Computation-Aware Robust Gaussian Processes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Gaussian Processes (GPs) are flexible nonparametric statistical models equipped with principled uncertainty quantification for both noise and model uncertainty. However, their cubic inference complexity requires them to be combined with approximation techniques when applied to large datasets. Recent work demonstrated that such approximations introduce an additional source of uncertainty, *computational uncertainty*, and that the latter could be quantified, leading to the *computation-aware* GP, also known as IterGP. In this short communication, we demonstrate that IterGP is not "robust", in the sense that a quantity of interest, the posterior influence function, is not bounded. Subsequently, drawing inspiration from recent work on Robust Conjugate GPs, we introduce a novel class of GPs: IterRCGPs. We carry out a number of theoretical analyses, demonstrating the robustness of IterRCGPs among other things.

1 Introduction

Gaussian Processes (GPs, Rasmussen and Williams (2006)) are a class of probabilistic models enjoying many properties such as universal approximation or closed-form computations. Due to their principled uncertainty quantification, they are becoming increasingly popular when applied in high-stakes domains like medical datasets (Cheng *et al.*, 2019; Chen *et al.*, 2023) or used as a surrogate model in Bayesian Optimization (Garnett, 2023). This being said, GPs suffer from a cubic inference complexity, hindering their use on large datasets. As a remedy, approximation techniques like Sparse Variational Gaussian Processes (Titsias, 2009) or the Nyström approximation are often used (Williams and Seeger, 2000; Wild *et al.*, 2023).

These approximations introduce bias in uncertainty quantification, which, as recently demonstrated, can be quantified and combined with mathematical uncertainty, leading to the development of *computation-aware* GPs (Wenger *et al.*, 2022), also known as IterGPs. This combined uncertainty is shown to be the correct measure for capturing overall uncertainty, as limited computation introduces computational error. While this analysis applies to standard GPs, many practical applications require variations, e.g., to deal with heteroscedasticity or outliers.

Recent work by Altamirano *et al.* (2024) introduced the robust conjugate GP (RCGP), which unifies three classes of GPs. RCGP retains conjugacy, enabling a closed-form posterior while exhibiting a robustness property. However, like standard GPs, RCGP faces significant inference complexity, necessitating approximation methods such as sparse variational RCGP, and therefore suggesting the use of the framework developed by Wenger *et al.* (2022).

Contributions. Our work can be seen as bridging the gap between computation-aware GPs and Robust Conjugate GPs. As such, our contributions are mainly theoretical and can be summarized as follows:

Submitted to Workshop on Bayesian Decision-making and Uncertainty, 38th Conference on Neural Information Processing Systems (BDU at NeurIPS 2024). Do not distribute.

- We present IterRCGP, a novel computation-aware Gaussian Process (GP) framework that extends IterGPs by accommodating a broader range of observation noise models.
- We demonstrate that IterRCGP inherits the robustness properties characteristic of RCGP.
- We establish that IterRCGP exhibits convergence behavior and worst-case errors analogous to IterGP.

2 Preliminaries

We first introduce notations for GP regression (Rasmussen and Williams, 2006). Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a dataset, with $(\mathbf{x}_j, y_j) \in \mathbb{R}^d \times \mathbb{R}$ such that $y_j = f(\mathbf{x}_j) + \epsilon$ and $\epsilon \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$ for all j. The latent function f is modeled with a GP prior:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{1}$$

This defines a distribution over functions f whose mean is $\mathbb{E}[f(\mathbf{x})] = m(\mathbf{x})$ and covariance $\text{cov}[f(\mathbf{x}), f(\mathbf{x}')] = k(\mathbf{x}, \mathbf{x}')$. k is a kernel function measuring the similarity between inputs. For any finite-dimensional collection of inputs $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the function values $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^\top \in \mathbb{R}^n$ follow a multivariate normal distribution $\mathbf{f} \sim \mathcal{N}(\mathbf{m}, \mathbf{K})$, where $\mathbf{m} = [m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n)]^\top$ and $\mathbf{K} \in \mathbb{R}^{n \times n} = [k(\mathbf{x}_j, \mathbf{x}_l)]_{1 \le j,l \le n}$ is the kernel matrix.

Given $\mathcal{D}$, the posterior predictive distribution $p(f(\mathbf{x}) \mid \mathcal{D})$ is Gaussian for all $\mathbf{x}$ with mean $\mu_*(\mathbf{x})$ and variance $k_*(\mathbf{x}, \mathbf{x})$, such that

$$\mu_*(\mathbf{x}) = m(\mathbf{x}) + \mathbf{k}_\mathbf{x}^\top (\mathbf{K} + \sigma^2_{\text{noise}} \mathbf{I})^{-1} (\mathbf{y} - \mathbf{m}),$$
$$k_*(\mathbf{x}, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_\mathbf{x}^\top (\mathbf{K} + \sigma^2_{\text{noise}} \mathbf{I})^{-1} \mathbf{k}_\mathbf{x},$$

where $\mathbf{y} = [y_1, \ldots, y_n] \in \mathbb{R}^n$ and $\mathbf{k}_\mathbf{x} = [k(\mathbf{x}, \mathbf{x}_1), \cdots, k(\mathbf{x}, \mathbf{x}_n)]^\top \in \mathbb{R}^n$.

Next, we introduce an extension of GPs: Robust Conjugate Gaussian Processes (RCGPs).

Robust conjugate Gaussian process. RCGP follows the generalized Bayesian inference framework, substituting the classical likelihood with the loss function L_n^w (Altamirano *et al.*, 2024) defined as

$$L_n^w(\mathbf{f}, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \left(\sum_{j=1}^{n} w^2(\mathbf{x}_j, y_j) s^2_{\text{model}}(\mathbf{x}_j, y_j) + 2 \nabla_y [w^2(\mathbf{x}_j, y_j) s_{\text{model}}(\mathbf{x}_j, y_j)] \right), \tag{2}$$

where $s_{\text{model}}(\mathbf{x}, y) = \sigma^{-2}_{\text{noise}}(f(\mathbf{x}) - y), \sigma^2_{\text{noise}} > 0$. The core component of L_n^w is the weighting function w, which depends on $\mathbf{x}$ and y. Altamirano *et al.* (2024)[Table 1] provides three weighting functions corresponding to homoscedastic, heteroscedastic, and outliers-robust GPs. Building on L_n^w, the authors further define the RCGP's predictive posterior distribution $p^w(f(\mathbf{x})|\mathcal{D})$ as follows:

$$\hat{\mu}_*(\mathbf{x}) = m(\mathbf{x}) + \mathbf{k}_\mathbf{x}^\top \overbrace{(\mathbf{K} + \sigma^2_{\text{noise}} \mathbf{J}_\mathbf{w})^{-1} (\mathbf{y} - \mathbf{m}_\mathbf{w})}^{\hat{\mathbf{v}}} \tag{3}$$
$$\hat{k}_*(\mathbf{x}, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_\mathbf{x}^\top \tilde{\mathbf{K}}^{-1} \mathbf{k}_\mathbf{x} \tag{4}$$

for $\mathbf{w} = (w(\mathbf{x}_1, y_1), \ldots, w(\mathbf{x}_n, y_n))^\top$, $\tilde{\mathbf{K}} = \mathbf{K} + \sigma^2_{\text{noise}} \mathbf{J}_\mathbf{w}$, $\mathbf{m}_\mathbf{w} = \mathbf{m} + \sigma^2_{\text{noise}} \nabla_y \log(\mathbf{w}^2)$, and $\mathbf{J}_\mathbf{w} = \text{diag}(\frac{\sigma^2_{\text{noise}}}{2} \mathbf{w}^{-2})$. A key advantage of RCGP is its robustness to outliers and non-Gaussian errors. While vanilla GPs exhibit an unbounded posterior influence function, RCGP, under certain conditions, maintains a bounded posterior influence function (Altamirano *et al.*, 2024)[Proposition 3.2].

