
Generalization Errors and Learning Curves for Regression with Multi-task Gaussian Processes

Kian Ming A. Chai

School of Informatics, University of Edinburgh,
10 Crichton Street, Edinburgh EH8 9AB, UK
k.m.a.chai@ed.ac.uk

Abstract

We provide some insights into how task correlations in multi-task Gaussian process (GP) regression affect the generalization error and the learning curve. We analyze the asymmetric two-tasks case, where a secondary task is to help the learning of a primary task. Within this setting, we give bounds on the generalization error and the learning curve of the primary task. Our approach admits intuitive understandings of the multi-task GP by relating it to single-task GPs. For the case of one-dimensional input-space under optimal sampling with data only for the secondary task, the limitations of multi-task GP can be quantified explicitly.

1 Introduction

Gaussian processes (GPs) (see e.g., [1]) have been applied to many practical problems. In recent years, a number of models for multi-task learning with GPs have been proposed to allow different tasks to leverage on one another [2–5]. While it is generally assumed that learning multiple tasks together is beneficial, we are not aware of any work that quantifies such benefits, other than PAC-based theoretical analysis for multi-task learning [6–8]. Following the tradition of the theoretical works on GPs in machine learning, our goal is to quantify the benefits using average-case analysis.

We concentrate on the asymmetric two-tasks case, where the secondary task is to help the learning of the primary task. Within this setting, the main parameters are (1) the degree of “relatedness” ρ between the two tasks, and (2) the ratio π_S of total training data for the secondary task. While higher $|\rho|$ and lower π_S is clearly more beneficial to the primary task, the extent and manner that this is so has not been clear. To address this, we measure the benefits using generalization error, learning curve and optimal error, and investigate the influence of ρ and π_S on these quantities.

We will give non-trivial lower and upper bounds on the generalization error and the learning curve. Both types of bounds are important in providing assurance on the quality of predictions: an upper bound provides an estimate of the amount of training data needed to attain a minimum performance level, while a lower bound provides an understanding of the limitations of the model [9]. Our approach relates multi-task GPs to single-task GPs and admits intuitive understandings of multi-task GPs. For one-dimensional input-space under optimal sampling with data only for the secondary task, we show the limit to which error for the primary task can be reduced. This dispels any misconception that abundant data for the secondary task can remedy no data for the primary task.

2 Preliminaries and problem statement

2.1 Multi-task GP regression model and setup

The multi-task Gaussian process regression model in [5] learns M related functions $\{f_m\}_{m=1}^M$ by placing a zero mean GP prior which directly induces correlations between tasks. Let y_m be an

observation of the m th function at \mathbf{x} . Then the model is given by

$$\langle f_m(\mathbf{x})f_{m'}(\mathbf{x}') \rangle \stackrel{\text{def}}{=} K_{mm}^f k^{\mathbf{x}}(\mathbf{x}, \mathbf{x}') \quad y_m \sim \mathcal{N}(f_m(\mathbf{x}), \sigma_m^2), \quad (1)$$

where $k^{\mathbf{x}}$ is a covariance function over inputs, and K^f is a positive semi-definite matrix of inter-task similarities, and σ_m^2 is the noise variance for the m th task.

The current focus is on the two tasks case, where the secondary task S is to help improve the performance of the primary task T ; this is the *asymmetric multi-task learning* as coined in [10]. We fix K^f to be a correlation matrix, and let the variance be explained fully by $k^{\mathbf{x}}$ (the converse has been done in [5]). Thus K^f is fully specified by the correlation $\rho \in [-1, 1]$ between the two tasks. We further fix the noise variances of the two tasks to be the same, say σ_n^2 . For the training data, there are n_T (resp. n_S) observations at locations X_T (resp. X_S) for task T (resp. S). We use $n \stackrel{\text{def}}{=} n_T + n_S$ for the total number of observations, $\pi_S \stackrel{\text{def}}{=} n_S/n$ for the proportion of observations for task S , and also $X \stackrel{\text{def}}{=} X_T \cup X_S$. The aim is to infer the noise-free response f_{T^*} for task T at \mathbf{x}_* . See Figure 1.

The covariance matrix of the noisy training data is $K(\rho) + \sigma_n^2 I$, where

$$K(\rho) \stackrel{\text{def}}{=} \begin{pmatrix} K_{TT}^{\mathbf{x}} & \rho K_{TS}^{\mathbf{x}} \\ \rho K_{ST}^{\mathbf{x}} & K_{SS}^{\mathbf{x}} \end{pmatrix}; \quad (2)$$

and $K_{TT}^{\mathbf{x}}$ (resp. $K_{SS}^{\mathbf{x}}$) is the matrix of covariances (due to $k^{\mathbf{x}}$) between locations in X_T (resp. X_S); $K_{TS}^{\mathbf{x}}$ is the matrix of cross-covariances from locations in X_T to locations in X_S ; and $K_{ST}^{\mathbf{x}}$ is $K_{TS}^{\mathbf{x}}$ transposed. The posterior variance at \mathbf{x}_* for task T is

$$\sigma_T^2(\mathbf{x}_*, \rho, \sigma_n^2, X_T, X_S) = k_{**} - \mathbf{k}_*^T (K(\rho) + \sigma_n^2 I)^{-1} \mathbf{k}_*, \quad \text{where } \mathbf{k}_*^T \stackrel{\text{def}}{=} ((\mathbf{k}_{T^*}^{\mathbf{x}})^T \rho (\mathbf{k}_{S^*}^{\mathbf{x}})^T); \quad (3)$$

and k_{**} is the prior variance at \mathbf{x}_* , and $\mathbf{k}_{T^*}^{\mathbf{x}}$ (resp. $\mathbf{k}_{S^*}^{\mathbf{x}}$) is the vector of covariances (due to $k^{\mathbf{x}}$) between locations in X_T (resp. X_S) and \mathbf{x}_* . Where appropriate and clear from context, we will suppress some of the parameters in $\sigma_T^2(\mathbf{x}_*, \rho, \sigma_n^2, X_T, X_S)$, or use X for (X_T, X_S) . Note that $\sigma_T^2(\rho) = \sigma_T^2(-\rho)$, so that $\sigma_T^2(1)$ is the same as $\sigma_T^2(-1)$; for brevity, we only write the former.

If the GP prior is correctly specified, then the posterior variance (3) is also the generalization error at \mathbf{x}_* [1, §7.3]. The latter is defined as $\langle (f_T^*(\mathbf{x}_*) - \hat{f}_T(\mathbf{x}_*))^2 \rangle_{f_T^*}$, where $\hat{f}_T(\mathbf{x}_*)$ is the posterior mean at \mathbf{x}_* for task T , and the expectation is taken over the distribution from which the true function f_T^* is drawn. In this paper, in order to distinguish succinctly from the generalization error introduced in the next section, we use posterior variance to mean the generalization error at \mathbf{x}_* . Note that the actual y -values observed at X do not effect the posterior variance at any test location.

