Bayesian model averaging is suboptimal for generalization under model misspecification

Andres R. Masegosa[1]

Abstract

Virtually any model we use in machine learning to make predictions does not perfectly represent reality. So, most of the learning happens under model misspecification. In this work, we present a novel analysis of the generalization performance of Bayesian model averaging under model misspecification and i.i.d. data using a new family of second-order PAC-Bayes bounds. This analysis shows, in simple and intuitive terms, that Bayesian model averaging provides suboptimal generalization performance when the model is misspecified. In consequence, we provide strong theoretical arguments showing that Bayesian methods are not optimal for learning predictive models, unless the model class is perfectly specified. Using novel second-order PAC-Bayes bounds, we derive a new family of Bayesian-like algorithms, which can be implemented as variational methods. The output of these algorithms is a new posterior distribution, different from the Bayesian posterior, which induces a predictive distribution with better generalization performance. Experiments with Bayesian neural networks illustrate these methods [1].

1. Introduction

The Bayesian approach to machine learning (Ghahramani, 2015) is based on the computation of the Bayesian posterior distribution, which defines a probability distribution over our model family. Then, the predictions are made through Bayesian model averaging (Hoeting et al., 1999), which combines the predictions of the individual models of the family weighted by their posterior probability.

We know that our model families are idealizations which only provide an approximation to the real-world distributions generating the data. But this has not been a strong impediment for this Bayesian approach. Empirically, Bayesian model averaging has been very successful in many different areas of machine learning (Blei et al., 2003; Hoffman et al., 2013). Theoretically, there are very well-established asymptotic results (Kleijn et al., 2012) that shows how the Bayesian posterior tends to concentrate around the model which is closest to the true data generating distribution in terms of Kullback-Leibler divergence.

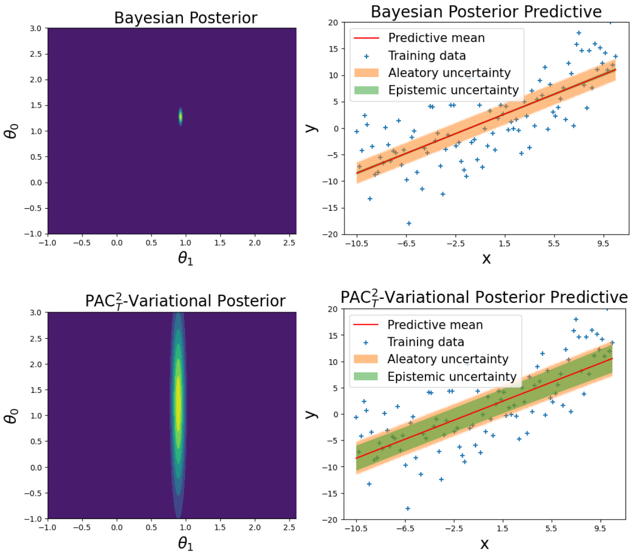

Figure 1. The exact Bayesian posterior and our new proposed (PAC^2_T-Variational) posterior, and their respective posterior predictive distributions, for a linear regression model with a misspecified constant noise term (the data noise is higher than the linear model's noise). The Bayesian posterior concentrates around the best single linear model, while our method estimates a posterior which introduces high variance in the intercept parameter θ_0 to induce a posterior predictive distribution with higher noise that better fits the data distribution (see Appendix H.2 for details).

But, at the same time, many other works have been continuously challenging the Bayesian approach to machine learning, showing that Bayesian model averaging is not a superior learning method. Ensemble models (Dietterich, 2000), an alternative approach for model combination, have

[*] Equal contribution [1]University of Almeria (Spain). Correspondence to: Andres R. Masegosa <andresma@ual.es>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning. Copyright 2020 by the author(s).

[1] An extended version of this work can be found in (Masegosa, 2019)

consistently provided very competitive generalization performance in a wide range of different problems, even in terms of well-calibrated probability predictions (Snoek et al., 2019). Although it is claimed that ensemble models could be an approximate way of capturing the multi-modality structure of the Bayesian posterior (Wilson & Izmailov, 2020), recent works (Wenzel et al., 2020) also provide strong evidence on how the generalization performance of Bayesian neural networks can be significantly improved by considering different posteriors distributions for model averaging that largely deviate from the Bayesian posterior. However, the reasons behind these results are yet unknown.