3 Computation-aware RCGPs

In the same spirit of Wenger *et al.* (2022), we treat the representer weights $\hat{\mathbf{v}}$ introduced in Equation 3 as a random variable with the prior $p(\hat{\mathbf{v}}) = \mathcal{N}(\hat{\mathbf{v}}; \mathbf{0}, \tilde{\mathbf{K}}^{-1})$. We then update $p(\hat{\mathbf{v}})$ by iteratively applying the tractable matrix-vector multiplication. For a particular iteration $i \in \{0, \dots, n\}$, we have the current belief distribution $p(\hat{\mathbf{v}}) = \mathcal{N}(\hat{\mathbf{v}}; \tilde{\mathbf{v}}_i, \tilde{\boldsymbol{\Sigma}}_i)$ where

$$\tilde{\mathbf{v}}_i = \tilde{\mathbf{v}}_{i-1} + \tilde{\boldsymbol{\Sigma}}_{i-1}\tilde{\mathbf{K}}\mathbf{s}_i(\mathbf{s}_i^\top \tilde{\mathbf{K}}\tilde{\boldsymbol{\Sigma}}_{i-1}\tilde{\mathbf{K}}\mathbf{s}_i)^{-1}\tilde{\alpha}_i = \tilde{\mathbf{C}}_i(\mathbf{y} - \mathbf{m_w}) \tag{5}$$

$$\tilde{\boldsymbol{\Sigma}}_i = \tilde{\boldsymbol{\Sigma}}_{i-1} - \tilde{\boldsymbol{\Sigma}}_{i-1}\tilde{\mathbf{K}}\mathbf{s}_i(\mathbf{s}_i^\top \tilde{\mathbf{K}}\tilde{\boldsymbol{\Sigma}}_{i-1}\tilde{\mathbf{K}}\mathbf{s}_i)^{-1}\mathbf{s}_i^\top \tilde{\mathbf{K}}\tilde{\boldsymbol{\Sigma}}_{i-1} \tag{6}$$

$$\tilde{\alpha}_i = \mathbf{s}_i^\top \underbrace{\tilde{\mathbf{K}}(\hat{\mathbf{v}} - \tilde{\mathbf{v}}_{i-1})}_{\mathbf{r}_{i-1}} \tag{7}$$

$$\tilde{\mathbf{C}}_i = \tilde{\mathbf{K}}^{-1} - \tilde{\boldsymbol{\Sigma}}_i \tag{8}$$

Here, $\mathbf{s}_i$ denotes the policy corresponding to a specific approximation method (Wenger *et al.*, 2022)[Table 1]. This policy serves as the projection of the residual $\mathbf{r}_{i-1}$ results in α_i. The belief regarding the representer weights encodes the computational error as an added source of uncertainty, which can be integrated with the inherent uncertainty of the mathematical posterior.

We obtain the predictive posterior of IterRCGP by integrating out the representer weights: $p(f(\mathbf{x})|\mathcal{D}) = \int p(f(\mathbf{x})|\hat{\mathbf{v}})p(\hat{\mathbf{v}})d\hat{\mathbf{v}} = \mathcal{N}(\mathbf{f}; \hat{\mu}_i(\mathbf{x}), \hat{k}_i(\mathbf{x}, \mathbf{x}))$ where

$$\hat{\mu}_i(\mathbf{x}) = m(\mathbf{x}) + \mathbf{k}_*^\top \tilde{\mathbf{v}}_i \tag{9}$$

$$\hat{k}_i(\mathbf{x}, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_\mathbf{x}^\top \tilde{\mathbf{K}}^{-1}\mathbf{k}_\mathbf{x} + \underbrace{\mathbf{k}_\mathbf{x}^\top \tilde{\boldsymbol{\Sigma}}_i \mathbf{k}_\mathbf{x}}_{k_i^{\text{comp.}}(\mathbf{x},\mathbf{x})} = k(\mathbf{x}, \mathbf{x}) - \underbrace{\mathbf{k}_\mathbf{x}^\top \tilde{\mathbf{C}}_i \mathbf{k}_\mathbf{x}}_{\text{combined uncertainty}} \tag{10}$$

IterRCGP follows [Algorithm 1] from Wenger *et al.* (2022) to compute an estimate weights $\tilde{\mathbf{v}}_i$ and the rank-i precision matrix approximation $\tilde{\mathbf{C}}_i$.

4 Theoretical results

In this section, we present the theoretical properties of IterRCGP, building upon the IterGP framework and the RCGP class. Our theoretical analysis primarily aims to establish the following key results:

- Robustness property of IterGP and IterRCGP (Proposition 1).
- Convergence of IterRCGP's posterior mean in reproducing kernel Hilbert space (RKHS) norm (Proposition 2) and pointwise (Corollary 4).
- Combined uncertainty of IterRCGP is a tight worst-case bound on the relative distance to all potential latent functions shifted by the function $\mathbf{m_w}$ consistent with computational observations, similar to its IterGP counterpart (Proposition 3).

We establish the robustness properties of IterGP and IterRCGP using the Posterior Influence Function (PIF) as the robustness criterion. Appendix 1 provides a detailed definition of PIF. The proposition presented below is closely related to Altamirano *et al.* (2024)[Proposition 3.2].

Proposition 1. *(Robustness property)*
Suppose $f \sim \mathcal{GP}(m, k)$, $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\text{noise}}\mathbf{I})$ and let $C'_k \in \mathbb{R}; k = 1, 2, 3$ be constants independent of y^c_m. For any given iteration $i \in \{0, \dots, n\}$, IterGP regression has the PIF

$$\text{PIF}_{\text{IterGP}}(y^c_m, \mathcal{D}, i) = C'_1(y_m - y^c_m)^2 \tag{11}$$

which is not robust: $\text{PIF}_{\text{IterGP}}(y^c_m, \mathcal{D}, i) \to \infty$ as $|y^c_m| \to \infty$. In contrast, for the IterRCGP with $\sup_{\mathbf{x},y} w(\mathbf{x}, y) < \infty$,

$$\text{PIF}_{\text{IterRCGP}}(y^c_m, \mathcal{D}, i) = C'_2(w(x_n, y^c_n)^2 y^c_n)^2 + C'_3 \tag{12}$$

Therefore, if $\sup_{\mathbf{x},y} y\, w(\mathbf{x}, y)^2 < \infty$, IterRCGP regression is robust since $\sup_{y^c_m} |\text{PIF}_{\text{IterRCGP}}(y^c_m, \mathcal{D}, i)| < \infty$.

The proposition demonstrates that IterGP and IterRCGP inherit the same robustness properties as their respective counterparts, GP and RCGP. Specifically, the condition $\sup_{\mathbf{x},y} w(\mathbf{x}, y) < \infty$ ensures each observation has a finite weight, which is the key factor underpinning robustness.

The following proposition is analogous to [Theorem 1] in Wenger *et al.* (2022).

Proposition 2. *(Convergence in RKHS norm of the robust posterior mean approximation)*
Let $\mathcal{H}_k$ be the RKHS w.r.t. kernel k, $\sigma^2_{\text{noise}} > 0$ and let $\hat{\boldsymbol{\mu}}_* - \mathbf{m} \in \mathcal{H}_k$ be the unique solution to following minimization problem

$$\operatorname{argmin}_{f \in \mathcal{H}_k} L_n^w(\mathbf{f}, \mathbf{x}, \mathbf{y}) + \frac{1}{2n} \|\mathbf{f}\|^2_{\mathcal{H}_k} \tag{13}$$

which is equivalent to the mathematical RCGP mean posterior shifted by prior mean $\mathbf{m}$. Then for $i \in \{0, \dots, n\}$ the IterRCGP posterior mean $\hat{\boldsymbol{\mu}}_i$ satisfies:

$$\|\hat{\boldsymbol{\mu}}_* - \hat{\boldsymbol{\mu}}_i\|_{\mathcal{H}_k} \leq \hat{\rho}(i) c(\mathbf{J_w}) \|\hat{\boldsymbol{\mu}}_* - \mathbf{m}\|_{\mathcal{H}_k} \tag{14}$$

where $\hat{\rho}$ is the relative bound errors corresponding to the number of iterations i and the constant $c(\mathbf{J_w}) = \sqrt{1 + \frac{\lambda_{\max}(\mathbf{J_w})}{\lambda_{\min}(\mathbf{K})}} \to 1$ as $\lambda_{\max}(\mathbf{J_w}) \to 0$.

Appendix B provides more details about the relative bound errors. Proposition 2 provides a bound on the RKHS-norm error between the posterior mean of IterRCGP and the mathematical posterior mean of RCGP.

The final proposition parallels [Theorem 2] in Wenger *et al.* (2022), demonstrating that the combined uncertainty is a tight bound for all functions g that could have yielded the same computational outcomes.

Proposition 3. *(Combined and computational uncertainty as worst-case errors)*
Let $\sigma^2_{\text{noise}} \geq 0$ and $\hat{k}_i(\cdot,\cdot) = \hat{k}_*(\cdot,\cdot) + k_i^{\textit{comp.}}(\cdot,\cdot)$ be the combined uncertainty of IterRCGP. Furthermore, let $\mathbf{g} = [g(\mathbf{x}_1), \cdots, g(\mathbf{x}_n)] \in \mathbb{R}^n$. Then, for any new $\mathbf{x} \in \mathcal{X}$ we have

$$\sup_{\|g - m_w\|_{\mathcal{H}_{k^{\sigma w}}} \leq 1} \underbrace{g(\mathbf{x}) - \hat{\mu}^g(\mathbf{x})}_{\textit{math. err.}} + \underbrace{\hat{\mu}^g(\mathbf{x}) - \hat{\mu}_i^g(\mathbf{x})}_{\textit{comp. err.}} = \sqrt{\hat{k}_i(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}}} \tag{15}$$

$$\sup_{\|g - m_w\|_{\mathcal{H}_{k^{\sigma w}}} \leq 1} \underbrace{\hat{\mu}^g(\mathbf{x}) - \hat{\mu}_i^g(\mathbf{x})}_{\textit{comp. err.}} = \sqrt{k_i^{\textit{comp.}}(\mathbf{x}, \mathbf{x})} \tag{16}$$

where $\hat{\mu}^g(\cdot) = k(\cdot, \mathbf{X})\tilde{\mathbf{K}}^{-1}(\mathbf{g} - \mathbf{m_w})$ is the RCGP's posterior and $\hat{\mu}_i^g(\cdot) = k(\cdot, \mathbf{X})\tilde{\mathbf{C}}_i(\mathbf{g} - \mathbf{m_w})$ IterRCGP's posterior mean for the latent function g and the function m_w lies in $\mathcal{H}_{k^{\sigma w}}$.