Problem statement Given the above setting, the aim is to investigate how training observations for task S can benefit the predictions for task T . We measure the benefits using generalization error, learning curve and optimal error, and investigate how these quantities vary with ρ and π_S .

2.2 Generalization errors, learning curves and optimal errors

We outline the general approach to obtain the generalization error and the learning curve [1, §7.3] under our setting, where we have two tasks and are concerned with the primary task T . Let $p(\mathbf{x})$ be the probability density, common to both tasks, from which test and training locations are drawn, and assume that the GP prior is correctly specified. The generalization error for task T is obtained by averaging the posterior variance for task T over \mathbf{x}_* , and the learning curve for task T is obtained by averaging the generalization error over training sets X :

$$\text{generalization error:} \quad \epsilon_T(\rho, \sigma_n^2, X_T, X_S) \stackrel{\text{def}}{=} \int \sigma_T^2(\mathbf{x}_*, \rho, \sigma_n^2, X_T, X_S) p(\mathbf{x}_*) d\mathbf{x}_* \quad (4)$$

$$\text{learning curve:} \quad \epsilon_T^{\text{avg}}(\rho, \sigma_n^2, \pi_S, n) \stackrel{\text{def}}{=} \int \epsilon_T(\rho, \sigma_n^2, X_T, X_S) p(X) dX, \quad (5)$$

where the training locations in X are drawn i.i.d, that is, $p(X)$ factorizes completely into a product of $p(\mathbf{x})$ s. Besides averaging ϵ_T to obtain the learning curve, one may also use the optimal experimental design methodology and minimize ϵ_T over X to find the optimal generalization error [11, chap. II]:

$$\text{optimal error:} \quad \epsilon_T^{\text{opt}}(\rho, \sigma_n^2, \pi_S, n) \stackrel{\text{def}}{=} \min_X \epsilon_T(\rho, \sigma_n^2, X_T, X_S). \quad (6)$$

Both $\epsilon_T(0, \sigma_n^2, X_T, X_S)$ and $\epsilon_T(1, \sigma_n^2, X_T, X_S)$ reduce to single-task GP cases; the former discards training observations at X_S , while the latter includes them. Similar analogues to single-task GP cases for $\epsilon_T^{\text{avg}}(0, \sigma_n^2, \pi_S, n)$ and $\epsilon_T^{\text{avg}}(1, \sigma_n^2, \pi_S, n)$, and $\epsilon_T^{\text{opt}}(0, \sigma_n^2, \pi_S, n)$ and $\epsilon_T^{\text{opt}}(1, \sigma_n^2, \pi_S, n)$ can be obtained. Note that ϵ_T^{avg} and ϵ_T^{opt} are well-defined since $\pi_S n = n_S \in \mathbb{N}_0$ by the definition of π_S .

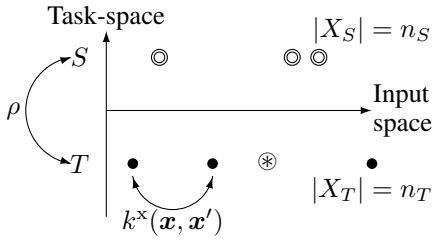


Figure 1: The two tasks S and T have task correlation ρ . The data set X_T (resp. X_S) for task T (resp. S) consists of the \bullet s (resp. \odot s). The test location \mathbf{x}_* for task T is denoted by \otimes .

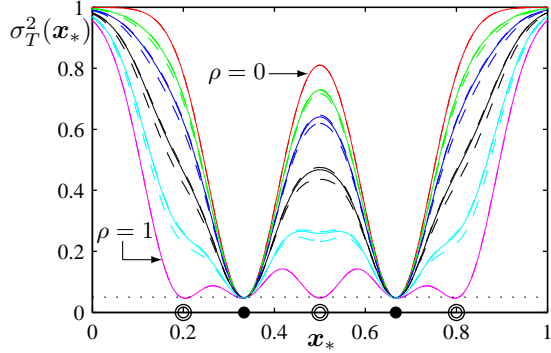


Figure 2: The posterior variances of each test location within $[0, 1]$ given data \bullet s at $1/3$ and $2/3$ for task T , and \odot s at $1/5$, $1/2$ and $4/5$ for task S .

2.3 Eigen-analysis

We now state known results of eigen-analysis used in this paper. Let $\bar{\kappa} \stackrel{\text{def}}{=} \kappa_1 > \kappa_2 > \dots$ and $\phi_1(\cdot), \phi_2(\cdot), \dots$ be the eigenvalues and eigenfunctions of the covariance function k^x under the measure $p(\mathbf{x})d\mathbf{x}$: they satisfy the integral equation $\int k^x(\mathbf{x}, \mathbf{x}')\phi_i(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \kappa_i\phi_i(\mathbf{x}')$. Let $\bar{\lambda} \stackrel{\text{def}}{=} \lambda_1 > \lambda_2 > \dots > \lambda_{n_S} \stackrel{\text{def}}{=} \lambda$ be the eigenvalues of $K_{S,S}^x$. If the locations in X_S are sampled from $p(\mathbf{x})$, then $\kappa_i = \lim_{n_S \rightarrow \infty} \lambda_i/n_S$, $i = 1 \dots n_S$; see e.g., [1, §4.3.2] and [12, Theorem 3.4]. However, for finite n_S used in practice, the estimate λ_i/n_S for κ_i is better for the larger eigenvalues than for the smaller ones. Additionally, in one-dimension with uniform $p(\mathbf{x})$ on the unit interval, if k^x satisfies the Sacks-Ylvisaker conditions of order r , then $\kappa_i \propto (\pi i)^{-2r-2}$ in the limit $i \rightarrow \infty$ [11, Proposition IV.10, Remark IV.2]. Broadly speaking, an order r process is exactly r times mean square differentiable. For example, the stationary Ornstein-Uhlenbeck process is of order $r = 0$.

3 Generalization error

In this section, we derive expressions for the generalization error (and the bounds thereon) for the two-tasks case in terms of the single-task one. To illustrate and further motivate the problem, Figure 2 plots the posterior variance $\sigma_T^2(\mathbf{x}_*, \rho)$ as a function of \mathbf{x}_* given two observations for task T and three observations for task S . We roughly follow [13, Fig. 2], and use squared exponential covariance function with length-scale 0.11 and noise variance $\sigma_n^2 = 0.05$. Six solid curves are plotted, corresponding, from top to bottom, to $\rho^2 = 0, 1/8, 1/4, 1/2, 3/4$ and 1. The two dashed curves enveloping each solid curve are the lower and upper bounds derived in this section; the dashed curves are hardly visible because the bounds are rather tight. The dotted line is the prior noise variance.