In this work, we perform a theoretical analysis of the generalization performance of Bayesian model averaging under model misspecification using a novel family of PAC-Bayesian bounds (McAllester, 1999). This analysis shows, in very simple and intuitive terms, that Bayesian model averaging provides suboptimal generalization performance because the Bayesian posterior is the minimum of a first-order PAC-Bayes bound, which can be quite loose when the model family is misspecified. Based on this analysis, we propose the use of tighter second-order PAC-Bayesian bounds and derive a new sound family of Bayesian-like algorithms based on the minimization of these new PAC-Bayes bounds. The result is a new posterior distribution, different from the Bayesian posterior, which induces new posterior predictive distributions with better generalization capacity. See Figure 1 for an illustrative example. Experiments on toy and real data sets with Bayesian neural networks illustrate these learning algorithms. Code is available in `https://github.com/PGM-Lab/PAC2BAYES`. Related work is discussed in Appendix A.

2. Previous Knowledge

For the shake of simplicity, our analysis is made under unsupervised settings or density estimation, but it readily applies to supervised classification settings too by considering labelled data and conditional probability distributions.

We denote the training data set $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, where $\boldsymbol{x} \in \mathcal{X}$. And the probability distribution over $\mathcal{X}$ is denoted by $p(\boldsymbol{x}|\boldsymbol{\theta})$, which indexed by a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. As a standard requirement in generalization analysis methods (McAllester, 1999), we assume that the samples in D are i.i.d. generated from some unknown data generating distribution denoted $\nu(\boldsymbol{x})$. For this analysis, we assume that $p(\boldsymbol{x}|\boldsymbol{\theta})$ is upper-bounded. This condition is always satisfied in supervised classifications settings (i.e, in this case we have a conditional distribution $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ whose maximum is always equal to 1). But, for example, when $p(\boldsymbol{x}|\boldsymbol{\theta})$ is a Normal density function, we have to restrict the parameter space $\boldsymbol{\Theta}$ to only consider variances higher than a given $\epsilon > 0$. Finally, we also assume we are learning under model misspecification, i.e., $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}\ p(\cdot|\boldsymbol{\theta}) \neq \nu$.

The *posterior predictive distribution* induced by a probability distribution $\rho(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$ is defined as

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})\rho(\boldsymbol{\theta})d\boldsymbol{\theta} = \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]. \tag{1}$$

ρ is referred as a posterior distribution because it depends on the data. When $\rho(\boldsymbol{\theta})$ is the Bayesian posterior, i.e., $\rho(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|D)$, the above posterior predictive corresponds to Bayesian model averaging (Hoeting et al., 1999).

The *learning problem* we study here is to find the optimal probability distribution $\rho(\boldsymbol{\theta})$ for performing *model averaging*. More precisely, the learning problem can be expressed as finding the probability distribution $\rho(\boldsymbol{\theta})$ which defines the *posterior predictive distribution* $p(\boldsymbol{x})$ with the smallest cross-entropy wrt to $\nu(\boldsymbol{x})$,

$$\rho^\star \quad = \quad \arg\min_{\rho} \underbrace{\mathbb{E}_{\nu(\boldsymbol{x})}[-\ln \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]]}_{CE(\rho)}, \tag{2}$$

where $CE(\rho)$ denotes this cross-entropy associated to ρ, which also depends on the data generating distribution $\nu(\boldsymbol{x})$. The distribution $\rho^\star$ also satisfies that $\rho^\star = \arg\min_\rho KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})])$.

2.1. Bayesian Learning

The key quantity in *Bayesian statics* (Bernardo & Smith, 2009) is the *Bayesian posterior*, $p(\boldsymbol{\theta}|D) \propto \pi(\boldsymbol{\theta}) \prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta})$ is known as the *prior* distribution. When a new observation $\boldsymbol{x}'$ arrives we compute the *Bayesian posterior predictive* distribution to make predictions about $\boldsymbol{x}'$, $p(\boldsymbol{x}'|D) = \mathbb{E}_{p(\boldsymbol{\theta}|D)}[p(\boldsymbol{x}'|\boldsymbol{\theta})]$.

