i$
  return $\boldsymbol{v}$

Theorem S5 (Preconditioned Conjugate Gradient Method)
Let $\hat{\boldsymbol{P}} \in \mathbb{R}^{n \times n}$ be a symmetric positive definite preconditioner. Assume the actions of Algorithm 1 are given by

$$\boldsymbol{s}_1^{\mathrm{CG}} = \hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_0$$
$$\boldsymbol{s}_i^{\mathrm{CG}} = \hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_{i-1} - \frac{(\hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_{i-1})^\intercal \hat{\boldsymbol{K}} \boldsymbol{s}_{i-1}}{\boldsymbol{s}_{i-1}^\intercal \hat{\boldsymbol{K}} \boldsymbol{s}_{i-1}} \boldsymbol{s}_{i-1} \tag{S19}$$

the preconditioned conjugate gradient method. Then Algorithm 1 recovers preconditioned CG initialized at $\boldsymbol{v}_0^{\mathrm{CG}} = \mathbf{0}$, i.e. it holds for $i \in \{1, \dots, n\}$ that

$$\boldsymbol{s}_i = \boldsymbol{d}_i = \boldsymbol{s}_i^{\mathrm{CG}} \tag{S20}$$
$$\boldsymbol{v}_i = \boldsymbol{v}_i^{\mathrm{CG}} \tag{S21}$$
$$\boldsymbol{r}_{i-1} = \boldsymbol{r}_{i-1}^{\mathrm{CG}}. \tag{S22}$$

Proof. First note that by assumption $\boldsymbol{s}_i = \boldsymbol{s}_i^{\mathrm{CG}}$ for all i. We prove the remaining claims by induction. For the base case we have by assumption $\boldsymbol{d}_0 = \boldsymbol{\Sigma}_0 \hat{\boldsymbol{K}} \boldsymbol{s}_0 = \boldsymbol{s}_0 = \boldsymbol{s}_0^{\mathrm{CG}}$ and $\boldsymbol{v}_0 = \mathbf{0} = \boldsymbol{v}_0^{\mathrm{CG}}$. Now for the induction step $i \to i+1$ assume the hypotheses (S20), (S21) and (S22) hold $\forall j \leq i$. Using the properties of CG it holds for $j' < i$ that

$$\boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{s}_{j'} = 0 \tag{S23}$$
$$(\hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_i)^\intercal \boldsymbol{s}_{j'} = 0 \tag{S24}$$
$$(\hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_i)^\intercal \boldsymbol{r}_{j'} = 0 \tag{S25}$$
$$\langle \boldsymbol{s}_1, \dots, \boldsymbol{s}_i \rangle = \langle \boldsymbol{r}_0, \hat{\boldsymbol{P}}^{-1} \hat{\boldsymbol{K}} \boldsymbol{r}_0, \dots, (\hat{\boldsymbol{P}}^{-1} \hat{\boldsymbol{K}})^{i-1} \boldsymbol{r}_0 \rangle = \langle \hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_0, \dots, \hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_{i-1} \rangle \tag{S26}$$

We now first show $\hat{\boldsymbol{K}}$-conjugacy of the actions in iteration $i+1$. We have for $j \leq i$ that

$$\begin{aligned}\boldsymbol{s}_{i+1}^\intercal \hat{\boldsymbol{K}} \boldsymbol{s}_j &= \big(\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i - \frac{(\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i)^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_i}{\boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_i}\boldsymbol{s}_i\big)^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_j \\ &= (\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i)^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_j - \frac{(\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i)^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_i}{\boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_i}\boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_j\end{aligned}$$

Now if $j = i$, clearly $\boldsymbol{s}_{i+1}^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_j = \boldsymbol{s}_{i+1}^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_i = 0$. If $j < i$, we have using (S26), that

$$\hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{K}}\boldsymbol{s}_j \in \langle \hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{K}}\boldsymbol{r}_0, \ldots, (\hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{K}})^j \boldsymbol{r}_0\rangle \subset \langle \hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_0, \ldots, \hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_j\rangle. \tag{S27}$$

Therefore we obtain for $j < i$, that

$$\boldsymbol{s}_{i+1}^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_j \overset{\text{(S23)}}{=} \boldsymbol{r}_i^\intercal \hat{\boldsymbol{P}}^{-1}\hat{\boldsymbol{K}}\boldsymbol{s}_j \overset{\text{(S27)}}{=} \boldsymbol{r}_i^\intercal \left(\sum_{\ell=1}^{j} \gamma_\ell \hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_\ell\right) \overset{\text{(S25)}}{=} 0. \tag{S28}$$

Thus in combination we have

$$\forall j \in \{1, \ldots, i\}: \quad \boldsymbol{s}_{i+1}^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_j = 0. \tag{S29}$$

Now for the search direction we have

$$\begin{aligned}\boldsymbol{d}_{i+1} = \boldsymbol{\Sigma}_i \hat{\boldsymbol{K}}\boldsymbol{s}_{i+1} &= \left(\boldsymbol{\Sigma}_0 - \sum_{j=1}^{i} \frac{\boldsymbol{d}_j \boldsymbol{d}_j^\intercal}{\eta_j}\right)\hat{\boldsymbol{K}}\boldsymbol{s}_{i+1} \\ &= \boldsymbol{s}_{i+1} - \sum_{j=1}^{i} \frac{\boldsymbol{d}_j^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_{i+1}}{\eta_j}\boldsymbol{d}_j \overset{\text{(S20)}}{=} \boldsymbol{s}_{i+1} - \sum_{j=1}^{i} \frac{\boldsymbol{s}_j^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_{i+1}}{\eta_j}\boldsymbol{d}_j \\ &\overset{\text{(S29)}}{=} \boldsymbol{s}_{i+1}.\end{aligned} \tag{S30}$$

Further, we have for the solution estimate, that $\boldsymbol{v}_{i+1} = \boldsymbol{v}_i + \boldsymbol{d}_{i+1}\frac{\alpha_{i+1}}{\eta_{i+1}}$. It holds that

$$\begin{aligned}\alpha_{i+1} = \boldsymbol{s}_{i+1}^\intercal \boldsymbol{r}_i &= \big(\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i - \frac{(\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i)^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_i}{\boldsymbol{s}_i^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_i}\boldsymbol{s}_i\big)^\intercal \boldsymbol{r}_i \\ &= (\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i)^\intercal \boldsymbol{r}_i - \sum_{j=}^{i} c_j (\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_{j-1})^\intercal \boldsymbol{r}_i \overset{\text{(S25)}}{=} (\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i)^\intercal \boldsymbol{r}_i\end{aligned}$$

as well as

$$\eta_{i+1} = \boldsymbol{s}_{i+1}^\intercal \hat{\boldsymbol{K}}\boldsymbol{\Sigma}_i \hat{\boldsymbol{K}}\boldsymbol{s}_{i+1} = \boldsymbol{d}_{i+1}^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_{i+1} \overset{\text{(S30)}}{=} \boldsymbol{s}_{i+1}^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_{i+1}$$

Combining the above and recalling Algorithm S3, we obtain

$$\boldsymbol{v}_{i+1} = \boldsymbol{v}_i + \boldsymbol{d}_{i+1}\frac{\alpha_{i+1}}{\eta_{i+1}} = \boldsymbol{v}_i + \boldsymbol{d}_{i+1}\frac{(\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i)^\intercal \boldsymbol{r}_i}{\boldsymbol{s}_{i+1}^\intercal \hat{\boldsymbol{K}}\boldsymbol{s}_{i+1}} = \boldsymbol{v}_{i+1}^{\mathrm{CG}}.$$

Finally, the residual is computed identically in Algorithm 1 as in Algorithm S3, giving

$$\boldsymbol{r}_i = (\boldsymbol{y} - \boldsymbol{\mu}) - \hat{\boldsymbol{K}}\boldsymbol{v}_i = (\boldsymbol{y} - \boldsymbol{\mu}) - \hat{\boldsymbol{K}}\boldsymbol{v}_i^{\mathrm{CG}} = \boldsymbol{r}_i^{\mathrm{CG}}.$$

This proves the claims. □

Corollary S2 (Preconditioned Gradient Actions as CG Actions)
Choosing actions

$$\boldsymbol{s}_i = \hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_{i-1} \tag{S31}$$

in Theorem S5 instead also reproduces the preconditioned conjugate gradient method, i.e. it holds for $i \in \{1, \ldots, n\}$ that

$$\boldsymbol{d}_i = \boldsymbol{s}_i^{\mathrm{CG}} \tag{S32}$$

$$\boldsymbol{v}_i = \boldsymbol{v}_i^{\mathrm{CG}} \tag{S33}$$

$$\boldsymbol{r}_{i-1} = \boldsymbol{r}_{i-1}^{\mathrm{CG}}. \tag{S34}$$

Proof. It suffices to show that $\boldsymbol{d}_i = \boldsymbol{s}_i^{\mathrm{CG}}$. The rest of the argument is then identical to the proof of Theorem S5. We prove the claim by induction. For the base case by assumption $\boldsymbol{s}_1 = \hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_0 = \boldsymbol{s}_1^{\mathrm{CG}}$. Now for the induction step $i \to i+1$, assume that $\boldsymbol{d}_j = \boldsymbol{s}_j$ for all $j \leq i$, then

$$\begin{aligned}
\boldsymbol{d}_{i+1} &= \boldsymbol{\Sigma}_i \hat{\boldsymbol{K}} \hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_i \\
&= (\boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}}) \hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_i \\
&= \hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i - \boldsymbol{D}_i(\boldsymbol{D}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{D}_i)^{-1} \boldsymbol{D}_i^\intercal \hat{\boldsymbol{K}} \hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_i && \text{By eq. (S37).} \\
&\overset{\text{IH}}{=} \hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i - \boldsymbol{S}_i^{\mathrm{CG}}((\boldsymbol{S}_i^{\mathrm{CG}})^\intercal \hat{\boldsymbol{K}} \boldsymbol{S}_i^{\mathrm{CG}})^{-1} (\boldsymbol{S}_i^{\mathrm{CG}})^\intercal \hat{\boldsymbol{K}} \hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_i
\end{aligned}$$

Now by the same argument as in eq. (S28) in the proof of Theorem S5 we have for all $j < i$ that $\boldsymbol{r}_i^\intercal \hat{\boldsymbol{P}}^{-1} \hat{\boldsymbol{K}} \boldsymbol{s}_j^{\mathrm{CG}} = 0$. Therefore

$$\begin{aligned}
&= \hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i - \boldsymbol{s}_i^{\mathrm{CG}}((\boldsymbol{s}_i^{\mathrm{CG}})^\intercal \hat{\boldsymbol{K}} \boldsymbol{s}_i^{\mathrm{CG}})^{-1} (\boldsymbol{s}_i^{\mathrm{CG}})^\intercal \hat{\boldsymbol{K}} \hat{\boldsymbol{P}}^{-1} \boldsymbol{r}_i \\
&= \boldsymbol{s}_{i+1}^{\mathrm{CG}} && \text{By eq. (S19).}
\end{aligned}$$

This proves the claim. □

Corollary S3 (Deflated Conjugate Gradient Method)
Let the first $0 < \ell < n$ actions $(\boldsymbol{s}_i)_{i=1}^{\ell}$ of Algorithm 1 be linearly independent and the remaining ones be given by $\boldsymbol{s}_i = \hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i$, where $\hat{\boldsymbol{P}} \approx \hat{\boldsymbol{K}}$ is a preconditioner. Then Algorithm 1 is equivalent to the preconditioned deflated CG algorithm [63, Alg. 3.6] with deflation subspace $\operatorname{span}\{\boldsymbol{S}_\ell\}$.

