PAC-Bayesian Theory Meets Bayesian Inference

Authored by:

Simon Lacoste-Julien
Pascal Germain
Francis Bach
Alexandre Lacoste

Abstract

We exhibit a strong link between frequentist PAC-Bayesian bounds and the Bayesian marginal likelihood. That is, for the negative log-likelihood loss function, we show that the minimization of PAC-Bayesian generalization bounds maximizes the Bayesian marginal likelihood. This provides an alternative explanation to the Bayesian Occam's razor criteria, under the assumption that the data is generated by an i.i.d. distribution. Moreover, as the negative log-likelihood is an unbounded loss function, we motivate and propose a PAC-Bayesian theorem tailored for the sub-gamma loss family, and we show that our approach is sound on classical Bayesian linear regression tasks.

1 Paper Body

Since its early beginning [24, 34], the PAC-Bayesian theory claims to provide ?PAC guarantees to Bayesian algorithms? (McAllester [24]). However, despite the amount of work dedicated to this statistical learning theory?many authors improved the initial results [8, 21, 25, 30, 35] and/or generalized them for various machine learning setups [4, 12, 15, 20, 28, 31, 32, 33]?it is mostly used as a frequentist method. That is, under the assumptions that the learning samples are i.i.d.-generated by a data-distribution, this theory expresses probably approximately correct (PAC) bounds on the generalization risk. In other words, with probability 1 , the generalization risk is at most " away from the training risk. The Bayesian side of PAC-Bayes comes mostly from the fact that these bounds are expressed on the averaging/aggregation/ensemble of multiple predictors (weighted by a posterior distribution) and incorporate prior knowledge. Although it is still sometimes referred as a theory that bridges the Bayesian and frequentist approach [e.g., 16], it has been merely used to justify Bayesian methods until now.1 In this work, we provide a direct connection between Bayesian inference techniques [summarized by 5, 13] and PAC-Bayesian risk bounds in a general setup. Our study is based on a simple but insightful connection between

the Bayesian marginal likelihood and PAC-Bayesian bounds (previously mentioned by Gr?nwald [14]) obtained by considering the negative log-likelihood loss function (Section 3). By doing so, we provide an alternative explanation for the Bayesian Occam?s razor criteria [18, 22] in the context of model selection, expressed as the complexity-accuracy trade-off appearing in most PAC-Bayesian results. In Section 4, we extend PAC-Bayes theorems to regression problems with unbounded loss, adapted to the negative log-likelihood loss function. Finally, we study the Bayesian model selection from a PAC-Bayesian perspective (Section 5), and illustrate our finding on classical Bayesian regression tasks (Section 6).

2

PAC-Bayesian Theory

We denote the learning sample (X, Y)={(xi , yi)}ni=1 2(X ?Y)n , that contains n input-output pairs. The main assumption of frequentist learning theories?including PAC-Bayes?is that (X, Y) is 1

Some existing connections [3, 6, 14, 19, 29, 30, 36] are discussed in Appendix A.1.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

randomly sampled from a data generating distribution that we denote D. Thus, we denote (X, Y)?Dn the i.i.d. observation of n elements. From a frequentist perspective, we consider in this work loss functions ‘ : F?X ?Y ! R, where F is a (discrete or continuous) set of predictors f : X ! Y, and we write the empirical risk on the sample (X, Y) and the generalization error on distribution D as n 1X ‘ b LX,Y (f) = ‘(f, xi , yi) ; LD‘ (f) = E ‘(f, x, y) . n i=1 (x,y)?D The PAC-Bayesian theory [24, 25] studies an averaging of the above losses according to a posterior distribution ?? over F. That is, it provides probably approximately correct generalization bounds on the (unknown) quantity Ef ??? LD‘ (f) = Ef ??? E(x,y)?D ‘(f, x, y) , given the empirical estimate ‘ Ef ??? LbX,Y (f) and some other parameters. Among these, most PAC-Bayesian theorems rely on the Kullback-Leibler divergence KL(? ?k?) = Ef ??? ln[? ?(f)/?(f)] between a prior distribution ? over F?specified before seeing the learning sample X, Y ?and the posterior ???typically obtained by feeding a learning process with (X, Y).