The main challenge in the application of Bayesian statics in machine learning is the computation of the normalization constant of the Bayesian posterior (Murphy, 2012). Usually, the computation of this quantity is not tractable and we have to resort to approximate methods. Variational Inference (VI) (Blei et al., 2017) is a popular method in Bayesian machine learning to compute an approximation of the Bayesian posterior. In standard VI settings, we choose a *tractable* family of probability distributions over $\boldsymbol{\Theta}$, denoted by $\mathcal{Q}$, and the learning problem consists in finding the probability distribution $q \in \mathcal{Q}$ which is closest to the Bayesian posterior in terms of the inverse KL divergence, $\arg\min_{q\in\mathcal{Q}} KL(q(\boldsymbol{\theta}), p(\boldsymbol{\theta}|D))$. Solving this minimization problem is equivalent to maximize the following function, which is known as the ELBO function,

$$q^\star(\boldsymbol{\theta}) = arg\max_{q\in\mathcal{Q}} \mathbb{E}_{q(\boldsymbol{\theta})}[\ln p(D|\boldsymbol{\theta})] - KL(q, \pi). \tag{3}$$

There are many methods for solving this maximization problem for different probabilistic models (Zhang et al., 2018).

2.2. PAC-Bayesian Theory and Bayesian statistics

Let us define the empirical log-loss for the model $\boldsymbol{\theta}$ on the sample D, denoted $\hat{L}(\boldsymbol{\theta}, D) = \frac{1}{n}\sum_{i=1}^{n} -\ln p(\boldsymbol{x}_i|\boldsymbol{\theta})$, and the expected log-loss for the model $\boldsymbol{\theta}$, denoted $L(\boldsymbol{\theta}) = \mathbb{E}_{\nu(\boldsymbol{x})}[-\ln p(\boldsymbol{x}|\boldsymbol{\theta})]$. PAC-Bayes theory (McAllester, 1999) provides probabilistic bounds over the averaging of the expected log-loss according to a probability distribution ρ over $\boldsymbol{\Theta}$, i.e., $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$. Although many different PAC-Bayes bounds are present in the literature, most of them only apply to bounded losses and do not cover the log-loss. (Alquier et al., 2016) introduced a PAC-Bayes bound for a restrictive set of unbounded losses, which were later extended to general unbounded losses (Germain et al., 2016; Shalaeva et al., 2019). We reproduce here this PAC-Bayes bound [2],

Theorem 1 (Germain et al. (2016))**.** *For any prior distribution π over $\boldsymbol{\Theta}$ and for any $\xi \in (0,1)$ and $c > 0$, with probability at least $1-\xi$ over draws of training data $D \sim \nu^n(\boldsymbol{x})$, for all distribution ρ over $\boldsymbol{\Theta}$ simultaneously,*

$$\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] \leq \mathbb{E}_{\rho(\boldsymbol{\theta})}[\hat{L}(\boldsymbol{\theta}, D)] + \frac{KL(\rho,\pi) + \ln\frac{1}{\xi} + \psi_{\pi,\nu}(c,n)}{c\,n},$$

where $\psi_{\pi,\nu}(c,n) = \ln \mathbb{E}_{\pi(\boldsymbol{\theta})}\mathbb{E}_{D\sim\nu^n(\boldsymbol{x})}[e^{c\,n(L(\boldsymbol{\theta})-\hat{L}(\boldsymbol{\theta},D))}]$.

But PAC-Bayes bounds also provide a well-founded approach to learning. As these bounds hold simultaneously for all densities ρ, the learning algorithm consists in choosing the distribution ρ which minimizes the upper bound over the *generalization risk*. Fortunately, we can compute the ρ distribution minimizing the PAC-Bayes bound of Theorem 1 for constant c, ξ, n and D values, because the $\psi_{\pi,\nu}(c,n)$ term is also constant wrt ρ. (Germain et al., 2016) noted that the Bayesian posterior distribution is the minimum of this PAC-Bayes bound over the expected log-loss $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$,

Lemma 1. *(Germain et al., 2016) The Bayesian posterior $p(\boldsymbol{\theta}|D)$ is the distribution over $\boldsymbol{\Theta}$ which minimizes the PAC-Bayes bound introduced in Theorem 1 for $c=1$ and constant ξ, n and D values.*

