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

Authored by:

Kian M. Chai

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-task 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 Paper Body

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 PACbased 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? ? between the two tasks, and (2) the ratio ?S of total training data for the secondary task. While higher —?— and lower ?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 ? and ?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 2.1

Preliminaries and problem statement Multi-task GP regression model and setup

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

observation of the mth function at x. Then the model is given by x 0 f hfm (x)fm0 (x0)i def = Kmm0 k (x, x)

2 ym ? N (fm (x), ?m),

(1)

where k x is a covariance function over inputs, and K f is a positive semi-definite matrix of inter-task 2 similarities, and ?m is the noise variance for the mth 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 x (the converse has been done in [5]). Thus K f is fully specified by the correlation ? ? [?1, 1] between the two tasks. We further fix the noise variances of the two tasks to be the same, say ?n2 . For the training data, there are nT (resp. nS) observations at locations XT (resp. XS) for task T (resp. S). We use n def = nT + nS for the total number of observations, ?S def = nS /n for the proportion of observations for task S, and also X def = XT ? XS . The aim is to infer the noise-free response fT ? for task T at x? . See Figure 1. The covariance matrix of the noisy training data is K(?) + ?n2 I, where x

KT T ?KTx S K(?) def ; = ?K x x KSS ST

(2)

x and KTx T (resp. KSS) is the matrix of covariances (due to k x) between locations in XT (resp. XS); x x KT S is the matrix of cross-covariances from locations in XT to locations in XS ; and KST is KTx S transposed. The posterior variance at x? for task T is

x T x T 2 ?1 def ?T2 (x? , ?, ?n2 , XT , XS) = k?? ? kT k? , where kT ; (3) ? (K(?) + ?n I) ? = (kT ?) ?(kS?)

and k?? is the prior variance at x? , and kxT ? (resp. kxS?) is the vector of covariances (due to k x) between locations in XT (resp. XS) and x? . Where appropriate and clear from context, we will suppress some of the parameters in ?T2 (x? , ?, ?n2 , XT , XS), or use X for (XT , XS). Note that ?T2 (?) = ?T2 (??), so that ?T2 (1) is the same as ?T2 (?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 x? [1, ?7.3]. The latter is defined as h(fT? (x?) ? f?T (x?))2 ifT? , where f?T (x?) is the posterior mean at x? for task T , and the expectation is taken over the distribution from which the true function fT? 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 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 ? and ?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(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 x? , and the learning curve for task T is obtained by averaging the generalization error over training sets X: R generalization error: T (?, ?n2 , XT , XS) def (4) = ?T2 (x? , ?, ?n2 , XT , XS)p(x?)dx? R avg 2 2 def learning curve: T (?, ?n , ?S , n) = T (?, ?n , XT , XS)p(X)dX, (5) where the training locations in X are drawn i.i.d, that is, p(X) factorizes completely into a product of p(x)s. Besides averaging T to obtain the learning curve, one may also use the optimal experimental design methodology and minimize T over X to find the optimal generalization error [11, chap. II]: optimal error: T (0, ?n2 , XT , XS)

2 2 def opt T (?, ?n , ?S , n) = minX T (?, ?n , XT , XS).

T (1, ?n2 , XT , XS)

(6)

Both and reduce to single-task GP cases; the former discards training observations at XS , while the latter includes them. Similar analogues to single-task GP avg opt opt 2 2 2 2 cases for avg T (0, ?n , ?S , n) and T (1, ?n , ?S , n), and T (0, ?n , ?S , n) and T (1, ?n , ?S , n) can be avg opt obtained. Note that T and T are well-defined since ?S n = nS ? N0 by the definition of ?S . 2

1

Task-space

—XS — = nS

S Input space

?