Two appealing aspects of PAC-Bayesian theorems are that they provide data-driven generalization bounds that are computed on the training sample (i.e., they do not rely on a testing sample), and that they are uniformly valid for all ?? over F. This explains why many works study them as model selection criteria or as an inspiration for learning algorithm conception. Theorem 1, due to Catoni [8], has been used to derive or study learning algorithms [10, 17, 26, 27]. Theorem 1 (Catoni [8]). Given a distribution D over X ? Y, a hypothesis set F, a loss function ‘0 : F ? X ? Y ! [0, 1], a prior distribution ? over F, a real number 2 (0, 1], and a real number ¿ 0, with probability at least 1 over the choice of (X, Y) ? Dn , we have ? 1 1 1 b ‘0 ? ln ‘0 8? ? on F : E LD (f) ? 1 e Ef ??? LX,Y (f) n KL(?k?)+ . (1) f ??? 1 e Theorem 1 is limited to loss functions mapping to the range [0, 1]. Through a straightforward rescaling we

can extend it to any bounded loss, i.e., ‘ : F ? X ? Y ! [a, b], where [a, b] ? R. This is done by using := b a and with the rescaled loss function ‘0 (f, x, y) := (‘(f, x, y) a)/(b a) 2 [0, 1] . After few arithmetic manipulations, we can rewrite Equation (1) as h ? ?i ‘ 8? ? on F : E LD‘ (f) ? a + b aa b 1 exp E LbX,Y (f)+a 1 KL(? ?k?)+ ln 1 . (2) f ???

1 e

n

f ???

From an algorithm design perspective, Equation (2) suggests optimizing a trade-off between the empirical expected loss and the Kullback-Leibler divergence. Indeed, for fixed ?, X, Y , n, and , minimizing Equation (2) is equivalent to find the distribution ?? that minimizes ‘ n E LbX,Y (f) + KL(? ?k?) . (3) f ???

It is well known [1, 8, 10, 21] that the optimal Gibbs posterior ??? is given by ??? (f) =

1 ZX,Y

?(f) e

b ‘ (f) nL X,Y

,

(4)

where ZX,Y is a normalization term. Notice that the constant of Equation (1) is now absorbed in the loss function as the rescaling factor setting the trade-off between the expected empirical loss and KL(? ?k?).

3

Bridging Bayes and PAC-Bayes

In this section, we show that by choosing the negative log-likelihood loss function, minimizing the PAC-Bayes bound is equivalent to maximizing the Bayesian marginal likelihood. To obtain this result, we first consider the Bayesian approach that starts by defining a prior p(?) over the set of possible model parameters ?. This induces a set of probabilistic estimators f? 2 F, mapping x to a probability distribution over Y. Then, we can estimate the likelihood of observing y given x and ?, i.e., p(y—x, ?) ? f? (y—x).2 Using Bayes? rule, we obtain the posterior p(?—X, Y): p(?) p(Y —X, ?) p(?—X, Y) = / p(?) p(Y —X, ?) , (5) p(Y —X) R Qn where p(Y —X, ?) = i=1 p(yi —xi , ?) and p(Y —X) = ? p(?) p(Y —X, ?) d?. 2

To stay aligned with the PAC-Bayesian setup, we only consider the discriminative case in this paper. One can extend to the generative setup by considering the likelihood of the form p(y, x—?) instead.

2

To bridge the Bayesian approach with the PAC-Bayesian framework, we consider the negative log-likelihood loss function [3], denoted ‘nll and defined by ‘nll (f? , x, y) ?

(6)

ln p(y—x, ?) .

‘ Then, we can relate the empirical loss LbX,Y of a predictor to its likelihood: n

1X ‘nll LbX,Y (?) = ‘nll (?, xi , yi) = n i=1
or, the other way around,
n
1X ln p(yi —xi , ?) = n i=1
p(Y —X, ?) = e
b ‘nll (?) nL X,Y
1 ln p(Y —X, ?) , n
.
(7)

Unfortunately, existing PAC-Bayesian theorems work with bounded loss functions or in very specific contexts [e.g., 9, 36], and ‘nll spans the whole real axis in its general form. In Section 4, we explore PAC-Bayes bounds for unbounded losses. Meanwhile, we consider priors with bounded likelihood. 1 This can be done by assigning a prior of zero to any ? yielding ln p(y—x,?) 2 / [a, b]. Now, using Equation (7) in the optimal posterior (Equation 4) simplifies to b ‘nll

?(?) e n LX,Y (?) p(?) p(Y —X, ?) ?? (?) = = = p(?—X, Y) , ZX,Y p(Y —X) ?

where the normalization constant ZX,Y corresponds to the Bayesian marginal likelihood: Z b ‘nll ZX,Y ? p(Y —X) = ?(?) e n LX,Y (?) d? .
(8)
(9)
?