The consequence of Proposition 3 is then formalized through the following corollary:

Corollary 4. *(Pointwise convergence of robust posterior mean)*
Assume the conditions of Proposition 3 hold and assume the latent function $g \in \mathcal{H}_{k^{\sigma w}}$. Let $\hat{\boldsymbol{\mu}}$ be the corresponding mathematical RCGP posterior mean and $\hat{\boldsymbol{\mu}}_i$ the IterRCGP posterior mean. It holds that

$$\frac{|g(\mathbf{x}) - \hat{\mu}_i(\mathbf{x})|}{\|g\|_{\mathcal{H}_{k^{\sigma w}}}} \leq \sqrt{\hat{k}_i(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}}} \tag{17}$$

$$\frac{\hat{\mu}(\mathbf{x}) - \hat{\mu}_i(\mathbf{x})}{\|g\|_{\mathcal{H}_{k^{\sigma w}}}} \leq \sqrt{k_i^{\textit{comp.}}(\mathbf{x}, \mathbf{x})} \tag{18}$$

5 Conclusion

In this paper, we demonstrated that computation-aware GPs as presented by Wenger *et al.* (2022) lack robustness in the PIF sense. Subsequently, we introduced Iter RCGPs, a novel class of provably robust computation-aware GPs. Since our work mainly involves theoretical analyses, our immediate perspective is to run numerical experiments using synthetic and real-world datasets. Next, one interesting avenue for applying Iter RCGPs is that of Bayesian Optimization (BO), a domain where uncertainty quantification is key to coming up with good exploration policies.

Indeed, the issue of refined uncertainty quantification has recently gained attention in BO. One approach addresses this by jointly optimizing the selection of the optimal data point along with the SVGP parameters and the locations of the inducing points (Maus *et al.*, 2024). Another study incorporates conformal prediction into BO by leveraging the conformal Bayes posterior and proposing generalized versions of the corresponding BO acquisition functions (Stanton *et al.*, 2023).

References

Altamirano, M., Briol, F.-X., and Knoblauch, J. (2024). Robust and conjugate gaussian process regression. In *The 41st International Conference on Machine Learning*.

Chen, Y., Prati, A., Montgomery, J., and Garnett, R. (2023). A multi-task gaussian process model for inferring time-varying treatment effects in panel data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*.

Cheng, L., Ramchandran, S., Vatanen, T., Lietzén, N., Lahesmaa, R., Vehtari, A., and Lähdesmäki, H. (2019). An additive gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nature Communications*.

Garnett, R. (2023). *Bayesian Optimization*. Cambridge University Press.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

Maus, N., Kim, K., Pleiss, G., Eriksson, D., Cunningham, J. P., and Gardner, J. R. (2024). Approximation-aware bayesian optimization.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Stanton, S., Maddox, W., and Wilson, A. G. (2023). Bayesian optimization with conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*.

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*.

Wenger, J., Pleiss, G., Pförtner, M., Hennig, P., and Cunningham, J. P. (2022). Posterior and computational uncertainty in gaussian processes. In *Advances in Neural Information Processing Systems*.

Wild, V., Kanagawa, M., and Sejdinovic, D. (2023). Connections and equivalences between the nyström method and sparse variational gaussian processes.

Williams, C. and Seeger, M. (2000). Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*.

A Proof of Proposition 1

Posterior influence function. Given the dataset $\mathcal{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^n$, we define the contamination of $\mathcal{D}$ indexed by $m \in \{1, \dots, n\}$ as $\mathcal{D}_m^c = (\mathcal{D} \setminus (\mathbf{x}_m, y_m)) \cup (\mathbf{x}_m, y_m^c)$. PIF in general, aims to measure the impact of y_m^c on inference through the divergence between the contaminated and uncontaminated posteriors $p(\mathbf{f}|\mathcal{D}_m^c)$ and $p(\mathbf{f}|\mathcal{D})$:

$$\mathrm{PIF}(y_m^c, \mathcal{D}) = \mathrm{KL}(p(\mathbf{f}|\mathcal{D}) \| p(\mathbf{f}|\mathcal{D}_m^c)) \tag{S1}$$

where we call a posterior robust if $\sup_{y \in \mathcal{Y}} |\mathrm{PIF}(y_m^c, \mathcal{D})| < \infty$.

We then establish the following lemma to prove Proposition 1.

Lemma 5. *For an arbitrary matrice $\hat{\mathbf{S}} \in \mathbb{R}^{m \times n}$ and positive semidefinite matrice $\hat{\mathbf{B}} \in \mathbb{R}^{n \times n}$, we have that*

$$\mathrm{Tr}((\hat{\mathbf{S}}\hat{\mathbf{B}}\hat{\mathbf{S}}^\top)^{-1}) = \hat{\mathbf{S}}^{+\top}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{G}}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{S}}^{+} \tag{S2}$$

where we define $\hat{\mathbf{G}} = \mathbf{I} - \hat{\mathbf{B}}^{-1/2}(\mathbf{I} - \hat{\mathbf{S}}^+\hat{\mathbf{S}})(\hat{\mathbf{B}}^{-1/2}(\mathbf{I} - \hat{\mathbf{S}}^+\hat{\mathbf{S}}))^+$ and $^+$ denotes the Moore-Penrose inverse.

Proof:

The whole proof is derived from an answer to a question posted on the Mathematics Stack Exchange Forums, which we write here for conciseness.

Denote $\hat{\mathbf{O}} = \mathbf{I} - \hat{\mathbf{S}}^+\hat{\mathbf{S}}$ and $\mathbf{H}(\alpha) = (\hat{\mathbf{S}}(\alpha\mathbf{I} + \hat{\mathbf{B}}^{-1})^{-1}\hat{\mathbf{S}}^\top)^{-1}$. Note that

$$(\hat{\mathbf{S}}\hat{\mathbf{B}}\hat{\mathbf{S}}^\top)^{-1} = \lim_{\alpha \to 0} \mathbf{H}(\alpha) \tag{S3}$$

By applying Woodbury matrix identity, we can rewrite $\mathbf{H}(\alpha)$ as follows:

$$\mathbf{H}(\alpha) = \left(\frac{1}{\alpha}\hat{\mathbf{S}}\hat{\mathbf{S}}^\top - \frac{1}{\alpha}\hat{\mathbf{S}}\hat{\mathbf{B}}^{-1/2}\left(\mathbf{I} + \frac{1}{\alpha}\hat{\mathbf{B}}^{-1}\right)^{-1}\frac{1}{\alpha}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{S}}^\top\right)^{-1} \tag{S4}$$

Since $\hat{\mathbf{S}}\hat{\mathbf{S}}^\top$ is invertible, we can apply the Woodbury matrix identity for the second time to obtain

$$\begin{aligned}
\mathbf{H}(\alpha) &= \alpha(\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1} - (\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1}\hat{\mathbf{S}}\hat{\mathbf{B}}^{-1/2} \\
&\quad (-(\mathbf{I} + \frac{1}{\alpha}\hat{\mathbf{B}}^{-1}) + \frac{1}{\alpha}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{S}}^\top(\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1}\hat{\mathbf{S}}\hat{\mathbf{B}}^{-1/2})^{-1}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{S}}^\top(\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1} \qquad \text{(S5)} \\
&= \alpha(\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1} + (\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1}\hat{\mathbf{S}}\hat{\mathbf{B}}^{-1/2}(\mathbf{I} + \frac{1}{\alpha}\hat{\mathbf{B}}^{-1/2}(\mathbf{I} - \hat{\mathbf{S}}^\top(\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1}\hat{\mathbf{S}})\hat{\mathbf{B}}^{-1/2})^{-1} \\
&\quad \hat{\mathbf{B}}^{-1/2}\hat{\mathbf{S}}^\top(\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1} \qquad \text{(S6)}
\end{aligned}$$

We note that

$$\hat{\mathbf{S}}^\top(\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1} = \hat{\mathbf{S}}^+ \tag{S7}$$

$$\mathbf{I} - \hat{\mathbf{S}}^\top(\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1}\hat{\mathbf{S}} = \hat{\mathbf{O}} \tag{S8}$$

Then, we rewrite $\mathbf{H}(\alpha)$ as follows:

$$\mathbf{H}(\alpha) = \alpha(\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1} + \hat{\mathbf{S}}^{+\top}\hat{\mathbf{B}}^{-1/2}\left(\mathbf{I} + \frac{1}{\alpha}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{O}}\hat{\mathbf{O}}\hat{\mathbf{B}}^{-1/2}\right)^{-1}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{S}}^{+} \tag{S9}$$