T

—XT — = nT

k x (x, x0) Figure 1: The two tasks S and T have task correlation ?. The data set XT (resp. XS) for task T (resp. S) consists of the ?s (resp. s). The test location x? for task T is denoted by . 2.3

?T2 (x?) 0.8

?=0

0.6 0.4

?=1 0.2 0

0
0.2
0.4
x? 0.6
0.8
1

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

Eigen-analysis

We now state known results of eigen-analysis used in this paper. Let ? ? def = ?1 ¿ ?2 ¿ . . . and ?1 (?), ?2 (?), . . . be the eigenvalues and eigenfunctionsRof the covariance function k x under the measure p(x)dx: they satisfy the integral equation k x (x, x0)?i (x)p(x)dx = ?i ?i (x0). Let x ? def . If the locations in XS are sampled ? = ? be the eigenvalues of KSS = ?1 ¿ ?2 ¿ . . . ¿ ?nS def ? ? /n , i = 1 . . . n ; see e.g., [1, ?4.3.2] and [12, Theorem 3.4]. from p(x), then ?i = limnS ?? i S S However, for finite nS used in practice, the estimate ?i /nS for ?i is better for the larger eigenvalues than for the smaller ones. Additionally, in one-dimension with uniform p(x) on the unit interval, if k x satisfies the Sacks-Ylvisaker conditions of order r, then ?i ? (?i)?2r?2 in the limit i ? ? [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 ?T2 (x? , ?) as a function of 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 ?n2 = 0.05. Six solid curves are plotted, corresponding, from top to bottom, to ?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 ?T2 (x? , ?) 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 ?, ?deeper? depressions for larger ?2 . From the figure, and also from definition, it is clear that the following trivial bounds on ?T2 (x? , ?) hold: Proposition 1. For all x? , ?T2 (x? , 1) 6 ?T2 (x? , ?) 6 ?T2 (x? , 0). Integrating wrt to x? then gives the following corollary: Corollary 2. T (1, ?n2 , XT , XS) 6 T (?, ?n2 , XT , XS) 6 T (0, ?n2 , XT , XS). 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 XS = ?, then ?T2 (x? 1) = ?T2 (x? , ?) = ?T2 (x? 0) for all x? and ?. In the converse case where there is no training data for the primary task, that is, XT = ?, we instead have the following proposition: 3

Proposition 3. For all x? , ?T2 (x? , ?, ?, XS) = ?2 ?T2 (x? , 1, ?, XS) + (1 ? ?2)k?? . Proof.

x ?T2 (x? , ?, ?, XS) = k?? ? ?2 (kxS?)T (KSS + ?n2 I)?1 kxS?

x = (1 ? ?2)k?? + ?2 k?? ? (kxS?)T (KSS + ?n2 I)?1 kxS?

= (1 ? ?2)k?? + ?2 ?T2 (x? , 1, ?, XS). Hence the posterior variance is a weighted average of the prior variance k?? and the posterior variance at perfect correlation. When the cardinality of XS increases under infill asymptotics [14, ?3.3], limnS ?? ?T2 (x? , 1, ?, XS) = 0

=?

limnS ?? ?T2 (x? , ?, ?, XS) = (1 ? ?2)k?? .

(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 ? between the tasks prevents any training location for task S from having correlation higher than ? with a test location for task T . Suppose correlations in the input-space are given by an isotropic covariance function k x (—x ? x0 —). 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(x?)dx? and using the fact that the mean prior variance is given by the sum of the process eigenvalues. P? Corollary 4. T (?, ?n2 , ?, XS) = ?2 T (1, ?n2 , ?, XS) + (1 ? ?2) i=1 ?i . 3.2

A lower bound

When XT 6= ?, the correlations between locations in XT and locations in XS complicate the situation. However, since ?T2 (?) is a continuous and monotonically decreasing function of ?, there exists an ? ? [0, 1], which depends on ?, x? and X, such that ?T2 (?) = ??T2 (1) + (1 ? ?)?T2 (0). That ? depends on x? obstructs further analysis. The next proposition gives a lower bound ?T2 (?) of the ? of x? . same form satisfying ?T2 (1) 6 ?T2 (?) 6 ?T2 (?), where the mixing proportion is independent ? 2 2 2 2 2 def Proposition 5. Let ?T (x? , ?) = ? ?T (x? , 1) + (1 ? ?)?T (x? , 0). Then for all x? : ? (a) ?T2 (x? , ?) 6 ?T2 (x? , ?) ? (b) ?T2 (x? , ?) ? ?T2 (x? , ?) 6 ?2 (?T2 (x? , 0) ? ?T2 (x? , 1))