Proof. By the form of preconditioned deflated CG given in Algorithm 3.6 of Saad et al. [63] and Corollary S2, it suffices to show that the residual $\boldsymbol{r}_\ell$ satisfies $\boldsymbol{S}_\ell^\intercal \boldsymbol{r}_\ell = \boldsymbol{0}$ and that for all $i > \ell$, it holds that

$$\boldsymbol{s}_i^{\mathrm{defCG}} = \boldsymbol{d}_i = (\boldsymbol{I} - \boldsymbol{C}_{i-1}\hat{\boldsymbol{K}})\boldsymbol{s}_i.$$

Now it holds by Lemma S2 and eq. (S37), that

$$\boldsymbol{S}_\ell^\intercal \boldsymbol{r}_\ell = \boldsymbol{S}_\ell^\intercal (\boldsymbol{I} - \hat{\boldsymbol{K}}\boldsymbol{C}_\ell)(\boldsymbol{y} - \boldsymbol{\mu}) = \underbrace{\boldsymbol{S}_\ell^\intercal (\boldsymbol{I} - \hat{\boldsymbol{K}}\boldsymbol{S}_\ell(\boldsymbol{S}_\ell^\intercal \hat{\boldsymbol{K}}\boldsymbol{S}_\ell)^{-1}\boldsymbol{S}_\ell^\intercal)}_{=\boldsymbol{0}}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}.$$

This proves the first claim. Now, by Saad et al. [63, Alg. 3.6], the search directions $(\boldsymbol{s}_i^{\mathrm{defCG}})_{i=\ell+1}^{n}$ of preconditioned deflated CG are given by

$$\begin{aligned}
\boldsymbol{s}_i^{\mathrm{defCG}} &= \boldsymbol{s}_i^{\mathrm{CG}} - \boldsymbol{S}_\ell(\boldsymbol{S}_\ell^\intercal \hat{\boldsymbol{K}}\boldsymbol{S}_\ell)^{-1}\boldsymbol{S}_\ell^\intercal \hat{\boldsymbol{K}}\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i \\
&= (\boldsymbol{I} - \boldsymbol{C}_{\ell+1:(i-1)}\hat{\boldsymbol{K}})\boldsymbol{s}_i - \boldsymbol{S}_\ell(\boldsymbol{S}_\ell^\intercal \hat{\boldsymbol{K}}\boldsymbol{S}_\ell)^{-1}\boldsymbol{S}_\ell^\intercal \hat{\boldsymbol{K}}\hat{\boldsymbol{P}}^{-1}\boldsymbol{r}_i && \text{Corollary S2} \\
&= (\boldsymbol{I} - \boldsymbol{C}_{\ell+1:(i-1)}\hat{\boldsymbol{K}})\boldsymbol{s}_i - \boldsymbol{C}_\ell \hat{\boldsymbol{K}}\boldsymbol{s}_i \\
&= (\boldsymbol{I} - (\boldsymbol{C}_{\ell+1:(i-1)} - \boldsymbol{C}_\ell)\hat{\boldsymbol{K}})\boldsymbol{s}_i \\
&= (\boldsymbol{I} - \boldsymbol{C}_{i-1}\hat{\boldsymbol{K}})\boldsymbol{s}_i \\
&= \boldsymbol{d}_i
\end{aligned}$$

This proves the claim. □

Remark S1 (Preconditioning and Algorithm 1)
Iterative methods typically have convergence rates depending on the condition number of the system matrix. One successful strategy in practice to accelerate convergence is to use a preconditioner $\hat{\boldsymbol{P}} \approx \hat{\boldsymbol{K}}$ [64]. A preconditioner needs to be cheap to compute and allow efficient matrix-vector multiplication $\boldsymbol{v} \mapsto \hat{\boldsymbol{P}}^{-1}\boldsymbol{v}$. Now, Algorithm 1 implicitly constructs and applies a *deflation-based preconditioner*, which are defined via a deflation subspace to be projected out [65]. In Algorithm 1 this is precisely the already explored space $\operatorname{span}\{\boldsymbol{S}_i\} = \operatorname{span}\{\boldsymbol{D}_i\}$ spanned by the actions. Therefore, if we run a mixed strategy, meaning first choosing actions that define a certain subspace and then choose residual actions, we recover the *deflated conjugate gradient method* [63] (see Corollary S3 for a proof). Alternatively, one can also use byproducts of the iteration of Algorithm 1 to construct a diagonal-plus-low-rank preconditioner of the form $\hat{\boldsymbol{P}} = \sigma^2\boldsymbol{I} + \boldsymbol{U}\boldsymbol{U}^\intercal \approx \hat{\boldsymbol{K}}$ where $\boldsymbol{U} = \boldsymbol{K}\boldsymbol{D}_i \operatorname{diag}(\eta_1, \ldots, \eta_i) \in \mathbb{R}^{n \times i}$. Therefore, again if running a mixed strategy, one can first construct a preconditioner and then accelerate the convergence of subsequent CG iterations. In this sense one can double-dip in terms of preconditioning (conjugate) gradient iterations by combining these two techniques *at essentially no overhead*.

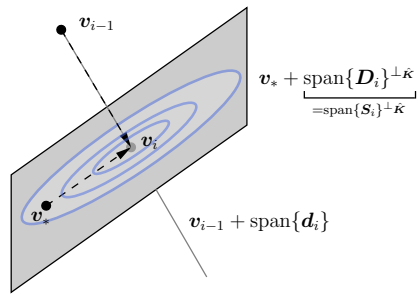


Figure S2: *Geometric perspective on the probabilistic linear solver learning representer weights* $\boldsymbol{v}_*$.

S1.4 Inducing Point Methods

Theorem S6 (Approximate Posterior Mean of Nyström, SoR, DTC and SVGP)
Let $\boldsymbol{Z} \in \mathbb{R}^{n \times m}$ *be a set of distinct inducing inputs such that* $\operatorname{rank}(\boldsymbol{K}_{\boldsymbol{XZ}}) = m \leq n$*. Then the posterior mean of the Nyström variants subset of regressors (SoR) and deterministic training conditional (DTC) is identical to the one of SVGP and given by*

$$\begin{aligned} \mu(\cdot) &= k(\cdot, \boldsymbol{Z})(\boldsymbol{K}_{\boldsymbol{ZX}}\boldsymbol{K}_{\boldsymbol{XZ}} + \sigma^2 \boldsymbol{K}_{\boldsymbol{ZZ}})^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu}) \\ &= q(\cdot, \boldsymbol{X})\boldsymbol{K}_{\boldsymbol{XZ}}(\boldsymbol{K}_{\boldsymbol{ZX}}(q(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I})\boldsymbol{K}_{\boldsymbol{XZ}})^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu}) \end{aligned} \tag{S35}$$

Proof. First, note that by eqns. (16b) and (20b) of Quiñonero-Candela and Rasmussen [20] the posterior mean of SoR and DTC is identical and given by

$$\mu(\cdot) = k(\cdot, \boldsymbol{Z})(\boldsymbol{K}_{\boldsymbol{ZX}}\boldsymbol{K}_{\boldsymbol{XZ}} + \sigma^2 \boldsymbol{K}_{\boldsymbol{ZZ}})^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu})$$

Now, by Theorem 5 of Wild et al. [43] the posterior mean of SVGP for a fixed set of inducing points is equivalent to the Nyström approximation, which takes the form above. Recognizing that $\boldsymbol{K}_{\boldsymbol{ZX}}\boldsymbol{K}_{\boldsymbol{XZ}} \in \mathbb{R}^{m \times m}$ is invertible, it holds that

$$\begin{aligned} \mu(\cdot) &= k(\cdot, \boldsymbol{Z})(\boldsymbol{K}_{\boldsymbol{ZX}}\boldsymbol{K}_{\boldsymbol{XZ}} + \sigma^2 \boldsymbol{K}_{\boldsymbol{ZZ}})^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu}) \\ &= k(\cdot, \boldsymbol{Z})(\boldsymbol{K}_{\boldsymbol{ZZ}}(\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}\boldsymbol{K}_{\boldsymbol{XZ}} + \sigma^2 \boldsymbol{I}))^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu}) \\ &= k(\cdot, \boldsymbol{Z})\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}((\boldsymbol{K}_{\boldsymbol{ZX}}\boldsymbol{K}_{\boldsymbol{XZ}})^{-1}(\boldsymbol{K}_{\boldsymbol{ZX}}\boldsymbol{K}_{\boldsymbol{XZ}}\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}\boldsymbol{K}_{\boldsymbol{XZ}} + \sigma^2 \boldsymbol{K}_{\boldsymbol{ZX}}\boldsymbol{K}_{\boldsymbol{XZ}}))^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu}) \\ &= k(\cdot, \boldsymbol{Z})\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}\boldsymbol{K}_{\boldsymbol{XZ}}(\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{K}_{\boldsymbol{XZ}}\boldsymbol{K}_{\boldsymbol{ZZ}}^{-1}\boldsymbol{K}_{\boldsymbol{ZX}} + \sigma^2 \boldsymbol{I})\boldsymbol{K}_{\boldsymbol{XZ}})^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu}) \\ &= q(\cdot, \boldsymbol{X})\boldsymbol{K}_{\boldsymbol{XZ}}(\boldsymbol{K}_{\boldsymbol{ZX}}(q(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I})\boldsymbol{K}_{\boldsymbol{XZ}})^{-1}\boldsymbol{K}_{\boldsymbol{ZX}}(\boldsymbol{y} - \boldsymbol{\mu}) \end{aligned}$$

