Prediction Performance After Learning in Gaussian Process Regression

Johan Wågberg **Dave Zachariah** **Thomas B. Schön** **Petre Stoica**
Department of Information Technology, Uppsala University, Sweden

Abstract

This paper considers the quantification of the prediction performance in Gaussian process regression. The standard approach is to base the prediction error bars on the theoretical predictive variance, which is a lower bound on the mean square-error (MSE). This approach, however, does not take into account that the statistical model is learned from the data. We show that this omission leads to a systematic underestimation of the prediction errors. Starting from a generalization of the Cramér-Rao bound, we derive a more accurate MSE bound which provides a measure of uncertainty for prediction of Gaussian processes. The improved bound is easily computed and we illustrate it using synthetic and real data examples.

1 Introduction

In this paper we consider the problem of learning a function $f(\mathbf{x})$ from a dataset $\mathcal{D}_N = \{\mathbf{x}_i, y_i\}_{i=1}^N$ where

$$y = f(\mathbf{x}) + \varepsilon \ \in \ \mathbb{R}. \tag{1}$$

The aim is to predict $f(\mathbf{x}_\star)$ at a test point $\mathbf{x}_\star$. In machine learning, spatial statistics and statistical signal processing, it is common to model $f(\mathbf{x})$ as a Gaussian process (GP) and ε as an uncorrelated zero-mean Gaussian noise (Bishop 2006; Murphy 2012; Pérez-Cruz et al. 2013; Stein 1999). This probabilistic framework shares several properties with kernel and spline-based methods (C. Rasmussen and C. Williams 2006; Schölkopf and Smola 2002; Suykens et al. 2002). One of the strengths of the GP framework is that both a predictor $\widehat{f}(\mathbf{x}_\star)$ and its error bars are readily obtained using the mean and variance of $f(\mathbf{x}_\star)$. This quantification of the prediction uncertainty is valuable in itself but also in applications that involve decision making, e.g. in the exploration-exploitation phase of active learning and control (Deisenroth, Fox, et al. 2015; Deisenroth and C. Rasmussen 2011; Likar and Kocijan 2007). Another recent example is Bayesian optimization techniques using Gaussian processes (Shahriari et al. 2016).

In general, however, the model for $f(\mathbf{x})$ is not fully specified but contains unknown hyperparameters, denoted $\boldsymbol{\theta}$, that can be learned from data. Plugging an estimate $\widehat{\boldsymbol{\theta}}$ into the predictor $\widehat{f}(\mathbf{x}_\star)$ will therefore inflate its errors due to the uncertainty of the learned model itself. In this case the standard error bounds will systematically underestimate the actual prediction errors. One possibility is to assign a prior distribution to $\boldsymbol{\theta}$ and marginalize out the parameters from the posterior distribution of $\check{f}_\star$ (C. K. I. Williams and C. E. Rasmussen 1996). While conceptually straightforward, this approach is challenging to implement in general as it requires the user to choose a reasonable prior distribution and computationally demanding numerical integration techniques.

Our contribution in this paper is the derivation of more accurate error bound for prediction after learning, using a generalization of the Cramér-Rao Bound (CRB) (Cramér 1946; Rao 1945). The bound is computionally inexpensive to implement using standard tools in the GP framework. We illustrate the bound using both synthetic and real data.

2 Problem formulation and related work

We consider a general input space $\mathbf{x} \in \mathcal{X}$. To establish the notation ahead we write the Gaussian process and

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

the vector of hyperparameters as

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m_\alpha(\mathbf{x}), k_\beta(\mathbf{x},\mathbf{x}')\right) \quad \text{and} \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \\ \sigma^2 \end{bmatrix}, \quad (2)$$

where σ^2 denotes the variance of ε in (1). The vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameterize the mean and covariance functions, $m_\alpha(\mathbf{x})$ and $k_\beta(\mathbf{x},\mathbf{x}')$, respectively. For an arbitrary test point $\mathbf{x}_\star$ we write $f_\star = f(\mathbf{x}_\star)$ and consider the mean-square error

$$\text{MSE}\left(\widehat{f}_\star\right) \triangleq \text{E}\left[|f_\star - \widehat{f}_\star|^2\right],$$

where the expectation is taken with respect to $f_\star$ and the data $\mathbf{y}$. When $\boldsymbol{\theta}$ is given, the optimal predictor is

$$\check{f}_\star(\boldsymbol{\theta}) = m_\star + \mathbf{w}^\mathsf{T}(\mathbf{y} - \mathbf{m}), \quad (3)$$

where $\mathbf{w} = \left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_\star$, $m_\star = m_\alpha(\mathbf{x}_\star)$ and $\mathbf{m} = [m(\mathbf{x}_1) \cdots m(\mathbf{x}_N)]^\mathsf{T}$. In addition, $\mathbf{k}_\star = [k_\beta(\mathbf{x}_\star,\mathbf{x}_1) \cdots k_\beta(\mathbf{x}_\star,\mathbf{x}_N)]^\mathsf{T}$ and $\mathbf{K} = \{k_\beta(\mathbf{x}_i,\mathbf{x}_j)\}_{i,j}$. Eq. (3) is equal to the mean of the predictive distribution $p(f_\star|\mathbf{y},\boldsymbol{\theta})$ and is a function of both $\mathbf{y}$ and $\boldsymbol{\theta}$ (C. Rasmussen and C. Williams 2006). The minimum MSE then follows directly from the predictive variance, denoted $\sigma^2_{\star|y}(\boldsymbol{\theta})$. Here, however, we provide an alternative derivation based on a generalization of the CRB (Gill and Levit 1995; Van Trees and Bell 2013 [1968]; Zachariah and Stoica 2015). This tool will also enable us to tackle the general problem considered later on.

Result 1. *When $\boldsymbol{\theta}$ is known,*

$$\text{MSE}\left(\widehat{f}_\star\right) \geq \underbrace{k_{\star\star} - \mathbf{k}_\star^\mathsf{T}\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_\star}_{=\sigma^2_{\star|y}(\boldsymbol{\theta})}, \quad (4)$$

where $k_{\star\star} = k_\beta(\mathbf{x}_\star,\mathbf{x}_\star)$.