3. The Bayesian posterior is suboptimal for generalization

In the previous section, we saw that the Bayesian posterior minimizes a PAC-Bayes upper bound over the expected log-loss. So, by minimizing the PAC-Bayes bound, we aim to minimize the expected log-loss $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$. In fact, the minimum of the PAC-Bayes bound (i.e., the Bayesian posterior) converges, in the large sample limit and in probability, to a distribution minimizing the expected log-loss $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$ due to well-known asymptotic results of the Bayesian posterior under model misspecification (Kleijn et al., 2012). And this distribution can be characterized as a Dirac-delta distribution, denoted $\delta_{\boldsymbol{\theta}^\star_{ML}}(\boldsymbol{\theta})$, centered around $\boldsymbol{\theta}^\star_{ML}$, which is the parameter that minimizes the KL divergence wrt the true distribution, $\boldsymbol{\theta}^\star_{ML} = \arg\min_\theta KL(\nu(\boldsymbol{x}), p(\boldsymbol{x}|\boldsymbol{\theta}))$. This also applies for the variational posterior (Wang & Blei, 2019), i.e the variational posterior also converges in the large sample limit to $\delta_{\boldsymbol{\theta}^\star_{ML}}(\boldsymbol{\theta})$, a minimum of the expected log-loss function. See Appendix D for a formal proof of these statements.

But the question is whether the minimization of the expected log-loss is a good strategy for minimizing the cross-entropy function (i.e., to find densities which optimally averages, in terms of generalization performance, our model class).

In principle, this is a good strategy because, by the Jensen inequality, the expected log-loss is an upper bound of the cross-entropy function,

$$\underbrace{\mathbb{E}_{\nu(\boldsymbol{x})}[-\ln \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]]}_{CE(\rho)} \leq \underbrace{\mathbb{E}_{\rho(\boldsymbol{\theta})}[\mathbb{E}_{\nu(\boldsymbol{x})}[-\ln p(\boldsymbol{x}|\boldsymbol{\theta})]]}_{\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]}. \tag{4}$$

This strategy would be *optimal* if the minimum of the expected log-loss was also the minimum of the cross-entropy loss. But, as shown in the following result, this only happens when the best model in isolation, $p(\boldsymbol{x}|\boldsymbol{\theta}^\star_{ML})$, provides better performance than any model averaging, $\mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]$,

Lemma 2. *A distribution minimizing $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$, denoted $\rho^\star_{ML}$, is also a minimizer of the cross-entropy loss $CE(\rho)$ if and only if for any distribution ρ over $\boldsymbol{\Theta}$ we have that,*

$$KL(\nu(\boldsymbol{x}), p(\boldsymbol{x}|\boldsymbol{\theta}^\star_{ML})) \leq KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]).$$

And $\rho^\star_{ML}$ can always be characterized as a Dirac-delta distribution center around $\boldsymbol{\theta}^\star_{ML}$, i.e., $\rho^\star_{ML}(\boldsymbol{\theta}) = \delta_{\boldsymbol{\theta}^\star_{ML}}(\boldsymbol{\theta})$.

According to this result, the Bayesian posterior is an optimal learning strategy under perfect model specification because we have that $KL(\nu(\boldsymbol{x}), p(\boldsymbol{x}|\boldsymbol{\theta}^\star_{ML})) = 0$, and $\rho^\star_{ML}$ will be a minimum of $CE(\rho)$. But perfect model specification rarely happens in practice. The problem with the Bayesian posterior lies in the inequality of Equation (4), which is the result of the application of a first-order Jensen inequality (Liao & Berg, 2019). And a first-order Jensen inequality induces a *linear* bound whose minimum is always at the border of the solution space, i.e., a Dirac-delta distribution. For this reason, we also refer to the expected log-loss $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$ as a first-order oracle bound, and to the PAC-Bayes bound of Theorem 1 as a first-order PAC-Bayes bound. But if we use a tighter second-order Jensen inequality to upper bound the cross-entropy loss, we will never end up in these extreme, no-model-averaging, solutions. Figure E.5 at the appendix graphically illustrates this situation.

[2] This bound is usually expressed in terms of any $\lambda > 0$, which we equivalently set here as $\lambda = c\,n$, for any $c > 0$.

4. Second-order PAC-Bayes bounds

Here, we exploit second-order Jensen bounds (Becker, 2012; Liao & Berg, 2019) for deriving a tighter bound over the cross-entropy function.

