Information Metrics for Model Selection in Function Estimation

Tansu Alpcan
Department of Electrical and Electronic Engineering
The University of Melbourne, Australia
Email: tansu.alpcan@unimelb.edu.au

***Abstract*—A model selection framework is presented for function estimation under limited information, where only a small set of (noisy) data points are available for inferring the nonconvex unknown function of interest. The framework introduces information-theoretic metrics which quantify model complexity and are used in a multi-objective formulation of the function estimation problem. The intricate relationship between information obtained through observations and model complexity is investigated. The framework is applied to the hyperparameter selection problem in Gaussian Process Regression. As a result of its generality, the framework introduced is applicable to a variety of settings and practical problems with information limitations such as channel estimation, black-box optimisation, and dual control.**

I. Introduction

In many real-world problems unknown functions of interest have to be estimated using limited information. The available data is often small due to resource limitations (e.g. high cost of observation) and/or fast-changing nature of the underlying system (time limitations). These constraints rule out several conventional methods for estimating the function such as Monte Carlo sampling, randomised algorithms [1], evolutionary algorithms, or simulated annealing [2].

The function of interest often plays a significant role in the overall system performance. It may represent fading in a communication channel, the objective function in a black-box optimisation problem or a dynamical system in dual control [3]. Hence, a principled approach is needed to interpret the collected data and make most of it in such joint learning and decision-making problems.

Recent works [4]–[6] have presented novel results on using information-theoretic metrics for quantifying and selecting new data points for black-box optimisation problems. However, model selection problem has not been discussed in those works, where it was assumed that the hyperparameters of the Gaussian regression models used for estimating objective functions were optimally chosen. Building upon these earlier results, this paper focuses solely on the model (meta- or hyperparameter) selection problem and investigates the intricate relationship between information availability and model complexity. Adopting a frequency domain approach, two information-theoretic metrics are defined and used in a multi-objective formulation. The resulting framework is applied to the hyperparameter selection in Gaussian Process Regression (GPR) [7].

There is a large literature on the applications of information-theoretic concepts to model selection problems. A comprehensive and accessible overview is provided in [8]. Information-based model selection criteria for regression are presented in [9]. Model selection in the context of GPR is discussed in [7, Chap. 5]. *This paper differentiates from earlier work* due to its frequency-based approach to model complexity and the resulting multi-objective formulation of the problem.

The rest of the paper is organised as follows. The next section presents the problem formulation. Section III introduces two information metrics motivated by frequency-domain analysis. Section IV applies the concepts introduced to GPR and illustrates the framework with a numerical example. The paper concludes with a brief discussion in Section V.

II. Problem Formulation

Let $\mathcal{X} \subset \mathbb{R}$ be the compact real domain of a continuous real-valued function $f : \mathcal{X} \to \mathbb{R}$. In order to keep to notation clean in the paper, the function f takes a scalar argument defined on the unit interval, $\mathcal{X} = [0, 1]$, without loss of any generality. Consequently, f is bounded and assumes its minimum and maximum on $\mathcal{X}$ [10, p. 47]. Assume that f is also differentiable and Lipschitz continuous with a constant L such that $|df(x)/dx| \leq L$.

The main objective here is to estimate the unknown function f which is known only on a finite set of data points

$$\mathcal{D} = \{[x_1, f(x_1)], \dots, [x_D, f(x_D)]\} \tag{1}$$

that are obtained through observations. Let $\hat{f}(x)$ be the best estimate of $f(x)$ given $\mathcal{D}$. Finding the "best" $\hat{f}$ is known as the *regression problem*.

Choosing the right model plays a very important role in addressing the function approximation problem defined.[1] To simplify the task, let α denote the unknown model hyperparameter and reduce model selection to the selection of the best α. There are three factors directly affecting this modelling problem. The first one is how well the estimated function $\hat{f}$ approximates f on $\mathcal{D}$ or how well it fits the observed data, e.g. $\min_{\hat{f}} \sum_{x \in \mathcal{D}} \left| \hat{f}(x) - f(x) \right|$. The second factor is the *descriptive model complexity*. Following Occam's razor [11], a simpler model is desirable as long as it performs

[1] Since a non-parametric regression method is assumed in this paper, the "model selection" translates to the choice of meta- or hyperparameters.

satisfactorily, i.e. it should be as simple as possible but not simpler [8]. This principle is closely related to the concepts of *generalisation* and *regularisation* which are well known in the pattern recognition literature [12].