This proves the claim. □

S2 Theoretical Results and Proofs

S2.1 Properties of Algorithm 1

Lemma S1 (Geometric Properties of Algorithm 1)
Let $i \in \{1, \ldots, n\}$*, and assume* $\boldsymbol{\Sigma}_0$ *is chosen such that* $\boldsymbol{\Sigma}_0 \hat{\boldsymbol{K}} \boldsymbol{s}_j = \boldsymbol{s}_j$ *for all* $j \leq i$ *(e.g.* $\boldsymbol{\Sigma}_0 = \hat{\boldsymbol{K}}^{-1}$*). Then it holds for the quantities computed by Algorithm 1 that*

$$\operatorname{span}\{\boldsymbol{S}_i\} = \operatorname{span}\{\boldsymbol{D}_i\} \tag{S36}$$

$$\boldsymbol{C}_i = \boldsymbol{D}_i(\boldsymbol{D}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{D}_i)^{-1}\boldsymbol{D}_i^\intercal = \boldsymbol{S}_i(\boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{S}_i)^{-1}\boldsymbol{S}_i^\intercal \tag{S37}$$

$$\boldsymbol{C}_i\hat{\boldsymbol{K}} \text{ is the } \hat{\boldsymbol{K}}\text{-orthogonal projection onto } \operatorname{span}\{\boldsymbol{D}_i\} \tag{S38}$$

$$\boldsymbol{\Sigma}_i\hat{\boldsymbol{K}} \text{ is the } \hat{\boldsymbol{K}}\text{-orthogonal projection onto } \operatorname{span}\{\boldsymbol{D}_i\}^{\perp_{\hat{\boldsymbol{K}}}} \tag{S39}$$

$$\boldsymbol{d}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{d}_j = 0 \qquad \text{for all } j < i \tag{S40}$$

where $\boldsymbol{S}_i = (\boldsymbol{s}_1 \cdots \boldsymbol{s}_i) \in \mathbb{R}^{n \times i}$ *and* $\boldsymbol{D}_i = (\boldsymbol{d}_1 \cdots \boldsymbol{d}_i) \in \mathbb{R}^{n \times i}$.

Proof. We prove the claims by induction. We begin with the base case $i = 1$.

By assumption it holds that $\boldsymbol{S}_1 = \boldsymbol{s}_1 = \boldsymbol{\Sigma}_0 \hat{\boldsymbol{K}} \boldsymbol{s}_1 = \boldsymbol{d}_1 = \boldsymbol{D}_1$. Now by Algorithm 1, we have $\boldsymbol{C}_1 = \frac{1}{\eta_1} \boldsymbol{d}_1 \boldsymbol{d}_1^\intercal$, which with the above proves (S37). By the batched form (S37) of $\boldsymbol{C}_i$, the statements (S38) and (S39) follow immediately. Finally, $\hat{\boldsymbol{K}}$-orthogonality for a single search direction holds trivially.

Now for the induction step $i \to i+1$. Assume that eqs. (S36) to (S40) hold for iteration i. Then we have that

$$\boldsymbol{d}_{i+1} = \boldsymbol{\Sigma}_i \hat{\boldsymbol{K}} \boldsymbol{s}_{i+1} = \boldsymbol{s}_{i+1} - \boldsymbol{C}_i \hat{\boldsymbol{K}} \boldsymbol{s}_{i+1} \overset{\text{(S37)}}{=} \boldsymbol{s}_{i+1} - \boldsymbol{S}_i (\boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{S}_i)^{-1} \boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{s}_{i+1} \in \operatorname{span}\{\boldsymbol{S}_{i+1}\}$$

By the induction hypothesis the above also implies $\operatorname{span}\{\boldsymbol{S}_{i+1}\} = \operatorname{span}\{\boldsymbol{D}_{i+1}\}$. This proves eq. (S36). Next, we have by the induction hypotheses (S37) and (S40) that

$$\begin{aligned}
\boldsymbol{C}_{i+1} &= \boldsymbol{C}_i + \frac{1}{\eta} \boldsymbol{d}_{i+1} \boldsymbol{d}_{i+1}^\intercal \\
&= \boldsymbol{D}_i (\boldsymbol{D}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{D}_i)^{-1} \boldsymbol{D}_i^\intercal + \frac{1}{\eta_{i+1}} \boldsymbol{d}_{i+1} \boldsymbol{d}_{i+1}^\intercal \\
&= \sum_{k=1}^{i+1} \frac{1}{\eta_k} \boldsymbol{d}_k \boldsymbol{d}_k^\intercal \\
&= \boldsymbol{D}_{i+1} (\boldsymbol{D}_{i+1}^\intercal \hat{\boldsymbol{K}} \boldsymbol{D}_{i+1})^{-1} \boldsymbol{D}_{i+1}^\intercal
\end{aligned}$$

This proves the first equality of eq. (S37). For the second, first recognize that an orthogonal projection onto a linear subspace $\operatorname{span}\{\boldsymbol{A}\}$ with respect to the $\boldsymbol{B}$-inner product is given by $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}^\intercal \boldsymbol{B} \boldsymbol{A})^{-1} \boldsymbol{A}^\intercal \boldsymbol{B}$. The projection onto its $\boldsymbol{B}$-orthogonal subspace is given by $\boldsymbol{P}_{\boldsymbol{A}^{\perp_{\boldsymbol{B}}}} = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{A}}$. Therefore eqs. (S38) and (S39) follow directly from the above argument. Now since projection onto a subspace is unique and independent of the choice of basis, we have by $\operatorname{span}\{\boldsymbol{D}_{i+1}\} = \operatorname{span}\{\boldsymbol{S}_{i+1}\}$ that

$$\boldsymbol{C}_i \hat{\boldsymbol{K}} = \boldsymbol{P}_{\boldsymbol{D}_{i+1}} = \boldsymbol{P}_{\boldsymbol{S}_{i+1}} = \boldsymbol{S}_i (\boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{S}_i)^{-1} \boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}}$$

Now since $\hat{\boldsymbol{K}}$ is non-singular, the second equality of eq. (S37) follows. Finally, we will prove $\hat{\boldsymbol{K}}$-orthogonality of the search directions. Let $j < i+1$, then it holds that

$$\boldsymbol{d}_{i+1}^\intercal \hat{\boldsymbol{K}} \boldsymbol{d}_j = (\underbrace{\boldsymbol{\Sigma}_i \hat{\boldsymbol{K}} \boldsymbol{s}_{i+1}}_{\in \operatorname{span}\{\boldsymbol{S}_i\}^{\perp_{\hat{\boldsymbol{K}}}}})^\intercal \hat{\boldsymbol{K}} \underbrace{\boldsymbol{d}_j}_{\in \operatorname{span}\{\boldsymbol{S}_i\}} = 0$$

by eqs. (S36) and (S39). This completes the proof. □

Corollary S4
Let $i \in \{1, \dots, n\}$. It holds for $\boldsymbol{C}_i \hat{\boldsymbol{K}}$, the $\hat{\boldsymbol{K}}$-orthogonal projection onto $\boldsymbol{S}_i$, that

$$(\boldsymbol{C}_i \hat{\boldsymbol{K}})^2 = \boldsymbol{C}_i \hat{\boldsymbol{K}} \tag{S41}$$

$$\boldsymbol{C}_i \hat{\boldsymbol{K}} \boldsymbol{C}_i = \boldsymbol{C}_i \tag{S42}$$

Further for $\boldsymbol{H}_i = \boldsymbol{\Sigma}_i \hat{\boldsymbol{K}} = \boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}}$ the $\hat{\boldsymbol{K}}$-orthogonal projection onto $\boldsymbol{S}_i^{\perp_{\hat{\boldsymbol{K}}}}$, we have

$$\boldsymbol{H}_i^2 = \boldsymbol{H}_i \tag{S43}$$

$$\boldsymbol{H}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{H}_i = \boldsymbol{H}_i^\intercal \hat{\boldsymbol{K}} = \hat{\boldsymbol{K}} \boldsymbol{H}_i \tag{S44}$$

Proof. By Lemma S1, it holds that $\boldsymbol{C}_i = \boldsymbol{S}_i (\boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{S}_i)^{-1} \boldsymbol{S}_i^\intercal$. Therefore

$$\boldsymbol{C}_i \hat{\boldsymbol{K}} \boldsymbol{C}_i = \boldsymbol{S}_i (\boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{S}_i)^{-1} \boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{S}_i (\boldsymbol{S}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{S}_i)^{-1} \boldsymbol{S}_i^\intercal = \boldsymbol{C}_i.$$

This proves (S42) and (S41). Define $\boldsymbol{H}_i = \boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}}$, then

$$\boldsymbol{H}_i \boldsymbol{H}_i = (\boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}})(\boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}}) = \boldsymbol{I} - 2\boldsymbol{C}_i \hat{\boldsymbol{K}} + (\boldsymbol{C}_i \hat{\boldsymbol{K}})^2 = \boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}} = \boldsymbol{H}_i$$

as well as

$$\begin{aligned}
\boldsymbol{H}_i^\intercal \hat{\boldsymbol{K}} \boldsymbol{H}_i &= (\boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}})^\intercal \hat{\boldsymbol{K}} (\boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}}) = (\hat{\boldsymbol{K}} - \hat{\boldsymbol{K}} \boldsymbol{C}_i \hat{\boldsymbol{K}})(\boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}}) \\
&= \hat{\boldsymbol{K}} - 2\hat{\boldsymbol{K}} \boldsymbol{C}_i \hat{\boldsymbol{K}} + \hat{\boldsymbol{K}} (\boldsymbol{C}_i \hat{\boldsymbol{K}})^2 \\
&= \hat{\boldsymbol{K}} - \hat{\boldsymbol{K}} \boldsymbol{C}_i \hat{\boldsymbol{K}} = \boldsymbol{H}_i^\intercal \hat{\boldsymbol{K}} = \hat{\boldsymbol{K}} \boldsymbol{H}_i.
\end{aligned}$$