This shows that the optimal PAC-Bayes posterior given by the generalization bound of Theorem 1 coincides with the Bayesian posterior, when one chooses ‘nll as loss function and := b a (as in Equation 2). Moreover, using the posterior of Equation (8) inside Equation (3), we obtain ‘nll n E ? LbX,Y (?) + KL(? ?? k?) ???? Z Z h b ‘nll (?) b ‘nll (?) b ‘nll (?) i nL nL nL X,Y X,Y X,Y ?(?) e ?(?) e ?(?) e ‘nll b = n LX,Y (?) d? + ln d? ZX,Y ZX,Y ?(?) ZX,Y ? Z ? ‘ h i nll b nL (?) X,Y Z ?(?) e 1 1 = ln ZX,Y d? = ZX,Y ln ZX,Y = ln ZX,Y . ZX,Y X,Y
(10)
?

In other words, minimizing the PAC-Bayes bound is equivalent to maximizing the marginal likelihood. Thus, from the PAC-Bayesian standpoint, the latter encodes a trade-off between the averaged negative log-likelihood loss function and the prior-posterior Kullback-Leibler divergence. Note that Equation (10) has been mentioned by Gr?nwald [14], based on an earlier observation of Zhang [36]. However, the PAC-Bayesian theorems proposed by the latter do not bound the generalization loss directly, as the ?classical? PAC-Bayesian results [8, 24, 29] that we extend to regression in forthcoming Section 4 (see the corresponding remarks in Appendix A.1). We conclude this section by proposing a compact form of Theorem 1 by expressing it in terms of the marginal likelihood, as a direct consequence of Equation (10). Corollary 2. Given a data distribution D, a parameter set ?, a prior distribution ? over ?, a 2 (0, 1], if ‘nll lies in [a, b],

we have, with probability at least 1 over the choice of (X, Y) ? Dn , h i p E ? L‘Dnll (?) ? a + 1 beaa b 1 ea n ZX,Y , ????

where ??? is the Gibbs optimal posterior (Eq. 8) and ZX,Y is the marginal likelihood (Eq. 9).

In Section 5, we exploit the link between PAC-Bayesian bounds and Bayesian marginal likelihood to expose similarities between both frameworks in the context of model selection. Beforehand, next Section 4 extends the PAC-Bayesian generalization guarantees to unbounded loss functions. This is mandatory to make our study fully valid, as the negative log-likelihood loss function is in general unbounded (as well as other common regression losses). 3

4

PAC-Bayesian Bounds for Regression

This section aims to extend the PAC-Bayesian results of Section 3 to real valued unbounded loss. These results are used in forthcoming sections to study ‘nll , but they are valid for broader classes of loss functions. Importantly, our new results are focused on regression problems, as opposed to the usual PAC-Bayesian classification framework. The new bounds are obtained through a recent theorem of Alquier et al. [1], stated below (we provide a proof in Appendix A.2 for completeness). Theorem 3 (Alquier et al. [1]). Given a distribution D over X ? Y, a hypothesis set F, a loss function ‘ : F ? X ? Y ! R, a prior distribution ? over F, a 2 (0, 1], and a real number ¿ 0, with probability at least 1 over the choice of (X, Y) ? Dn , we have ? 1 1 ‘ ‘ b 8? ? on F : E LD (f) ? E LX,Y (f) + KL(? ?k?) + ln + ‘,?,D (, n) , (11) f ???

where

f ???

‘,?,D (

, n) = ln E

E 0

f ?? X 0,Y ?D

exp n

h ?

LD‘ (f)

‘ LbX 0,Y 0 (f)

?i

.

(12)

Alquier et al. used Theorem 3 to design a learning algorithm for {0, 1}-valued classification losses. Indeed, a bounded loss function ‘ : F ? X ? Y ! [a, b] can be used along with Theorem 3 by applying the Hoeffding?s lemma to Equation (12), that gives ‘,?,D (, n) ? 2 (b a)2 /(2n). More specifically, with := n, we obtain the following bound ? ? ‘ 8? ? on F : E LD‘ (f) ? E LbX,Y (f) + 1 KL(? ?k?) + ln 1 + 1 (b a)2 . (13) f ???

n

f ???

2

Note that the latter bound leads to the same trade-off as Theorem 1 (expressed by Equation 3). However, the choice := n has the inconvenience that the bound value is at least 12 (b a)2 , even p at the limit n ! 1. With := n the bound converges (a result similar to Equation (14) is also formulated by Pentina and Lampert [28]): ? ? ‘ 8? ? on F : E LD‘ (f) ? E LbX,Y (f) + p1 KL(? ?k?) + ln 1 + 1 (b a)2 . (14) f ???

2

n

f ???