These two factors are in direct conflict with each other. A perfect match between $\hat{f}$ and f on $\mathcal{D}$ can be obtained by increasing model complexity which is also known as "overfitting" [12]. The third and often ignored factor is how much "information" the existing data $\mathcal{D}$ provides to the model at hand or how much information is needed to learn a model of certain *learning complexity*.

A principled way of approaching the model selection task is to quantify the second and third factors using information-theoretic metrics and pose it as a multi-objective optimisation problem. Let $C_M(\hat{f})$ measure the non-negative descriptive complexity. Further define a measure on model learning complexity $C_L(\hat{f})$ and a bound on it $B(\mathcal{D})$ based on the available data $\mathcal{D}$. Then, the second objective is $\min_{\hat{f}} C_M(\hat{f})$ and the third one can be formulated as $C_L(\hat{f}) \leq B(\mathcal{D})$. One possible combination of all three objectives [13] leads to the following optimisation problem:

$$\min_{\hat{f}} w_1 \sum_{x \in \mathcal{D}} \frac{1}{|\mathcal{D}|} \left| \hat{f}(x) - f(x) \right| + w_2\, C_M(\hat{f})$$
$$\text{subject to} \;\; C_L(\hat{f}) \leq B(\mathcal{D}), \tag{2}$$

where w_1, w_2 are weighting and normalisation parameters, and $|\mathcal{D}|$ is the cardinality of the set $\mathcal{D}$.

An alternative formulation based on the "entropy maximisation principle" in the literature maximises entropy subject to a constraint imposed by the information contained in the observed data [8, Chap. 2.12.1]. Using the definitions introduced above this corresponds to

$$\max_{\hat{f}} C_M(\hat{f}) \;\; \text{subject to} \;\; C_L(\hat{f}) \leq B(\mathcal{D}). \tag{3}$$

Note that the first objective is automatically taken care of by the nonparameteric learning method considered here.

The problem formulations (2) and (3) are closely related to each other. While the former searches for the simplest model justified by the data set which fits the observed data points, the latter aims to find the most complex (maximum entropy) model justified by the data set.

III. Information Metrics for Model Selection

A typical way of describing (encoding) a function is taking its (e.g. Fourier) transform and then using the resulting frequency domain parameters as commonly done in image and sound compression. The same approach can also be used to derive the relevant information metrics to formulate and address the problem in (2).

The function f defined in Section II is assumed to be differentiable and Lipschitz continuous with constant L on $[0,1]$. Let $F(\omega)$ denote the Fourier transform of $f(x)$. Then, its energy is finite and given by $E_f = \int_{-\infty}^{\infty} |F(\omega)|^2\, dw$ from Perseval's theorem [14].

The function $f(x)$ is of *bounded variation*, $\int_0^1 |df(x)/dx|\, dx \leq L < \infty$, due to its differentiability and Lipschitz-continuity. It then follows directly from the Riemann lemma [14, p. 95] that

$$F(\omega) = \int_0^1 f(x) e^{-j\omega t} dt = O(\frac{1}{\omega^2}) \quad |\omega| \to \infty,$$

where $O(\cdot)$ denotes the following: for any real valued functions $g(x)$ and $h(x)$, $g(x) = O(h(x))$ $x \to \infty$, means there exists a positive $M \in \mathbb{R}$ such that $g(x) \leq M\, h(x)$ as $x \to \infty$.

Since f is duration-limited, it can be perfectly reconstructed from its countably many Fourier series coefficients, $Fs(n) = \frac{1}{T}\int_0^1 f(x) e^{-2\pi j x n/T} dx$, where n is an integer. Since $T = 1$ in this special case, we have $Fs(n) = F(n/T) = F(n)$, i.e. the Fourier series coefficients of f can be interpreted as uniform samples of its Fourier transform. Let $S(\omega) := |F(\omega)|^2$ be the power spectrum of the function $f(x)$ and $Ss(n/T) = Ss(n) := |Fs(n)|^2$ be its discrete counterpart corresponding to a uniformly sampled version of $S(\omega)$.