Proof. The Bayesian Cramér-Rao Bound (BCRB) is given by

$$\text{MSE}\left(\widehat{f}_\star\right) \geq J_\star^{-1},$$

where $J_\star \triangleq \text{E}\left[\left(\frac{\partial}{\partial f_\star}\ln p(\mathbf{y}, f_\star\,|\,\boldsymbol{\theta})\right)^2\right]$ is the Bayesian information of $f_\star$ (Van Trees and Bell 2013 [1968]). Using the chain rule, $\ln p(\mathbf{y}, f_\star\,|\,\boldsymbol{\theta}) = \ln p(f_\star\,|\,\mathbf{y},\boldsymbol{\theta}) + \ln p(\mathbf{y}\,|\,\boldsymbol{\theta})$, we obtain

$$\begin{aligned}\frac{\partial}{\partial f_\star}\ln p(\mathbf{y}, f_\star|\boldsymbol{\theta}) &= \frac{\partial}{\partial f_\star}\ln p(f_\star|\mathbf{y},\boldsymbol{\theta}) + 0 \\ &= \frac{\partial}{\partial f_\star}\left(-\frac{1}{2}\ln(2\pi\sigma^2_{\star|y}) - \frac{1}{2\sigma^2_{\star|y}}(f_\star - \check{f}_\star(\boldsymbol{\theta}))^2\right) \quad (5) \\ &= -\sigma^{-2}_{\star|y}\left(f_\star - \check{f}_\star(\boldsymbol{\theta})\right),\end{aligned}$$

under the assumptions made. Then the Bayesian information equals

$$J_\star = \text{E}\left[\sigma^{-4}_{\star|y}\left(f_\star - \check{f}_\star(\boldsymbol{\theta})\right)^2\right] = \frac{1}{\sigma^2_{\star|y}}. \quad (6)$$

□

Remarks: The lower bound (4) on the MSE, and the corresponding minimum error bars for a predictor $\widehat{f}_\star$, reflects the uncertainty of $f_\star$ alone. The bound is attained when $\widehat{f}_\star$ coincides with (3) which depends on $\boldsymbol{\theta}$. In general, however, $\boldsymbol{\theta}$ is unknown and typically learned from the data. Then the bound (4) will not reflect the additional errors of $\widehat{f}_\star$ arising from the unknown model parameters $\boldsymbol{\theta}$. The effect is a systematic underestimation of the prediction errors. For illustrative purposes we present an example with one-dimensional inputs, see Example 1 below.

Example 1. *Consider the Gaussian process (1) for $x \in \mathbb{R}$ with a linear mean function and a squared-exponential covariance function. That is, $m_\alpha(x) = \alpha x$ and $k_\beta(x,x') = \beta_0^2\exp\left(-\frac{1}{2\beta_1^2}\|x - x'\|^2\right)$ in (2). The process is sampled at $N = 10$ different points and the unknown hyperparameters are learned from the dataset by maximizing the marginal likelihood, $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \int p(\mathbf{y},\mathbf{f}|\boldsymbol{\theta})d\mathbf{f}$, where the vector $\mathbf{f}$ contains the N latent function values in the data.*

Figure 1 illustrates a realization of $f(x)$ along with the predicted values $\widehat{f}(x)$. The error bars are obtained from (4) which was derived under the assumption of $\boldsymbol{\theta}$ being known. The bars severely underestimate the uncertainty of the predictor since they remain nearly constant along the input space and do not contain the example realization of $f(x)$.

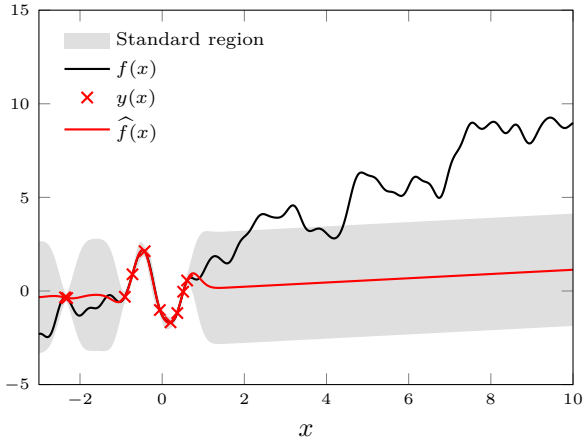

Figure 1: Predictions of $f(x)$ using hyperparameters that have been learned from data. The shaded error bars are credibility regions corresponding to $\widehat{f}(x) \pm 3\sigma_{\star|y}$.

A tighter MSE bound than (4) has been derived for the special case in which $\boldsymbol{\beta}$ is assumed to be known and the mean function is linear in the parameters, i.e., $m_\alpha(\mathbf{x}) = \boldsymbol{\alpha}^\mathsf{T}\mathbf{u}(\mathbf{x})$ where $\mathbf{u}(\mathbf{x})$ is a given basis function (Stein 1999). In the statistics literature, there have been attempts to extend to the analysis to models in which the covariance parameters $\boldsymbol{\beta}$ are unknown. The results do, however, not generalize to nonlinear $m_\alpha(\mathbf{x})$ and are either based on computationally demanding Taylor-series expansions or bootstrap techniques (Den Hertog et al. 2006; Zimmerman and Cressie 1992).

The goal of this paper is to derive a computationally inexpensive lower bound on the MSE that will provide more accurate error bars on $\widehat{f}_\star$ when $\boldsymbol{\theta}$ is unknown.

3 Prediction errors after learning the hyperparameters

In the general setting when $\boldsymbol{\theta}$ is unknown we have the following lower bound on the MSE.

Result 2. *When $\boldsymbol{\theta}$ is learned from $\mathbf{y}$ using an unbiased estimator, we have that:*

$$\boxed{\mathit{MSE}(\widehat{f}_\star) \geq \sigma^2_{\star|y} + \mathbf{g}^\mathsf{T}\mathbf{M}^{-1}\mathbf{g},} \tag{7}$$

where

$$\begin{aligned} \mathbf{g} &= \frac{\partial}{\partial \boldsymbol{\alpha}}(m_\star - \mathbf{m}^\mathsf{T}\mathbf{w}) \quad \textit{and} \\ \mathbf{M} &= \frac{\partial \mathbf{m}^\mathsf{T}}{\partial \boldsymbol{\alpha}}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\frac{\partial \mathbf{m}}{\partial \boldsymbol{\alpha}^\mathsf{T}}. \end{aligned} \tag{8}$$

Comparing with (4)*, the nonnegative term $\mathbf{g}^\mathsf{T}\mathbf{M}^{-1}\mathbf{g} \geq 0$ is the additional error incurred due to the lack of information about $\boldsymbol{\theta}$.*

Remarks: Eq. (7) is the Hybrid Cramér-Rao Bound which we abbreviate as $\mathrm{HCRB}(\boldsymbol{\theta}) \triangleq \sigma^2_{\star|y} + \mathbf{g}^\mathsf{T}\mathbf{M}^{-1}\mathbf{g}$ (Rockah and Schultheiss 1987; Van Trees and Bell 2013 [1968]).

First, note that $\mathbf{g}^\mathsf{T}\mathbf{M}^{-1}\mathbf{g}$ will be non-zero even in the simplest models where the data has an unknown constant mean, i.e., $m_\alpha(\mathbf{x}) \equiv \alpha$.

Second, eq. (7) depends on the unknown covariance parameters $\boldsymbol{\beta}$ only via $\mathbf{M}$ and not through any gradients as would be expected. As we show in the proofs below, this is follows from the properties of the Gaussian data distribution. In the special case of linear mean functions, $m_\alpha(\mathbf{x}) = \boldsymbol{\alpha}^\mathsf{T}\mathbf{u}(\mathbf{x})$, (7) coincides with MSE of the universal kriging estimator which assumes $\boldsymbol{\beta}$ to be known (Stein 1999).