Sub-Gaussian losses. In a regression context, it may be restrictive to consider strictly bounded loss functions. Therefore, we extend Theorem 3 to sub-Gaussian losses. We say that a loss function ‘ is sub-Gaussian with variance factor s2 under a prior ? and a data-distribution D if it can be described by a sub-Gaussian random variable V =LD‘ (f) ‘(f, x, y), i.e., its moment generating function is upper bounded by the one of a normal distribution of variance s2 (see Boucheron et al. [7, Section 2.3]): ? ? 2 2 ( ) = ln E e V = ln E E exp LD‘ (f) ‘(f, x, y) ? 2s , 8 2 R . (15) V f ?? (x,y)?D

The above sub-Gaussian assumption corresponds to the Hoeffding assumption of Alquier et al. [1], and allows to obtain the following result. Corollary 4. Given D, F, ‘, ? and defined in the statement of Theorem 3, if the loss is sub-Gaussian with variance factor s2 , we have, with probability at least 1 over the choice of (X, Y) ? Dn , ? ? 8? ? on F : E L ‘ (f) ? E Lb ‘ (f) + 1 KL(? ?k?) + ln 1 + 1 s2 . f ???

D

f ???

X,Y

n

2

Proof. For i = 1 . . . n, we denote ‘i a i.i.d. realization of the random variable LD‘ (f) ‘(f, x, y). ? Pn ? ? ? Pn Qn 2 2 2 2 s s ‘,?,D (, n) = ln E exp n i=1 ‘i = ln i=1 E exp n ‘i = i=1 ‘i (n) ? n 2n2 = 2n , where the inequality comes from the sub-Gaussian loss assumption (Equation 15). The result is then obtained from Theorem 3, with := n. Sub-gamma losses. We say that an unbounded loss function ‘ is sub-gamma with a variance factor s2 and scale parameter c, under a prior ? and a data-distribution D, if it can be described by a sub-gamma random variable V (see Boucheron et al. [7, Section 2.4]), that is V

( ) ?

s2 c2 (

ln(1

c)

c) ?

2 2 s 2(1 c)

,

8 2 (0, 1c) .

(16)

Under this sub-gamma assumption, we obtain the following new result, which is necessary to study linear regression in the next sections. 4

Corollary 5. Given D, F, ‘, ? and defined in the statement of Theorem 3, if the loss is sub-gamma with variance factor s2 and scale c ¡ 1, we have, with probability at least 1 over (X, Y) ? Dn , ? ? ‘ 8? ? on F : E LD‘ (f) ? E LbX,Y (f) + n1 KL(? ?k?) + ln 1 + 2(11 c) s2 . (17) f ???

f ???

As a special case, with ‘ := ‘nll and ?? := ??? (Equation 8), we have E L‘Dnll (?) ?

?????

s2 2(1 c)

1 n

(18)

ln (ZX,Y) .

Proof. Following the same path as in the proof of Corollary 4 (with Pn Qn Pn ‘,?,D (n, n) = ln E exp [i=1 ‘i] = ln i=1 E exp [‘i] = i=1

:= n), we have ‘i (1)

? n 2(1s

2

c)

=

n s2 2(1 c)

,

where the inequality comes from the sub-gamma loss assumption, with 1 2 (0, 1c). Squared loss. The parameters s and c of Corollary 5 rely on the chosen loss function and prior, and the assumptions concerning the data distribution. As an example, consider a regression problem where X ?Y ? Rd ?R, a family of linear predictors fw (x) = w ? x, with w 2 Rd , and a Gaussian prior N (0, ?2 I). Let us assume that the input examples are generated by x ? N (0, x2 I) with label y = w?? x + ?, where w? 2 Rd and ? ? N (0, ?2) is a Gaussian noise. Under the squared loss function y)2 ,

(19) ? we show in Appendix A.4 that Corollary 5 is valid with s2 2 x2 (?2 d + kw? k2) + ?2 (1 c) and c 2 x2 ?2 . As expected, the bound degrades when the noise increases ‘sqr (w, x, y) = (w ? x

?

Regression versus classification. The classical PAC-Bayesian theorems are stated in a classification context and bound the generalization error/loss of the stochastic Gibbs predictor G??. In order to predict the label of an example x 2 X , the Gibbs predictor first draws a hypothesis h 2 F according to ??, and then returns h(x). Maurer [23] shows that we can generalize PAC-Bayesian bounds on the generalization risk of the Gibbs classifier to any loss function with output between zero and one. Provided that y 2 { 1, 1} and h(x) 2 [1, 1], a common choice is to use the 1 linear loss function ‘001 (h, x, y) = 12 2 y h(x). The Gibbs generalization loss is then given by 0 RD (G??) = E(x,y)?D Eh??? ‘01 (h, x, y) . Many PAC-Bayesian works use RD (G??) as a surrogate loss to study the

zero-one classification loss of the majority vote classifier RD (B??): ? ? h i RD (B??) = Pr y E h(x) ¡ 0 = E I y E h(x) ¡ 0 , (20) (x,y)?D

h???