Definition III.1. *Let the function $f(x)$ defined on $x \in \mathcal{X} = [0,1]$ be real-valued, differentiable, and Lipschitz-continuous with the constant L. Let $S(\omega)$ be the spectral density of f and $Ss(n)$ be its discrete counterpart. The entropy-like functional capturing the descriptive model complexity of $f(x)$ is defined as*

$$C_M(f) := -\sum_{n=-\infty}^{\infty} Ss(n) \log_2(Ss(n)). \tag{4}$$

Couple of interesting observations can be made based on the definition of $C_M(f)$. Firstly, the functional $C_M(f)$ clearly has an entropy interpretation, if ω is considered as a (frequency) random variable with probability mass function $Ss(n)$. Secondly, the value $C_M(f)$ provides naturally an upper-bound on the Kolmogorov complexity of f up to a fixed constant [11].

The descriptive complexity measure in Definition III.1 makes intuitive sense. Let $f_1(x) = (1/\sqrt{2}) sin(\omega_1 x)$ be sinusoidal function and $f_2(x)$ be a band-limited white noise function on $\mathcal{X}$ such that $S_2(\omega) = 1/(2W)$, $|\omega| \leq W$. Then, $C_M(f_1) = 1$ and $C_M(f_2) = \log_2(2W)$. Furthermore, the complexity of white noise becomes infinite as $W \to \infty$. In other words, a single sinusoid (a basis function of the Fourier transform) has unit and the white noise has infinite descriptive model complexity.

The entropy-like measure C_M of Definition III.1 quantifies the frequency spread or information content of an already known function and exhibits intuitive behaviour as discussed above. However, it assigns high and low frequency functions the same complexity. For example, the function $f_1(x) = (1/\sqrt{2}) sin(\omega_1 x)$ has unit model complexity regardless of whether ω_1 is 1 or 10^6. This is in contrast to the fact that learning a slowly varying unknown function is easier, i.e. requires less number of observations, than a function that has most of its energy in high frequencies. It is therefore necessary to distinguish these two separate concepts and define another metric measuring *model learning complexity* which differs from the *descriptive model complexity* in Definition III.1.

Fortunately, the very-well known concept of bandwidth can be used for this purpose.

As the first step, define the band-limited function

$$f_W(x) := \frac{1}{2\pi}\int_{-W}^{W} F(\omega)e^{j\omega t}d\omega, \tag{5}$$

on $\mathcal{X} = [0,1]$ that approximates f arbitrarily well for increasing W. The residual energy of the approximation error, e_a, quadratically decreases to zero as $|W| \to \infty$:

$$e_a(f - f_W) = \int_0^1 |f(x) - f_W(x)|^2\, dx = O(\frac{1}{W^2}),$$

which is a consequence of the Riemann lemma and Parseval's theorem [14]. The spectral density of f_W is then

$$S_W(\omega) = \begin{cases} |F(\omega)|^2 & \text{, for } |\omega| \le W \\ 0 & \text{, else} \end{cases}. \tag{6}$$

Definition III.2. *Under the same assumptions of Definition III.1, define the energy-based bandwidth $W(\lambda)$ of f such that the energy of the bandlimited version of the function $f_{W(\lambda)}(x)$ (5) is $E_{f_{W(\lambda)}} = \int S_W(\omega)d\omega = \lambda E_f$, where $0 \le \lambda \le 1$. The learning complexity functional, based on this bandwidth criterion, is defined as*

$$C_L(f, \lambda) := W(\lambda).$$

Remark III.3. The definition of learning complexity can be extended to any meaningful definition of bandwidth. Note, however, that the classical definition of bandwidth equating it to the support of the spectral density $S(\omega)$ leads to a trivial result where $C_L(f, 1) = \infty$ for any duration-limited f.

The choice of λ presents a trade-off between how closely the estimated function $\hat{f}$ is desired to approximate f in the frequency domain by including higher frequency components, which increases the number of data points needed to justify such an estimate. Clearly, the closer λ is to one, the more observations are needed to estimate the high-frequency components in $\hat{f}$, which affects hyperparameter selection of the model. In this paper, λ is chosen somewhat arbitrarily as 0.9 corresponding to retain 90% of the signal energy.

The amount of information contained in the current set of observations $\mathcal{D}$, defined in (1), provides an upper-bound on learning model complexity, $C_L(f)$, in the Definition III.2. The principle here is *choosing a model which has a learning complexity consistent with the amount of available data*. The results presented next aim to express this principle in mathematical terms using Nyquist-Shannon sampling theorem [14].