Third, under standard regularity conditions, the maximum likelihood approach will yield estimates of $\boldsymbol{\theta}$ that are asymptotically unbiased and attain their corresponding error bounds (Van Trees and Bell 2013 [1968]).

Proof. The HCRB for $f_\star$ is given by

$$\mathrm{MSE}\left(\widehat{f}_\star\right) \geq (J_\star - \mathbf{J}_{\theta,\star}^\mathsf{T}\mathbf{J}_\theta^{-1}\mathbf{J}_{\theta,\star})^{-1}, \tag{9}$$

where the matrices are given by the hybrid information matrix

$$\begin{aligned} \mathbf{J} &\triangleq \mathrm{E}\left[\begin{bmatrix} \frac{\partial \ln p(\mathbf{y}, f_\star|\boldsymbol{\theta})}{\partial f_\star} \\ \frac{\partial \ln p(\mathbf{y}, f_\star|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \end{bmatrix} \begin{bmatrix} \frac{\partial \ln p(\mathbf{y}, f_\star|\boldsymbol{\theta})}{\partial f_\star} \\ \frac{\partial \ln p(\mathbf{y}, f_\star|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \end{bmatrix}^\mathsf{T}\right] \\ &= \begin{bmatrix} J_\star & \mathbf{J}_{\theta,\star}^\mathsf{T} \\ \mathbf{J}_{\theta,\star} & \mathbf{J}_\theta \end{bmatrix}. \end{aligned} \tag{10}$$

To prepare for the subsequent steps, we introduce $\mathbf{u} = [\mathbf{y}^\mathsf{T}\ f_\star]^\mathsf{T}$ and let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the joint mean and covariance matrix respectively, i.e. $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Next, we define the linear combiner

$$\widetilde{\mathbf{w}}^\mathsf{T} = \begin{bmatrix}\mathbf{0} & 1\end{bmatrix} - \mathbf{w}^\mathsf{T}\begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix}$$

and note that

$$\widetilde{\mathbf{w}}^\mathsf{T}(\mathbf{u} - \boldsymbol{\mu}) = \widetilde{\mathbf{w}}^\mathsf{T}\left(\begin{bmatrix}\mathbf{y} \\ f_\star\end{bmatrix} - \begin{bmatrix}\mathbf{m} \\ m_\star\end{bmatrix}\right) = f_\star - \widehat{f}_\star. \tag{11}$$

Similarly, $\widetilde{\mathbf{w}}^\mathsf{T}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\alpha}^\mathsf{T}} = \frac{\partial}{\partial \boldsymbol{\alpha}^\mathsf{T}}\widetilde{\mathbf{w}}^\mathsf{T}\boldsymbol{\mu} = \mathbf{g}^\mathsf{T}$.

To compute the block $\mathbf{J}_{\theta,\star}$ in (10), we first establish the following derivatives:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\alpha}}\ln p(\mathbf{u}|\boldsymbol{\theta}) &= \frac{\partial \boldsymbol{\mu}^\mathsf{T}}{\partial \boldsymbol{\alpha}}\boldsymbol{\Sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu}), \\ \frac{\partial}{\partial \beta_i}\ln p(\mathbf{u}|\boldsymbol{\theta}) &= -\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \beta_i}\right\} + \\ &\quad + \frac{1}{2}(\mathbf{u} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \beta_i}\boldsymbol{\Sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu}), \\ \frac{\partial}{\partial \sigma^2}\ln p(\mathbf{u}|\boldsymbol{\theta}) &= -\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma^2}\right\} + \\ &\quad + \frac{1}{2}(\mathbf{u} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma^2}\boldsymbol{\Sigma}^{-1}(\mathbf{u} - \boldsymbol{\mu}) \\ \frac{\partial}{\partial f_\star}\ln p(\mathbf{u}|\boldsymbol{\theta}) &= -\sigma^{-2}_{\star|y}\widetilde{\mathbf{w}}^\mathsf{T}(\mathbf{u} - \boldsymbol{\mu}), \end{aligned}$$

where the last equality follows from (5) and (11). Then we obtain

$$\begin{aligned} &\mathrm{E}\left[\frac{\partial}{\partial f_\star}\ln p(\mathbf{u}|\boldsymbol{\theta})\frac{\partial}{\partial \boldsymbol{\alpha}^\mathsf{T}}\ln p(\mathbf{u}|\boldsymbol{\theta})\right] = \\ &= -\sigma^{-2}_{\star|y}\widetilde{\mathbf{w}}^\mathsf{T}\,\mathrm{E}\left[(\mathbf{u} - \boldsymbol{\mu})(\mathbf{u} - \boldsymbol{\mu})^\mathsf{T}\right]\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\alpha}^\mathsf{T}} \\ &= -\sigma^{-2}_{\star|y}\mathbf{g}^\mathsf{T} \end{aligned}$$

$$\mathrm{E}\left[\frac{\partial}{\partial f_\star}\ln p(\mathbf{u}|\boldsymbol{\theta})\frac{\partial}{\partial \beta_i}\ln p(\mathbf{u}|\boldsymbol{\theta})\right] =$$
$$= \frac{1}{2}\sigma_{\star|y}^{-2}\widetilde{\mathbf{w}}^\mathsf{T}\underbrace{\mathrm{E}[(\mathbf{u}-\boldsymbol{\mu})]}_{=\mathbf{0}}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\beta_i}\right\}$$
$$- \frac{1}{2}\sigma_{\star|y}^{-2}\widetilde{\mathbf{w}}^\mathsf{T}\underbrace{\mathrm{E}\left[(\mathbf{u}-\boldsymbol{\mu})(\mathbf{u}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\beta_i}\boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})\right]}_{=\mathbf{0}}$$
$$= 0.$$

Similarly, $\mathrm{E}\left[\frac{\partial}{\partial f_\star}\ln p(\mathbf{u}|\boldsymbol{\theta})\frac{\partial}{\partial\sigma^2}\ln p(\mathbf{u}|\boldsymbol{\theta})\right] = 0$. Therefore

$$\mathbf{J}_{\theta,\star}^\mathsf{T} = \begin{bmatrix}-\sigma_{\star|y}^{-2}\mathbf{g}^\mathsf{T} & \mathbf{0} & 0\end{bmatrix}. \tag{12}$$

Next, using the distribution of $\mathbf{u}$, $\mathbf{J}_\theta$ is obtained via Slepian-Bangs formula (Bangs 1971; Slepian 1954; Stoica and Moses 2005):

$$\{\mathbf{J}_\theta\}_{i,j} = \frac{\partial\boldsymbol{\mu}^\mathsf{T}}{\partial\theta_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_j} + \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\theta_j}\right\}.$$