Applying the Woodbury matrix identity for the third time provides

$$\mathbf{H}(\alpha) = \alpha(\hat{\mathbf{S}}\hat{\mathbf{S}}^\top)^{-1} + \hat{\mathbf{S}}^{+\top}\hat{\mathbf{B}}^{-1/2}(\mathbf{I} - \hat{\mathbf{B}}^{-1/2}\hat{\mathbf{O}}(\alpha\mathbf{I} + \hat{\mathbf{O}}\hat{\mathbf{B}}^{-1}\hat{\mathbf{O}})^{-1}\hat{\mathbf{O}}\hat{\mathbf{B}}^{-1/2})\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{S}}^{+} \tag{S10}$$

Since the Moore-Penrose inverse of a matrice $\mathbf{A}$ is a limit:

$$\mathbf{A}^+ = \lim_{\alpha\to 0}(\mathbf{A}^\top\mathbf{A} + \alpha\mathbf{I})^{-1}\mathbf{A}^\top = \lim_{\alpha\to 0}\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \alpha\mathbf{I})^{-1} \tag{S11}$$

We can take the limit of $\mathbf{H}(\alpha)$ as $\alpha \to 0$ and apply the limit relation above to obtain the following result:

$$(\hat{\mathbf{S}}\hat{\mathbf{B}}\hat{\mathbf{S}}^\top)^{-1} = \hat{\mathbf{S}}^{+\top}\hat{\mathbf{B}}^{-1/2}\underbrace{(\mathbf{I} - \hat{\mathbf{B}}^{-1/2}\hat{\mathbf{O}}(\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{O}})^{+})}_{\hat{\mathbf{G}}}\hat{\mathbf{B}}^{-1/2}\hat{\mathbf{S}}^{+} \tag{S12}$$

PIF for the IterGP. IterGP regression has the PIF for some constant $C_1' \in \mathbb{R}$.

$$\mathrm{PIF}_{\mathrm{IterGP}}(y_m^c, \mathcal{D}, i) = C_1'(y_m - y_m^c)^2 \tag{S13}$$

and is not robust: $\mathrm{PIF}_{\mathrm{IterGP}}(y_m^c, \mathcal{D}, i) \to \infty$ as $|y_m^c| \to \infty$.

Proof:

Let $p(\mathbf{f}|\mathcal{D}) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_i, \mathbf{K}_i)$ and $p(\mathbf{f}|\mathcal{D}_m^c) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_i^c, \mathbf{K}_i^c)$ be the uncontaminated and contaminated computation-aware GP, respectively. Here,

$$\boldsymbol{\mu}_i = \mathbf{m} + \mathbf{K}\mathbf{v}_i \tag{S14}$$
$$\mathbf{K}_i = \mathbf{K}\mathbf{C}_i\sigma^2_{\text{noise}}\mathbf{I}_n \tag{S15}$$
$$\boldsymbol{\mu}_i^c = \mathbf{m} + \mathbf{K}\mathbf{v}_i^c \tag{S16}$$
$$\mathbf{K}_i^c = \mathbf{K}\mathbf{C}_i\sigma^2_{\text{noise}}\mathbf{I}_n \tag{S17}$$

Note that both $\mathbf{K}_i$ and $\mathbf{K}_i^c$ share the same matrice $\mathbf{C}_i$. Then, the PIF has the following form:

$$\mathrm{PIF}_{\mathrm{IterGP}}(y_m^c, \mathcal{D}, i) = \frac{1}{2}(\mathrm{Tr}(\mathbf{K}_i^c\mathbf{K}_i) - n + (\boldsymbol{\mu}_i^c - \boldsymbol{\mu}_i)^\top(\mathbf{K}_i^c)^{-1}(\boldsymbol{\mu}_i^c - \boldsymbol{\mu}_i) + \ln\left(\frac{\det(\mathbf{K}_i^c)}{\det(\mathbf{K}_i)}\right) \tag{S18}$$

Based on Altamirano *et al.* (2024), the PIF leads to the following form:

$$\mathrm{PIF}_{\mathrm{IterGP}}(y_m^c, \mathcal{D}, i) = \frac{1}{2}\left((\boldsymbol{\mu}_i^c - \boldsymbol{\mu}_i)^\top(\mathbf{K}_i^c)^{-1}(\boldsymbol{\mu}_i^c - \boldsymbol{\mu}_i)\right) \tag{S19}$$

Notice that the term $\boldsymbol{\mu}_i^c - \boldsymbol{\mu}_i$ can be written as

$$\begin{aligned}
\boldsymbol{\mu}_i^c - \boldsymbol{\mu}_i &= (\mathbf{m} + \mathbf{K}\mathbf{v}_i^c) - (\mathbf{m} + \mathbf{K}\mathbf{v}_i) && \text{(S20)}\\
&= \mathbf{K}(\mathbf{v}_i^c - \mathbf{v}_i) && \text{(S21)}\\
&= \mathbf{K}(\mathbf{C}_i(\mathbf{y}^c - \mathbf{m}) - \mathbf{C}_i(\mathbf{y} - \mathbf{m})) && \text{(S22)}\\
&= \mathbf{K}(\mathbf{C}_i(\mathbf{y}^c - \mathbf{y})) && \text{(S23)}
\end{aligned}$$

Substituting the RHS of Eq. (S23) to $\boldsymbol{\mu}_i^c - \boldsymbol{\mu}_i$ in Eq. (S19), we obtain

$$\mathrm{PIF}_{\mathrm{IterGP}}(y_m^c, \mathcal{D}, i) = \frac{1}{2}(\mathbf{C}_i(\mathbf{y}^c - \mathbf{y}))^\top \mathbf{K}\left(\mathbf{K}\mathbf{C}_i\sigma^2\mathbf{I}\right)^{-1}\mathbf{K}(\mathbf{C}_i(\mathbf{y}^c - \mathbf{y})) \tag{S24}$$
$$= \frac{1}{2}(\mathbf{y}^c - \mathbf{y})^\top \mathbf{C}_i^\top \mathbf{K}\sigma_{\text{noise}}^{-2}\mathbf{I}(\mathbf{y}^c - \mathbf{y}) \tag{S25}$$

Note that $\mathbf{y}$ and $\mathbf{y}^c$ have only one exception for the $m-$th element. Thus, we have

$$\mathrm{PIF}_{\mathrm{IterGP}}(y_m^c, \mathcal{D}, i) = \frac{1}{2}[\mathbf{C}_i^\top \mathbf{K}\sigma^{-2}\mathbf{I}]_{mm}(y_m^c - y_m)^2 \tag{S26}$$

PIF for the IterRCGP. For the IterRCGP with $\sup_{\mathbf{x},y} w(\mathbf{x}, y) < \infty$, the following holds

$$\mathrm{PIF}_{\mathrm{IterRCGP}}(y_m^c, \mathcal{D}, i) \leq C_2'(w(\mathbf{x}_m, y_m^c)^2 y_m^c)^2 + C_3' \tag{S27}$$

for some constants $C_2', C_3' \in \mathbb{R}$. Therefore, if $\sup_{\mathbf{x},y} y\, w(\mathbf{x}, y)^2 < \infty$, the computation-aware RCGP is robust since $|\mathrm{PIF}_{\mathrm{IterRCGP}}(y_m^c, \mathcal{D}, i)| < \infty$.

Proof:

Without loss of generality, we aim to prove the bound for $m = n$. We can extend the proof for an arbitrary $m \in \{1, \ldots, n\}$. Let $p^w(\mathbf{f}|\mathcal{D}) = \mathcal{N}(\mathbf{f}; \hat{\boldsymbol{\mu}}_i, \hat{\mathbf{K}}_i)$ and $p^w(\mathbf{f}|\mathcal{D}_m^c) = \mathcal{N}(\mathbf{f}; \hat{\boldsymbol{\mu}}_i^c, \hat{\mathbf{K}}_i^c)$ be the uncontaminated and contaminated computation-aware RCGP, respectively. Here,

$$\hat{\boldsymbol{\mu}}_i = \mathbf{m} + \mathbf{K}\tilde{\mathbf{C}}_i\tilde{\mathbf{v}}_i \tag{S28}$$
$$\hat{\mathbf{K}}_i = \mathbf{K}\tilde{\mathbf{C}}_i\sigma_{\text{noise}}^2\mathbf{J}_{\mathbf{w}} \tag{S29}$$
$$\hat{\boldsymbol{\mu}}_i^c = \mathbf{m} + \mathbf{K}\tilde{\mathbf{C}}_i^c\tilde{\mathbf{v}}_i^c \tag{S30}$$
$$\hat{\mathbf{K}}_i^c = \mathbf{K}\tilde{\mathbf{C}}_i^c\sigma_{\text{noise}}^2\mathbf{J}_{\mathbf{w}^c} \tag{S31}$$

where $\mathbf{w}^c = (w(\mathbf{x}_1, y_1), \ldots, w(\mathbf{x}_n, y_n^c))^\top$. The PIF has the following form

$$\mathrm{PIF}_{\mathrm{IterRCGP}}(y_m^c, \mathcal{D}, i) = \frac{1}{2}\left(\underbrace{\mathrm{Tr}((\hat{\mathbf{K}}_i^c)^{-1}\hat{\mathbf{K}}_i) - n}_{(1)} + \underbrace{(\hat{\boldsymbol{\mu}}_i^c - \hat{\boldsymbol{\mu}}_i)^\top(\hat{\mathbf{K}}_i^c)^{-1}(\hat{\boldsymbol{\mu}}_i^c - \hat{\boldsymbol{\mu}}_i)}_{(2)} + \underbrace{\ln\left(\frac{\det(\hat{\mathbf{K}}_i^c)}{\det(\hat{\mathbf{K}}_i)}\right)}_{(3)}\right) \tag{S32}$$