Consider the case where the data points in $\mathcal{D}$ are equispaced on $\mathcal{X} = [0,1]$. Then, the data set $\mathcal{D}$ can be interpreted as a sampling of function $f(x)$, which can be approximated by its band-limited version, $f_{ua}(x)$, through interpolation

$$f_{ua}(x) := \sum_{k=0}^{2W_s} f(\frac{k}{2W_s})\frac{\sin(W_s x - k)}{W_s x - k},$$

where W_s is the bandwidth (support set) of $f_{ua}(x)$. In fact, under the assumptions made on $f(x)$, the uniform approximation error is bounded by:

$$|e_{unif}(t)| := |f(x) - f_{ua}(x)| \le \frac{2}{\pi}\int_{W_s}^{\infty} F(\omega)d\omega. \tag{7}$$

Moreover, this bound can be refined to

$$|e_{unif}(t)| \le \frac{2L}{W_s}, \;\; W_s > 1, \tag{8}$$

where L is the Lipschitz constant. Both error bounds (7) and (8), which are corollaries of Theorems 2 and 3 in [15], are satisfied if the number of samples is chosen as $N = 2W_s$.

When the uniform sampling assumption is relaxed to allow data points in $\mathcal{D}$ be chosen according to any distribution on $\mathcal{X}$, then the following result is obtained [16]. First, define a counterpart of f_{ua} as

$$f_{nua}(x) := \frac{1}{N}\sum_k f(t_k)\frac{sin(W(t - t_k))}{\pi(t - t_k)},$$

where $t_k,\; k = 1, \ldots, N$ is any ordered set of points on $\mathcal{X}$. Then, the general approximation error between $f_W(x)$ defined in (5) and $f_{nua}(x)$ is given by

$$|e_{gen}(t)| = |f_W(x) - f_{nua}(x)| = \frac{O(W\log(W))}{N}. \tag{9}$$

It is now possible to derive how many data points are needed to satisfy a given error bound, ε, when approximating $f(x)$ or $f_W(x)$ using an interpolation based on a given $\mathcal{D}$.

Proposition III.4. *Let $\varepsilon > 0$ bound the error in approximating $f(x)$ or $f_W(x)$ using an interpolation based on $\mathcal{D}$. Let N_{unif} and N_{gen} be the number of uniformly and arbitrarily distributed observations, respectively, i.e. the cardinality of $\mathcal{D}$ for each case. Then, the following hold:*

$$N_{unif} \ge \frac{4L}{\varepsilon}, \qquad N_{gen} \ge \frac{O(W\log(W))}{\varepsilon}.$$

Proof: The result in (8) provides the minimum number of data points in $\mathcal{D}$ that are uniformly distributed on $\mathcal{X}$, N_{unif}, needed to achieve the given error bound, whereas the one in (9) provides the counterpart, N_{gen}, for a general distribution of data points. ■

Note that, the number N_{gen} is independent of distribution and increases with the bandwidth of $f_W(x)$, i.e. as f_{nua} approximates f better.

IV. Model Selection in GP Regression

A. GP Regression Overview

A short overview of Gaussian Process (GP) regression is presented next for completeness [7], [17]. A GP is formally defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. In general, it is completely specified by its mean function $m(x)$ which is assumed to be zero here for simplicity, and covariance function

$$c : (\mathcal{X}, \mathcal{X}) \to \mathbb{R}, \;\; c(x, \tilde{x}) := E[f(x)f(\tilde{x})],\; \forall x, \tilde{x} \in \mathcal{X}.$$

Hence, the GP is characterised in this special case entirely by its covariance function $c(x, \tilde{x})$.

Given a set of data $\mathcal{D}$ and assuming fixed Gaussian observation noise, the covariance matrix is defined as the sum of a *kernel matrix* Q and noise variance σ:

$$C_{ij}(\alpha) := Q_{ij}(\alpha) + \sigma, \ \forall i, j = 1, \ldots, \text{card}(\mathcal{D}) \qquad (10)$$

where $\text{card}(\mathcal{D})$ is the cardinality of the data set $\mathcal{D}$ and α is a kernel hyperparameter. While it is possible to choose here any (positive definite) kernel function $q_\alpha(x, \tilde{x}) : (\mathcal{X}, \mathcal{X}, \mathbb{R}^+) \to \mathbb{R}$, one classical choice is the Gaussian kernel,

$$q_\alpha(x, \tilde{x}) = \exp\left[-\frac{|x - \tilde{x}|^2}{2\alpha^2}\right]. \qquad (11)$$

which leads to the kernel matrix $Q_{ij}(\alpha) = q_\alpha(x_i, x_j)$, where $x_i, \, x_j \in \mathcal{D}$.