This yields a block-diagonal matrix

$$\mathbf{J}_\theta = \begin{bmatrix}\frac{\partial\boldsymbol{\mu}^\mathsf{T}}{\partial\boldsymbol{\alpha}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\alpha}^\mathsf{T}} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & * & *\\ \mathbf{0} & * & *\end{bmatrix}. \tag{13}$$

where the right-lower block does not affect (9) due to the zeros in (12). Inserting (12), (13) and (6) into (9) then yields

$$\mathrm{MSE}\left(\widehat{f}_\star\right) \geq \left(\sigma_{\star|y}^{-2} - \sigma_{\star|y}^{-2}\mathbf{g}^\mathsf{T}\left(\frac{\partial\boldsymbol{\mu}^\mathsf{T}}{\partial\boldsymbol{\alpha}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\alpha}^\mathsf{T}}\right)^{-1}\mathbf{g}\sigma_{\star|y}^{-2}\right)^{-1}$$
$$= \sigma_{\star|y}^2 + \mathbf{g}^\mathsf{T}\left(\frac{\partial\boldsymbol{\mu}^\mathsf{T}}{\partial\boldsymbol{\alpha}}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\alpha}^\mathsf{T}} - \sigma_{\star|y}^{-2}\mathbf{g}\mathbf{g}^\mathsf{T}\right)^{-1}\mathbf{g}, \tag{14}$$

where the last equality follows from the matrix inversion lemma. Using the properties of the block-inverse of $\boldsymbol{\Sigma}$, we show that the inner parenthesis equals $\frac{\partial\mathbf{m}^\mathsf{T}}{\partial\boldsymbol{\alpha}}(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\frac{\partial\mathbf{m}}{\partial\boldsymbol{\alpha}^\mathsf{T}}$ in Appendix B. □

Remark: Result 2 is based on the framework in Rockah and Schultheiss (1987), se also Van Trees and Bell (2013 [1968]). This assumes that the bias of the learning method $\widehat{\boldsymbol{\theta}}$ is zero. In Appendix A we provide an alternative proof of Result 2 that relaxes this assumption.

Example 1. *(cont'd) To illustrate the difference between (4) and (7), consider Figure 2. It shows the same realization of $f(x)$ as in Figure 1, along with the predicted values $\widehat{f}(x)$. The error bars are now obtained from (7) which takes into account that $\boldsymbol{\theta}$ has to be learned from the data. These bars clearly quantify the errors more accurately and contain the realization $f(x)$, in contrast to the standard approach.*

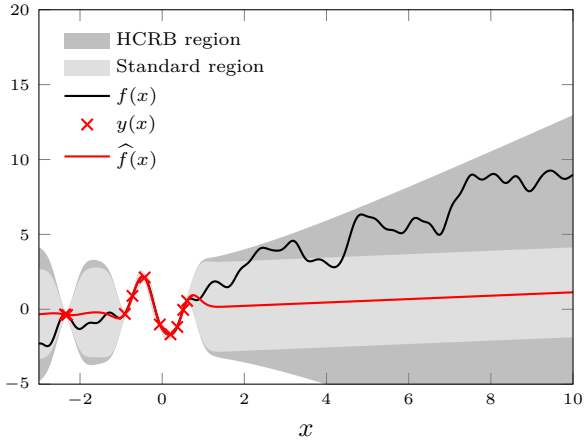

Figure 2: Predictions of $f(x)$ using hyperparameters that have been learned from data. The dark shaded error bars are regions corresponding to $\widehat{f}(x) \pm 3\sqrt{\mathrm{HCRB}}$.

Example 2. *(Time series prediction) Here we consider a temporal process with an unknown linear trend and periodicity (per) modeled by mean function $m_\alpha(x) = \alpha_1 + \alpha_2 x$, covariance kernel $k_\beta^{per}(x,x') = \beta_1^2\exp\left(-\frac{2}{\beta_2^2}\sin^2\left(\frac{\pi}{\beta_3}\|x-x'\|\right) + \frac{1}{\beta_4^2}\|x-x'\|^2\right)$ and unit noise level. In Figure 3 we show a single realization of the process together with the prediction error bars computed using both the predictive variance and the HCRB. As can be seen, $f(x)$ falls outside of the credibility region provided by the standard method.*

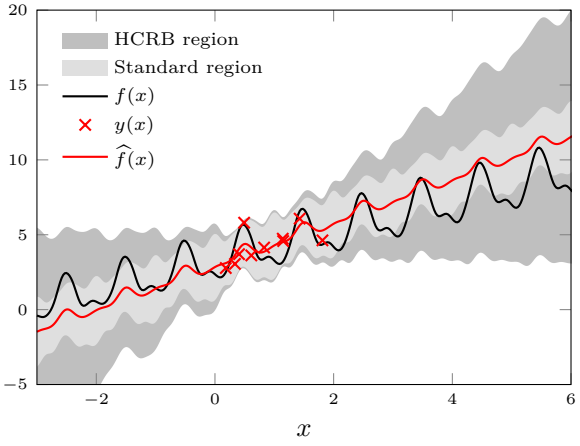

Figure 3: Predictions of $f(x)$ using hyperparameters that have been learned from data. The dark shaded error bars are regions corresponding to $\widehat{f}(x) \pm 3\sqrt{\mathrm{HCRB}}$.

4 Examples

We illustrate the HCRB and its practical utility by means of several examples. For the sake of visualization, we have considered problems with one-dimensional inputs, but the HCRB is of course valid for any dimension of $\mathcal{X}$. The first set of examples use synthetically generated datasets in order to assess the accuracy of the error bound. The final example uses real CO_2 concentration data. We used the maximum likelihood approach to learn $\boldsymbol{\theta}$ in all examples but alternative methods, such as cross-validation, could be considered as well.

4.1 Synthetic data

First, we consider a process $f(x)$ with the popular squared-exponential covariance (SE) function

$$k_{\beta}^{\mathrm{SE}}(x, x') = \beta_1^2 \exp\left(-\frac{1}{2\beta_2^2}\|x - x'\|^2\right), \tag{15}$$

where we assume both the signal variance β_1^2 and length scale β_2 to be unknown. As mean function, we assign the most basic model, a constant mean $m_\alpha(x) = \alpha$, where α is unknown. The unknown hyperparameters generating the data are denoted

$$\boldsymbol{\theta}_0 = \begin{bmatrix} \alpha_0 \\ \boldsymbol{\beta}_0 \\ \sigma_0^2 \end{bmatrix}, \text{ where } \begin{cases} \alpha_0 = 20, \\ \boldsymbol{\beta}_0 = \begin{bmatrix} 2 & 0.8 \end{bmatrix}^{\mathsf{T}}, \\ \sigma_0^2 = 2^2. \end{cases} \tag{16}$$

Figure 4 shows the empirical MSE of (3) after learning the hyperparameters from $N = 25$ observations (obtained from 10^3 Monte Carlo iterations), where $\widehat{f}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)$ denotes evaluating the predictor using mean parameter $\boldsymbol{\alpha}$, covariance parameter $\boldsymbol{\beta}$ and noise level σ^2. We compare this error with the theoretical bounds given by (4) and (7). We see that the bound $\mathrm{HCRB}(\boldsymbol{\theta}_0)$ is tight as expected in this example and that $\sigma^2_{\star|y}(\boldsymbol{\theta}_0)$ systematically underestimates the errors. Similarly, when using estimated bounds by inserting the learned hyperparameters $\widehat{\boldsymbol{\theta}}$ into (4) and (7) the gap between $\sigma^2_{\star|y}(\widehat{\boldsymbol{\theta}})$ and $\mathrm{HCRB}(\widehat{\boldsymbol{\theta}})$ not extreme, but still present, with the latter giving a better representation of the true error than the estimated predictive variance.