We first derive the bound for (1):

$$(1) = \mathrm{Tr}((\hat{\mathbf{K}}_i^c)^{-1}\hat{\mathbf{K}}_i) - n \tag{S33}$$
$$= \mathrm{Tr}\left((\mathbf{K}\tilde{\mathbf{C}}_i^c\sigma_{\text{noise}}^2\mathbf{J}_{\mathbf{w}^c})^{-1}\mathbf{K}\tilde{\mathbf{C}}_i\sigma_{\text{noise}}^2\mathbf{J}_{\mathbf{w}}\right) - n \tag{S34}$$
$$= \mathrm{Tr}(\sigma_{\text{noise}}^{-2}\mathbf{J}_{\mathbf{w}^c}^{-1}(\tilde{\mathbf{C}}_i^c)^{-1}\tilde{\mathbf{C}}_i\sigma_{\text{noise}}^2\mathbf{J}_{\mathbf{w}}) - n \tag{S35}$$
$$\leq \mathrm{Tr}(\sigma_{\text{noise}}^{-2}\mathbf{J}_{\mathbf{w}^c}^{-1}(\tilde{\mathbf{C}}_i^c)^{-1})\mathrm{Tr}(\tilde{\mathbf{C}}_i\sigma_{\text{noise}}^2\mathbf{J}_{\mathbf{w}}) - n \tag{S36}$$
$$\leq \mathrm{Tr}(\sigma_{\text{noise}}^{-2}\mathbf{J}_{\mathbf{w}^c}^{-1})\mathrm{Tr}(\tilde{\mathbf{C}}_i^c)^{-1})\mathrm{Tr}(\tilde{\mathbf{C}}_i\sigma_{\text{noise}}^2\mathbf{J}_{\mathbf{w}}) - n \tag{S37}$$

The first and second inequality come from the fact that $\mathrm{Tr}(\mathbf{A}\mathbf{F}) \leq \mathrm{Tr}(\mathbf{A})\mathrm{Tr}(\mathbf{F})$ for two positive semidefinite matrices $\mathbf{A}$ and $\mathbf{F}$. Since $\mathrm{Tr}(\tilde{\mathbf{C}}_i\sigma_{\text{noise}}^2\mathbf{J}_{\mathbf{w}})$ does not contain the contamination term, we can write $\bar{C}_1 = \mathrm{Tr}(\tilde{\mathbf{C}}_i \sigma^2_{\text{noise}} \mathbf{J}_{\mathbf{w}})$. Let $\mathbf{B} = (\mathbf{S}_i^\top \tilde{\mathbf{K}}^c \mathbf{S}_i)^{-1}$ such that $\mathbf{C}_i^c = \mathbf{S}_i^\top \mathbf{B} \mathbf{S}_i^\top$. Observe that matrice $\mathbf{B}$ is positive semidefinite. Thus, we can apply Lemma 5 to obtain the bound of $\mathrm{Tr}((\tilde{\mathbf{C}}_i^c)^{-1})$:

$$
\begin{aligned}
\mathrm{Tr}((\tilde{\mathbf{C}}_i^c)^{-1}) &= \mathrm{Tr}((\mathbf{S}_i^\top \mathbf{B} \mathbf{S}_i^\top)^{-1}) && \text{(S38)}\\
&= \mathrm{Tr}(\mathbf{S}_i^{+\top} \mathbf{B}^{-1/2} \mathbf{G} \mathbf{B}^{-1/2} \mathbf{S}_i^{+}) && \text{(S39)}\\
&\leq \mathrm{Tr}(\mathbf{S}_i^{+} \mathbf{S}_i^{+\top}) \mathrm{Tr}(\mathbf{B}^{-1/2} \mathbf{B}^{-1/2}) \mathrm{Tr}(\mathbf{G}) && \text{(S40)}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathrm{Tr}(\mathbf{G}) &= \mathrm{Tr}(\mathbf{I} - \mathbf{B}^{-1/2}(\mathbf{I} - \mathbf{S}_i^{+}\mathbf{S}_i)(\mathbf{B}^{-1/2}(\mathbf{I} - \mathbf{S}_i^{+}\mathbf{S}_i))^{+}) && \text{(S41)}\\
&= n - \mathrm{Tr}(\mathbf{B}^{-1/2}(\mathbf{I} - \mathbf{S}_i^{+}\mathbf{S}_i)(\mathbf{I} - \mathbf{S}_i^{+}\mathbf{S}_i)^{+} \mathbf{B}^{-1/2+}) && \text{(S42)}\\
&\leq n - \mathrm{Tr}(\mathbf{B}^{-1/2+} \mathbf{B}^{-1/2}) \mathrm{Tr}((\mathbf{I} - \mathbf{S}_i^{+}\mathbf{S}_i)(\mathbf{I} - \mathbf{S}_i^{+}\mathbf{S}_i)^{+}) && \text{(S43)}
\end{aligned}
$$

The inequality S40 stems from the trace circular property and the inequality of the product of two semidefinite matrices. Note that $\mathrm{Tr}(\mathbf{G}) \leq n$ since $\mathbf{B}^{-1/2+}\mathbf{B}^{-1/2}$ and $(\mathbf{I} - \mathbf{S}_i^{+}\mathbf{S}_i)(\mathbf{I} - \mathbf{S}_i^{+}\mathbf{S}_i)^{+}$ in S43 are positive semidefinite matrice; thus both have non-negative trace value. Therefore, we find that

$$
\begin{aligned}
\mathrm{Tr}((\tilde{\mathbf{C}}_i^c)^{-1}) &\leq n \mathrm{Tr}(\mathbf{S}_i^{+} \mathbf{S}_i^{+\top}) \mathrm{Tr}(\mathbf{B}^{-1}) && \text{(S44)}\\
&\leq n \mathrm{Tr}(\mathbf{S}_i^{+} \mathbf{S}_i^{+\top}) \mathrm{Tr}(\mathbf{S}_i \mathbf{S}_i^\top) \mathrm{Tr}(\tilde{\mathbf{K}}^c) && \text{(S45)}\\
&= \bar{C}_2 \mathrm{Tr}(\mathbf{K} + \sigma^2_{\text{noise}} \mathbf{J}_{\mathbf{w}^c}) && \text{(S46)}
\end{aligned}
$$

where we define $\bar{C}_2 = n \mathrm{Tr}(\mathbf{S}_i^{+} \mathbf{S}_i^{+\top}) \mathrm{Tr}(\mathbf{S}_i \mathbf{S}_i^\top)$. We then plug S46 into S37 to obtain

$$
\begin{aligned}
(1) &\leq \mathrm{Tr}(\sigma^{-2}_{\text{noise}} \mathbf{J}_{\mathbf{w}^c}^{-1}) \mathrm{Tr}(\mathbf{K} + \sigma^2_{\text{noise}} \mathbf{J}_{\mathbf{w}^c}) \bar{C}_1 \bar{C}_2 - n && \text{(S47)}\\
&= \left(\sum_{j=1}^{n} \left(\sigma^{-2}_{\text{noise}} w^2(\mathbf{x}_j, y_j)\right) \sum_{k=1}^{n} \left(\mathbf{K}_{kk} + \sigma^2_{\text{noise}} w^{-2}(\mathbf{x}_k, y_k)\right) \right) \bar{C}_1 \bar{C}_2 - n && \text{(S48)}\\
&\leq \left(n^2 \sup_{\mathbf{x}, y} w^2(\mathbf{x}, y) \sup_{\hat{\mathbf{x}}, \hat{y}} w^{-2}(\hat{\mathbf{x}}, \hat{y}) \right) \bar{C}_1 \bar{C}_2 - n = \bar{C}_3 && \text{(S49)}
\end{aligned}
$$

Next, we derive the bound for (2). Following Altamirano *et al.* (2024), we have that

$$
(2) \leq \lambda_{\max}((\hat{\mathbf{K}}_i^c)^{-1}) \|\hat{\boldsymbol{\mu}}_i^c - \hat{\boldsymbol{\mu}}_i\|_1^2 \qquad \text{(S50)}
$$