The Fourier transform of the stationary covariance function gives the spectral density $S(\omega)$ which follows directly from Wiener-Khinchin theorem [7]. The spectral density of the covariance function with Gaussian kernel (11), when there is no observation noise, is

$$S_\alpha(\omega) = \alpha\sqrt{2\pi}\, e^{-2\pi^2\alpha^2\omega^2}. \qquad (12)$$

Given the data set $\mathcal{D}$, define the vector

$$k(\mathcal{D}) := [q_\alpha(x_1, f(x_1)), \ldots q_\alpha(x_D, f(x_D))] \qquad (13)$$

and scalar

$$\kappa := q_\alpha(x, x) + \sigma = 1 + \sigma. \qquad (14)$$

Then, the *predictive distribution* at a given point x, $p_{\hat{x}}(\mathcal{D}, x)$, is a Gaussian random variable, $\mathcal{N}(\hat{f}, v)$, with the mean $\hat{f}$ and variance v:

$$\hat{f}(\mathcal{D}, x) := k^T C^{-1} \bar{f}(\mathcal{D}) \text{ and } v(\mathcal{D}, x) := \kappa - k^T C^{-1} k, \qquad (15)$$

where $\bar{f}(\mathcal{D}) = [f(x_1), f(x_2), \ldots, f(x_D)]^T$. Note that the variance is independent of the individual state dimension. This is a key result that defines GP regression. The mean function $\hat{f}(x)$ of the Gaussian distribution provides a prediction of the objective function $f(x)$. Furthermore, the variance function $v(x)$ can be used to measure the uncertainty level of the predictions provided by $\hat{f}$.

B. Model Selection

In the GP regression context, the model selection problem (2) becomes one of hyperparameter selection due to the non-parametric nature of GP. Specifically, let α denote the unknown model hyperparameters in the chosen GP kernel.[2] Then, the estimated function $\hat{f}_\alpha$ is parameterised by α. Using the respective definitions of $C_M(\alpha)$ in Definition III.1 and $C_L(\alpha)$ in Definition III.2, the model selection problem (2) is converted to selection of the hyperparameter α as follows:

$$\min_\alpha w_1 \sum_{x\in\mathcal{D}} \frac{1}{|\mathcal{D}|} \left|\hat{f}_\alpha(x) - f(x)\right| + w_2\, C_M(\hat{f}_\alpha)$$
$$\text{subject to } \ C_L(\alpha) \leq B(\mathcal{D}). \qquad (16)$$

This formulation clearly provides multiple degrees of freedom in the choice of the parameters such as the weights w_1, w_2 balancing the objectives and the upper-bound on bandwidth B. The selection of these parameters is rather problem and context dependent in practice.

The power spectral density $S(\alpha)$ (12) of the widely-used Gaussian kernel function (11) has unit energy and is parameterised by α. The entropy of $S(\alpha)$, $E_S(\alpha)$ is $E_S(\alpha) = -0.5 \ln 2\pi e - \ln\alpha$. Thus, the model complexity of the GP with Gaussian kernel (11) based on the Definition III.1 becomes

$$C_M(\alpha) = -\ln\alpha - \frac{1}{2}\ln 2\pi e. \qquad (17)$$

Likewise, using the 90% energy-based bandwidth in Definition III.2 and (12), the learning model complexity is

$$C_L(\alpha) \approx \frac{1}{\alpha\pi\sqrt{2}}\,\text{erf}^{-1}(0.8) = \frac{0.204}{\alpha}, \qquad (18)$$

where $\text{erf}^{-1}(\cdot)$ is the inverse error function.

The discussion in Section III, which can be seen as an extension of the well-known Nyquist criterion [14] to duration-limited signals, provides the background for deriving the value $B(\mathcal{D})$ in (16). First, assume that the data points in $\mathcal{D}$ are equispaced. Then, given N, which corresponds to the sampling frequency on $\mathcal{X} = [0, 1]$, one choice for the bound B is simply $B(\mathcal{D}) = N/2$. However, the approximation errors in (7) and (8) for $W_s = N/2$ should be taken into account. For the non-uniform case, a heuristic approximation such as using the mean or maximum distance between samples as a basis for sampling frequency can be used. In that case, the result on the error bound in (9) holds.