? (c) arg max?2 ?T2 (x? , ?) ? ?T2 (x? , ?) ¿ 1/2. ? The proofs are in supplementary material ?S.2. The lower bound ?T2 (?) depends explicitly on ?2 . ? for task S, through the gap It depends implicitly on ?S , which is the proportion of observations between ?T2 (1) and ?T2 (0). If there is no training data for the primary task, i.e., if ?S = 1, the bound reduces to Proposition 3, and becomes exact for all values of ?. If ?S = 0, the bound is also exact. For

?S 6? {0, 1}, the bound is exact when ? ? {?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 ?2 of the gap between the trivial bounds ?T2 (0) and ?T2 (1). Part (c) of the proposition says that the bound is least tight for a value of ?2 greater than 1/2. We provide an intuition on Proposition 5a. Let f?1 (resp. f?0) be the posterior mean of the single-task GP when ? = 1 (resp. ? = 0). Contrasted with the multi-task predictor f?T , f?1 directly involves the noisy observations for task T at XS , so it has more information on task T . Hence, predicting f?1 (x?) gives the trivial lower bound ?T2 (1) on ?T2 (?). The tighter bound ?T2 (?) is obtained by ?throwing ? away? information and predicting f?1 (x?) with probability ?2 and f?0 (x?) with probability (1 ? ?2). Finally, the next corollary is readily obtained from Proposition 5a by integrating wrt p(x?)dx? . This is possible because ? is independent of x? . Corollary 6. Let T (?, ?n2 , XT , XS) def = ?2 T (1, ?n2 , XT , XS) + (1 ? ?2)T (0, ?n2 , XT , XS). Then ? 2 2 T (?, ?n , XT , XS) 6 T (?, ?n , XT , XS). ? 3.3

An upper bound via equivalent isotropic noise at XS

The following question motivates our upper bound: if the training locations in XS had been observed for task T rather than for task S, what is the variance ? ? 2n of the equivalent isotropic noise at XS so 4

that the posterior variance remains the same? To answer this question, we first refine the definition of ?T2 (?) to include a different noise variance parameter s2 for the XS observations: h 2 i ?1 T ? I 0 ?T2 (x? , ?, ?n2 , s2 , XT , XS) def k? ; (8) = k?? ? k? K(?) + 0n s2 I cf. (3). We may suppress the parameters x? , XT and XS when writing ?T2 (?). The variance ? ? 2n of the equivalent isotropic noise is a function of x? defined by the equation ?T2 (x? , 1, ?n2 , ? ? 2n) = ?T2 (x? ?, ?n2 , ?n2). (9) 2 For any x? there is always a ? ? n that satisfies the equation because the difference 2 2 2 2 2 2 ?(?, ?n2 , s2) def (10) = ?T (?, ?n , ?n) ? ?T (1, ?n , s) 2 is a continuous and monotonically decreasing function of s . To make progress, we seek an upper bound ? ? 2n for ? ? 2n that is independent of the choice of x? : ?(?, ?n2 , ? ? 2n) 6 0 for all test locations. Of ? 2n , which is the minimum possible ? interest is the tight upper bound ? ? 2n , given in the next proposition. ? + ?2) + ?2 . ? be the maximum eigenvalue of K x , ? def ? 2n def Proposition 7. Let ? = ?(? = ??2 ? 1 and ? n n SS ? 2n). The bound is tight in this sense: for any Then for all x? , ?T2 (x? , ?, ?n2 , ?n2) 6 ?T2 (x? , 1, ?n2 , ? ? 2n) 6 ?T2 (x? , 1, ?n2 , ? ? ? 2n , if ?x? ?T2 (x? , ?, ?n2 , ?n2) 6 ?T2 (x? , 1, ?n2 , ? ? 2n), then ?x? ?T2 (x? , ?, ?n2 , ? ? 2n). Proof sketch. Matrix K(?) may be factorized as

x KT T I 0 K(?) = x 0 ?I KST

KTx S ?2 x ? KSS

I 0

0 . ?I I 0 0 ?I

By using this factorization in the posterior variance (8) and taking out the (kx?)T

def

?(?, ?n2 , s2)

def
where
=
=
?T2 (?, ?n2 , s2) = k?? ?
(kx)T , (kxS?)T and 2 Tx? KTx S KT T ?n I + x x ??2 KSS KST 0
(11)
factors, we obtain
(kx?)T [?(?, ?n2 , s2)]?1 kx? , 0 ??2 s2 I
=
?(1, ?n2 , s2)
(12)
0 0 . +? x + s2 I 0 KSS