We expand $\lambda_{\max}((\hat{\mathbf{K}}_i^c)^{-1})$ and derive the following bound:

$$
\begin{aligned}
\lambda_{\max}((\hat{\mathbf{K}}_i^c)^{-1}) &= \lambda_{\max}(\sigma^{-2} \mathbf{J}_{\mathbf{w}^c}^{-1} (\tilde{\mathbf{C}}_i^c)^{-1} \mathbf{K}^{-1}) && \text{(S51)}\\
&\leq \lambda_{\max}(\sigma^{-2}_{\text{noise}} \mathbf{J}_{\mathbf{w}^c}^{-1}) \lambda_{\max}((\tilde{\mathbf{C}}_i^c)^{-1}) \lambda_{\max}(\mathbf{K}^{-1}) && \text{(S52)}\\
&= \lambda_{\max}(\sigma^{-2}_{\text{noise}} \mathbf{J}_{\mathbf{w}^c}^{-1}) \lambda_{\min}(\tilde{\mathbf{C}}_i^c) \lambda_{\max}(\mathbf{K}^{-1}) && \text{(S53)}\\
&\leq \lambda_{\max}(\sigma^{-2}_{\text{noise}} \mathbf{J}_{\mathbf{w}^c}^{-1}) \left(\lambda_{\min}((\tilde{\mathbf{K}}^c)^{-1})\right) \lambda_{\max}(\mathbf{K}^{-1}) && \text{(S54)}\\
&\leq \lambda_{\max}(\sigma^{-2}_{\text{noise}} \mathbf{J}_{\mathbf{w}^c}^{-1}) \lambda_{\min}((\tilde{\mathbf{K}}^c)^{-1}) \lambda_{\max}(\mathbf{K}^{-1}) && \text{(S55)}\\
&\leq \lambda_{\max}(\sigma^{-2}_{\text{noise}} \mathbf{J}_{\mathbf{w}^c}^{-1}) (\lambda_{\max}(\mathbf{K}) + \lambda_{\max}(\sigma^2_{\text{noise}} \mathbf{J}_{\mathbf{w}^c})) \lambda_{\max}(\mathbf{K}^{-1}) && \text{(S56)}
\end{aligned}
$$

The first inequality follows from the maximum eigenvalue of the product of two positive semidefinite matrices. The fact that the maximum eigenvalue of a matrice is equal to the minimum eigenvalue of the inverse leads to the second equality. Recall that $\tilde{\mathbf{C}}_i^c = (\tilde{\mathbf{K}}^c)^{-1} - \mathbf{\Sigma}_i$. Since $\tilde{\mathbf{C}}_i^c$, $(\tilde{\mathbf{K}}^c)^{-1}$ and $\mathbf{\Sigma}_i$ are positive semidefinite matrices, the third inequality holds. The fourth inequality stems from the equivalence of the maximum eigenvalue and the addition property of the maximum eigenvalue of two positive semidefinite matrices.

Since $\mathbf{J}_{\mathbf{w}^c}^{-1} = \mathrm{diag}((\mathbf{w}^c)^2)$, and $\sup_{\mathbf{x},y} w(\mathbf{x}, y) < \infty$, it holds that $\lambda_{\max}(\sigma_{\text{noise}}^{-2}\mathbf{J}_{\mathbf{w}^c}^{-1}) = \bar{C}_4 < +\infty$ and $\lambda_{\max}(\sigma_{\text{noise}}^2 \mathbf{J}_{\mathbf{w}^c}) = \bar{C}_5 < +\infty$, such that

$$\lambda_{\max}((\hat{\mathbf{K}}_i^c)^{-1}) \leq \bar{C}_4(\lambda_{\max}(\mathbf{K}) + \bar{C}_5)\lambda_{\max}(\mathbf{K}^{-1}) = \bar{C}_6 \tag{S57}$$

We substitute $\bar{C}_6$ into (2) to obtain

$$\begin{aligned}
(2) &\leq \bar{C}_6 \|\hat{\boldsymbol{\mu}}_i^c - \hat{\boldsymbol{\mu}}_i\|_1^2 && \text{(S58)}\\
&= \bar{C}_6 \|(\mathbf{m} + \mathbf{K}\tilde{\mathbf{v}}_i^c) - (\mathbf{m} + \mathbf{K}\tilde{\mathbf{v}}_i)\|_1^2 && \text{(S59)}\\
&= \bar{C}_6 \|\mathbf{K}(\tilde{\mathbf{C}}_i^c(\mathbf{y} - \mathbf{m}_{\mathbf{w}^c}) - \tilde{\mathbf{C}}_i(\mathbf{y} - \mathbf{m}_{\mathbf{w}}))\|_1^2 && \text{(S60)}\\
&\leq \bar{C}_6 \|\mathbf{K}\|_F \|\tilde{\mathbf{C}}_i^c(\mathbf{y} - \mathbf{m}_{\mathbf{w}^c}) - \tilde{\mathbf{C}}_i(\mathbf{y} - \mathbf{m}_{\mathbf{w}})\|_1^2 && \text{(S61)}\\
&\leq \bar{C}_6 \|\mathbf{K}\|_F (\|(\tilde{\mathbf{K}}^c)^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{w}^c}) - (\tilde{\mathbf{K}})^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{w}})\|_1^2 + \|\tilde{\mathbf{\Sigma}}_i^c(\mathbf{y} - \mathbf{m}_{\mathbf{w}^c}) - \tilde{\mathbf{\Sigma}}_i(\mathbf{y} - \mathbf{m}_{\mathbf{w}})\|_1^2) && \text{(S62)}\\
&\leq q\bar{C}_6 \|\mathbf{K}\|_F (\|(\tilde{\mathbf{K}}^c)^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{w}^c}) - (\tilde{\mathbf{K}})^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{w}})\|_1^2 && \text{(S63)}\\
&= q\bar{C}_6 \|\mathbf{K}\|_F ((\mathbf{K} + \sigma_{\text{noise}}^2 \mathbf{J}_{\mathbf{w}^c})^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{w}^c}) - (\mathbf{K} + \sigma_{\text{noise}}^2 \mathbf{J}_{\mathbf{w}})(\mathbf{y} - \mathbf{m}_{\mathbf{w}})\|_1^2 && \text{(S64)}
\end{aligned}$$

for a constant $q > 0$. The second equality follows from Wenger *et al.* (2022)[Eq. (S45)]. The first inequality follows the Cauchy-Schwarz inequality. The second inequality stems from the definition of $\tilde{\mathbf{C}}_i, \tilde{\mathbf{C}}_i^c$, and the triangle inequality. Finally, the last inequality holds since $(\tilde{\mathbf{K}}_i^{-1} - \tilde{\mathbf{\Sigma}}_i), \tilde{\mathbf{K}}_i^{-1}, \tilde{\mathbf{\Sigma}}_i \succeq 0$.

Applying results from Altamirano *et al.* (2024), we obtain

$$\begin{aligned}
(2) &\leq q\bar{C}_6 \|\mathbf{K}\|_F ((\mathbf{K} + \sigma_{\text{noise}}^2 \mathbf{J}_{\mathbf{w}^c})^{-1}(\mathbf{y} - \mathbf{m}_{\mathbf{w}^c}) - (\mathbf{K} + \sigma_{\text{noise}}^2 \mathbf{J}_{\mathbf{w}})(\mathbf{y} - \mathbf{m}_{\mathbf{w}})\|_1^2 && \text{(S65)}\\
&\leq q\bar{C}_6 \|\mathbf{K}\|_F 2((\bar{C}_7 + \bar{C}_8)^2 + (\bar{C}_9 + \bar{C}_{10})^2 (w(x_n, y_n^c)^2 y_n^c)^2) && \text{(S66)}\\
&\leq \bar{C}_{11} + \bar{C}_{12}(w(x_n, y_n^c)^2 y_n^c)^2 && \text{(S67)}
\end{aligned}$$

where $\bar{C}_{11} = q\bar{C}_6\|\mathbf{K}\|_F 2(\bar{C}_7 + \bar{C}_8)^2$ and $\bar{C}_{12} = q\bar{C}_6\|\mathbf{K}\|_F 2(\bar{C}_9 + \bar{C}_{10})^2$. The terms $\bar{C}_7, \bar{C}_8, \bar{C}_9, \bar{C}_{10}$ equal to $\tilde{C}_6, \tilde{C}_8, \tilde{C}_7, \tilde{C}_9$ in Altamirano *et al.* (2024).