Using the above results, the model (hyperparameter) selection problem (16) can be reformulated in the special case of GP regression with Gaussian kernel as

$$\min_\alpha w_1 \sum_{x\in\mathcal{D}} \frac{1}{|\mathcal{D}|} \left|\hat{f}_\alpha(x) - f(x)\right| - w_2 \ln\alpha$$
$$\text{subject to } \ \alpha \geq \frac{0.408}{N}, \qquad (19)$$

where N is the cardinality of $\mathcal{D}$. Here, data uniformity is assumed to simplify the formulation.

For the specific GP kernel chosen (11), the smaller α the higher is the model complexity. For the same set of given data, choosing a smaller α leads to overfitting and increases the uncertainty of the estimation (see [7, Chap. 5.4] for a nice discussion on this). Therefore, the lower bound on α in (19) limits the the tendency of choosing a more complicated model than the available data warrants. In effect, the formulation (19) finds the simplest possible model which fits the observed data well-enough under the learning constraint.

The alternative problem formulation (3) for the special case of GP regression becomes

$$\min_\alpha \ln\alpha \ \text{ subject to } \ \alpha \geq \frac{0.408}{N}, \qquad (20)$$

which clearly admits the boundary solution $\alpha^* = 0.408/N$. In other words, the "entropy maximisation principle" chooses the

[2] This paper assumes that a kernel function is already chosen for simplicity. The kernel choice can also be posed as a model selection problem itself.

most complex model justified by the available data (learning constraint). In comparison, the formulation (19) allows explicitly tuning the balance between the data fitting and model complexity objectives.

C. Numerical Example

The presented framework is illustrated with a numerical example. The unknown function to be estimated is the arbitrarily chosen polynomial:

$$f(x) := 20.48x^4 + 2.56x^3 - 3.84x^2 - 0.49x - 0.02, \quad x \in [0,1].$$

Based on the discussion in the previous section, the problem formulation in (19) is used to choose the hyperparameter α and derive the estimate $\hat{f}(x)$. The objectives in (19) are normalised using standard methods and given equal weights.

The function estimates $\hat{f}(x)$ are obtained for different number of data points and plotted against the real values of $f(x)$ in Figures 1-4. The corresponding lower bounds on α and the optimal α^* values are listed in Table I. The estimated functions qualitatively match intuitive expectations.

TABLE I
OPTIMAL α VALUES AND LOWER BOUNDS

Figure	α^*	Lower bound
1	0.988	0.136
2	0.129	0.102
3	0.085	0.082
4	0.085	0.068

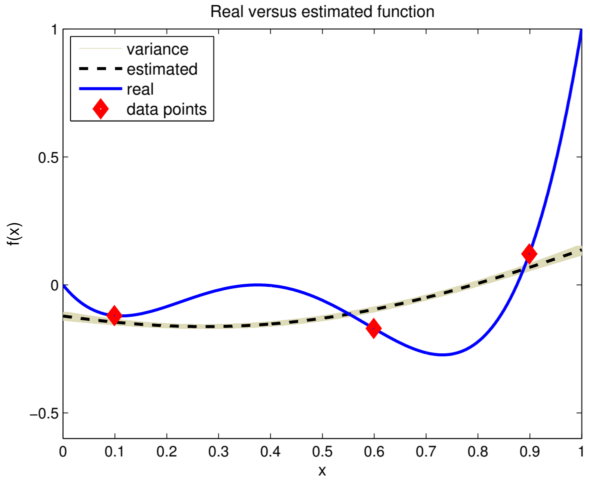

Fig. 1. Estimation with 3 data points; $\alpha \geq 0.136$, $\alpha^* = 0.988$.