Similar to the case of single-task learning, each training point creates a depression on the $\sigma_T^2(\mathbf{x}_*, \rho)$ surface [9, 13]. However, while each training point for task T creates a “full” depression that reaches the prior noise variance (horizontal dotted line at 0.05), the depression created by each training point for task S depends on ρ , “deeper” depressions for larger ρ^2 . From the figure, and also from definition, it is clear that the following trivial bounds on $\sigma_T^2(\mathbf{x}_*, \rho)$ hold:

Proposition 1. For all \mathbf{x}_* , $\sigma_T^2(\mathbf{x}_*, 1) \leq \sigma_T^2(\mathbf{x}_*, \rho) \leq \sigma_T^2(\mathbf{x}_*, 0)$.

Integrating wrt to \mathbf{x}_* then gives the following corollary:

Corollary 2. $\epsilon_T(1, \sigma_n^2, X_T, X_S) \leq \epsilon_T(\rho, \sigma_n^2, X_T, X_S) \leq \epsilon_T(0, \sigma_n^2, X_T, X_S)$.

Sections 3.2 and 3.3 derive lower and upper bounds that are tighter than the above trivial bounds. Prior to the bounds, we consider a degenerate case to illustrate the limitations of multi-task learning.

3.1 The degenerate case of no training data for primary task

It is clear that if there is no training data for the secondary task, that is, if $X_S = \emptyset$, then $\sigma_T^2(\mathbf{x}_*, 1) = \sigma_T^2(\mathbf{x}_*, \rho) = \sigma_T^2(\mathbf{x}_*, 0)$ for all \mathbf{x}_* and ρ . In the converse case where there is no training data for the primary task, that is, $X_T = \emptyset$, we instead have the following proposition:

Proposition 3. For all \mathbf{x}_* , $\sigma_T^2(\mathbf{x}_*, \rho, \emptyset, X_S) = \rho^2 \sigma_T^2(\mathbf{x}_*, 1, \emptyset, X_S) + (1 - \rho^2)k_{**}$.

Proof.
$$\begin{aligned} \sigma_T^2(\mathbf{x}_*, \rho, \emptyset, X_S) &= k_{**} - \rho^2 (\mathbf{k}_{S^*}^x)^\top (K_{S^*}^x + \sigma_n^2 I)^{-1} \mathbf{k}_{S^*}^x \\ &= (1 - \rho^2)k_{**} + \rho^2 [k_{**} - (\mathbf{k}_{S^*}^x)^\top (K_{S^*}^x + \sigma_n^2 I)^{-1} \mathbf{k}_{S^*}^x] \\ &= (1 - \rho^2)k_{**} + \rho^2 \sigma_T^2(\mathbf{x}_*, 1, \emptyset, X_S). \quad \square \end{aligned}$$

Hence the posterior variance is a weighted average of the prior variance k_{**} and the posterior variance at perfect correlation. When the cardinality of X_S increases under infill asymptotics [14, §3.3],

$$\lim_{n_S \rightarrow \infty} \sigma_T^2(\mathbf{x}_*, 1, \emptyset, X_S) = 0 \quad \implies \quad \lim_{n_S \rightarrow \infty} \sigma_T^2(\mathbf{x}_*, \rho, \emptyset, X_S) = (1 - \rho^2)k_{**}. \quad (7)$$

This is the limit for the posterior variance at *any* test location for task T , if one has training data only for the secondary task S . This is because a correlation of ρ between the tasks prevents any training location for task S from having correlation higher than ρ with a test location for task T . Suppose correlations in the input-space are given by an isotropic covariance function $k^x(|\mathbf{x} - \mathbf{x}'|)$. If we translate correlations into distances between data locations, then any training location from task S is beyond a certain radius from any test location for task T . In contrast, a training location from task T may lay arbitrarily close to a test location for task T , subject to the constraints of noise.

We obtain the generalization error in this degenerate case, by integrating Proposition 3 wrt $p(\mathbf{x}_*)d\mathbf{x}_*$ and using the fact that the mean prior variance is given by the sum of the process eigenvalues.

Corollary 4. $\epsilon_T(\rho, \sigma_n^2, \emptyset, X_S) = \rho^2 \epsilon_T(1, \sigma_n^2, \emptyset, X_S) + (1 - \rho^2) \sum_{i=1}^{\infty} \kappa_i$.

3.2 A lower bound

When $X_T \neq \emptyset$, the correlations between locations in X_T and locations in X_S complicate the situation. However, since $\sigma_T^2(\rho)$ is a continuous and monotonically decreasing function of ρ , there exists an $\alpha \in [0, 1]$, which depends on ρ , \mathbf{x}_* and X , such that $\sigma_T^2(\rho) = \alpha \sigma_T^2(1) + (1 - \alpha) \sigma_T^2(0)$. That α depends on \mathbf{x}_* obstructs further analysis. The next proposition gives a lower bound $\underline{\sigma}_T^2(\rho)$ of the same form satisfying $\sigma_T^2(1) \leq \underline{\sigma}_T^2(\rho) \leq \sigma_T^2(\rho)$, where the mixing proportion is independent of \mathbf{x}_* .

Proposition 5. Let $\underline{\sigma}_T^2(\mathbf{x}_*, \rho) \stackrel{\text{def}}{=} \rho^2 \sigma_T^2(\mathbf{x}_*, 1) + (1 - \rho^2) \sigma_T^2(\mathbf{x}_*, 0)$. Then for all \mathbf{x}_* :

- (a) $\underline{\sigma}_T^2(\mathbf{x}_*, \rho) \leq \sigma_T^2(\mathbf{x}_*, \rho)$
- (b) $\sigma_T^2(\mathbf{x}_*, \rho) - \underline{\sigma}_T^2(\mathbf{x}_*, \rho) \leq \rho^2 (\sigma_T^2(\mathbf{x}_*, 0) - \sigma_T^2(\mathbf{x}_*, 1))$
- (c) $\arg \max_{\rho^2} [\sigma_T^2(\mathbf{x}_*, \rho) - \underline{\sigma}_T^2(\mathbf{x}_*, \rho)] \geq 1/2$.