Another simple but common mean function is the linear mean $m_\alpha(x) = \alpha x$. Again, using a process with the squared exponential covariance function,(15), with hyperparameters $\boldsymbol{\beta}_0$ and noise level σ_0^2 as in (16) but with a linear mean function with $\alpha_0 = 2$, we evaluate the empirical MSE (obtained by 10^3 Monte Carlo samples) and the theoretical bounds in Figure 5. The

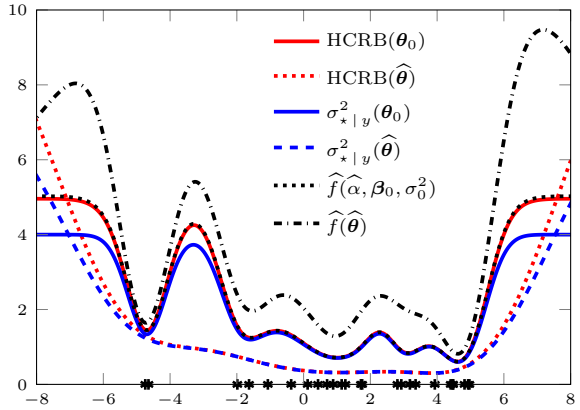

Figure 4: MSE of predictors along $x \in \mathcal{X}$ using learned hyperparameters and the bounds (4) and (7). The red curves show the true and estimated HCRB, based on $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}$, respectively. In blue, we show predictive variances $\sigma^2_{\star\,|\,y}$ corresponding to $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}$, respectively. The black dots indicate the input sample locations x. $f(x)$ has a constant mean function, $m_\alpha(x) = \alpha$ and a squared exponential covariance function (15) with hyperparameters as in (16).

hyperparameters were learned using $N = 25$ observations. The gap between the bounds become even more pronounced in predictions outside the sampled region.

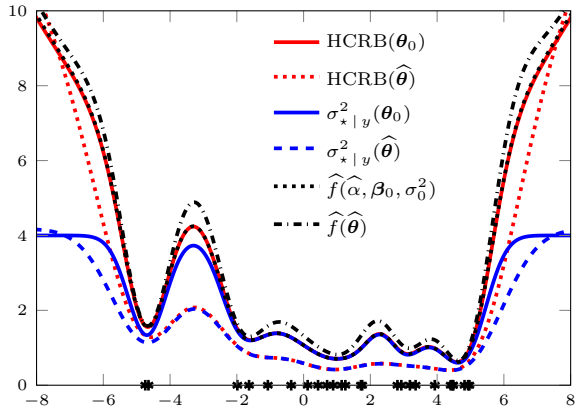

Figure 5: MSE of predictors using learned hyperparameters and the bounds (4) and (7). Here $f(x)$ has a linear mean function, $m_\alpha(x) = \alpha x$ and a squared exponential covariance function (15) with $\alpha_0 = 2$, and $\boldsymbol{\beta}_0$ and σ_0^2 as in (16).

Next, we consider an example inspired by frequency estimation in colored noise, which is a challenging problem. We model a process $f(x)$ using a sinusoid mean function

$$m_\alpha(x) = \alpha_1 \sin(\alpha_2 x + \alpha_3). \tag{17}$$

with unknown amplitude, frequency and phase, and a squared exponential covariance function (15). Here, we let $\boldsymbol{\alpha}_0 = \begin{bmatrix} 3 & 2 & \pi/4 \end{bmatrix}^{\mathsf{T}}$, $\boldsymbol{\beta}_0 = \begin{bmatrix} 0.5 & 3 \end{bmatrix}^{\mathsf{T}}$ and

$\sigma_0^2 = 0.5^2$. The process was sampled at 25 non-uniformly spaced input points, cf. Firuge 6. Note how the conditional variance $\sigma^2_{\star\,|\,y}$ severly underestimates the uncertainty in the predictions between -5 and -3, but how the HCRB, even when estimated from data, provides a much more accurate bound.

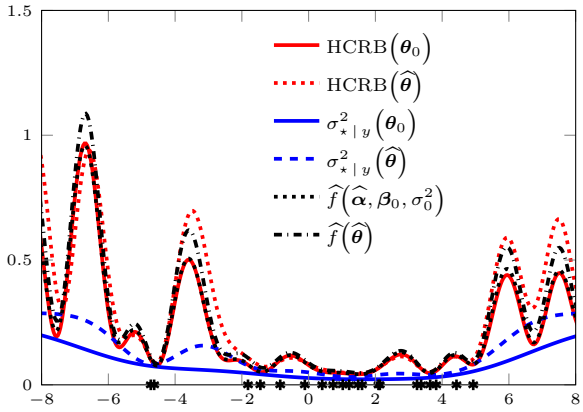

Figure 6: Empirical MSE and theoretical and estimated lower bound on MSE for a process $f(x)$ with a sinusoidal mean function $m_\alpha(x) = \alpha_1 \sin(\alpha_2 x + \alpha_3)$ and squared exponential covairance function (15), with $\boldsymbol{\alpha}_0 = \begin{bmatrix} 3 & 2 & \frac{\pi}{4} \end{bmatrix}^\mathsf{T}$, $\boldsymbol{\beta}_0 = \begin{bmatrix} 0.5^2 & 3 \end{bmatrix}^\mathsf{T}$ and $\sigma_0^2 = .5^2$.

4.2 Marginalizing the mean parameters

For the special case in which the mean function is linear in the parameters, that is, $m_\alpha(\mathbf{x}) = \boldsymbol{\alpha}^\mathsf{T}\mathbf{u}(\mathbf{x})$, it is possible to consider an alternative parameterization: A Gaussian hyperprior $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\widetilde{\boldsymbol{\beta}}))$ can be assigned with positive covariance parameters $\widetilde{\boldsymbol{\beta}}$. By marginalizing out $\boldsymbol{\alpha}$ from $f(\mathbf{x})$, we then obtain an additional term to the covariance function $k_{\widetilde{\beta}}(\mathbf{x}, \mathbf{x}') = \mathbf{u}^\mathsf{T}(\mathbf{x})\mathrm{diag}(\widetilde{\boldsymbol{\beta}})\mathbf{u}(\mathbf{x}')$ where $\widetilde{\boldsymbol{\beta}}$ is augmented to the hyperparameters. Correspondly, the mean function becomes $m(\mathbf{x}) \equiv \mathbf{0}$, which is a common assumption in the Gaussian process literature (C. Rasmussen and C. Williams 2006). This model parameterization will therefore have an alternative predictive variance $\sigma^2_{\star|y}$ that captures the uncertainty of the linear mean parameters.