Next, in order to illustrate the role of learning complexity the lower bound on α is ignored in solving (19), which yields an optimal hyperparameter of $\alpha^* = 0.002$. The resulting function estimate, shown in Figure 5, is clearly overfitting. During the simulations, it was noted that the objective function in (19) is not convex and has multiple minima. Therefore, the

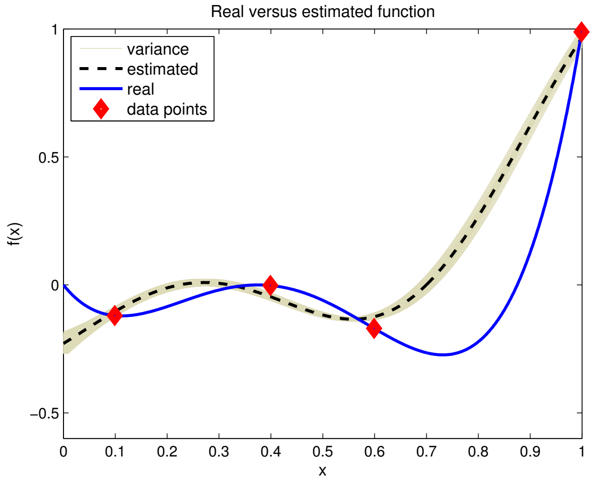

Fig. 2. Estimation with 4 data points; $\alpha \geq 0.102$, $\alpha^* = 0.129$.

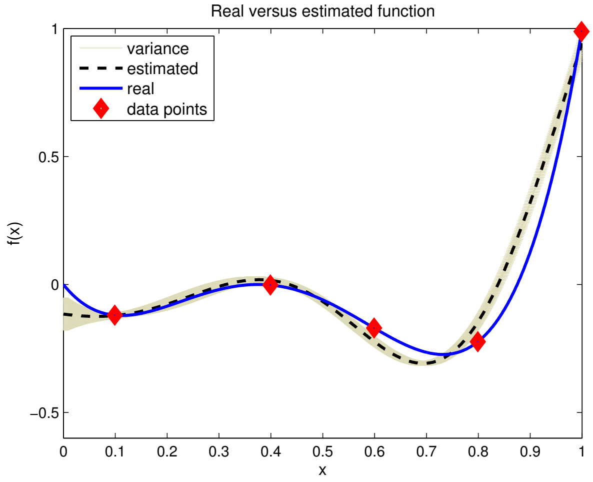

Fig. 3. Estimation with 5 data points; $\alpha \geq 0.082$, $\alpha^* = 0.085$.

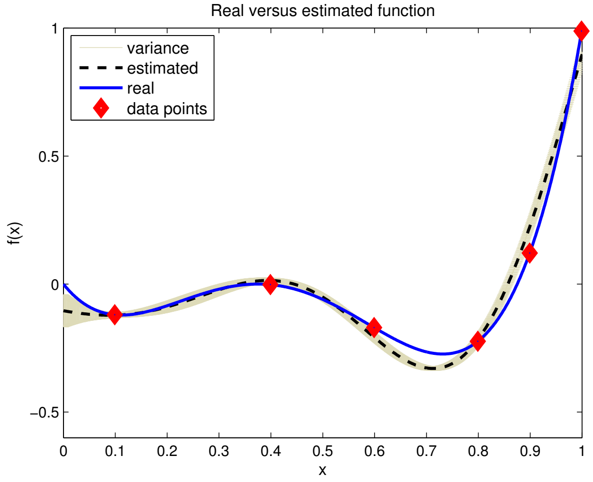

Fig. 4. Estimation with 6 data points; $\alpha \geq 0.068$, $\alpha^* = 0.085$.

constraint on α, and hence the learning complexity clearly plays a significant role in choosing the right model along with descriptive complexity.

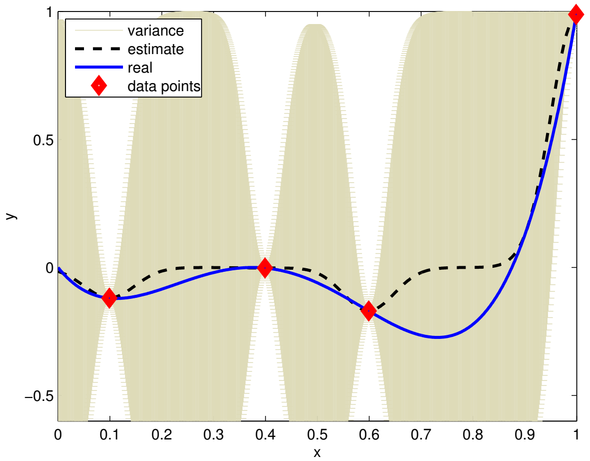

Fig. 5. Estimation with 4 data points ignoring learning complexity bound; $\alpha \geq 0$, $\alpha^* = 0.002$.