The second expression for ? makes clear that, in the terms of ?T2 (?, ?n2 , ?n2), having data XS 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 ??2 ¿ 1, and hence ? ¿ 0. The increase in posterior variance due to having XS at task S with noise variance ?n2 rather than having them at task T with noise variance s2 is given by ?(?, ?n2 , s2), which we may now write as

?(?, ?n2 , s2) = (kx?)T (?(1, ?n2 , s2))?1 ? (?(?, ?n2 , ?n2))?1 kx? . (13) 2 2 2 2 Recall that we seek an upper bound ? ? n for ? ? n such that ?(?, ?n , ? ? n) 6 0 for all test locations. In ? + ?2) + ?2 6 ? ? 2n def ? 2n ; details can be found in supplementary material general, this requires ? = ?(? n n ? 2n is evident from the construction. ?S.3. The tightness ? ? 2n) is the tight upper bound because it inflates the noise (co)variance at XS Intuitively, ?T2 (x? , 1, ?n2 , ? x ? 2n I. Analogously, the tight lower bound on ? ? 2n is given + ?n2 I/?2) to ? just sufficiently, from (?KSS 2 2 2 2 2 2 ?2 2 2 def ?n 6 ? ?n 6 ? ? n , where the first inequality by =? n = ?(? + ?n) + ?n . In summary, ? ?n 6 =? n 6 ? ? is obtained by substituting in zero for ? in =? 2n . Hence observing XS at S is at most as ?noisy? as ? ? + ? 2) noise variance, an additional ?(? and at least as ?noisy? as an additional ?(? + ?n2) noise n ? —?— is larger, variance. Since ? decreases with —?—, the additional noise variances are smaller when i.e., when the task S is more correlated with task T . We give a description of how the above bounds scale with nS , using the results stated ? ? nS ? in section 2.3. For large enough nS , we may write ? ? and ? ? nS ?nS . Further? more, for uniformly distributed inputs in the one-dimension unit interval, if the covariance

function satisfies Sacks-Ylvisaker conditions of order r, then ? = ? (?n)?2r?2 , so that n S S

? and ?, we have ? ? 2n and =? 2n are linear in ? ? 2n = ??2 ?n2 + ? ?(nS) ? = ? (?nS)?2r?1 . Since ?

? ? ?2r?1 ? 2n , note that although it scales linearly and =? 2n = ??2 ?n2 + ? ? nS . For the upper bound ? 2 ? 2n) depends on ?S def with nS , the eigenvalues of K(1) scales with n, thus ?T (1, ?n2 , ? = nS /n. In contrast the lower bound =? 2n is dominated by ??2 ?n2 , so that ?T2 (1, ?n2 , =? 2n) does not depend on ?S even for moderate sizes nS . Therefore, the lower bound is not as useful as the upper bound. Finally, if we refine T as we have done for ?T2 in (8), we obtain the following corollary: ? 2n , XT , XS). Then Corollary 8. Let ?T (?, ?n2 , ?n2 , XT , XS) def = T (1, ?n2 , ? ?T (?, ?n2 , ?n2 , XT , XS) ¿ T (?, ?n2 , ?n2 , XT , XS). 5

3.4

Exact computation of generalization error

The factorization of ?T2 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 x? to give P?