The proofs are in supplementary material §S.2. The lower bound $\underline{\sigma}_T^2(\rho)$ depends explicitly on ρ^2 . It depends implicitly on π_S , which is the proportion of observations for task S , through the gap between $\sigma_T^2(1)$ and $\sigma_T^2(0)$. If there is no training data for the primary task, i.e., if $\pi_S = 1$, the bound reduces to Proposition 3, and becomes exact for all values of ρ . If $\pi_S = 0$, the bound is also exact. For $\pi_S \notin \{0, 1\}$, the bound is exact when $\rho \in \{-1, 0, 1\}$. As from Figure 2 and later from our simulation results in section 5.3, this bound is rather tight. Part (b) of the proposition states the tightness of the bound: it is no more than factor ρ^2 of the gap between the trivial bounds $\sigma_T^2(0)$ and $\sigma_T^2(1)$. Part (c) of the proposition says that the bound is least tight for a value of ρ^2 greater than 1/2.

We provide an intuition on Proposition 5a. Let \bar{f}_1 (resp. \bar{f}_0) be the posterior mean of the single-task GP when $\rho = 1$ (resp. $\rho = 0$). Contrasted with the multi-task predictor \bar{f}_T , \bar{f}_1 directly involves the noisy observations for task T at X_S , so it has more information on task T . Hence, predicting $\bar{f}_1(\mathbf{x}_*)$ gives the trivial lower bound $\sigma_T^2(1)$ on $\sigma_T^2(\rho)$. The tighter bound $\underline{\sigma}_T^2(\rho)$ is obtained by “throwing away” information and predicting $\bar{f}_1(\mathbf{x}_*)$ with probability ρ^2 and $\bar{f}_0(\mathbf{x}_*)$ with probability $(1 - \rho^2)$.

Finally, the next corollary is readily obtained from Proposition 5a by integrating wrt $p(\mathbf{x}_*)d\mathbf{x}_*$. This is possible because ρ is independent of \mathbf{x}_* .

Corollary 6. Let $\underline{\epsilon}_T(\rho, \sigma_n^2, X_T, X_S) \stackrel{\text{def}}{=} \rho^2 \epsilon_T(1, \sigma_n^2, X_T, X_S) + (1 - \rho^2) \epsilon_T(0, \sigma_n^2, X_T, X_S)$. Then $\underline{\epsilon}_T(\rho, \sigma_n^2, X_T, X_S) \leq \epsilon_T(\rho, \sigma_n^2, X_T, X_S)$.

3.3 An upper bound via equivalent isotropic noise at X_S

The following question motivates our upper bound: if the training locations in X_S had been observed for task T rather than for task S , *what is the variance $\tilde{\sigma}_n^2$ of the equivalent isotropic noise at X_S so*

that the posterior variance remains the same? To answer this question, we first refine the definition of $\sigma_T^2(\cdot)$ to include a different noise variance parameter s^2 for the X_S observations:

$$\sigma_T^2(\mathbf{x}_*, \rho, \sigma_n^2, s^2, X_T, X_S) \stackrel{\text{def}}{=} \mathbf{k}_{**} - \mathbf{k}_*^T \left[K(\rho) + \begin{pmatrix} \sigma_n^2 I & 0 \\ 0 & s^2 I \end{pmatrix} \right]^{-1} \mathbf{k}_*; \quad (8)$$

cf. (3). We may suppress the parameters \mathbf{x}_* , X_T and X_S when writing $\sigma_T^2(\cdot)$. The variance $\tilde{\sigma}_n^2$ of the equivalent isotropic noise is a function of \mathbf{x}_* defined by the equation

$$\sigma_T^2(\mathbf{x}_*, 1, \sigma_n^2, \tilde{\sigma}_n^2) = \sigma_T^2(\mathbf{x}_*, \rho, \sigma_n^2, \sigma_n^2). \quad (9)$$

For any \mathbf{x}_* there is always a $\tilde{\sigma}_n^2$ that satisfies the equation because the difference

$$\Delta(\rho, \sigma_n^2, s^2) \stackrel{\text{def}}{=} \sigma_T^2(\rho, \sigma_n^2, \sigma_n^2) - \sigma_T^2(1, \sigma_n^2, s^2) \quad (10)$$

is a continuous and monotonically decreasing function of s^2 . To make progress, we seek an upper bound $\bar{\sigma}_n^2$ for $\tilde{\sigma}_n^2$ that is independent of the choice of \mathbf{x}_* : $\Delta(\rho, \sigma_n^2, \bar{\sigma}_n^2) \leq 0$ for all test locations. Of interest is the tight upper bound $\bar{\sigma}_n^2$, which is the minimum possible $\tilde{\sigma}_n^2$, given in the next proposition.

Proposition 7. *Let $\bar{\lambda}$ be the maximum eigenvalue of K_{SS}^x , $\beta \stackrel{\text{def}}{=} \rho^{-2} - 1$ and $\bar{\sigma}_n^2 \stackrel{\text{def}}{=} \beta(\bar{\lambda} + \sigma_n^2) + \sigma_n^2$. Then for all \mathbf{x}_* , $\sigma_T^2(\mathbf{x}_*, \rho, \sigma_n^2, \sigma_n^2) \leq \sigma_T^2(\mathbf{x}_*, 1, \sigma_n^2, \bar{\sigma}_n^2)$. The bound is tight in this sense: for any $\bar{\sigma}_n^2$, if $\forall \mathbf{x}_* \sigma_T^2(\mathbf{x}_*, \rho, \sigma_n^2, \sigma_n^2) \leq \sigma_T^2(\mathbf{x}_*, 1, \sigma_n^2, \bar{\sigma}_n^2)$, then $\forall \mathbf{x}_* \sigma_T^2(\mathbf{x}_*, \rho, \sigma_n^2, \bar{\sigma}_n^2) \leq \sigma_T^2(\mathbf{x}_*, 1, \sigma_n^2, \bar{\sigma}_n^2)$.*

Proof sketch. Matrix $K(\rho)$ may be factorized as

$$K(\rho) = \begin{pmatrix} I & 0 \\ 0 & \rho I \end{pmatrix} \begin{pmatrix} K_{TT}^x & K_{TS}^x \\ K_{ST}^x & \rho^{-2} K_{SS}^x \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & \rho I \end{pmatrix}. \quad (11)$$