(x,y)?D

h???

where I[?] being the indicator function. Given a distribution ??, an upper bound on the Gibbs risk is converted to an upper bound on the majority vote risk by RD (B??) ? 2RD (G??) [20]. In some situations, this factor of two may be reached, i.e., RD (B??) ' 2RD (G??). In other situations, we may have RD (B??) = 0 even if RD (G??) = 12 ? (see Germain et al. [11] for an extensive study). Indeed, these bounds obtained via the Gibbs risk are exposed to be loose and/or unrepresentative of the majority vote generalization error.3 In the current work, we study regression losses instead of classification ones. That is, the provided results express upper bounds on Ef ??? LD‘ (f) for any (bounded, sub-Gaussian, or sub-gamma) losses. Of course, one may want to bound the regression loss of the averaged regressor F??(x) = Ef ??? f (x). In this case, if the loss function ‘ is convex (as the squared loss), Jensen?s inequality gives LD‘ (F??) ? Ef ??? LD‘ (f) . Note that a strict inequality replaces the factor two mentioned above for the classification case, due to the non-convex indicator function of Equation (20). Now that we have generalization bounds for real-valued loss functions, we can continue our study linking PAC-Bayesian results to Bayesian inference. In the next section, we focus on model selection. 3 It is noteworthy that the best PAC-Bayesian empirical bound values are so far obtained by considering a majority vote of linear classifiers, where the prior and posterior are Gaussian [2, 10, 20], similarly to the Bayesian linear regression analyzed in Section 6.

5

5

Analysis of Model Selection

We consider L distinct models {Mi }L i=1 , each one defined by a set of parameters ?i . The PACBayesian theorems naturally suggest selecting the model that is best adapted for the given task by evaluating the bound for each model {Mi }L i=1 and selecting the one with the lowest bound [2, 25, 36]. This is closely linked with the Bayesian model selection procedure, as we showed in Section 3 that minimizing the PAC-Bayes bound amounts to maximizing the marginal likelihood. Indeed, given a collection of L optimal Gibbs posteriors?one for each model?given by Equation (8), p(?—X, Y, Mi) ? ???i (?) =

‘

b nll (?) nL 1 X,Y , ZX,Y,i ?i (?) e

for ? 2 ?i ,

the Bayesian Occam?s razor criteria [18, 22] chooses the one with the higher model evidence Z b‘ p(Y —X, Mi) ? ZX,Y,i = ?i (?) e n LX,Y (?) d? .

(21) (22)

?i

Corollary 6 below formally links the PAC-Bayesian and the Bayesian model selection. To obtain this result, we simply use the bound of Corollary 5 L times,

together with ‘nll and Equation (10). From the union bound (a.k.a. Bonferroni inequality), it is mandatory to compute each bound with a confidence parameter of /L, to ensure that the final conclusion is valid with probability at least 1 . Corollary 6. Given a data distribution D, a family of model parameters {?i }L i=1 and associated priors {?i }L ?where ? is defined over ? ? , a 2 (0, 1], if the loss is sub-gamma with parameters i i i=1 s2 and c ¡ 1, then, with probability at least 1 over (X, Y) ? Dn , 8i 2 {1, . . . , L} : E ? L‘Dnll (?) ? 2(11 c) s2 n1 ln ZX,Y,i L . ????i

where ???i is the Gibbs optimal posterior (Eq. 21) and ZX,Y,i is the marginal likelihood (Eq. 22). Hence, under the uniform prior over the L models, choosing the one with the best model evidence is equivalent to choosing the one with the lowest PAC-Bayesian bound. Hierarchical Bayes. To perform proper inference on hyperparameters, we have to rely on the Hierarchical Bayes approach. This is done by considering an hyperprior p(?) over the set of hyperparameters H. Then, the prior p(?—?) can be conditioned on a choice of hyperparameter ?. The p(Y —X,?) Bayes rule of Equation (5) becomes p(?, ?—X, Y) = p(?) p(?—?) . p(Y —X)

Under the negative log-likelihood loss function, we can rewrite the results of Corollary 5 as a generalization bound on E????0 E?????? L‘Dnll (?), where ??0 (?) / ?0 (?) ZX,Y,? is the hyperposterior on H and ?0 the hyperprior. Indeed, Equation (18) becomes ? ? E ? L‘Dnll (?) = E ? E ? L‘Dnll (?) ? 2(11 c) s2 n1 ln E ZX,Y,? . (23) ????

???0

????0 ?????