To study the effect of this alternative parameterization on the bounds and prediction, we consider a linear trend $m_\alpha(x) = \alpha_1 + \alpha_2 x$ along with $k_\beta(x, x') = k^{\mathrm{SE}}_{\beta^{\mathrm{SE}}}(x, x')$ as given in (15). The data was generated according to this model and the HCRB is evaluated in Figure 7. The marginalized model is here $m(x) \equiv 0$ and $k^{\mathrm{SE}}_{\beta^{\mathrm{SE}}}(x, x') + k_{\beta^{\mathrm{aff}}}(x, x')$, where

$$k_{\beta^{\mathrm{aff}}}(x, x') = \beta_1 + \beta_2 x x'$$

corresponds to the unknown linear mean function. The predictive variance of this model was evaluated learning its hyperparameters using data from the original model and inserting them into $\sigma^2_{\star|y}$.

For the special case in which only $\boldsymbol{\beta}^{\mathrm{aff}}$ is learned, the correspondence between $\sigma^2_{\star|y}$ and HCRB is striking in Figure 7. In this case, also the the predictors, using the original and marginalized models, respectively, perform nearly identically. When all hyperparameters are learned in the original and marginal models, respectively, the empirical $\sigma^2_{\star|y}$ turns out to be more accurate than the empirical HCRB in the extremes. The results suggest that for linear mean functions there is a potential advantage in using the marginalized model to assess the prediction accuracy. However, in this case we also note that the performance of the predictor based on the marginalized model is degraded in comparison to that based on the original model.

4.3 CO_2 concentration data

With the previous examples in mind, we now consider real CO_2 concentration data[1] analyzed in C. Rasmussen and C. Williams (2006). The data exhibits a trend as well as periodicities. These features can be modeled using the mean and covariance functions considered in the previous example. In addition, to capture smooth variations as well as erratic patterns, we consider using a squared-exponential kernel $k^{\mathrm{SE}}(x, x')$ and a rational quadratic (RQ) kernel $k^{\mathrm{RQ}}_\beta(x, x') = \beta_1^2 \left(1 + \frac{1}{2\beta_2\beta_3^2}\|x - x'\|^2\right)^{-\beta_3}$. The final covariance function can be written as:

$$k_\beta(x, x') = k^{\mathrm{SE}}_\beta(x, x') + k^{\mathrm{per}}_\beta(x, x') + k^{\mathrm{RQ}}_\beta(x, x').$$

In this example, the hyperparameters are learned using monthly data from the years 1995 to 2003. The prediction error bars using the predictive variance and HCRB are plotted in Figure 8. Using validation data from 2004 to March 2016 we assess the error bars. As can be seen several data points fall outside of standard approach fall outside of the 99.7% credibility region but are contained in the HCRB region.

5 Discussion

We used the Hybrid Cramér-Rao Bound as a tool to analyze the prediction performance of Gaussian process regression after learning. When comparing the new bound with the commonly used predictive variance we showed that the latter will systematically underestimate the minimum MSE, even for the simplest datasets with unknown constant mean. This leads to

[1] `ftp://ftp.cmdl.noaa.gov/ccg/co2/trends/co2_mm_mlo.txt`

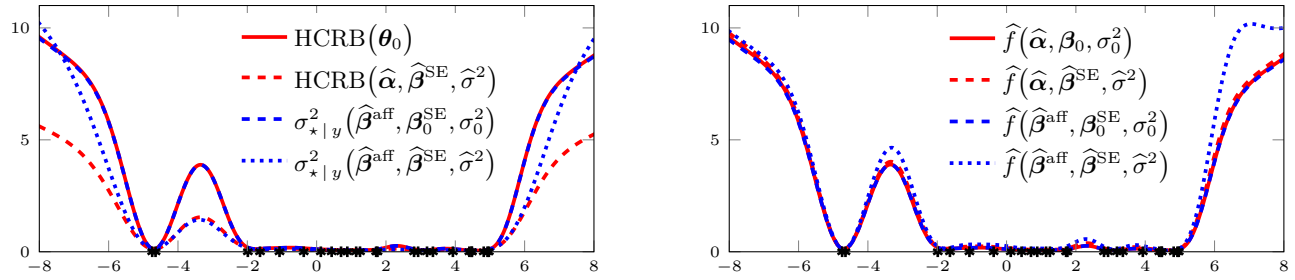

Figure 7: A comparison between bounds and MSE for an original and marginalized data model. Left: The corresponding HCRB and predictive variance. Right: MSE of corresponding predictors.

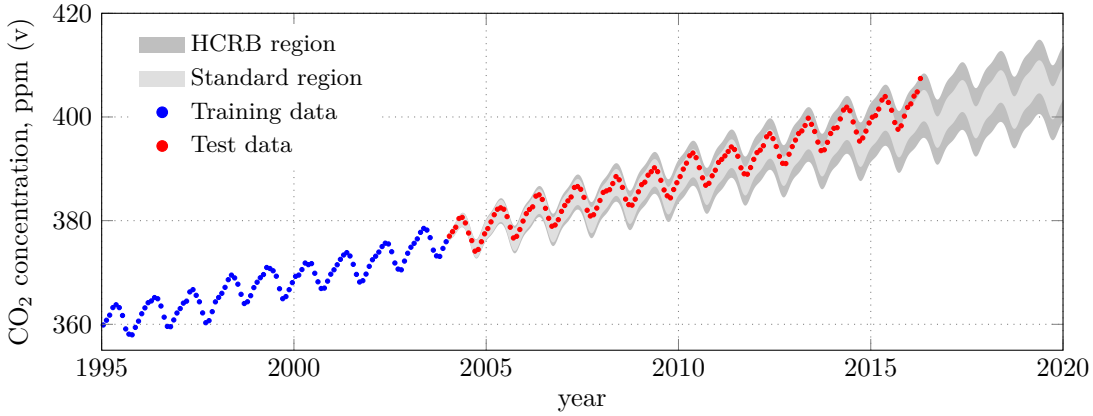

Figure 8: Monthly average atmospheric CO_2 concentration measured at Mauna Loa. GP model fit on data until December 2003. Error bars based on $\widehat{f}(x) \pm 3\sigma_{\star|y}$ and $\widehat{f}(x) \pm 3\sqrt{\mathrm{HCRB}}$.

incorrect prediction error bars. The underestimation gap arises from uncertainty of the hyperparameters and we provide an explicit and general characterization of it. The resulting HCRB is a simple closed-form expression and computationally cheap to implement.

In the examples we showed that the HCRB provides a tighter lower bound of the MSE for the standard predictor than the nominal predictive variance. The HCRB is easily computed using the quantities in the predictor itself and provides more accurate error bars, even when using estimated hyperparameters. In future work, we will investigate the accuracy of the estimated HCRB further. For the special case of linear mean functions, the results indicate a possible advantage of using an alternative marginalized model and assessing its corresponding BCRB using learned model parameters.

Acknowledgments

The authors would like to thank Dr. Marc Deisenroth and Prof. Carl E. Rasmussen for fruitful discussions.