□

Lemma S2
Let $\boldsymbol{\Sigma}_0 = \hat{\boldsymbol{K}}^{-1}$, then it holds that

$$\boldsymbol{C}_i(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{v}_i, \tag{S45}$$
$$\boldsymbol{\Sigma}_i(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{v}_* - \boldsymbol{v}_i. \tag{S46}$$

Proof. We prove the statement by induction. By assumption $\boldsymbol{C}_0(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{v}_0$. Now assume (S45) holds. Then for $i \to i+1$, we have

$$\boldsymbol{C}_{i+1}(\boldsymbol{y} - \boldsymbol{\mu}) = (\boldsymbol{C}_i + \frac{1}{\eta_{i+1}}\boldsymbol{d}_{i+1}\boldsymbol{d}_{i+1}^\intercal)(\boldsymbol{y} - \boldsymbol{\mu}) \overset{\text{IH}}{=} \boldsymbol{v}_i + \frac{\boldsymbol{d}_{i+1}^\intercal(\boldsymbol{y} - \boldsymbol{\mu})}{\eta_{i+1}}\boldsymbol{d}_{i+1}$$

Now by the update to the representer weights in Algorithm 1 it suffices to show that $\alpha_{i+1} = \boldsymbol{d}_{i+1}^\intercal(\boldsymbol{y} - \boldsymbol{\mu})$. We have

$$\begin{aligned}\boldsymbol{d}_{i+1}^\intercal(\boldsymbol{y} - \boldsymbol{\mu}) &= (\boldsymbol{\Sigma}_i\hat{\boldsymbol{K}}\boldsymbol{s}_{i+1})^\intercal(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{s}_{i+1}^\intercal\hat{\boldsymbol{K}}\boldsymbol{\Sigma}_i(\boldsymbol{y} - \boldsymbol{\mu}) \\ &= \boldsymbol{s}_{i+1}^\intercal\hat{\boldsymbol{K}}(\hat{\boldsymbol{K}}^{-1} - \boldsymbol{C}_i)(\boldsymbol{y} - \boldsymbol{\mu}) \overset{\text{IH}}{=} \boldsymbol{s}_{i+1}^\intercal((\boldsymbol{y} - \boldsymbol{\mu}) - \hat{\boldsymbol{K}}\boldsymbol{v}_i) = \boldsymbol{s}_{i+1}^\intercal\boldsymbol{r}_i = \alpha_i.\end{aligned}$$

□

Lemma S3
Let $\boldsymbol{\Sigma}_0 = \hat{\boldsymbol{K}}^{-1}$, $\boldsymbol{C}_0 = \boldsymbol{0}$ and consequently $\boldsymbol{v}_0 = \boldsymbol{0}$, then it holds for the residual at iteration $i \in \{1, \dots, n\}$ that

$$\boldsymbol{r}_{i-1} = \hat{\boldsymbol{K}}(\boldsymbol{v}_* - \boldsymbol{v}_{i-1}) \tag{S47}$$
$$= \hat{\boldsymbol{K}}\boldsymbol{\Sigma}_{i-1}\hat{\boldsymbol{K}}\boldsymbol{v}_* \tag{S48}$$
$$= (\hat{\boldsymbol{K}} - \boldsymbol{Q}_{i-1})\boldsymbol{v}_*. \tag{S49}$$

Proof. It holds by definition, that

$$\boldsymbol{r}_{i-1} = (\boldsymbol{y} - \boldsymbol{\mu}) - \hat{\boldsymbol{K}}\boldsymbol{v}_{i-1} = \hat{\boldsymbol{K}}\boldsymbol{v}_* - \hat{\boldsymbol{K}}\boldsymbol{v}_{i-1} = \hat{\boldsymbol{K}}(\boldsymbol{v}_* - \boldsymbol{v}_{i-1}).$$

Further we have by eq. (S46), that

$$= \hat{\boldsymbol{K}}\boldsymbol{\Sigma}_{i-1}(\boldsymbol{y} - \boldsymbol{\mu}) = \hat{\boldsymbol{K}}\boldsymbol{\Sigma}_{i-1}\hat{\boldsymbol{K}}\boldsymbol{v}_*,$$

and finally, by the definition of the kernel matrix approximation in Algorithm 1, we obtain

$$= \hat{\boldsymbol{K}}(\hat{\boldsymbol{K}}^{-1} - \boldsymbol{C}_{i-1})\hat{\boldsymbol{K}}\boldsymbol{v}_* = (\hat{\boldsymbol{K}} - \boldsymbol{Q}_{i-1})\boldsymbol{v}_*.$$

□

Proposition S3 (Batch of Observations)
Let $\boldsymbol{\Sigma}_0$ such that $\boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{s}_j = \boldsymbol{s}_j$ for all $j \in \{1, \dots, i\}$. Then after i iterations the posterior over the representer weights in (4) *is equivalent to the one computed for a batch of observations, i.e.*

$$\boldsymbol{v}_i = \boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i(\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i)^{-1}\boldsymbol{S}_i^\intercal(\boldsymbol{y} - \boldsymbol{\mu})$$
$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i(\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i)^{-1}\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{\Sigma}_0$$

Proof. This can be seen as a direct consequence of recursively applying Bayes' theorem

$$p(\boldsymbol{v}_* \mid \{\alpha_i\}_{i=1}^m, \{\boldsymbol{s}_i\}_{i=1}^m) = \frac{p(\alpha_m \mid \boldsymbol{s}_m, \boldsymbol{v}_*)p(\boldsymbol{v}_* \mid \{\alpha_i\}_{i=1}^{m-1}, \{\boldsymbol{s}_i\}_{i=1}^{m-1})}{\int p(\alpha_m \mid \boldsymbol{s}_m, \boldsymbol{v}_*)p(\boldsymbol{v}_* \mid \{\alpha_i\}_{i=1}^{m-1}, \{\boldsymbol{s}_i\}_{i=1}^{m-1})d\boldsymbol{v}_*}.$$

However, here we also give a geometric proof based on the projection property of the precision matrix approximation $\boldsymbol{C}_i$. By using eq. (S37) and the assumption on $\boldsymbol{\Sigma}_0$ we have that

$$\begin{aligned}\boldsymbol{C}_i = \boldsymbol{S}_i(\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{S}_i)^{-1}\boldsymbol{S}_i^\intercal &= \boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i(\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i)^{-1}\boldsymbol{S}_i^\intercal \\ &= \boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i(\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i)^{-1}\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{\Sigma}_0\end{aligned}$$

This proves that

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_0 - \boldsymbol{C}_i = \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i(\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{\Sigma}_0\hat{\boldsymbol{K}}\boldsymbol{S}_i)^{-1}\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{\Sigma}_0$$

Now by eq. (S45) it holds that $\boldsymbol{C}_i(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{v}_i$. This proves the claim. □

Proposition S4 (Posterior Contraction)
Let $\boldsymbol{S}_i \in \mathbb{R}^{n\times i}$ be the actions chosen by Algorithm 1, then its posterior contracts as

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_0^{-1}\right) = \operatorname{tr}(\boldsymbol{\Sigma}_i\hat{\boldsymbol{K}}) = n - \operatorname{rank}(\boldsymbol{S}_i).$$

Proof. Since $\boldsymbol{\Sigma}_0 = \hat{\boldsymbol{K}}^{-1}$, we have by eq. (S37), that

$$\begin{aligned}
\operatorname{tr}\left(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_0^{-1}\right) &= \operatorname{tr}((\boldsymbol{\Sigma}_0 - \boldsymbol{C}_i)\hat{\boldsymbol{K}}) \\
&= \operatorname{tr}\left(\boldsymbol{I}_n - \boldsymbol{S}_i(\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{S}_i)^\dagger\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\right) \\
&= \operatorname{tr}(\boldsymbol{I}_n) - \operatorname{tr}(\underbrace{\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{S}_i(\boldsymbol{S}_i^\intercal\hat{\boldsymbol{K}}\boldsymbol{S}_i)^\dagger}_{\in\mathbb{R}^{i\times i}}) \\
&= n - \operatorname{rank}(\boldsymbol{S}_i)
\end{aligned}$$

Now, if the actions $\boldsymbol{S}_i$ are chosen linearly independent, then $\operatorname{rank}(\boldsymbol{S}_i) = i$. □

Theorem S7 (Online GP Approximation with Algorithm 1)
Let $n \leq n' \in \mathbb{N}$ and consider training data sets $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{X}' \in \mathbb{R}^{n'\times d}, \boldsymbol{y}' \in \mathbb{R}^{n'}$. Consider two sequences of actions $(\boldsymbol{s}_i)_{i=1}^n \in \mathbb{R}^n$ and $(\tilde{\boldsymbol{s}}_i)_{i=1}^{n+n'} \in \mathbb{R}^{n+n'}$ such that for all $i \in \{1,\dots,n\}$, it holds that

$$\tilde{\boldsymbol{s}}_i = \begin{pmatrix}\boldsymbol{s}_i \\ \boldsymbol{0}\end{pmatrix} \tag{S50}$$

Then the posterior returned by Algorithm 1 for the dataset $(\boldsymbol{X}, \boldsymbol{y})$ using actions $\boldsymbol{s}_i$ is identical to the posterior returned by Algorithm 1 for the extended dataset using actions $\tilde{\boldsymbol{s}}_i$, i.e. it holds for any $i \in \{1,\dots,n\}$, that

$$\textsc{IterGP}(\mu, k, \boldsymbol{X}, \boldsymbol{y}, (\boldsymbol{s}_i)_i) = (\mu_i, k_i) = (\tilde{\mu}_i, \tilde{k}_i) = \textsc{IterGP}\left(\mu, k, \begin{pmatrix}\boldsymbol{X} \\ \boldsymbol{X}'\end{pmatrix}, \begin{pmatrix}\boldsymbol{y} \\ \boldsymbol{y}'\end{pmatrix}, (\tilde{\boldsymbol{s}}_i)_i\right).$$