Theorem 2. *Any distributions ρ over $\boldsymbol{\Theta}$ satisfy that,*

$$CE(\rho) \leq \mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho),$$

where $\mathbb{V}(\rho)$ is the normalized variance of $p(\boldsymbol{x}|\boldsymbol{\theta})$ wrt $\rho(\boldsymbol{\theta})$,

$$\mathbb{V}(\rho) = \mathbb{E}_{\nu(\boldsymbol{x})}[\frac{1}{2\max_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})^2}\mathbb{E}_{\rho(\boldsymbol{\theta})}[(p(\boldsymbol{x}|\boldsymbol{\theta}) - p(\boldsymbol{x}))^2]].$$

This second-order Jensen bound differs from the expected log-loss (or first-order Jensen bound) in this new variance term $\mathbb{V}(\rho)$, which is positive when the ρ distribution is not a Dirac-delta distribution. So, this second-order Jensen bound is tighter than the the expected log-loss function and, also, induces high variance solutions when it is minimized.

But the key point is that a distribution minimizing this new second-order Jensen bound induces better model averaging solutions than a distribution minimizing the expected log-loss function,

Lemma 3. *Let us denote $\rho^\star_{J2}$ and $\rho^\star_{ML}$ a distribution minimizing the second-order oracle bound of Theorem 2 and $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$, respectively. The following inequality holds*

$$KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho^\star_{J2}}[p(\boldsymbol{x}|\boldsymbol{\theta})]) \leq KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho^\star_{ML}}[p(\boldsymbol{x}|\boldsymbol{\theta})]),$$

and the equality holds if we are under perfect model specification.

The above results show that a learning strategy based on the minimization of this second-order Jensen bound is better than a learning strategy based on the minimization of the expected log-loss, which is the case of the Bayesian posterior. In fact, in case of perfect model specification, the minimum of the second-order Jensen bound equals the minimum of the expected log-loss function.

The following result introduces a (second-order) PAC-Bayes bound over the second-order Jensen bound, which also provides generalization guarantees over the performance of the posterior predictive distribution by upper bounding the cross-entropy function.

Theorem 3. *For any prior distribution ρ over $\boldsymbol{\Theta}$ independent of D and for any $\xi \in (0,1)$ and $c > 0$, with probability at least $1-\xi$ over draws of training data $D \sim \nu^n(\boldsymbol{x})$, for all distribution ρ over $\boldsymbol{\Theta}$, simultaneously,*

$$CE(\rho) \leq \mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho) \leq$$

$$\mathbb{E}_{\rho(\boldsymbol{\theta})}[\hat{L}(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D) + \frac{KL(\rho,\pi) + \frac{1}{2}\ln\frac{1}{\xi} + \frac{1}{2}\psi'_{\pi,\nu}(c,n)}{cn},$$

where $\psi'_{\pi,\nu}(c,n)$ is the same term as in Theorem 1 adapted to this setting and $\hat{\mathbb{V}}(\rho, D)$ is the empirical version of $\mathbb{V}(\rho)$.

Again, the $\psi'_{\pi,\nu}(c,n)$ term depends on $\nu(\boldsymbol{x})$ and can not be computed, but we can minimize this PAC-Bayes bound for constant c, ξ, n and D values, because $\psi'_{\pi,\nu}(c,n)$ is also constant wrt ρ.

The key part of this new PAC-Bayes bound is the variance term $\mathbb{V}(\rho, D)$, which measures the diversity in the predictions made by each of the models in isolation.

5. PAC^2-Variational Learning

Our learning strategy is then to minimize the second-order PAC-Bayes bound introduced in Theorem 3 because it is a *probabilistic approximate correct* bound over the generalization performance of the posterior predictive distribution.

Like in variational inference (see Section 2.1), we can choose a tractable family of densities $\rho(\boldsymbol{\theta}|\boldsymbol{\lambda}) \in \mathcal{Q}$, parametrized by some parameter vector $\boldsymbol{\lambda}$, to solve the minimization of the second-order PAC-Bayes bound of Theorem 3. By discarding constant terms of this bound wrt ρ and setting $c = 1$ in order to keep the connection with Bayesian approaches[3], the minimization problem can be written as,

$$\arg\min_{\lambda} \mathbb{E}_{\rho(\boldsymbol{\theta}|\lambda)}[\hat{L}(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D) + \frac{KL(\rho,\pi)}{n}. \quad (5)$$

Appendix B provides more details about this method and other novel learning algorithms.