If the alternative formulation (3) based on the maximum entropy principle is used, the lower bounds in the Table I become the optimal hyperparameters, α^*, of (20). The results are then visually similar to those in the Figures 1-4. The difference between the two sets of results is directly affected by the weighting parameters w_1, w_2 in (19), which provide an additional freedom of choice to the designer in formulating the problem.

V. Conclusion

A model selection framework is presented for function estimation under limited information. The framework introduces information-theoretic metrics for quantifying descriptive and learning model complexity, which are then used in a multi-objective formulation. As a concrete example, the framework is applied to the hyperparameter selection problem in Gaussian Process Regression (GPR) and illustrated with a numerical example. The results have interesting implications for online learning and provide a novel method for adjusting model complexity during the estimation process based on data availability. As a result of its generality, the framework introduced is applicable to a variety of settings and practical problems with information limitations such as channel estimation, black-box optimisation, and dual control.

Future research directions include further analysis of the concepts introduced in relation to the existing information criteria, extension to multi-variate functions, and applications to parametric learning frameworks.

Acknowledgment

The author thanks Iman Shames, Girish Nair, and Erik Weyer for helpful discussions and suggestions. This research was supported in part by the Australian Research Council's Discovery Projects funding scheme (DP140100819).

References

[1] R. Tempo, G. Calafiore, and F. Dabbene, *Randomized Algorithms for Analysis and Control of Uncertain Systems*. London, UK: Springer-Verlag, 2005.

[2] R. Rutenbar, "Simulated annealing algorithms: an overview," *IEEE Circuits and Devices Magazine*, vol. 5, no. 1, pp. 19–26, January 1989.

[3] B. Wittenmark, "Adaptive dual control methods: An overview," in *5th IFAC symposium on Adaptive Systems in Control and Signal Processing*, 1995, pp. 67–72.

[4] T. Alpcan, "A framework for optimization under limited information," *Journal of Global Optimization*, pp. 1–26, 2012.

[5] ——, "A risk-based approach to optimisation under limited information," in *Proc. of the 20th Intl. Symp. on Mathematical Theory of Networks and Systems (MTNS)*, July 2012.

[6] ——, "A framework for optimization under limited information," in *5th International ICST Conference on Performance Evaluation Methodologies and Tools*, ser. VALUETOOLS '11. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011, pp. 234–243.

[7] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[8] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference : a practical information-theoretic approach*, 2nd ed. New York: Springer, 2002.

[9] A. D. R. McQuarrie and C.-L. Tsai, *Regression and Time Series Model Selection*. World Scientific, 1998.

[10] H. L. Royden, *Real Analysis*, 3rd ed. New Jersey, USA: Prentice-Hall, 1988.

[11] M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd ed., ser. Texts in Computer Science. New York, NY, USA: Springer, 1997.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[13] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Structural and Multidisciplinary Optimization*, vol. 26, no. 6, pp. 369–395, 2004. [Online]. Available: http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s00158-003-0368-6

[14] A. Papoulis, *Signal analysis*. New York: McGraw-Hill, 1977.

[15] P. L. Butzer and W. Splettstösser, "A sampling theorem for duration-limited functions with error estimates," *Information and Control*, vol. 34, no. 1, pp. 55–65, 1977. [Online]. Available: http://dx.doi.org/10.1016/S0019-9958(77)90264-9

[16] P. Ferreira, "Nonuniform sampling of nonbandlimited signals," *Signal Processing Letters, IEEE*, vol. 2, no. 5, pp. 89–91, may 1995.

[17] D. J. C. MacKay, "Introduction to Gaussian Processes," in *Neural Networks and Machine Learning*, ser. NATO ASI Series, C. M. Bishop, Ed. Kluwer Academic Press, 1998, pp. 133–166.

Author/s:
Alpcan, T

Title:
Information metrics for model selection in function estimation

Date:
2014-01-01

Citation:
Alpcan, T. (2014). Information metrics for model selection in function estimation. 2014 Australian Communications Theory Workshop, AusCTW 2014, pp.45-50. IEEE. https://doi.org/10.1109/AusCTW.2014.6766426.

Persistent Link:
http://hdl.handle.net/11343/241550