T (?, ?n2 , XT , XS) = hk?? i ? tr ??1 hkx? (kx?)T i = i=1 ?i ? tr ??1 M , where ? denotes ?(?, ?n2 , ?n2), the expectations are taken over x? , and M is an n-by-n matrix with R P? Mpq def = k x (xp , x?) k x (xq , x?) p(x?)dx? = i=1 ?2i ?i (xp)?i (xq), where xp , xq ? X. When the eigenfunctions ?i (?)s are not bounded, the infinite-summation expression for Mpq is often difficult to use. Nevertheless, analytical results for Mpq are still possible in some cases using the integral expression. An example is the case of the squared exponential covariance function with normally distributed 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 SacksP? ?(2r+1)/(2r+2) 2 Ylvisaker conditions of order r, then opt) + (1 ? ?2) i=1 ?i . T (?, ? , 1, n) = ?(nS P? 2 2 2 opt 2 Proof. We obtain opt T (?, ? , 1, n) = ? T (1, ?n , 1, n) + (1 ? ?) i=1 ?i by minimizing Corollary 4 wrt XS . Under the same conditions as the proposition, the optimal generalization error using the single-task GP decays with training set size n as ?(n?(2r+1)/(2r+2)) [11, Proposition V.3]. Thus ?(2r+1)/(2r+2) ?(2r+1)/(2r+2) 2 2) = ?(nS). ?2 opt T (1, ?n , 1, n) = ? ?(nS P A directly corollary of the above result is that one cannot expect to do better than (1 ? ?2) ?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: avg avg 2 2 2 Corollary 10. avg T (1, ?n , ?S , n) 6 T (?, ?n , ?S , n) 6 T (0, ?n , ?S , n). The gap between the trivial bounds can be analyzed as follows. Recall that ?S n ? N0 by definition, avg avg 2 2 2 so that avg T (1, ?n , ?S , (1 ? ?S)n) = T (0, ?n , ?S , n). Therefore T (1, ?n , ?S , n) is equivalent to avg 2 T (0, ?n , ?S , n) scaled along the n-axis by the factor (1 ? ?S) ? [0, 1], and hence the gap between the trivial bounds becomes wider with ?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

P? For the single-task GP, a lower bound on its learning curve is ?n2 i=1 ?i /(?n2 + n?i) [15]. We shall call this the single-task OV bound. This lower bound can be combined with Corollary 6. ? ? X X ?i ?i 2 2 2 2 2 2 Proposition 11. avg (?, ? , ? , n) ¿ ? ? + (1 ? ?)? , n S n n T 2 + n? 2 + (1 ? ?)n? ? ? i i S i=1 n i=1 n 2 2 or equivalently, avg T (?, ?n , ?S , n) ¿ ?n

? X i=1

2 2 or equivalently, avg T (?, ?n , ?S , n) ¿ ?n

? X i=1

1

b1i ?i , ?n2 + n?i

with b1i def =

?n2 + (1 ? ?2 ?S)n?i , ?n2 + (1 ? ?S)n?i

2 2 b0i ?i 0 def ?n + (1 ? ? ?S)n?i , with b . = i ?n2 + (1 ? ?S)n?i ?n2 + n?i

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

6

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 ith summand is obtained by combining the corresponding pair of ith summands in the first inequality. The third inequality is obtained from the second by swapping the denominator of b1i with that of ?i /(?n2 + n?i) for every i. For fixed ?n2 , ?S and n, denote the above bound by OV? . Then OV0 and OV1 are both single task 2 bounds. In particular, from Corollary 10, we have that the OV1 is a lower bound on avg T (?, ?n , ?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? is always better than OV1 . As ?2 decreases, the two-tasks bound moves towards the OV0 ; and as ?S increases, the gap between OV0 and OV1 increases. In addition, the gap is also larger for rougher processes, which are harder to learn. Therefore, the relative tightness of OV? over OV1 is more noticeable for lower ?2 , higher ?S and rougher processes. The second expression in the Proposition 11 is useful for comparing with the OV1 . Each summand for the two-tasks case is a factor b1i of the corresponding summand for the single-task case. Since b1i ? [1, (1 ? ?2 ?S)/(1 ? ?S)[, OV? is more than OV1 by at most (1 ? ?2)?S /(1 ? ?S) times. Similarly, the third expression of the proposition is useful for comparing with OV0 : each summand for the the two-tasks case is a factor b0i ?](1 ? ?2 ?S), 1] of the corresponding single-task one. Hence, OV? is less than OV0 by up to ?2 ?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 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 (?, ?), and eigenvalues ?i and eigenfunctions ?i (?) under the measure p(x)dx; and suppose that the noise (co)variances of the observations are given by ? 2 (?, ?). For n obdef servations {xi }ni=1 , let H and ? be matrices such that Hij P + ? 2 (xi , xj) and = k x (xi , xj)P ? ? 2 def ?ij = ?j (xi). Then the learning curve at n is upper-bounded by i=1 ?i ? n i=1 ?i /ci , where T def ci = (? H?)ii /n, and the expectation in ci is taken over the set of n input locations drawn independently from p(x). Unlike [16], we do not assume that the noise variance ? 2 (xi , xj) is of the form ?n2 ?ij . Instead ? 2n), we proceed directly from the exact posterior of proceeding from the upper bound ?T2 (1, ?n2 , ? variance given by (12). Thus we set the observation noise (co)variance ? 2 (xi , xj) to