By using this factorization in the posterior variance (8) and taking out the $\begin{pmatrix} I & 0 \\ 0 & \rho I \end{pmatrix}$ factors, we obtain

$$\sigma_T^2(\rho, \sigma_n^2, s^2) = \mathbf{k}_{**} - (\mathbf{k}_*^x)^T [\Sigma(\rho, \sigma_n^2, s^2)]^{-1} \mathbf{k}_*^x, \quad (12)$$

where $(\mathbf{k}_*^x)^T \stackrel{\text{def}}{=} ((\mathbf{k}_{T*}^x)^T, (\mathbf{k}_{S*}^x)^T)$ and

$$\Sigma(\rho, \sigma_n^2, s^2) \stackrel{\text{def}}{=} \begin{pmatrix} K_{TT}^x & K_{TS}^x \\ K_{ST}^x & \rho^{-2} K_{SS}^x \end{pmatrix} + \begin{pmatrix} \sigma_n^2 I & 0 \\ 0 & \rho^{-2} s^2 I \end{pmatrix} = \Sigma(1, \sigma_n^2, s^2) + \beta \begin{pmatrix} 0 & 0 \\ 0 & K_{SS}^x + s^2 I \end{pmatrix}.$$

The second expression for Σ makes clear that, in the terms of $\sigma_T^2(\rho, \sigma_n^2, \sigma_n^2)$, having data X_S for task S is equivalent to an additional correlated noise at these observations for task T . This expression motivates the question that began this section. Note that $\rho^{-2} \geq 1$, and hence $\beta \geq 0$.

The increase in posterior variance due to having X_S at task S with noise variance σ_n^2 rather than having them at task T with noise variance s^2 is given by $\Delta(\rho, \sigma_n^2, s^2)$, which we may now write as

$$\Delta(\rho, \sigma_n^2, s^2) = (\mathbf{k}_*^x)^T [(\Sigma(1, \sigma_n^2, s^2))^{-1} - (\Sigma(\rho, \sigma_n^2, \sigma_n^2))^{-1}] \mathbf{k}_*^x. \quad (13)$$

Recall that we seek an upper bound $\bar{\sigma}_n^2$ for $\tilde{\sigma}_n^2$ such that $\Delta(\rho, \sigma_n^2, \bar{\sigma}_n^2) \leq 0$ for all test locations. In general, this requires $\bar{\sigma}_n^2 \stackrel{\text{def}}{=} \beta(\bar{\lambda} + \sigma_n^2) + \sigma_n^2 \leq \tilde{\sigma}_n^2$; details can be found in supplementary material §S.3. The tightness $\bar{\sigma}_n^2$ is evident from the construction. \square

Intuitively, $\sigma_T^2(\mathbf{x}_*, 1, \sigma_n^2, \bar{\sigma}_n^2)$ is the tight upper bound because it inflates the noise (co)variance at X_S just sufficiently, from $(\beta K_{SS}^x + \sigma_n^2 I / \rho^2)$ to $\bar{\sigma}_n^2 I$. Analogously, the tight lower bound on $\tilde{\sigma}_n^2$ is given by $\underline{\sigma}_n^2 \stackrel{\text{def}}{=} \beta(\underline{\lambda} + \sigma_n^2) + \sigma_n^2$. In summary, $\rho^{-2} \sigma_n^2 \leq \underline{\sigma}_n^2 \leq \tilde{\sigma}_n^2 \leq \bar{\sigma}_n^2 \leq \bar{\sigma}_n^2$, where the first inequality is obtained by substituting in zero for λ in $\underline{\sigma}_n^2$. Hence observing X_S at S is at most as “noisy” as an additional $\beta(\bar{\lambda} + \sigma_n^2)$ noise variance, and at least as “noisy” as an additional $\beta(\underline{\lambda} + \sigma_n^2)$ noise variance. Since β decreases with $|\rho|$, the additional noise variances are smaller when $|\rho|$ is larger, i.e., when the task S is more correlated with task T .

We give a description of how the above bounds scale with n_S , using the results stated in section 2.3. For large enough n_S , we may write $\bar{\lambda} \approx n_S \bar{\kappa}$ and $\underline{\lambda} \approx n_S \kappa_{n_S}$. Furthermore, for uniformly distributed inputs in the one-dimension unit interval, if the covariance function satisfies Sacks-Ylvisaker conditions of order r , then $\kappa_{n_S} = \Theta((\pi n_S)^{-2r-2})$, so that $\bar{\lambda} = \Theta((\pi n_S)^{-2r-1})$. Since $\bar{\sigma}_n^2$ and $\underline{\sigma}_n^2$ are linear in $\bar{\lambda}$ and $\underline{\lambda}$, we have $\bar{\sigma}_n^2 = \rho^{-2} \sigma_n^2 + \beta \Theta(n_S)$ and $\underline{\sigma}_n^2 = \rho^{-2} \sigma_n^2 + \beta \Theta(n_S^{-2r-1})$. For the upper bound $\bar{\sigma}_n^2$, note that although it scales linearly with n_S , the eigenvalues of $K(1)$ scales with n , thus $\sigma_T^2(1, \sigma_n^2, \bar{\sigma}_n^2)$ depends on $\pi_S \stackrel{\text{def}}{=} n_S/n$. In contrast the lower bound $\underline{\sigma}_n^2$ is dominated by $\rho^{-2} \sigma_n^2$, so that $\sigma_T^2(1, \sigma_n^2, \underline{\sigma}_n^2)$ does not depend on π_S even for moderate sizes n_S . Therefore, the lower bound is not as useful as the upper bound.

Finally, if we refine ϵ_T as we have done for σ_T^2 in (8), we obtain the following corollary:

Corollary 8. *Let $\bar{\epsilon}_T(\rho, \sigma_n^2, \sigma_n^2, X_T, X_S) \stackrel{\text{def}}{=} \epsilon_T(1, \sigma_n^2, \bar{\sigma}_n^2, X_T, X_S)$. Then $\bar{\epsilon}_T(\rho, \sigma_n^2, \sigma_n^2, X_T, X_S) \geq \epsilon_T(\rho, \sigma_n^2, \sigma_n^2, X_T, X_S)$.*

3.4 Exact computation of generalization error

The factorization of σ_T^2 expressed by (12) allows the generalization error to be computed exactly in certain cases. We replace the quadratic form in (12) by matrix trace and then integrate out \mathbf{x}_* to give

$$\epsilon_T(\rho, \sigma_n^2, X_T, X_S) = \langle k_{**} \rangle - \text{tr}(\Sigma^{-1} \langle \mathbf{k}_*^x (\mathbf{k}_*^x)^T \rangle) = \sum_{i=1}^{\infty} \kappa_i - \text{tr}(\Sigma^{-1} M),$$

where Σ denotes $\Sigma(\rho, \sigma_n^2, \sigma_n^2)$, the expectations are taken over \mathbf{x}_* , and M is an n -by- n matrix with

$$M_{pq} \stackrel{\text{def}}{=} \int k^x(\mathbf{x}_p, \mathbf{x}_*) k^x(\mathbf{x}_q, \mathbf{x}_*) p(\mathbf{x}_*) d\mathbf{x}_* = \sum_{i=1}^{\infty} \kappa_i^2 \phi_i(\mathbf{x}_p) \phi_i(\mathbf{x}_q), \quad \text{where } \mathbf{x}_p, \mathbf{x}_q \in X.$$

When the eigenfunctions $\phi_i(\cdot)$ s are not bounded, the infinite-summation expression for M_{pq} is often difficult to use. Nevertheless, analytical results for M_{pq} are still possible in some cases using the integral expression. An example is the case of the squared exponential covariance function with normally distributed \mathbf{x} , when the integrand is a product of three Gaussians.