Proof. Define $\tilde{\boldsymbol{X}} = \begin{pmatrix}\boldsymbol{X} \\ \boldsymbol{X}'\end{pmatrix}$ and $\tilde{\boldsymbol{y}} = \begin{pmatrix}\boldsymbol{y} \\ \boldsymbol{y}'\end{pmatrix}$. We begin by showing that the search directions of both methods satisfy

$$\boldsymbol{d}_i' = \begin{pmatrix}\boldsymbol{d}_i \\ \boldsymbol{0}\end{pmatrix}. \tag{S51}$$

We proceed by induction. For $i = 0$ it holds by definition of Algorithm 1 and eq. (S50) that

$$\tilde{\boldsymbol{d}}_0 = \tilde{\boldsymbol{s}}_0 = \begin{pmatrix}\boldsymbol{s}_0 \\ \boldsymbol{0}\end{pmatrix} = \begin{pmatrix}\boldsymbol{d}_0 \\ \boldsymbol{0}\end{pmatrix}. \tag{S52}$$

Now for the induction step $i \to i+1$, assume that (S51) holds for $j \in \{1,\dots,i\}$. Then, we have

$$\begin{aligned}
\tilde{\boldsymbol{d}}_{i+1} &= \tilde{\boldsymbol{\Sigma}}_{i-1}(k(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}) + \sigma^2\boldsymbol{I}_{n+n'})\tilde{\boldsymbol{s}}_{i+1} \\
&= (\boldsymbol{I}_{n+n'} - \tilde{\boldsymbol{C}}_i(k(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}) + \sigma^2\boldsymbol{I}_{n+n'}))\tilde{\boldsymbol{s}}_{i+1} \\
&= \tilde{\boldsymbol{s}}_{i+1} - \sum_{j=1}^{i}\frac{1}{\tilde{\eta}_j}\tilde{\boldsymbol{d}}_j(\tilde{\boldsymbol{d}}_j)^\intercal(k(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}) + \sigma^2\boldsymbol{I}_{n+n'})\tilde{\boldsymbol{s}}_{i+1} \\
&\overset{\text{IH}}{=} \begin{pmatrix}\boldsymbol{s}_{i+1} \\ \boldsymbol{0}\end{pmatrix} - \sum_{j=1}^{i}\frac{1}{\tilde{\eta}_j}\begin{pmatrix}\boldsymbol{d}_j \\ \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{d}_j^\intercal & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}k(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{I}_n & k(\boldsymbol{X}, \boldsymbol{X}') \\ k(\boldsymbol{X}', \boldsymbol{X}) & k(\boldsymbol{X}', \boldsymbol{X}') + \boldsymbol{I}_{n'}\end{pmatrix}\begin{pmatrix}\boldsymbol{s}_{i+1} \\ \boldsymbol{0}\end{pmatrix} \\
&= \begin{pmatrix}\boldsymbol{s}_{i+1} - \sum_{j=1}^{i}\frac{1}{\eta_j}\boldsymbol{d}_j(\boldsymbol{d}_j)^\intercal\hat{\boldsymbol{K}}\boldsymbol{s}_{i+1} \\ \boldsymbol{0}\end{pmatrix} \\
&= \begin{pmatrix}\boldsymbol{d}_{i+1} \\ \boldsymbol{0}\end{pmatrix}
\end{aligned}$$

where we used that $\tilde{\eta}_j = \tilde{\boldsymbol{s}}_j^\intercal (k(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}) + \sigma^2 \boldsymbol{I}_{n+n'}) \tilde{\boldsymbol{d}}_j = \boldsymbol{s}_j^\intercal \hat{\boldsymbol{K}} \boldsymbol{d}_j = \eta_j$. This proves eq. (S51). Now recognize that

$$\begin{aligned}
\tilde{\alpha}_j = \tilde{\boldsymbol{s}}_j^\intercal \tilde{\boldsymbol{r}}_j &= \tilde{\boldsymbol{s}}_j^\intercal (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\mu}} - \tilde{\boldsymbol{K}} \tilde{\boldsymbol{C}}_i (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\mu}})) \\
&= \tilde{\boldsymbol{s}}_j^\intercal (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\mu}} - (\tilde{\boldsymbol{K}} + \sigma^2 \boldsymbol{I}) \sum_{\ell=1}^{j} \frac{1}{\tilde{\eta}_\ell} \tilde{\boldsymbol{d}}_\ell \tilde{\boldsymbol{d}}_\ell^\intercal (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\mu}})) \\
&= \boldsymbol{s}_j^\intercal (\boldsymbol{y} - \boldsymbol{\mu}) - \sum_{\ell=1}^{j} \frac{1}{\eta_\ell} \boldsymbol{s}_j^\intercal \hat{\boldsymbol{K}} \boldsymbol{d}_\ell \boldsymbol{d}_\ell^\intercal (\boldsymbol{y} - \boldsymbol{\mu}) \\
&= \boldsymbol{s}_j^\intercal (\boldsymbol{y} - \boldsymbol{\mu} - \hat{\boldsymbol{K}} \boldsymbol{C}_j (\boldsymbol{y} - \boldsymbol{\mu})) \\
&= \boldsymbol{s}_j^\intercal \boldsymbol{r}_j \\
&= \alpha_j
\end{aligned}$$

Therefore, we finally have that

$$\begin{aligned}
\tilde{\mu}_i(\cdot) &= \mu(\cdot) + k(\cdot, \tilde{\boldsymbol{X}}) \tilde{\boldsymbol{v}}_i = \mu(\cdot) + k(\cdot, \tilde{\boldsymbol{X}}) \sum_{j=1}^{i} \frac{\tilde{\alpha}_j}{\tilde{\eta}_j} \tilde{\boldsymbol{d}}_j \\
&= \mu(\cdot) + k(\cdot, \boldsymbol{X}) \boldsymbol{v}_i
\end{aligned}$$

as well as

$$\begin{aligned}
\tilde{k}_i(\cdot, \cdot) &= k(\cdot, \cdot) - k(\cdot, \tilde{\boldsymbol{X}}) \tilde{\boldsymbol{C}}_i k(\tilde{\boldsymbol{X}}, \cdot) = k(\cdot, \cdot) - k(\cdot, \tilde{\boldsymbol{X}}) \sum_{j=1}^{i} \frac{1}{\tilde{\eta}_j} \tilde{\boldsymbol{d}}_j (\tilde{\boldsymbol{d}}_j)^\intercal k(\tilde{\boldsymbol{X}}, \cdot) \\
&= k(\cdot, \cdot) - k(\cdot, \boldsymbol{X}) \sum_{j=1}^{i} \frac{1}{\eta_j} \boldsymbol{d}_j (\boldsymbol{d}_j)^\intercal k(\boldsymbol{X}, \cdot) = k(\cdot, \cdot) - k(\cdot, \boldsymbol{X}) \boldsymbol{C}_i k(\boldsymbol{X}, \cdot) = k_i(\cdot, \cdot).
\end{aligned}$$

□

Remark S2 (Streaming Gaussian Processes)
Theorem S7 shows that any variant of IterGP can be used in the online setting where data arrives sequentially *while* the algorithm is running. Now, if we assume data points arrive one at a time, we choose unit vector actions (IterGP-Chol) and perform one iteration of Algorithm 1 after each data point, then Algorithm 1 simply computes the mathematical GP posterior.

S2.2 Approximation of Representer Weights

Proposition 2 (Relative Error Bound for the Representer Weights)
For any choice of actions a relative error bound $\rho(i)$, s.t. $\|\boldsymbol{v}_* - \boldsymbol{v}_i\|_{\hat{\boldsymbol{K}}} \leq \rho(i) \|\boldsymbol{v}_*\|_{\hat{\boldsymbol{K}}}$ is given by

$$\rho(i) = (\bar{\boldsymbol{v}}_*^\intercal \underbrace{(\boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}})}_{\textit{projection onto } \operatorname{span}\{\boldsymbol{S}_i\}^{\perp_{\hat{\boldsymbol{K}}}}} \bar{\boldsymbol{v}}_*)^{\frac{1}{2}} \leq \lambda_{\max}(\boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}}) \leq 1 \tag{9}$$

where $\bar{\boldsymbol{v}}_* = \boldsymbol{v}_* / \|\boldsymbol{v}_*\|_{\hat{\boldsymbol{K}}}$. If the actions $\{\boldsymbol{s}_i\}_{i=1}^n$ are linearly independent, then $\rho(i) \leq \delta_{n=i}$.