To relate to the bound obtained in Corollary 6, we consider the case of a discrete hyperparameter set 1 H = {?i }L i=1 , with a uniform prior ?0 (?i) = L (from now on, we regard each hyperparameter ?i as the specification of a model ?i). Then, Equation (23) becomes ?P ? L E ? L‘Dnll (?) = E ? E ? L‘Dnll (?) ? 2(11 c) s2 n1 ln i=1 ZX,Y,?i L . ????

????0 ?????

PL This bound is now a function of i=1 ZX,Y,?i instead of maxi ZX,Y,?i as in the bound given by the ?best? model in Corollary 6. This yields a tighter bound, corroborating the Bayesian wisdom that model averaging performs best. Conversely, when selecting a single hyperparameter ? ? 2 H, the hierarchical representation is equivalent to choosing a deterministic hyperposterior, satisfying ??0 (? ?) = 1 and 0 for every other values. We then have KL(? ?——?) = KL(? ?0 ——?0) + E KL(? ?? ——??) = ln(L) + KL(? ??? ——???) . ????0

With the optimal posterior for the selected ? ? , we have ‘nll ‘nll n E LbX,Y (?) + KL(? ?——?) = n E ? LbX,Y (?) + KL(? ???? ——???) + ln(L) ???? ????? ? ? ? Z = ln(ZX,Y,??) + ln(L) = ln X,Y,? . L

Inserting this result into Equation (17), we fall back on the bound obtained in Corollary 6. Hence, by comparing the values of the bounds, one can get an estimate on the consequence of performing model selection instead of model averaging. 6

6

Linear Regression

In this section, we perform Bayesian linear regression using the parameterization of Bishop [5]. The output space is Y := R and, for an arbitrary input space X , we use a mapping function : X !Rd . The model. Given (x, y) 2 X ? Y and model parameters ? := hw, i 2 Rd ? R+ , we consider the likelihood p(y—x, hw, i) = N (y—w ? (x), 2). Thus, the negative log-likelihood loss is ‘nll (h w, i, x, y) =

ln p(y—x, h w, i) =

1 2

ln(2?

2

)+

1 2

2

(y

w ? (x))2 .

(24)

For a fixed 2 , minimizing Equation (24) is equivalent to minimizing the squared loss function of Equation (19). We also consider an isotropic Gaussian prior of mean 0 and variance ?2 : p(w— ?) = N (w—0, ?2 I). For the sake of simplicity, we consider fixed parameters 2 and ?2 . The Gibbs optimal posterior (see Equation 8) is then given by ??? (w) ? p(w—X, Y, ,

?)

=

p(w— ?) p(Y —X,w,) p(Y —X, , ?)

b A = N (w — w,

1

(25)

),

b := 12 A 1 T y ; is a n?d matrix such that the ith line is (xi) ; where A := 12 T + 12 I ; w ? y := [y1 , . . . yn] is the labels-vector ; and the negative log marginal likelihood is ln p(Y —X, , =

‘nll n LbX,Y

—

?)

=

b + (w)

1

2 1 2

2

2

{z

ky

tr(

b ‘nll n Ew???? L X,Y

T

(w)
b 2+ wk A
n 2
ln(2?
1
2
)+
) + 2 12 tr(A } — ?
1
)
1
2
2 ?
b 2+ kwk
d 2
+
1
2
2 ?
b KL N (w,A
1 2
log —A— + d ln 2
b + kwk {z
1 2
?
log —A— + d ln
1) k N (0, 2 I) ?
?
}
.

‘nll b 2 + n2 ln(2? 2) = n LbX,Y b and insert To obtain the second equality, we substitute 2 1 2 ky wk (w) 1 1 1 1 1 1 T 1 1 T 1 1 1 tr(A) + tr(A) = tr(A + A) = tr(A A) = d2 . 2 2 2 2 2 2 ? 2 2 ? This exhibits how the Bayesian regression optimization problem is related to the mini‘nll mization of a PAC-Bayesian bound, expressed by a trade-off between Ew???? LbX,Y (w) and b A 1) k N (0, ?2 I) . See Appendix A.5 for detailed calculations. KL N (w,