The term (3) can be written as follows:

$$\begin{aligned}
(3) &= \ln\left(\frac{\det(\hat{\mathbf{K}}_i^c)}{\det(\hat{\mathbf{K}}_i)}\right) && \text{(S68)}\\
&= \ln\left(\frac{\det(\tilde{\mathbf{C}}_i^c \sigma_{\text{noise}}^2 \mathbf{J}_{\mathbf{w}^c})}{\det(\tilde{\mathbf{C}}_i \sigma_{\text{noise}}^2 \mathbf{J}_{\mathbf{w}})}\right) && \text{(S69)}\\
&= \ln(\det(\sigma_{\text{noise}}^{-2}\mathbf{J}_{\mathbf{w}}^{-1}\tilde{\mathbf{C}}_i^{-1})\det(\tilde{\mathbf{C}}_i^c)\det(\sigma_{\text{noise}}^2 \mathbf{J}_{\mathbf{w}^c})) && \text{(S70)}
\end{aligned}$$

Observe that we can write $\bar{C}_{13} = \ln(\det(\sigma_{\text{noise}}^{-2}\mathbf{J}_{\mathbf{w}}^{-1}\tilde{\mathbf{C}}_i^{-1})$ since it does not contain the contimation term. Furthermore, we obtain

$$
\begin{aligned}
(3) &= \ln(\bar{C}_{13}\det(\tilde{\mathbf{C}}_i^c)\det(\sigma^2_{\text{noise}}\mathbf{J}_{\mathbf{w}^c})) && \text{(S71)}\\
&\leq \ln(\bar{C}_{13}\det((\tilde{\mathbf{K}}^c)^{-1})\det(\sigma^2_{\text{noise}}\mathbf{J}_{\mathbf{w}^c})) && \text{(S72)}\\
&= \ln\left(\bar{C}_{13}\frac{\det(\sigma^2_{\text{noise}}\mathbf{J}_{\mathbf{w}^c})}{\det(\mathbf{K}+\sigma^2_{\text{noise}}\mathbf{J}_{\mathbf{w}^c})}\right) && \text{(S73)}\\
&\leq \ln\left(\bar{C}_{13}\frac{\det(\sigma^2_{\text{noise}}\mathbf{J}_{\mathbf{w}^c})}{\det(\mathbf{K})+\det(\sigma^2_{\text{noise}}\mathbf{J}_{\mathbf{w}^c})}\right) && \text{(S74)}
\end{aligned}
$$

The first inequality holds since $((\tilde{\mathbf{K}}_i^c)^{-1} - \tilde{\mathbf{\Sigma}}_i^c), (\tilde{\mathbf{K}}_i^c)^{-1}, \tilde{\mathbf{\Sigma}}_i^c \succeq 0$, so $\det((\tilde{\mathbf{K}}_i^c)^{-1}) \geq \det(\tilde{\mathbf{\Sigma}}_i^c)$. The last inequality leverages the fact that $\det(\mathbf{A}+\mathbf{F}) \geq \det(\mathbf{A}) + \det(\mathbf{F})$ for $\mathbf{A}$ and $\mathbf{F}$ are positive semidefinite matrices. Since $\det(\mathbf{K}), \det(\sigma^2_{\text{noise}}\mathbf{J}_{\mathbf{w}^c}) \geq 0$, we find that

$$
\ln\left(\frac{\det(\sigma^2_{\text{noise}}\mathbf{J}_{\mathbf{w}^c})}{\det(\mathbf{K})+\det(\sigma^2_{\text{noise}}\mathbf{J}_{\mathbf{w}^c})}\right) \leq 1 \tag{S75}
$$

Leading to the following inequality:

$$
(3) \leq \ln(\bar{C}_{13}) = \bar{C}_{14} \tag{S76}
$$

Finally, putting the three terms together, we obtain the following bound:

$$
\begin{aligned}
\text{PIF}_{\text{IterRCGP}}(y_m^c, \mathcal{D}, i) &\leq \bar{C}_3 + \bar{C}_{11} + \bar{C}_{12}(w(x_n, y_n^c)^2 y_n^c)^2 + \bar{C}_{14} && \text{(S77)}\\
&= C_2'(w(x_n, y_n^c)^2 y_n^c)^2 + C_3' && \text{(S78)}
\end{aligned}
$$

where $C_2' = \bar{C}_{12}$ and $C_3' = \bar{C}_3 + \bar{C}_{11} + \bar{C}_{14}$.

B Proof of Proposition 2

Unique solution of the empirical-risk minimization problem. We first show the existence of a unique solution to the empirical risk minimization problem corresponding to RCGP. For this purpose, we set $\mathbf{m} = \mathbf{0}$. Following Altamirano *et al.* (2024) (proof of [Proposition 3.1]), we can rewrite L_n^w and formulate the RCGP objective as the following empirical-risk minimization problem:

$$
\hat{\mathbf{f}} = \text{argmin}_{\mathbf{f}\in\mathcal{H}_k}\frac{1}{2n}\left(\underbrace{\mathbf{f}^\top\lambda^{-1}\mathbf{J}_\mathbf{w}^{-1}\mathbf{f} - 2\mathbf{f}^\top\lambda^{-1}\mathbf{J}_\mathbf{w}^{-1}(\mathbf{y}-\mathbf{m}_\mathbf{w}) + Q(\mathbf{x},\mathbf{y},\lambda)}_{L_n^w} + \|\mathbf{f}\|^2_{\mathcal{H}_k}\right) \tag{S79}
$$

where

$$
Q(\mathbf{x},\mathbf{y},\lambda) = \mathbf{y}^\top\lambda^{-1}\text{diag}(2\lambda^{-1}\mathbf{w}^2)\mathbf{y} - 4\lambda\nabla_y\mathbf{y}^\top\mathbf{w}^2 \tag{S80}
$$

for $\lambda > 0$. We then show the unique solution to S79 through the following lemma:

Lemma 6. *If $\lambda > 0$ and the kernel k is invertible, the solution to S79 is a unique, and is given by*

$$
\hat{f}(\mathbf{x}) = \mathbf{k}_\mathbf{x}(\mathbf{K}+\lambda\mathbf{J}_\mathbf{w})^{-1}(\mathbf{y}-\mathbf{m}_\mathbf{w}) = \sum_{j=1}^n \alpha_j k(\mathbf{x},\mathbf{x}_j), \mathbf{x}\in\mathcal{X} \tag{S81}
$$

where

$$(\alpha_i, \dots, \alpha_n) = (\mathbf{K} + \lambda \mathbf{J_w})^{-1}(\mathbf{y} - \mathbf{m_w}) \in \mathbb{R}^n \tag{S82}$$

Proof:

The optimization problem in S79 allows us to apply the representer theorem (Schölkopf *et al.*, 2001). It implies that the solution of S79 can be written as a weighted sum, i.e.,

$$\hat{\mathbf{f}} = \sum_{j=1}^{n} \alpha_j k(., \mathbf{x}_j) \tag{S83}$$

for $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. Let $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_n]^\top \in \mathbb{R}^n$. Substituting S83 into S79 provides

$$\text{argmin}_{\alpha \in \mathbb{R}^n} \frac{1}{2n}(\lambda^{-1}\boldsymbol{\alpha}^\top \mathbf{K}\mathbf{J}_\mathbf{w}^{-1}\mathbf{K}\boldsymbol{\alpha} - 2\lambda^{-1}\boldsymbol{\alpha}^\top \mathbf{K}\mathbf{J}_\mathbf{w}^{-1}(\mathbf{y} - \mathbf{m_w}) + Q(\mathbf{x}, \mathbf{y}, \lambda) + \|\hat{\mathbf{f}}\|^2_{\mathcal{H}_k}) \tag{S84}$$

where $\|\hat{\mathbf{f}}\|^2_{\mathcal{H}_k} = \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha}$, following the reproducing property. Taking the differentiation of the objective w.r.t. $\boldsymbol{\alpha}$, setting it equal to zero, and arranging the result yields the following equation:

$$\mathbf{K}(\mathbf{K} + \lambda \mathbf{J_w})\boldsymbol{\alpha} = \mathbf{K}(\mathbf{y} - \mathbf{m_w}) \tag{S85}$$

Since the objective in S84 is a convex function of $\boldsymbol{\alpha}$, we find that $\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{J_w})^{-1}(\mathbf{y} - \mathbf{m_w})$ provides the minimum of the objective (S79 and S84). Furthermore, we can verify that L_n^w is a convex function w.r.t. $\mathbf{f}$. Therefore, we conclude that $\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{J_w})^{-1}(\mathbf{y} - \mathbf{m_w})$ provides the unique solution to S79. As a remark, Proposition 6 closely connects with [Theorem 3.4] in Kanagawa *et al.* (2018).

Relative bound errors. We also provide the equivalence of Proposition 2 in Wenger *et al.* (2022):

Proposition 7. *For any choice of actions a relative bound error $\hat{\rho}(i)$ s.t. $\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_i\|_{\tilde{\mathbf{K}}} \leq \hat{\rho}(i)\|\hat{\mathbf{v}}\|_{\tilde{\mathbf{K}}}$ is given by*

$$\hat{\rho}(i) = (\bar{\mathbf{v}}^\top(\mathbf{I} - \tilde{\mathbf{C}}_i\tilde{\mathbf{K}})\bar{\mathbf{v}})^{1/2} \leq \lambda_{\max}(\mathbf{I} - \tilde{\mathbf{C}}_i\tilde{\mathbf{K}}) \leq 1 \tag{S86}$$

where $\bar{\mathbf{v}} = \hat{\mathbf{v}}/\|\tilde{\mathbf{v}}\|_{\tilde{\mathbf{K}}}$.

The proof is direct since we only need to substitute $\mathbf{C}_i, \hat{\mathbf{K}}, \mathbf{v}_*$ in Wenger *et al.* (2022) with $\tilde{\mathbf{C}}_i, \tilde{\mathbf{K}}, \hat{\mathbf{v}}$, respectively.