Proof. Define $\boldsymbol{H}_i = \boldsymbol{\Sigma}_i \hat{\boldsymbol{K}} = \boldsymbol{I} - \boldsymbol{C}_i \hat{\boldsymbol{K}}$. We have by Lemma S2, that

$$\|\boldsymbol{v}_* - \boldsymbol{v}_i\|_{\hat{\boldsymbol{K}}}^2 = \|\boldsymbol{H}_i \boldsymbol{v}_*\|_{\hat{\boldsymbol{K}}}^2 = (\boldsymbol{H}_i \boldsymbol{v}_*)^\intercal \hat{\boldsymbol{K}} \boldsymbol{H}_i \boldsymbol{v}_* \overset{\text{(S44)}}{=} \boldsymbol{v}_*{}^\intercal \boldsymbol{H}_i \boldsymbol{v}_* = \bar{\boldsymbol{v}}_*^\intercal \boldsymbol{H}_i \bar{\boldsymbol{v}}_* \|\boldsymbol{v}_*\|_{\hat{\boldsymbol{K}}}^2$$

This proves the first equality of Proposition 2. Further it holds that

$$\begin{aligned}
\|\boldsymbol{H}_i \boldsymbol{v}_*\|_{\hat{\boldsymbol{K}}} &= \|\hat{\boldsymbol{K}}^{\frac{1}{2}} \boldsymbol{H}_i \boldsymbol{v}_*\|_2 = \|(\boldsymbol{I} - \hat{\boldsymbol{K}}^{\frac{1}{2}} \boldsymbol{C}_i \hat{\boldsymbol{K}}^{\frac{1}{2}}) \hat{\boldsymbol{K}}^{\frac{1}{2}} \boldsymbol{v}_*\|_2 \leq \|\boldsymbol{I} - \hat{\boldsymbol{K}}^{\frac{1}{2}} \boldsymbol{C}_i \hat{\boldsymbol{K}}^{\frac{1}{2}}\|_2 \|\boldsymbol{v}_*\|_{\hat{\boldsymbol{K}}} \\
&= \lambda_{\max}(\boldsymbol{I} - \hat{\boldsymbol{K}}^{\frac{1}{2}} \boldsymbol{C}_i \hat{\boldsymbol{K}}^{\frac{1}{2}}) \|\boldsymbol{v}_*\|_{\hat{\boldsymbol{K}}}.
\end{aligned}$$

Now by Weyl's inequality and the fact that $\hat{\boldsymbol{K}}^{\frac{1}{2}} \boldsymbol{C}_i \hat{\boldsymbol{K}}^{\frac{1}{2}}$ is positive semi-definite, it holds that

$$\lambda_{\max}(\boldsymbol{H}_i) = \lambda_{\max}(\boldsymbol{I} - \hat{\boldsymbol{K}}^{\frac{1}{2}} \boldsymbol{C}_i \hat{\boldsymbol{K}}^{\frac{1}{2}}) \leq \lambda_{\max}(\boldsymbol{I}) - \lambda_{\min}(\hat{\boldsymbol{K}}^{\frac{1}{2}} \boldsymbol{C}_i \hat{\boldsymbol{K}}^{\frac{1}{2}}) \leq 1.$$

Now, recall that similar matrices $\boldsymbol{A}$ and $\boldsymbol{B} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$ have the same eigenvalues. Therefore

$$\boldsymbol{I} - \hat{\boldsymbol{K}}^{\frac{1}{2}}\boldsymbol{C}_i\hat{\boldsymbol{K}}^{\frac{1}{2}} = \hat{\boldsymbol{K}}^{\frac{1}{2}}(\boldsymbol{I} - \boldsymbol{C}_i\hat{\boldsymbol{K}})\hat{\boldsymbol{K}}^{-\frac{1}{2}}$$

and $\boldsymbol{I} - \boldsymbol{C}_i\hat{\boldsymbol{K}}$ have the same eigenvalues. Finally, since by eq. (S39) $\boldsymbol{H}_i$ is a projection onto $\operatorname{span}\{\boldsymbol{S}_i\}^{\perp_{\hat{\boldsymbol{K}}}}$, it has full rank at iteration n if the actions are linearly independent and therefore $\lambda_{\max}(\boldsymbol{H}_n) = 1$. This proves the claim. □

S2.3 Convergence Analysis of the Posterior Mean Approximation

Theorem 1 (Convergence in RKHS Norm of the Posterior Mean Approximation)
Let $\mathcal{H}_k$ be the RKHS associated with kernel $k(\cdot,\cdot)$, $\sigma^2 > 0$ and let $\mu_* - \mu \in \mathcal{H}_k$ be the unique solution to the regularized empirical risk minimization problem

$$\arg\min_{f\in\mathcal{H}_k} \tfrac{1}{n}\big(\textstyle\sum_{j=1}^n (f(\boldsymbol{x}_j) - y_j + \mu(\boldsymbol{x}_j))^2 + \sigma^2\|f\|_{\mathcal{H}_k}^2\big) \tag{11}$$

which is equivalent to the mathematical posterior mean up to shift by the prior μ [e.g. 1, Sec. 6.2]. Then for $i \in \{0,\dots,n\}$ the posterior mean $\mu_i(\cdot)$ computed by Algorithm 1 satisfies

$$\boxed{\|\mu_* - \mu_i\|_{\mathcal{H}_k} \le \rho(i)c(\sigma^2)\|\mu_* - \mu_0\|_{\mathcal{H}_k}} \tag{12}$$

where $\mu_0 = \mu$ is the prior mean and the constant $c(\sigma^2) = \sqrt{1 + \frac{\sigma^2}{\lambda_{\min}(\boldsymbol{K})}} \to 1$ as $\sigma^2 \to 0$.

Proof. Let $\rho(i)$ such that $\|\boldsymbol{v}_* - \boldsymbol{v}_i\|_{\hat{\boldsymbol{K}}} \le \rho(i)\|\boldsymbol{v}_* - \boldsymbol{v}_0\|_{\hat{\boldsymbol{K}}}$, where $\boldsymbol{v}_0 = \boldsymbol{0}$. Then, we have for $i \in \{0,\dots,n\}$, that

$$\begin{aligned}
\|\boldsymbol{v}_* - \boldsymbol{v}_i\|_{\boldsymbol{K}}^2 &\le \|\boldsymbol{v}_* - \boldsymbol{v}_i\|_{\hat{\boldsymbol{K}}}^2 \le \rho(i)^2\|\boldsymbol{v}_* - \boldsymbol{v}_0\|_{\hat{\boldsymbol{K}}}^2 \\
&= \rho(i)^2\big(\|\boldsymbol{v}_* - \boldsymbol{v}_0\|_{\boldsymbol{K}}^2 + \sigma^2\frac{1}{\lambda_{\min}(\boldsymbol{K})}\underbrace{\lambda_{\min}(\boldsymbol{K})\|\boldsymbol{v}_* - \boldsymbol{v}_0\|_2^2}_{\le\|\boldsymbol{v}_* - \boldsymbol{v}_0\|_{\boldsymbol{K}}^2}\big) \\
&\le \rho(i)^2\left(1 + \frac{\sigma^2}{\lambda_{\min}(\boldsymbol{K})}\right)\|\boldsymbol{v}_* - \boldsymbol{v}_0\|_{\boldsymbol{K}}^2
\end{aligned}$$

Now by assumption $\mu_i(\cdot) = \mu(\cdot) + \sum_{j=1}^n (\boldsymbol{v}_i)_j k(\cdot,\boldsymbol{x}_j) = \mu(\cdot) + k(\cdot,\boldsymbol{X})\boldsymbol{C}_i\boldsymbol{y}$. By the reproducing property we obtain for $\Delta = \boldsymbol{v}_* - \boldsymbol{v}_i$ that

$$\begin{aligned}
\|\boldsymbol{v}_* - \boldsymbol{v}_i\|_{\boldsymbol{K}}^2 &= \Delta^\intercal\boldsymbol{K}\Delta \\
&= \sum_{\ell=1}^n\sum_{j=1}^n \Delta_\ell\Delta_j k(\boldsymbol{x}_\ell,\boldsymbol{x}_j) \\
&= \sum_{\ell=1}^n\sum_{j=1}^n \Delta_\ell\Delta_j \langle k(\cdot,\boldsymbol{x}_\ell), k(\cdot,\boldsymbol{x}_j)\rangle_{\mathcal{H}_k} && k \text{ is the reproducing kernel of } \mathcal{H}_k \\
&= \langle\sum_{\ell=1}^n \Delta_\ell k(\cdot,\boldsymbol{x}_\ell), \sum_{j=1}^n \Delta_j k(\cdot,\boldsymbol{x}_j)\rangle_{\mathcal{H}_k} \\
&= \left\|\sum_{\ell=1}^n \Delta_\ell k(\cdot,\boldsymbol{x}_\ell)\right\|_{\mathcal{H}_k}^2 \\
&= \left\|\sum_{\ell=1}^n (\boldsymbol{v}_*)_\ell k(\cdot,\boldsymbol{x}_\ell) - \sum_{\ell=1}^n (\boldsymbol{v}_i)_\ell k(\cdot,\boldsymbol{x}_\ell)\right\|_{\mathcal{H}_k}^2 \\
&= \|\mu_* - \mu_i\|_{\mathcal{H}_k}^2 && \text{See Theorem 3.4 in Kanagawa et al. [36]}
\end{aligned}$$

Combining the above and setting $c(\sigma^2) = 1 + \frac{\sigma^2}{\lambda_{\min}(\boldsymbol{K})}$ we obtain

$$\|\mu_* - \mu_i\|_{\mathcal{H}_k} = \|\boldsymbol{v}_* - \boldsymbol{v}_i\|_{\boldsymbol{K}} \le \rho(i)c(\sigma^2)\|\boldsymbol{v}_* - \boldsymbol{v}_0\|_{\boldsymbol{K}} = \rho(i)c(\sigma^2)\|\mu_* - \mu_0\|_{\mathcal{H}_k}.$$

□

S2.4 Combined Uncertainty as Worst Case Error

Theorem 2 (Combined and Computational Uncertainty as Worst Case Errors)
Let $\sigma^2 \geq 0$ and let $k_i(\cdot,\cdot) = k_*(\cdot,\cdot) + k_i^{\text{comp}}(\cdot,\cdot)$ be the combined uncertainty computed by Algorithm 1. Then, for any $\boldsymbol{x} \in \mathcal{X}$ (assuming $\boldsymbol{x} \notin \boldsymbol{X}$ if $\sigma^2 > 0$) we have

$$\sup_{g \in \mathcal{H}_{k^\sigma} : \|g\|_{\mathcal{H}_{k^\sigma}} \leq 1} \overbrace{\underbrace{g(\boldsymbol{x}) - \mu_*^g(\boldsymbol{x})}_{\text{error of math. post. mean}} + \underbrace{\mu_*^g(\boldsymbol{x}) - \mu_i^g(\boldsymbol{x})}_{\text{computational error}}}^{\text{error of approximate posterior mean}} = \sqrt{k_i(\boldsymbol{x},\boldsymbol{x}) + \sigma^2}, \quad \textit{and} \tag{13}$$

$$\sup_{g \in \mathcal{H}_{k^\sigma} : \|g\|_{\mathcal{H}_{k^\sigma}} \leq 1} \underbrace{\mu_*^g(\boldsymbol{x}) - \mu_i^g(\boldsymbol{x})}_{\text{computational error}} = \sqrt{k_i^{\text{comp}}(\boldsymbol{x},\boldsymbol{x})} \tag{14}$$

where $\mu_*^g(\cdot) = k(\cdot,\boldsymbol{X})\hat{\boldsymbol{K}}^{-1} g(\boldsymbol{X})$ is the mathematical and $\mu_i^g(\cdot) = k(\cdot,\boldsymbol{X})\boldsymbol{C}_i g(\boldsymbol{X})$ IterGP's posterior mean for the latent function $g \in \mathcal{H}_{k^\sigma}$. If $\sigma^2 = 0$, then the above also holds for $\boldsymbol{x} \in \boldsymbol{X}$.