Model selection experiment. To produce Figures 1a and 1b, we reimplemented the toy experiment of Bishop [5, Section 3.5.1]. That is, we generated a learning sample of 15 data points according to y = sin(x) + ?, where x is uniformly sampled in the interval [0, 2?] and ? ? N (0, 14) is a Gaussian noise. We then learn seven different polynomial models applying Equation (25). More precisely, for a polynomial model of degree d, we map input x 2 R to a vector (x) = [1, x1 , x2 , . . . , xd] 2 Rd+1 , 1 and we fix parameters ?2 = 0.005 and 2 = 12 . Figure 1a illustrates the seven learned models. Figure 1b shows the negative log marginal likelihood computed for each polynomial model, and

is designed to reproduce Bishop [5, Figure 3.14], where it is explained that the marginal likelihood correctly indicates that the polynomial model of degree d = 3 is ?the simplest model which gives a good explanation for the observed data?. We show that this claim is well quantified by the trade-off intrinsic to our PAC-Bayesian approach: the complexity KL term keeps increasing with the parameter d 2 {1, 2, . . . , 7}, while the empirical risk drastically decreases from d = 2 to d = 3, and only slightly afterward. Moreover, we show that the generalization risk (computed on a test sample of size 1000) tends to increase with complex models (for d 4). Empirical comparison of bound values. Figure 1c compares the values of the PAC-Bayesian bounds presented in this paper on a synthetic dataset, where each input x2R20 is generated by a Gaussian x?N (0, I). The associated output y2R is given by y=w? ? x + ?, with kw? k= 12 , ??N (0, ?2), and ?2 = 19 . We perform Bayesian linear regression in the input space, i.e., (x)=x, 1 fixing ?2 = 100 and 2 =2. That is, we compute the posterior of Equation (25) for training samples of sizes from 10 to 106 . For each learned model, we compute the empirical negative log-likelihood loss 1 of Equation (24), and the three PAC-Bayes bounds, with confidence parameter of = 20 . Note that this loss function is an affine transformation of the squared loss studied in Section 4 (Equation 19), i.e., ‘nll (hw, ? i, x, y)= 12 ln(2? 2)+ 2 1 2 ‘sqr (w, out that ‘nll is sub-gamma with parameters ? x, y). It1 turns 1 2 2 2 ? 2 2 2 2 s 2 2 (x ?), as shown in Appendix A.6. The bounds x (? d+kw k)+ ? (1 c) and c of Corollary 5 are computed using the above mentioned values of kw? k, d, , x , ? , ? , leading 7

1.5
model model model model
1.0 0.5
d=1 d=2 d=3 d=4
60
model d=5 model d=6 model d=7 sin(x)
50 40
0.0
ln ZX,Y
30
KL(???k?) ‘nll n E?????LbX,Y (?)
0.5
20
1.0
n E?????LD‘nll(?)
10
1.5 2.0 0
1 2?
3 2?
?
0 1
2?
2

x
3
4 model degree d
5
6
7

(a) Predicted models. Black dots are the 15 training (b) Decomposition of the marginal likelihood into the samples. empirical loss and KL-divergence. 4.0
Alquier et al?s [a, b] bound (Theorem 3 + Eq 14) 3.5
Catoni?s [a, b] bound (Corollary 2)
3.0
sub-gamma bound (Corollary 5) E????? LD'nll (?) (test loss)
2.5
'nll E????? LbX,Y (?) (train loss)
2.0 1.5 1.0 1 10
102
103
104
105
n

(c) Bound values on a synthetic dataset according to the number of training samples.

Figure 1: Model selection experiment (a-b); and comparison of bounds values (c). to s2 ' 0.280 and c ' 0.005. As the two other bounds of Figure 1c are not suited for unbounded loss, we compute their value using a cropped loss [a, b] = [1, 4]. Different parameter values could have been chosen, sometimes leading to another picture: a large value of s degrades our sub-gamma bound, as a larger [a, b] interval does for the other bounds. In the studied setting, the bound of Corollary 5?that we have developed for (unbounded) subgamma losses?gives tighter guarantees than the two results for [a, b]-bounded losses (up to n=106). However, our new bound always maintains a gap of 2(11 c) s2 between its value and the generalization loss. The result of Corollary 2 (adapted from Catoni [8]) for bounded losses suffers from a similar gap, while having higher values than p our sub-gamma result. Finally, the result of Theorem 3 (Alquier et al. [1]), combined with = 1/ n (Eq. 14), converges to the expected loss, but it provides good guarantees only for large training sample (n & 105). Note that the latter bound is not directly minimized by our ?optimal posterior?, as opposed to the one with = 1/n (Eq. 13), for which we observe values between 5.8 (for n=106) and 6.4 (for n=10)?not displayed on Figure 1c.