4 Optimal error for the degenerate case of no training data for primary task

If training examples are provided only for task S , then task T has the following optimal performance.

Proposition 9. *Under optimal sampling on a 1-d space, if the covariance function satisfies Sacks-Ylvisaker conditions of order r , then $\epsilon_T^{\text{opt}}(\rho, \sigma^2, 1, n) = \Theta(n_S^{-(2r+1)/(2r+2)}) + (1 - \rho^2) \sum_{i=1}^{\infty} \kappa_i$.*

Proof. We obtain $\epsilon_T^{\text{opt}}(\rho, \sigma^2, 1, n) = \rho^2 \epsilon_T^{\text{opt}}(1, \sigma_n^2, 1, n) + (1 - \rho^2) \sum_{i=1}^{\infty} \kappa_i$ by minimizing Corollary 4 wrt X_S . Under the same conditions as the proposition, the optimal generalization error using the single-task GP decays with training set size n as $\Theta(n^{-(2r+1)/(2r+2)})$ [11, Proposition V.3]. Thus $\rho^2 \epsilon_T^{\text{opt}}(1, \sigma_n^2, 1, n) = \rho^2 \Theta(n_S^{-(2r+1)/(2r+2)}) = \Theta(n_S^{-(2r+1)/(2r+2)})$. \square

A directly corollary of the above result is that one cannot expect to do better than $(1 - \rho^2) \sum \kappa_i$ on the average. As this is a lower bound, the same can be said for incorrectly specified GP priors.

5 Theoretical bounds on learning curve

Using the results from section 3, lower and upper bounds on the learning curve may be computed by averaging over the choice of X using Monte Carlo approximation.¹ For example, using Corollary 2 and integrating wrt $p(X)dX$ gives the following trivial bounds on the learning curve:

Corollary 10. $\epsilon_T^{\text{avg}}(1, \sigma_n^2, \pi_S, n) \leq \epsilon_T^{\text{avg}}(\rho, \sigma_n^2, \pi_S, n) \leq \epsilon_T^{\text{avg}}(0, \sigma_n^2, \pi_S, n)$.

The gap between the trivial bounds can be analyzed as follows. Recall that $\pi_S n \in \mathbb{N}_0$ by definition, so that $\epsilon_T^{\text{avg}}(1, \sigma_n^2, \pi_S, (1 - \pi_S)n) = \epsilon_T^{\text{avg}}(0, \sigma_n^2, \pi_S, n)$. Therefore $\epsilon_T^{\text{avg}}(1, \sigma_n^2, \pi_S, n)$ is equivalent to $\epsilon_T^{\text{avg}}(0, \sigma_n^2, \pi_S, n)$ scaled along the n -axis by the factor $(1 - \pi_S) \in [0, 1]$, and hence the gap between the trivial bounds becomes wider with π_S .

In the rest of this section, we derive non-trivial theoretical bounds on the learning curve before providing simulation results. Theoretical bounds are particularly attractive for high-dimensional input-spaces, on which Monte Carlo approximation is harder.

5.1 Lower bound

For the single-task GP, a lower bound on its learning curve is $\sigma_n^2 \sum_{i=1}^{\infty} \kappa_i / (\sigma_n^2 + n\kappa_i)$ [15]. We shall call this the single-task OV bound. This lower bound can be combined with Corollary 6.

Proposition 11. $\epsilon_T^{\text{avg}}(\rho, \sigma_n^2, \pi_S, n) \geq \rho^2 \sigma_n^2 \sum_{i=1}^{\infty} \frac{\kappa_i}{\sigma_n^2 + n\kappa_i} + (1 - \rho^2) \sigma_n^2 \sum_{i=1}^{\infty} \frac{\kappa_i}{\sigma_n^2 + (1 - \pi_S)n\kappa_i}$,

or equivalently, $\epsilon_T^{\text{avg}}(\rho, \sigma_n^2, \pi_S, n) \geq \sigma_n^2 \sum_{i=1}^{\infty} \frac{b_i^1 \kappa_i}{\sigma_n^2 + n\kappa_i}$, with $b_i^1 \stackrel{\text{def}}{=} \frac{\sigma_n^2 + (1 - \rho^2 \pi_S)n\kappa_i}{\sigma_n^2 + (1 - \pi_S)n\kappa_i}$,

or equivalently, $\epsilon_T^{\text{avg}}(\rho, \sigma_n^2, \pi_S, n) \geq \sigma_n^2 \sum_{i=1}^{\infty} \frac{b_i^0 \kappa_i}{\sigma_n^2 + (1 - \pi_S)n\kappa_i}$, with $b_i^0 \stackrel{\text{def}}{=} \frac{\sigma_n^2 + (1 - \rho^2 \pi_S)n\kappa_i}{\sigma_n^2 + n\kappa_i}$.

¹Approximate lower bounds are also possible, by combining Corollary 6 and approximations in, e.g., [13].

Proof sketch. To obtain the first inequality, we integrate Corollary 6 wrt to $p(X)dX$, and apply the single-task OV bound twice. For the second inequality, its i th summand is obtained by combining the corresponding pair of i th summands in the first inequality. The third inequality is obtained from the second by swapping the denominator of b_i^1 with that of $\kappa_i/(\sigma_n^2 + n\kappa_i)$ for every i . \square

For fixed σ_n^2 , π_S and n , denote the above bound by OV_ρ . Then OV_0 and OV_1 are both single task bounds. In particular, from Corollary 10, we have that the OV_1 is a lower bound on $\epsilon_T^{\text{avg}}(\rho, \sigma_n^2, \pi_S, n)$. From the first expression of the above proposition, it is clear from the ‘‘mixture’’ nature of the bound that the two-tasks bound OV_ρ is always better than OV_1 . As ρ^2 decreases, the two-tasks bound moves towards the OV_0 ; and as π_S increases, the gap between OV_0 and OV_1 increases. In addition, the gap is also larger for rougher processes, which are harder to learn. Therefore, the relative tightness of OV_ρ over OV_1 is more noticeable for lower ρ^2 , higher π_S and rougher processes.