Proof. Let $\boldsymbol{x}_0 = \boldsymbol{x}$, $c_0 = 1$ and $c_j = -(\boldsymbol{C}_i k^\sigma(\boldsymbol{X},\boldsymbol{x}))_j$ for $j = 1, \dots n$, where $k^\sigma(\cdot,\cdot) \coloneqq k(\cdot,\cdot) + \sigma^2\delta(\cdot,\cdot)$. Then by Lemma 3.9 of Kanagawa et al. [36], it holds that

$$\begin{aligned}
\left(\sup_{g \in \mathcal{H}_{k^\sigma} : \|g\|_{\mathcal{H}_{k^\sigma}} \leq 1} (g(\boldsymbol{x}) - \mu_i^g(\boldsymbol{x}))\right)^2 &= \left(\sup_{g \in \mathcal{H}_{k^\sigma} : \|g\|_{\mathcal{H}_{k^\sigma}} \leq 1} \sum_{j=0}^n c_j g(\boldsymbol{x}_j)\right)^2 \\
&= \left\| k^\sigma(\cdot,\boldsymbol{x}_0) - \sum_{j=1}^n k(\boldsymbol{x},\boldsymbol{x}_j)\boldsymbol{C}_i k^\sigma(\cdot,\boldsymbol{x}_j) \right\|_{\mathcal{H}_{k^\sigma}}^2 \\
&= \|k^\sigma(\cdot,\boldsymbol{x}) - k(\boldsymbol{x},\boldsymbol{X})\boldsymbol{C}_i k^\sigma(\boldsymbol{X},\cdot)\|_{\mathcal{H}_{k^\sigma}}^2 \\
&= \langle k^\sigma(\cdot,\boldsymbol{x}), k^\sigma(\cdot,\boldsymbol{x})\rangle_{\mathcal{H}_{k^\sigma}} - 2\langle k^\sigma(\cdot,\boldsymbol{x}), k(\boldsymbol{x},\boldsymbol{X})\boldsymbol{C}_i k^\sigma(\boldsymbol{X},\cdot)\rangle_{\mathcal{H}_{k^\sigma}} \\
&\quad + \langle k(\boldsymbol{x},\boldsymbol{X})\boldsymbol{C}_i k^\sigma(\boldsymbol{X},\cdot), k(\boldsymbol{x},\boldsymbol{X})\boldsymbol{C}_i k^\sigma(\boldsymbol{X},\cdot)\rangle_{\mathcal{H}_{k^\sigma}}
\end{aligned}$$

Now by the reproducing property, it follows that

$$= k^\sigma(\boldsymbol{x},\boldsymbol{x}) - 2k^\sigma(\boldsymbol{x},\boldsymbol{X})\boldsymbol{C}_i k^\sigma(\boldsymbol{X},\boldsymbol{x}) + k^\sigma(\boldsymbol{x},\boldsymbol{X})\boldsymbol{C}_i k^\sigma(\boldsymbol{X},\boldsymbol{X})\boldsymbol{C}_i k^\sigma(\boldsymbol{X},\boldsymbol{x})$$

If $\sigma^2 > 0$ and $\boldsymbol{x} \neq \boldsymbol{x}_j$ or if $\sigma^2 = 0$, it holds that $k^\sigma(\boldsymbol{x},\boldsymbol{X}) = k(\boldsymbol{x},\boldsymbol{X})$. Further by definition $k^\sigma(\boldsymbol{X},\boldsymbol{X}) = \hat{\boldsymbol{K}}$ and finally by (S42), it holds that $\boldsymbol{C}_i\hat{\boldsymbol{K}}\boldsymbol{C}_i = \boldsymbol{C}_i$. Therefore we have

$$\begin{aligned}
&= k(\boldsymbol{x},\boldsymbol{x}) + \sigma^2 - 2k(\boldsymbol{x},\boldsymbol{X})\boldsymbol{C}_i k(\boldsymbol{X},\boldsymbol{x}) + k(\boldsymbol{x},\boldsymbol{X})\boldsymbol{C}_i\hat{\boldsymbol{K}}\boldsymbol{C}_i k(\boldsymbol{X},\boldsymbol{x}) \\
&= k(\boldsymbol{x},\boldsymbol{x}) - k(\boldsymbol{x},\boldsymbol{X})\boldsymbol{C}_i k(\boldsymbol{X},\boldsymbol{x}) + \sigma^2 \\
&= k_i(\boldsymbol{x},\boldsymbol{x}) + \sigma^2
\end{aligned}$$

We prove eq. (14) by an analogous argument. Choose $c_j \coloneqq ((\hat{\boldsymbol{K}}^{-1} - \boldsymbol{C}_i)k^\sigma(\boldsymbol{X},\boldsymbol{x}))_j$. We have

$$\begin{aligned}
&\left(\sup_{g \in \mathcal{H}_{k^\sigma} : \|g\|_{\mathcal{H}_{k^\sigma}} \leq 1} (\mu_*^g(\boldsymbol{x}) - \mu_i^g(\boldsymbol{x}))\right)^2 = \left(\sup_{g \in \mathcal{H}_{k^\sigma} : \|g\|_{\mathcal{H}_{k^\sigma}} \leq 1} \sum_{j=0}^n c_j g(\boldsymbol{x}_j)\right)^2 \\
&= \left\| \sum_{j=1}^n k(\boldsymbol{x},\boldsymbol{x}_j)(\hat{\boldsymbol{K}}^{-1} - \boldsymbol{C}_i)k^\sigma(\cdot,\boldsymbol{x}_j) \right\|_{\mathcal{H}_{k^\sigma}}^2 \\
&= \left\|k(\boldsymbol{x},\boldsymbol{X})(\hat{\boldsymbol{K}}^{-1} - \boldsymbol{C}_i)k^\sigma(\boldsymbol{X},\cdot)\right\|_{\mathcal{H}_{k^\sigma}}^2 \\
&= k^\sigma(\boldsymbol{x},\boldsymbol{X})\hat{\boldsymbol{K}}^{-1}\hat{\boldsymbol{K}}\hat{\boldsymbol{K}}^{-1}k^\sigma(\boldsymbol{X},\boldsymbol{x}) - 2k^\sigma(\boldsymbol{x},\boldsymbol{X})\hat{\boldsymbol{K}}^{-1}\hat{\boldsymbol{K}}\boldsymbol{C}_i k^\sigma(\boldsymbol{X},\boldsymbol{x}) + k^\sigma(\boldsymbol{x},\boldsymbol{X})\boldsymbol{C}_i\hat{\boldsymbol{K}}\boldsymbol{C}_i k^\sigma(\boldsymbol{X},\boldsymbol{x})
\end{aligned}$$

Again, we use that $k^\sigma(\boldsymbol{x},\boldsymbol{X}) = k(\boldsymbol{x},\boldsymbol{X})$ by assumption and (S42). Therefore

$$\begin{aligned}
&= k(\boldsymbol{x},\boldsymbol{X})(\hat{\boldsymbol{K}}^{-1} - \boldsymbol{C}_i)k(\boldsymbol{X},\boldsymbol{x}) \\
&= k_i^{\text{comp}}(\boldsymbol{x},\boldsymbol{x})
\end{aligned}$$