7

Conclusion

The first contribution of this paper is to bridge the concepts underlying the Bayesian and the PACBayesian approaches; under proper parameterization, the minimization of the PAC-Bayesian bound maximizes the marginal likelihood. This study motivates the second contribution of this paper, which is to prove PAC-Bayesian generalization bounds for regression with unbounded sub-gamma

loss functions, including the squared loss used in regression tasks. In this work, we studied model selection techniques. On a broader perspective, we would like to suggest that both Bayesian and PAC-Bayesian frameworks may have more to learn from each other than what has been done lately (even if other works paved the way [e.g., 6, 14, 30]). Predictors learned from the Bayes rule can benefit from strong PAC-Bayesian frequentist guarantees (under the i.i.d. assumption). Also, the rich Bayesian toolbox may be incorporated in PAC-Bayesian driven algorithms and risk bounding techniques. Acknowledgments We thank Gabriel Dub? and Maxime Tremblay for having proofread the paper and supplemental. 8

2 References

[1] Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. JMLR, 17(239):1?41, 2016. [2] Amiran Ambroladze, Emilio Parrado-Hern?ndez, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In NIPS, 2006. [3] Arindam Banerjee. On Bayesian bounds. In ICML, pages 81?88, 2006. [4] Luc B?gin, Pascal Germain, Fran?ois Laviolette, and Jean-Francis Roy. PAC-Bayesian theory for transductive learning. In AISTATS, pages 105?113, 2014. [5] Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. [6] P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2016. [7] St?phane Boucheron, G?bor Lugosi, and Pascal Massart. Concentration inequalities : a nonasymptotic theory of independence. Oxford university press, 2013. ISBN 978-0-19-953525-5. [8] Olivier Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning, volume 56. Inst. of Mathematical Statistic, 2007. [9] Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. Machine Learning, 72(1-2):39?61, 2008. [10] Pascal Germain, Alexandre Lacasse, Fran?ois Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In ICML, pages 353?360, 2009. [11] Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. JMLR, 16, 2015. [12] Pascal Germain, Amaury Habrard, Fran?ois Laviolette, and Emilie Morvant. A new PAC-Bayesian perspective on domain adaptation. In ICML, pages 859?868, 2016. [13] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. Nature, 521:452?459, 2015. [14] Peter Gr?nwald. The safe Bayesian - learning the learning rate via the mixability gap. In ALT, 2012. [15] Peter D. Gr?nwald and Nishant A. Mehta. Fast rates with unbounded losses. CoRR, abs/1605.00252, 2016. [16] Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin C. Cawley. Model selection: Beyond the Bayesian/frequentist divide. JMLR, 11:61?87, 2010. [17] Tamir Hazan, Subhransu Maji, Joseph Keshet, and

Tommi S. Jaakkola. Learning efficient random maximum a-posteriori predictors with non-decomposable loss functions. In NIPS, pages 1887?1895, 2013. [18] William H. Jeffreys and James O. Berger. Ockham?s razor and Bayesian analysis. American Scientist, 1992. [19] Alexandre Lacoste. Agnostic Bayes. PhD thesis, Universit? Laval, 2015. [20] John Langford and John Shawe-Taylor. PAC-Bayes & margins. In NIPS, pages 423?430, 2002. [21] Guy Lever, Fran?ois Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distributiondependent priors. Theor. Comput. Sci., 473:4?28, 2013. [22] David J. C. MacKay. Bayesian interpolation. Neural Computation, 4(3):415?447, 1992. [23] Andreas Maurer. A note on the PAC-Bayesian theorem. CoRR, cs.LG/0411099, 2004. [24] David McAllester. Some PAC-Bayesian theorems. Machine Learning, 37(3):355?363, 1999. [25] David McAllester. PAC-Bayesian stochastic model selection. Machine Learning, 51(1):5?21, 2003. [26] David McAllester and Joseph Keshet. Generalization bounds and consistency for latent structural probit and ramp loss. In NIPS, pages 2205?2212, 2011. [27] Asf Noy and Koby Crammer. Robust forward algorithms via PAC-Bayes and Laplace distributions. In AISTATS, 2014. [28] Anastasia Pentina and Christoph H. Lampert. A PAC-Bayesian bound for lifelong learning. In ICML, 2014. [29] Matthias Seeger. PAC-Bayesian generalization bounds for Gaussian processes. JMLR, 3:233?269, 2002. [30] Matthias Seeger. Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations. PhD thesis, University of Edinburgh, 2003. [31] Yevgeny Seldin and Naftali Tishby. PAC-Bayesian analysis of co-clustering and beyond. JMLR, 11, 2010. [32] Yevgeny Seldin, Peter Auer, Fran?ois Laviolette, John Shawe-Taylor, and Ronald Ortner. PAC-Bayesian analysis of contextual bandits. In NIPS, pages 1683?1691, 2011. [33] Yevgeny Seldin, Fran?ois Laviolette, Nicol? Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PACBayesian inequalities for martingales. In UAI, 2012. [34] John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In COLT, 1997. [35] Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-Bernstein inequality. In NIPS, 2013. [36] Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. IEEE Trans. Information Theory, 52(4):1307?1321, 2006.

9