This research is financially supported by the Swedish Foundation for Strategic Research (SSF) via the project *ASSEMBLE* (Contract number: RIT15-0012) . The work was also supported by the Swedish research Council (VR) via the projects *Probabilistic modeling of dynamical systems* (Contract number: 621-2013-5524) and (Contract number: 621-2014-5874).

A Alternative derivation of the bound

Unlike Rockah and Schultheiss (1987) and Van Trees and Bell (2013 [1968]), we will here prove Result 2 assuming only that the bias of the estimator with respect to $\check{f}_\star$ is invariant to $\boldsymbol{\theta}$. That is,

$$b(\boldsymbol{\theta}) \triangleq \mathrm{E}_y\left[\check{f}_\star(\boldsymbol{\theta}) - \widehat{f}_\star\right] \equiv b,$$

where b is a constant.

We begin by decomposing the MSE of an estimator $\widehat{f}_\star$:

$$\begin{aligned}\mathrm{MSE}\left(\widehat{f}_\star\right) &= \mathrm{E}[|f_\star - \widehat{f}_\star|^2] \\ &= \mathrm{E}_y\left[\mathrm{E}_{f|y}\left[|f_\star - \check{f}_\star + \check{f}_\star - \widehat{f}_\star|^2\right]\right] \\ &= \sigma^2_{\star|y} + \mathrm{E}_y\left[|\check{f}_\star - \widehat{f}_\star|^2\right],\end{aligned} \tag{18}$$

where $\check{f}_\star = \check{f}_\star(\boldsymbol{\theta})$ is the conditional mean (3) of $f_\star$. Since the first term in (18), $\sigma^2_{\star|y}$, is independent of the

estimator we will focus on finding a lower bound for the second term.

For notational simplicity, define the score function of the training data pdf as:

$$\boldsymbol{\phi} = \frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{y}|\boldsymbol{\theta}).$$

Then the correlation between the score function and the estimation error is

$$\begin{aligned}
\widetilde{\mathbf{g}} &= \mathrm{E}_y\left[\boldsymbol{\phi}(\check{f}_\star - \widehat{f}_\star)\right] \\
&= \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta})\right](\check{f}_\star - \widehat{f}_\star)\, d\mathbf{y} \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}}\left[p(\mathbf{y}|\boldsymbol{\theta})(\check{f}_\star - \widehat{f}_\star)\right] - p(\mathbf{y}|\boldsymbol{\theta})\left[\frac{\partial}{\partial \boldsymbol{\theta}}(\check{f}_\star - \widehat{f}_\star)\right] d\mathbf{y} \\
&= 0 - \mathrm{E}_y\left[\frac{\partial}{\partial \boldsymbol{\theta}}\check{f}_\star\right] \\
&= -\begin{bmatrix}\mathbf{g}^\mathsf{T} & \mathbf{0} & 0\end{bmatrix}^\mathsf{T}.
\end{aligned}$$

The first set of zeros follows from

$$\mathrm{E}_y\left[\frac{\partial}{\partial \boldsymbol{\beta}}\check{f}_\star\right] = \left(\frac{\partial}{\partial \boldsymbol{\beta}}\mathbf{w}^\mathsf{T}\right)\mathrm{E}_y\left[(\mathbf{y} - \mathbf{m})\right] = \mathbf{0}.$$

The final zero follows in a similar manner.

The Fisher information matrix is given by Slepian-Bangs formula:

$$\widetilde{\mathbf{J}} \triangleq \mathrm{E}_y\left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{y}|\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{y}|\boldsymbol{\theta})^\mathsf{T}\right] = \begin{bmatrix}\mathbf{M} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & * & * \\ \mathbf{0} & * & *\end{bmatrix}$$

The zeros therefore follow from the properties of the Gaussian distribution. We now form the product $\widetilde{\mathbf{g}}^\mathsf{T}\widetilde{\mathbf{J}}^{-1}\boldsymbol{\phi}$ and the nonnegative quadratic function

$$\begin{aligned}
0 &\le \mathrm{E}_y\left[|(\check{f}_\star - \widehat{f}_\star) - \widetilde{\mathbf{g}}^\mathsf{T}\widetilde{\mathbf{J}}^{-1}\boldsymbol{\phi}|^2\right] \\
&= \mathrm{E}_y\left[|\check{f}_\star - \widehat{f}_\star|^2\right] + \widetilde{\mathbf{g}}^\mathsf{T}\widetilde{\mathbf{J}}^{-1}\widetilde{\mathbf{g}} - 2\widetilde{\mathbf{g}}^\mathsf{T}\widetilde{\mathbf{J}}^{-1}\,\mathrm{E}_y\left[\boldsymbol{\phi}(\check{f}_\star - \widehat{f}_\star)\right] \\
&= \mathrm{E}_y\left[|\check{f}_\star - \widehat{f}_\star|^2\right] - \widetilde{\mathbf{g}}^\mathsf{T}\widetilde{\mathbf{J}}^{-1}\widetilde{\mathbf{g}}.
\end{aligned}$$

It follows that

$$\mathrm{E}_y\left[|\check{f}_\star - \widehat{f}_\star|^2\right] \;\ge\; \widetilde{\mathbf{g}}^\mathsf{T}\widetilde{\mathbf{J}}^{-1}\widetilde{\mathbf{g}}$$

Thus (18) is lower bounded by

$$\mathrm{MSE}\left(\widehat{f}_\star\right) \ge \sigma^2_{\star|y} + \mathbf{g}^\mathsf{T}\mathbf{M}^{-1}\mathbf{g},$$

which is Result 2.

B Proof of equality

Recall that $\boldsymbol{\mu} = \begin{bmatrix}\mathbf{m}^\mathsf{T} & m_\star\end{bmatrix}^\mathsf{T}$, $\Sigma = \begin{bmatrix}\mathbf{K} + \sigma^2\mathbf{I} & \mathbf{k}_\star \\ \mathbf{k}_\star^\mathsf{T} & k_{\star\star}\end{bmatrix}$, $\mathbf{w} = \left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_\star$, $\mathbf{g} = \frac{\partial}{\partial \boldsymbol{\alpha}}(m_\star - \mathbf{w}^\top\mathbf{m})$, and $\sigma^2_{\star\,|\,y} = k_{\star\star} - \mathbf{k}_\star^\mathsf{T}\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{k}_\star$. Then the following holds.

$$\frac{\partial \boldsymbol{\mu}^\mathsf{T}}{\partial \boldsymbol{\alpha}}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\alpha}^\mathsf{T}} - \sigma^{-2}_{\star|y}\mathbf{g}\mathbf{g}^\mathsf{T} = \frac{\partial \mathbf{m}^\mathsf{T}}{\partial \boldsymbol{\alpha}}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\frac{\partial \mathbf{m}}{\partial \boldsymbol{\alpha}^\mathsf{T}}$$

Proof. Let $\boldsymbol{\Sigma}_y = \mathbf{K} + \sigma^2\mathbf{I}$.