This concludes the proof. □

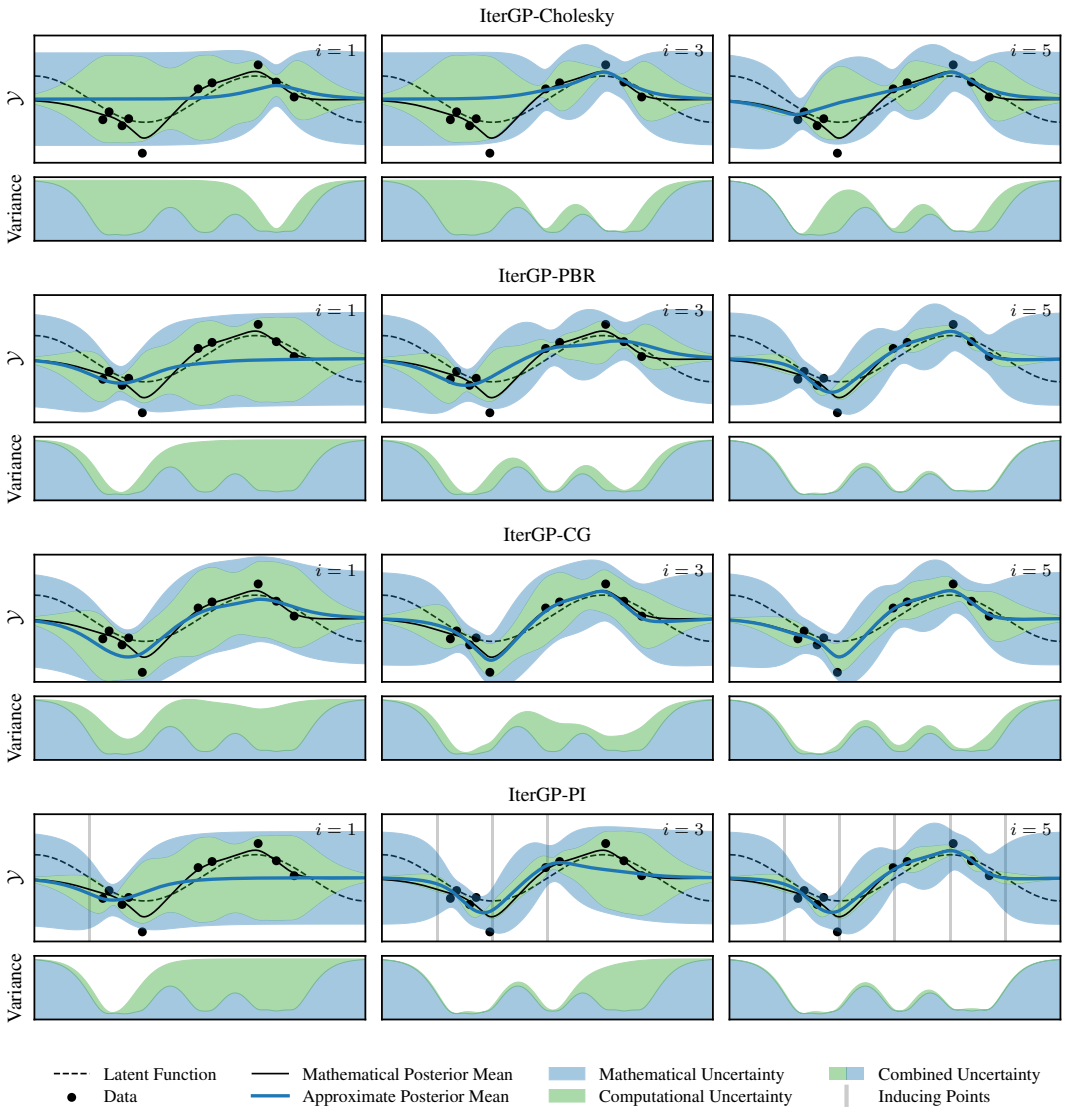


Figure S3: *Illustration of IterGP analogs of commonly used GP approximations.*

S3 Implementation of Algorithm 1

S3.1 Policy Choice

As illustrated in Figure 2, the choice of policy of Algorithm 1 determines where computation in input space is targeted and therefore where the combined posterior contracts first. However, the policy also determines whether the error in the posterior mean or (co-)variance are predominantly reduced first, as Figure S3 shows (cf. IterGP-Chol and IterGP-PBR). Therefore the policy choice is application-dependent. If I am primarily interested in the predictive mean, I may select residual actions (IterGP-CG). If downstream I am making use of the predictive uncertainty, I may want to contract uncertainty globally as quickly as possible at the expense of predictive accuracy (IterGP-PI). Such a choice is not unique to IterGP, but necessary whenever we select a GP approximation. What IterGP adds is computation-aware, meaningful uncertainty quantification in the sense of Corollary 1 no matter the choice of policy.

S3.2 Stopping Criterion

In our implementation of Algorithm 1 we use the following two stopping criteria. Our computational budget can be directly controlled by specifying a *maximum number of iterations*, since each iteration of IterGP needs the same number of matrix-vector multiplies. Alternatively, we terminate if the *absolute or relative norm of the residual* are sufficiently small, i.e. if

$$\|\boldsymbol{r}_i\|_2 < \delta_{\text{abstol}} \qquad \text{or} \qquad \|\boldsymbol{r}_i\|_2 < \delta_{\text{reltol}} \|\boldsymbol{y}\|_2. \tag{S53}$$

Of course other choices are possible. From a probabilistic numerics standpoint one may want to terminate once the combined marginal uncertainty at the training data is sufficiently small relative to the observation noise.

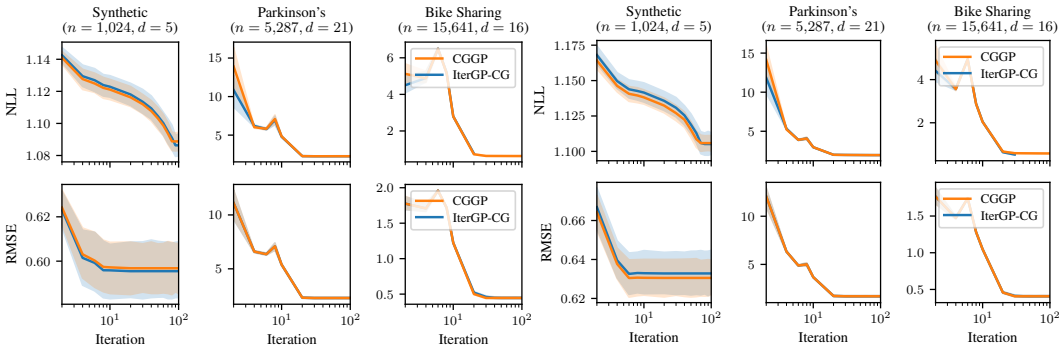
S3.3 Efficient Sampling from the Combined Posterior

Sampling from an exact GP posterior has cubic cost $\mathcal{O}(n_\diamond^3)$ in the number of evaluation points $n_\diamond$, which is prohibitive for many useful downstream applications such as numerical integration over the posterior using Monte-Carlo methods. Wilson et al. [46, 47] recently showed how to make use of *Matheron's rule* [45, 66, 67] to efficiently sample from a GP posterior by sampling from the prior and then performing a pathwise update. We can directly make use of this strategy since Algorithm 1 computes a low-rank approximation to the precision matrix. Assume we are given a draw $f'_{\text{prior}} \in \mathcal{H}_k^\theta$ from the prior[3] such that $\boldsymbol{y}' \sim \mathcal{N}(f'_{\text{prior}}(\boldsymbol{X}), \sigma^2 \boldsymbol{I})$ constitutes a draw from the prior predictive. Then

$$f'(\cdot) = f'_{\text{prior}}(\cdot) + k(\cdot, \boldsymbol{X}) \boldsymbol{C}_i (\boldsymbol{y} - \boldsymbol{y}') \tag{S54}$$

is a draw from the combined posterior by Matheron's rule, which we can evaluate in $\mathcal{O}(n_\diamond n i)$ for $n_\diamond$ evaluation points, since $\boldsymbol{C}_i$ has rank i.

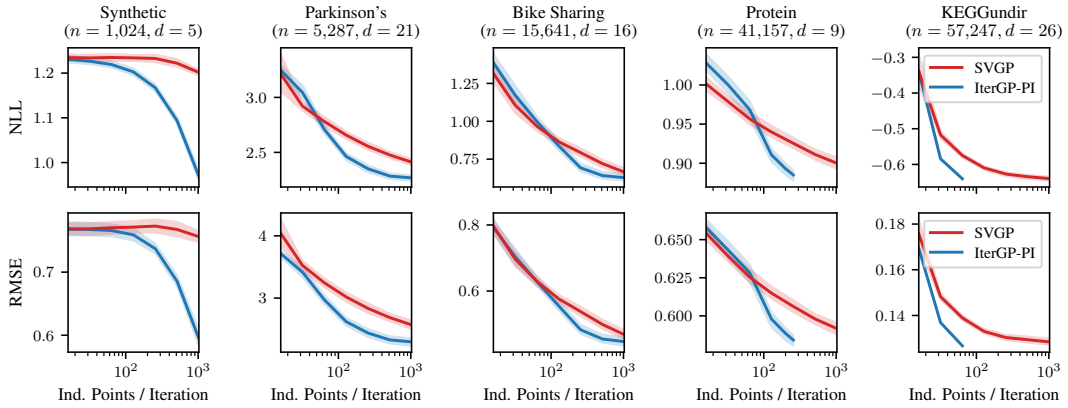
S4 Additional Experimental Results



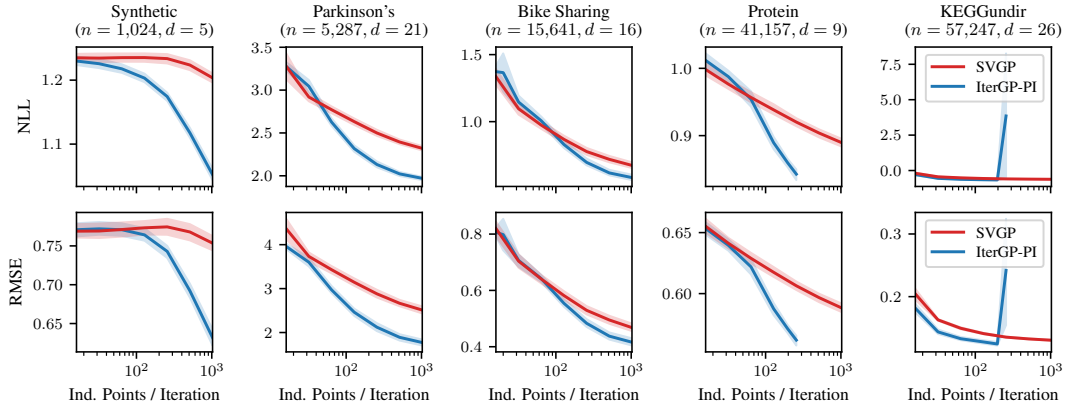
(a) RBF kernel (b) Matérn($\frac{3}{2}$) kernel

Figure S4: *Generalization of CGGP and its closest IterGP analog.* GP regression using an RBF and Matérn($\frac{3}{2}$) kernel on UCI datasets. The plot shows the average generalization error in terms of NLL and RMSE for an increasing number of solver iterations. The posterior mean of IterGP-CG and CGGP is identical, which explains the identical RMSE.

[3] In infinite dimensional reproducing kernel Hilbert spaces samples $f \sim \mathcal{GP}(\mu, k)$ from a Gaussian process almost surely do not lie in the RKHS $\mathcal{H}_k$ [Cor. 4.10, 36]. However, there exists $f' \in \mathcal{H}_k^\theta$ in a larger RKHS $\mathcal{H}_k^\theta \supset \mathcal{H}_k$ such that $f'(\boldsymbol{x}) = f(\boldsymbol{x})$ with probability 1 [Thm. 4.12, 36].



(a) RBF kernel



(b) Matérn($\frac{3}{2}$) kernel

Figure S5: *Generalization of SVGP and its closest IterGP analog.* GP regression using an RBF and Matérn($\frac{3}{2}$) kernel on UCI datasets. The plot shows the average generalization error in terms of NLL and RMSE for an increasing number of identical inducing points. After a small number of inducing points relative to the size of the training data, IterGP has significantly lower generalization error than SVGP. For the "KEGGundir" dataset after ≈ 128 iterations we observe numerical instability in some runs when computing the combined posterior of IterGP using a Matérn($\frac{3}{2}$) kernel.