The second expression in the Proposition 11 is useful for comparing with the OV_1 . Each summand for the two-tasks case is a factor b_i^1 of the corresponding summand for the single-task case. Since $b_i^1 \in [1, (1 - \rho^2\pi_S)/(1 - \pi_S)[$, OV_ρ is more than OV_1 by at most $(1 - \rho^2)\pi_S/(1 - \pi_S)$ times. Similarly, the third expression of the proposition is useful for comparing with OV_0 : each summand for the the two-tasks case is a factor $b_i^0 \in](1 - \rho^2\pi_S), 1]$ of the corresponding single-task one. Hence, OV_ρ is less than OV_0 by up to $\rho^2\pi_S$ times. In terms of the lower bound, this is the limit to which multi-task learning can outperform the single-task learning that ignores the secondary task.

5.2 Upper bound using equivalent noise

An upper bound on the learning curve of a single-task GP is given in [16]. We shall refer to this as the single-task FWO bound and combine it with the approach in section 3.3 to obtain an upper bound on the learning curve of task T . Although the single-task FWO bound was derived for observations with isotropic noise, with some modifications (see supplementary material §S.4), the derivations are still valid for observations with heteroscedastic and correlated noise. Below is a version of the FWO bound that has yet to assume isotropic noise:

Theorem 12. (*[16], modified second part of Theorem 6*) *Consider a zero-mean GP with covariance function $k^x(\cdot, \cdot)$, and eigenvalues κ_i and eigenfunctions $\phi_i(\cdot)$ under the measure $p(\mathbf{x})d\mathbf{x}$; and suppose that the noise (co)variances of the observations are given by $\gamma^2(\cdot, \cdot)$. For n observations $\{\mathbf{x}_i\}_{i=1}^n$, let H and Φ be matrices such that $H_{ij} \stackrel{\text{def}}{=} k^x(\mathbf{x}_i, \mathbf{x}_j) + \gamma^2(\mathbf{x}_i, \mathbf{x}_j)$ and $\Phi_{ij} \stackrel{\text{def}}{=} \phi_j(\mathbf{x}_i)$. Then the learning curve at n is upper-bounded by $\sum_{i=1}^\infty \kappa_i - n \sum_{i=1}^\infty \kappa_i^2/c_i$, where $c_i \stackrel{\text{def}}{=} \langle (\Phi^T H \Phi)_{ii} \rangle / n$, and the expectation in c_i is taken over the set of n input locations drawn independently from $p(\mathbf{x})$.*

Unlike [16], we do not assume that the noise variance $\gamma^2(\mathbf{x}_i, \mathbf{x}_j)$ is of the form $\sigma_n^2 \delta_{ij}$. Instead of proceeding from the upper bound $\sigma_T^2(1, \sigma_n^2, \bar{\sigma}_n^2)$, we proceed directly from the exact posterior variance given by (12). Thus we set the observation noise (co)variance $\gamma^2(\mathbf{x}_i, \mathbf{x}_j)$ to

$$\delta(\mathbf{x}_i \in X_T)\delta(\mathbf{x}_j \in X_T) \delta_{ij}\sigma_n^2 + \delta(\mathbf{x}_i \in X_S)\delta(\mathbf{x}_j \in X_S) [\beta k^x(\mathbf{x}_i, \mathbf{x}_j) + \rho^{-2}\delta_{ij}\sigma_n^2], \quad (14)$$

so that, through the definition of c_i in Theorem 12, we obtain

$$c_i = (1 + \beta\pi_S) \left\{ [(1 + \beta\pi_S^2)n/(1 + \beta\pi_S) - 1] \kappa_i + \int k^x(\mathbf{x}, \mathbf{x}) [\phi_i(\mathbf{x})]^2 p(\mathbf{x})d\mathbf{x} + \sigma_n^2 \right\}; \quad (15)$$

details are in the supplementary material §S.5. This leads to the following proposition:

Proposition 13. *Let $\beta \stackrel{\text{def}}{=} \rho^{-2} - 1$. Then, using the c_i s defined in (15), we have*

$$\epsilon_T^{\text{avg}}(\rho, \sigma_n^2, \pi_S, n) \leq \sum_{i=1}^\infty \kappa_i - n \sum_{i=1}^\infty \kappa_i^2/c_i.$$

Denote the above upper bound by FWO_ρ . When $\rho = \pm 1$ or $\pi_S = 0$, the single-task FWO upper bound is recovered. However, FWO_ρ with $\rho = 0$ gives the prior variance $\sum \kappa_i$ instead. A trivial upper bound can be obtained using Corollary 10, by replacing n with $(1 - \pi_S)n$ in the single-task FWO bound. The FWO_ρ bound is better than this trivial single-task bound for small n and high $|\rho|$.

5.3 Comparing bounds by simulations of learning curve

We compare our bounds with simulated learning curves. We follow the third scenario in [13]: the input space is one dimensional with Gaussian distribution $\mathcal{N}(0, 1/12)$, the covariance function is the

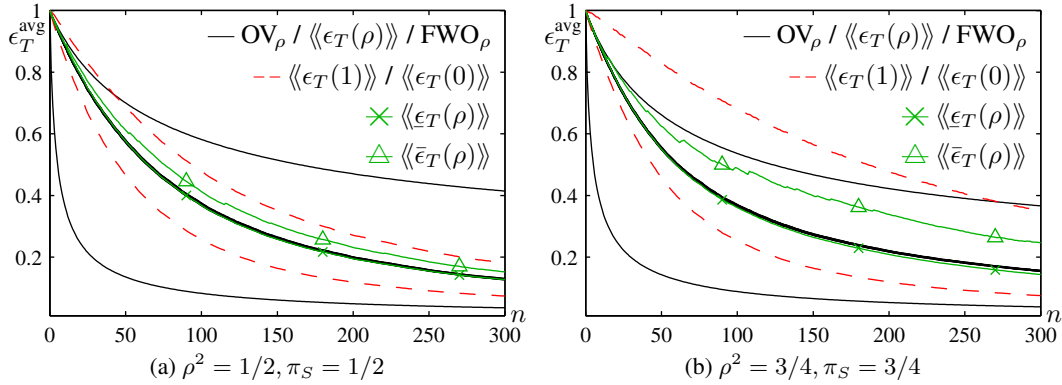


Figure 3: Comparison of various bounds for two settings of (ρ, π_S) . Each graph plots ϵ_T^{avg} against n and consists of the “true” multi-task learning curve (middle —), the theoretical lower/upper bounds of Propositions 11/13 (lower/upper —), the empirical trivial lower/upper bounds using Corollary 10 (lower/upper --), and the empirical lower/upper bounds using Corollaries 6/8 (\times / \triangle). The thickness of the “true” multi-task learning curve reflects 95% confidence interval.

unit variance squared exponential $k^x(x, x') = \exp[-(x - x')^2 / (2l^2)]$ with length-scale $l = 0.01$, the observation noise variance is $\sigma_n^2 = 0.05$, and the learning curves are computed for up to $n = 300$ training data points. When required, the average over x_* is computed analytically (see section 3.4). The empirical average over $X \stackrel{\text{def}}{=} X_T \cup X_S$, denoted by $\langle \cdot \rangle$, is computed over 100 randomly sampled training sets. The process eigenvalues κ_i s needed to compute the theoretical bounds are given in [17]. Supplementary material §S.6 gives further details.