?(xi ? XT)?(xj ? XT) ?ij ?n2 + ?(xi ? XS)?(xj ? XS) ?k x (xi , xj) + ??2 ?ij ?n2 , (14) so that, through the definition of ci in Theorem 12, we obtain n o

R 2 ci = (1 + ??S) (1 + ??S2)n/(1 + ??S) ? 1 ?i + k x (x, x) [?i (x)] p(x)dx + ?n2 ; (15) details are in the supplementary material ?S.5. This leads to the following proposition: Proposition 13. Let ? def in (15), we have = ??2 ? 1. Then, usingPthe ci s defined P? 2 ? 2 avg (?, ? , ? , n) 6 ? ? n i n S T i=1 i=1 ?i /ci . Denote the above upper bound by FWO? . When ? = ?1 or ?S = 0, the single-task FWO upper P bound is recovered. However, FWO? with ? = 0 gives the prior variance ?i instead. A trivial upper bound can be obtained using Corollary 10, by replacing n with (1 ? ?S)n in the single-task FWO bound. The FWO? bound is better than this trivial single-task bound for small n and high —?—. 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 N (0, 1/12), the covariance function is the 7

1

1

avg T

OV? / hhT (?)ii / FWO?

avg T

OV? / hhT (?)ii / FWO?

0.8

hhT (1)ii / hhT (0)ii ? hh?T (?)ii 4 hh? T (?)ii

0.8

hhT (1)ii / hhT (0)ii ? hh?T (?)ii 4 hh? T (?)ii

0.6
0.6
0.4
0.4
0.2
0.2
n 0
50
100
150
200
250
n
300
0
(a) ?2 = 1/2, ?S = 1/2
50
100
150
200
250
300
(b) ?2 = 3/4, ?S = 3/4

Figure 3: Comparison of various bounds for two settings of (?, ?S). Each graph plots avg T 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 (?/ 4). The thickness of the ?true? multi-task learning curve reflects 95% confidence interval. unit variance squared exponential k x (x, x0) = exp[?(x ? x0)2 /(2l2)] with length-scale l = 0.01, the observation noise variance is ?n2 = 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 def = XT ? XS , denoted by hh?ii, is computed over 100 randomly sampled training sets. The process eigenvalues ?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 ?2 ? {1/8, 1/4, 1/2, 3/4} and ?S ? {1/4, 1/2, 3/4} are computed. We compare the following: (a) the ?true? multi-task learning curve hhT (?)ii obtained by averaging ?T2 (?) over x? and X; (b) the theoretical bounds OV? and FWO? of Propositions 11 and 13; (c) the trivial upper and lower bounds that are single-task learning curves hhT (0)ii and hhT (1)ii obtained by averaging ?T2 (0) and ?T2 (1); and (d) the empirical lower bound hhT (?)ii and upper bound hh? T (?)ii 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 hhT (0)ii and hhT (1)ii increases with ?S , as described at the start of section 5. (b) We

find the lower bound hhT (?)ii a rather close approximation to the multi-task learning curve hhT (?)ii, as evidenced by ? much overlap between the the ? lines and the middle lines in Figure 3. (c) The curve for the empirical upper bound hh? T (?)ii using the equivalent noise method has jumps, e.g., the 4 lines in ? 2n increases whenever a datum for XS is sampled. Figure 3, because the equivalent noise variance ? (d) For small n, hhT (?)ii is closer to FWO? , but becomes closer to OV? as n increases, as shown by the unmarked solid lines in Figure 3. This is because the theoretical lower bound OV? is based on the asymptotically exact single-task OV bound and the T (?) 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 ? between the two tasks, and the proportion ?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 ? and ?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. 8

2 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. Sch?olkopf, 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 Bax-

ter. 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.

9