$$\begin{aligned}
\frac{\partial \boldsymbol{\mu}^\mathsf{T}}{\partial \boldsymbol{\alpha}}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\alpha}^\mathsf{T}} &= \frac{\partial \boldsymbol{\mu}^\mathsf{T}}{\partial \boldsymbol{\alpha}}\begin{bmatrix}\boldsymbol{\Sigma}_y & \mathbf{k}_\star \\ \mathbf{k}_\star^\mathsf{T} & k_{\star\star}\end{bmatrix}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\alpha}^\mathsf{T}} \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}}\begin{bmatrix}\mathbf{m}^\mathsf{T} & m_\star\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_y^{-1} + \boldsymbol{\Sigma}_y^{-1}\mathbf{k}_\star\sigma^{-2}_{\star\,|\,y}\mathbf{k}_\star^\mathsf{T}\boldsymbol{\Sigma}_y^{-1} & -\boldsymbol{\Sigma}_y^{-1}\mathbf{k}_\star\sigma^{-2}_{\star\,|\,y} \\ -\sigma^{-2}_{\star\,|\,y}\mathbf{k}_\star^\mathsf{T}\boldsymbol{\Sigma}_y^{-1} & \sigma^{-2}_{\star\,|\,y}\end{bmatrix}\frac{\partial}{\partial \boldsymbol{\alpha}^\mathsf{T}}\begin{bmatrix}\mathbf{m} \\ m_\star\end{bmatrix} \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}}\begin{bmatrix}\mathbf{m}^\mathsf{T} & m_\star\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_y^{-1} & 0 \\ 0 & 0\end{bmatrix}\frac{\partial}{\partial \boldsymbol{\alpha}^\mathsf{T}}\begin{bmatrix}\mathbf{m} \\ m_\star\end{bmatrix} + \sigma^{-2}_{\star\,|\,y}\frac{\partial}{\partial \boldsymbol{\alpha}}\begin{bmatrix}\mathbf{m}^\mathsf{T} & m_\star\end{bmatrix}\begin{bmatrix}\mathbf{w}\mathbf{w}^\mathsf{T} & -\mathbf{w} \\ -\mathbf{w}^\mathsf{T} & 1\end{bmatrix}\frac{\partial}{\partial \boldsymbol{\alpha}^\mathsf{T}}\begin{bmatrix}\mathbf{m} \\ m_\star\end{bmatrix} \\
&= \frac{\partial \mathbf{m}^\mathsf{T}}{\partial \boldsymbol{\alpha}}\boldsymbol{\Sigma}_y^{-1}\frac{\partial \mathbf{m}}{\partial \boldsymbol{\alpha}^\mathsf{T}} + \sigma^{-2}_{\star\,|\,y}\frac{\partial}{\partial \boldsymbol{\alpha}}(m_\star - \mathbf{w}^\mathsf{T}\mathbf{m})\frac{\partial}{\partial \boldsymbol{\alpha}^\mathsf{T}}(m_\star - \mathbf{w}^\mathsf{T}\mathbf{m})^\mathsf{T} \\
&= \frac{\partial \mathbf{m}^\mathsf{T}}{\partial \boldsymbol{\alpha}}\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\frac{\partial \mathbf{m}}{\partial \boldsymbol{\alpha}^\mathsf{T}} + \sigma^{-2}_{\star\,|\,y}\mathbf{g}\mathbf{g}^\mathsf{T}
\end{aligned}$$

□

References

Bangs, W. J. (1971). *Array Processing with Generalized Beamformers*. Ph.D. Thesis. Yale University.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Cramér, H. (1946). "A contribution to the theory of statistical estimation". In: *Scandinavian Actuarial Journal* 1946.1, pp. 85–94.

Deisenroth, M., D. Fox, and C. E. Rasmussen (2015). "Gaussian Processes for Data-Efficient Learning in Robotics and Control". In: *Transactions on Pattern Analysis and Machine Intelligence* 37.2, pp. 408–423.

Deisenroth, M. and C.E. Rasmussen (2011). "PILCO: A model-based and data-efficient approach to policy search". In: *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pp. 465–472.

Den Hertog, Dick, Jack PC Kleijnen, and AYD Siem (2006). "The correct Kriging variance estimated by bootstrapping". In: *Journal of the Operational Research Society* 57.4, pp. 400–409.

Gill, R.D. and B.Y. Levit (1995). "Applications of the van Trees inequality: a Bayesian Cramér-Rao bound". In: *Bernoulli*, pp. 59–79.

Likar, B. and J. Kocijan (2007). "Predictive control of a gas–liquid separation plant based on a Gaussian process model". In: *Computers & chemical engineering* 31.3, pp. 142–152.

Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive computation and machine learning series. MIT Press.

Pérez-Cruz, F., S. Van Vaerenbergh, JJ.J. Murillo-Fuentes, M. Lázaro-Gredilla, and I. Santamaria (2013). "Gaussian processes for nonlinear signal processing: An overview of recent advances". In: *IEEE Signal Processing Magazine* 30.4, pp. 40–50.

Rao, C.R. (1945). "Information and the accuracy attainable in the estimation of statistical parameters". In: *Bulletin of Calcutta Mathematical Society* 37.1, pp. 81–89.

Rasmussen, C.E. and C.K.I. Williams (2006). *Gaussian Processes for Machine Learning*. Adaptative computation and machine learning series. MIT Press.

Rockah, Y. and P.M. Schultheiss (1987). "Array shape calibration using sources in unknown locations–Part I: Far-field sources". In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 35.3, pp. 286–299.

Schölkopf, B. and A.J. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive computation and machine learning. MIT Press.

Shahriari, Bobak, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando de Freitas (2016). "Taking the human out of the loop: A review of bayesian optimization". In: *Proceedings of the IEEE* 104.1, pp. 148–175.

Slepian, D. (1954). "Estimation of signal parameters in the presence of noise". In: *Information Theory, Transactions of the IRE Professional Group on* 3.3, pp. 68–89.

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer New York.

Stoica, P. and R.L. Moses (2005). *Spectral analysis of signals*. Pearson/Prentice Hall.

Suykens, J.A.K., T. Van Gestel, and J. De Brabanter (2002). *Least Squares Support Vector Machines*. World Scientific.

Van Trees, H.L. and K.L. Bell (2013 [1968]). *Detection Estimation and Modulation Theory, Pt.I*. second. Detection Estimation and Modulation Theory. Wiley.

Williams, Christopher K. I. and Carl Edward Rasmussen (1996). "Gaussian Processes for Regression". In: *Advances in Neural Information Processing Systems 8*. Ed. by D. S. Touretzky and M. E. Hasselmo. MIT Press, pp. 514–520.

Zachariah, D. and P. Stoica (2015). "Cramer-Rao Bound Analog of Bayes' Rule [Lecture Notes]". In: *Signal Processing Magazine, IEEE* 32.2, pp. 164–168.

Zimmerman, Dale L and Noel Cressie (1992). "Mean squared prediction error in the spatial linear model with estimated covariance parameters". In: *Annals of the institute of statistical mathematics* 44.1, pp. 27–43.