Learning curves for pairwise combinations of $\rho^2 \in \{1/8, 1/4, 1/2, 3/4\}$ and $\pi_S \in \{1/4, 1/2, 3/4\}$ are computed. We compare the following: (a) the “true” multi-task learning curve $\langle \epsilon_T(\rho) \rangle$ obtained by averaging $\sigma_T^2(\rho)$ over x_* and X ; (b) the theoretical bounds OV_ρ and FWO_ρ of Propositions 11 and 13; (c) the trivial upper and lower bounds that are single-task learning curves $\langle \epsilon_T(0) \rangle$ and $\langle \epsilon_T(1) \rangle$ obtained by averaging $\sigma_T^2(0)$ and $\sigma_T^2(1)$; and (d) the empirical lower bound $\langle \epsilon_T(\rho) \rangle$ and upper bound $\langle \bar{\epsilon}_T(\rho) \rangle$ using Corollaries 6 and 8. Figure 3 gives some indicative plots of the curves.

We summarize with the following observations: (a) The gap between the trivial bounds $\langle \epsilon_T(0) \rangle$ and $\langle \epsilon_T(1) \rangle$ increases with π_S , as described at the start of section 5. (b) We find the lower bound $\langle \epsilon_T(\rho) \rangle$ a rather close approximation to the multi-task learning curve $\langle \epsilon_T(\rho) \rangle$, as evidenced by the much overlap between the \times lines and the middle — lines in Figure 3. (c) The curve for the empirical upper bound $\langle \bar{\epsilon}_T(\rho) \rangle$ using the equivalent noise method has jumps, e.g., the \triangle lines in Figure 3, because the equivalent noise variance $\bar{\sigma}_n^2$ increases whenever a datum for X_S is sampled. (d) For small n , $\langle \epsilon_T(\rho) \rangle$ is closer to FWO_ρ , but becomes closer to OV_ρ as n increases, as shown by the unmarked solid lines in Figure 3. This is because the theoretical lower bound OV_ρ is based on the asymptotically exact single-task OV bound and the $\epsilon_T(\rho)$ bound, which is observed to approximate the multi-task learning curve rather closely (point (b)).

Conclusions We have measured the influence of the secondary task on the primary task using the generalization error and the learning curve, parameterizing these with the correlation ρ between the two tasks, and the proportion π_S of observations for the secondary task. We have provided bounds on the generalization error and learning curves, and these bounds highlight the effects of ρ and π_S . This is a step towards understanding the role of the matrix K^f of inter-task similarities in multi-task GPs with more than two tasks. Analysis on the degenerate case of no training data for the primary task has uncovered an intrinsic limitation of multi-task GP. Our work contributes to an understanding of multi-task learning that is orthogonal to the existing PAC-based results in the literature.

Acknowledgments

I thank E Bonilla for motivating this problem, CKI Williams for helpful discussions and for proposing the equivalent isotropic noise approach, and DSO National Laboratories, Singapore, for financial support. This work is supported in part by the EU through the PASCAL2 Network of Excellence.

References

- [1] Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts, 2006.
- [2] Yee Whye Teh, Matthias Seeger, and Michael I. Jordan. Semiparametric latent factor models. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, pages 333–340. Society for Artificial Intelligence and Statistics, January 2005.
- [3] Edwin V. Bonilla, Felix V. Agakov, and Christopher K. I. Williams. Kernel Multi-task Learning using Task-specific Features. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*. Omni Press, March 2007.
- [4] Kai Yu, Wei Chu, Shipeng Yu, Volker Tresp, and Zhao Xu. Stochastic Relational Models for Discriminative Link Prediction. In B. Schölkopf, J. Platt, and T. Hofmann, editors, *Advances in Neural Information Processing Systems 19*, Cambridge, MA, 2007. MIT Press.
- [5] Edwin V. Bonilla, Kian Ming A. Chai, and Christopher K.I. Williams. Multi-task Gaussian process prediction. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.
- [6] Jonathan Baxter. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research*, 12:149–198, March 2000.
- [7] Andreas Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, January 2006.
- [8] Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine Learning*, 73(3):273–287, 2008.
- [9] Christopher K. I. Williams and Francesco Vivarelli. Upper and lower bounds on the learning curve for Gaussian processes. *Machine Learning*, 40(1):77–102, 2000.
- [10] Ya Xue, Xuejun Liao, Lawrence Carin, and Balaji Krishnapuram. Multi-task learning for classification with Dirichlet process prior. *Journal of Machine Learning Research*, 8:35–63, January 2007.
- [11] Klaus Ritter. *Average-Case Analysis of Numerical Problems*, volume 1733 of *Lecture Notes in Mathematics*. Springer, 2000.
- [12] Christopher T. H. Baker. *The Numerical Treatment of Integral Equations*. Clarendon Press, 1977.
- [13] Peter Sollich and Anason Halees. Learning curves for Gaussian process regression: Approximations and bounds. *Neural Computation*, 14(6):1393–1428, 2002.
- [14] Noel A. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993.
- [15] Manfred Opper and Francesco Vivarelli. General bounds on Bayes errors for regression with Gaussian processes. In Kearns et al. [18], pages 302–308.
- [16] Giancarlo Ferrari Trecate, Christopher K. I. Williams, and Manfred Opper. Finite-dimensional approximation of Gaussian processes. In Kearns et al. [18], pages 218–224.
- [17] Huaiyu Zhu, Christopher K. I. Williams, Richard Rohwer, and Michal Morciniec. Gaussian regression and optimal finite dimensional linear models. In Christopher M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *NATO ASI Series F: Computer and Systems Sciences*, pages 167–184. Springer-Verlag, Berlin, 1998.
- [18] Michael J. Kearns, Sara A. Solla, and David A. Cohn, editors. *Advances in Neural Information Processing Systems 11*, 1999. The MIT Press.