Proof of Proposition 2. Lemma 6 implies there exists a unique solution to the corresponding RCGP risk minimization problem. Choosing $\hat{\rho}(i)$ as described in Proposition 7, we have that $\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_i\|^2_{\tilde{\mathbf{K}}} \leq \hat{\rho}(i)\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_0\|_{\tilde{\mathbf{K}}}$, where $\tilde{\mathbf{v}}_0 = \mathbf{0}$. Then, for $i \in \{0, \dots, n\}$ we find that

$$\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_i\|^2_{\mathbf{K}} \leq \|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_i\|^2_{\tilde{\mathbf{K}}} \leq \hat{\rho}^2(i)\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_0\|^2_{\tilde{\mathbf{K}}} \tag{S87}$$

$$\leq \hat{\rho}(i)^2 \left(\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_0\|^2_{\mathbf{K}} + \frac{\lambda_{\max}(\mathbf{J_w})}{\lambda_{\min}(\mathbf{K})}\lambda_{\min}(\mathbf{K})\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_0\|^2_2\right) \tag{S88}$$

$$\leq \hat{\rho}(i)^2 \left(\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_0\|^2_{\mathbf{K}} + \frac{\lambda_{\max}(\mathbf{J_w})}{\lambda_{\min}(\mathbf{K})}\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_0\|^2_{\mathbf{K}}\right) \tag{S89}$$

$$\leq \hat{\rho}(i)^2 \left(1 + \frac{\lambda_{\max}(\mathbf{J_w})}{\lambda_{\min}(\mathbf{K})}\right)\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_0\|^2_{\mathbf{K}} \tag{S90}$$

The third inequality stems from the definition of $\mathbf{J_w}$ and the fact that the maximum eigenvalue of a diagonal matrice is the largest component of its diagonal. Applying result from Wenger *et al.* (2022), we have that

$$\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_i\|^2_{\mathbf{K}} = \|\hat{\boldsymbol{\mu}}_* - \hat{\boldsymbol{\mu}}_i\|^2_{\mathcal{H}_k} \tag{S91}$$

Combining both results and defining $c(\mathbf{J_w}) = \left(1 + \frac{\lambda_{\max}(\mathbf{J_w})}{\lambda_{\min}(\mathbf{K})}\right)$, we obtain

$$\|\hat{\boldsymbol{\mu}}_* - \hat{\boldsymbol{\mu}}_i\|_{\mathcal{H}_k} = \|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_i\|_{\mathbf{K}} \leq \hat{\rho}(i)c(\mathbf{J_w})\|\hat{\mathbf{v}} - \tilde{\mathbf{v}}_0\|_{\mathbf{K}} = \hat{\rho}(i)c(\mathbf{J_w})\|\hat{\boldsymbol{\mu}}_* - \mathbf{m}\|_{\mathcal{H}_k} \tag{S92}$$

C Proof of Proposition 3

Here, we refer to σ^2_{noise} as σ^2 to simplify the notation. Let $c_j = (\tilde{\mathbf{C}}_i k^{\sigma w}(\mathbf{X}, \mathbf{x}))_j$ for $j = 1, \ldots, n$, where we define $k^{\sigma w}(.,.) = k(.,.) + \frac{\sigma^2}{2}\delta_w(.,.)$, where

$$\delta_w(\mathbf{x}, \mathbf{x}') = \left\{ \begin{array}{cc} w^{-2}(\mathbf{x}, y) & \mathbf{x} = \mathbf{x}' \text{ and } \mathbf{x} \in \mathcal{D} \\ 2 & \mathbf{x} = \mathbf{x}' \text{ and } \mathbf{x} \notin \mathcal{D} \\ 0 & \mathbf{x} \neq \mathbf{x}' \end{array} \right. \tag{S93}$$

Since $g, m \in \mathcal{H}_{k^{\sigma w}}$, it implies that $g - m \in \mathcal{H}_{k^{\sigma w}}$. Then, applying [Lemma 3.9] in Kanagawa *et al.* (2018) provides

$$\left(\sup_{\|g-m_w\|_{\mathcal{H}_{k^{\sigma w}}} \leq 1} g(\mathbf{x}) - \hat{\mu}^g_i(\mathbf{x})\right)^2 = \left(\sup_{\|g-m_w\|_{\mathcal{H}_{k^{\sigma w}}} \leq 1} g(\mathbf{x}) - \sum_{j=1}^n c_j(g(\mathbf{x}_j) - m_w(\mathbf{x}_j))\right)^2 \tag{S94}$$

$$= \|k^{\sigma w}(., \mathbf{x}) - k(\mathbf{x}, \mathbf{X})\tilde{\mathbf{C}}_i k^{\sigma w}(\mathbf{X}, .)\|^2_{\mathcal{H}_{k^{\sigma w}}} \tag{S95}$$

$$= \langle k^{\sigma w}(., \mathbf{x}), k^{\sigma w}(., \mathbf{x})\rangle_{\mathcal{H}_k^{\sigma w}} - 2\langle k^{\sigma w}(., \mathbf{x}), k(\mathbf{x}, \mathbf{X})\tilde{\mathbf{C}}_i k^{\sigma w}(\mathbf{X}, .)\rangle_{\mathcal{H}_k^{\sigma w}} + \langle k(\mathbf{x}, \mathbf{X})\tilde{\mathbf{C}}_i k^{\sigma w}(\mathbf{X}, .), k(\mathbf{x}, \mathbf{X})\tilde{\mathbf{C}}_i k^{\sigma w}(\mathbf{X}, .)\rangle_{\mathcal{H}_k^{\sigma w}} \tag{S96}$$

By reproducing property, we have

$$= k^{\sigma w}(\mathbf{x}, \mathbf{x}) - 2k^{\sigma w}(\mathbf{x}, \mathbf{X})\tilde{\mathbf{C}}_i k^{\sigma w}(\mathbf{X}, \mathbf{x}) + k(\mathbf{x}, \mathbf{X})\tilde{\mathbf{C}}_i k^{\sigma w}(\mathbf{X}, \mathbf{X})\tilde{\mathbf{C}}_i k^{\sigma w}(\mathbf{X}, \mathbf{x}) \tag{S97}$$

if $\mathbf{x} \neq \mathbf{x}_j$ or $\sigma^2 = 0$, it holds that $k^{\sigma w}(\mathbf{x}, \mathbf{X}) = k(\mathbf{x}, \mathbf{X})$. By definition, we have $k^{\sigma w}(\mathbf{X}, \mathbf{X}) = \tilde{\mathbf{K}}$ and by Wenger *et al.* (2022)[Eq. (S42)], it holds that $\tilde{\mathbf{C}}_i\tilde{\mathbf{K}}\tilde{\mathbf{C}}_i = \tilde{\mathbf{C}}_i$. Therefore, we obtain

$$= k(\mathbf{x}, \mathbf{x}) + \sigma^2 - 2k(\mathbf{x}, \mathbf{X})\tilde{\mathbf{C}}_i k(\mathbf{X}, \mathbf{x}) + k(\mathbf{x}, \mathbf{X})\tilde{\mathbf{C}}_i\tilde{\mathbf{K}}\tilde{\mathbf{C}}_i k(\mathbf{X}, \mathbf{x}) \tag{S98}$$

$$= k(\mathbf{x}, \mathbf{x}) + \sigma^2 - k(\mathbf{x}, \mathbf{X})\tilde{\mathbf{C}}_i k(\mathbf{X}, \mathbf{x}) \tag{S99}$$

$$= \hat{k}_i(\mathbf{x}, \mathbf{x}) + \sigma^2 \tag{S100}$$

For the last result, we analogously choose $c_j = ((\tilde{\mathbf{K}}^{-1} - \tilde{\mathbf{C}}_i)k^{\sigma w}(\mathbf{X}, \mathbf{x}))_j$. Then, we obtain

$$\left(\sup_{\|g-m_w\|_{\mathcal{H}_{k^{\sigma w}}}\leq 1}\hat{\mu}^g(\mathbf{x})-\hat{\mu}_i^g(\mathbf{x})\right)^2=\left(\sup_{\|g-m_w\|_{\mathcal{H}_{k^{\sigma w}}}\leq 1}\sum_{j=0}^{n}c_j g(\mathbf{x}_j)\right)^2 \tag{S101}$$

$$= \|k(\mathbf{x},\mathbf{X})(\tilde{\mathbf{K}}^{-1}-\tilde{\mathbf{C}}_i)k^{\sigma w}(\mathbf{X},.)\|^2_{\mathcal{H}_{k^{\sigma w}}} \tag{S102}$$

$$= k^{\sigma w}(\mathbf{x},\mathbf{X})\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{K}}\tilde{\mathbf{K}}^{-1}k^{\sigma w}(\mathbf{X},\mathbf{x})-2k^{\sigma w}(\mathbf{x},\mathbf{X})\tilde{\mathbf{K}}^{-1}\tilde{\mathbf{K}}\tilde{\mathbf{C}}_i k^{\sigma w}(\mathbf{X},\mathbf{x})+ k^{\sigma w}(\mathbf{x},\mathbf{X})\tilde{\mathbf{C}}_i\tilde{\mathbf{K}}\tilde{\mathbf{C}}_i k^{\sigma w}(\mathbf{X},\mathbf{x}) \tag{S103}$$

$$= k(\mathbf{x},\mathbf{X})(\tilde{\mathbf{K}}^{-1}-\tilde{\mathbf{C}}_i)k(\mathbf{X},\mathbf{x}) \tag{S104}$$

$$= k_i^{\text{comp.}}(\mathbf{x},\mathbf{x}) \tag{S105}$$