6. Empirical Evaluation

We perform the empirical evaluation on two data sets, Fashion-Mnist (Xiao et al., 2017) and CIFAR-10 (Krizhevsky, 2009) using a multi-layer perceptron with 20 hidden units and a mean-field Normal distribution for modeling the distribution over the weights of the network. Full details are given in Appendix C.

The empirical results obtained here (see Figure C.4) validate the main hypothesis of our work: the use of tighter Jensen bounds addressing the gap introduced when upper bounding the cross-entropy function $CE(\rho)$ leads to learning algorithms that generalize better.

7. Conclusions

This work performs a novel theoretical analysis of the generalization capacity of Bayesian model averaging under model misspecification and provides strong theoretical arguments showing that Bayesian methods are suboptimal for learning predictive models when the model family is misspecified. These theoretical insights can be of help to better understand the generalization performance of Bayesian approaches.

[3] Appendix H.3 further discusses how to set this parameter.

Acknowledgments

This work is funded by the Spanish Ministry of Science, Innovation and Universities under the project TIN2015- 74368-JIN and by a Jose Castillejo scholarship CAS19/00279. I also thank Yevgeny Seldin from University of Copenhagen for fruitful discussions regarding the application of PAC-Bayesian theory.

References

Alquier, P. and Guedj, B. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Alquier, P., Ridgway, J., and Chopin, N. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

Becker, R. A. The variance drain and Jensen's inequality. *CAEPR Working Paper No. 2012-004.*, 2012. URL http://dx.doi.org/10.2139/ssrn.2027471.

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., et al. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

Bernardo, J. M. and Smith, A. F. *Bayesian Theory*, volume 405. John Wiley & Sons, 2009.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Chérief-Abdellatif, B.-E. and Alquier, P. Mmd-Bayes: Robust Bayesian estimation via maximum mean discrepancy. *arXiv preprint arXiv:1909.13339*, 2019.

Dietterich, T. G. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Friedman, B. *Principles and techniques of applied mathematics*. Courier Dover Publications, 1990.

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. PAC-Bayesian theory meets Bayesian inference. In *Advances in Neural Information Processing Systems*, pp. 1884–1892, 2016.

Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

Grünwald, P. The safe Bayesian: learning the learning rate via the mixability gap. In *International Conference on Algorithmic Learning Theory*, pp. 169–183. Springer, 2012.

Grünwald, P. Safe probability. *Journal of Statistical Planning and Inference*, 195:47–63, 2018.

Grünwald, P. and Van Ommen, T. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. Bayesian model averaging: a tutorial. *Statistical science*, pp. 382–401, 1999.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Insua, D. R. and Ruggeri, F. *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media, 2012.

Jankowiak, M., Pleiss, G., and Gardner, J. R. Parametric Gaussian process regressors. *arXiv preprint arXiv:1910.07123*, 2019.

Jewson, J., Smith, J. Q., and Holmes, C. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

Kleijn, B. J. K., Van der Vaart, A. W., et al. The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

Knoblauch, J. Robust deep Gaussian processes. *arXiv preprint arXiv:1904.02303*, 2019.

Knoblauch, J., Jewson, J., and Damoulas, T. Generalized variational inference. *arXiv preprint arXiv:1904.02063*, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Kuncheva, L. I. and Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

Langford, J. and Seeger, M. Bounds for averaging classifiers. Technical report, 2001.

Letarte, G., Germain, P., Guedj, B., and Laviolette, F. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 6869–6879, 2019.

Liao, J. and Berg, A. Sharpening Jensen's inequality. *The American Statistician*, 73(3):278–281, 2019.

Masegosa, A. R. Learning under model misspecification: Applications to variational and ensemble methods. *arXiv preprint arXiv:1912.08335*, 2019.

McAllester, D. A. PAC-Bayesian model averaging. In *COLT*, volume 99, pp. 164–170. Citeseer, 1999.

Mohamed, S., Rosca, M., Figurnov, M., and Mnih, A. Monte Carlo gradient estimation in machine learning. *arXiv preprint arXiv:1906.10652*, 2019.

Murphy, K. P. *Machine learning: A probabilistic perspective*. MIT press, 2012.

Seeger, M. Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations. Technical report, University of Edinburgh, 2003.

Shalaeva, V., Esfahani, A. F., Germain, P., and Petreczky, M. Improved PAC-Bayesian bounds for linear regression. *arXiv preprint arXiv:1912.03036*, 2019.

Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J., Ren, J., and Nado, Z. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, pp. 13969–13980, 2019.

Walker, S. G. Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 143(10): 1621–1633, 2013.

Wang, C. and Blei, D. M. A general method for robust Bayesian modeling. *Bayesian Analysis*, 13(4):1163–1191, 2018.

Wang, Y. and Blei, D. Variational Bayes under model misspecification. In *Advances in Neural Information Processing Systems*, pp. 13357–13367, 2019.

Wang, Y., Kucukelbir, A., and Blei, D. M. Robust probabilistic modeling with Bayesian data reweighting. In *International Conference on Machine Learningd*, pp. 3646–3655. JMLR. org, 2017.

Wenzel, F., Roth, K., Veeling, B. S., Świątkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. How good is the Bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

Wilson, A. G. and Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Zhang, T. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

Zhang, T. et al. From epsilon-entropy to kl-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006.

A. Related Work

PAC-Bayesian theory (McAllester, 1999) provides probably approximately correct (PAC) bounds on the generalization risk (i.e., with probability $1-\xi$, the generalization risk is at most ϵ away from the training risk.) Although PAC-Bayesian theory is mostly a *frequentist* method, connections between PAC-Bayes and Bayesian methods have been explored since the beginnings of the theory (Langford & Seeger, 2001; Seeger, 2003). But it was in (Germain et al., 2016) were a neat connection was established between Bayesian learning and PAC-Bayesian theory. However, they did not directly study the generalization performance of Bayesian model averaging and did not consider model misspecification.

There is a large literature showing that that Bayesian inference can behave suboptimally if the model is wrong (Grünwald, 2012; 2018; Grünwald & Van Ommen, 2017; Jewson et al., 2018; Walker, 2013). The *Safe Bayesian* method is probably the best-known framework (Grünwald, 2012). The main point of this approach is to guarantee the concentration of the Bayesian posterior around the best possible model. But this work shows that the concentration of the Bayesian posterior around the best possible model is the main reason behind the suboptimal generalization performance of Bayesian methods under model misspecification.

Other related works (Bissiri et al., 2016; Chérief-Abdellatif & Alquier, 2019; Jankowiak et al., 2019; Knoblauch et al., 2019; Letarte et al., 2019) propose Bayesian-like algorithms based on the use of alternative belief updating schemes which differs from the Bayesian approach. Again, the final goal of these works is not to study the generalization risk of Bayesian model averaging. Some of them (Bissiri et al., 2016; Knoblauch et al., 2019) are based on the use of alternative loss functions, different from the log-likelihood function, to derive new Bayesian-like algorithms. In this sense, our proposed learning algorithms employ a special loss which includes a correcting term to account for model misspecification.

Zhang (Zhang, 2006; Zhang et al., 2006) introduces information theoretical bounds which consider the log-loss and model misspecification. But the bounded quantity is not the generalization error of Bayesian model averaging, and their focus is to find the best single model, not the best model averaging distribution.

Robust Bayesian methods (Berger et al., 1994; Insua & Ruggeri, 2012; Knoblauch, 2019; Wang & Blei, 2018; Wang et al., 2017) also address the problem of model misspecification. But their focus is mainly in how *fix* the inference procedure under *small deviations from the assumptions* (e.g. outliers, error measurements, etc) rather than systematically study the generalization performance.

B. Learning by minimizing second-order PAC-Bayes bounds

Our learning strategy is then to minimize the second-order PAC-Bayes bound introduced in Theorem 3 because it is a *probabilistic approximate correct* bound over the generalization performance of the resulting posterior predictive distribution. In this case, we do not have a closed-form solution to find the distribution ρ minimizing this second-order PAC-Bayes bound. But, in the next subsections, we introduce several scalable methods for (approximately) solving this minimization problem.

B.1. PAC^2-Variational Learning

Like in variational inference (see Section 2.1), we can choose a tractable family of densities $\rho(\boldsymbol{\theta}|\boldsymbol{\lambda}) \in \mathcal{Q}$, parametrized by some parameter vector $\boldsymbol{\lambda}$, to solve the minimization of the second-order PAC-Bayes bound of Theorem 3. By discarding constant terms of this bound wrt ρ and setting $c=1$ in order to keep the connection with Bayesian approaches[4], the minimization problem can be written as,

$$\arg\min_{\lambda} \mathbb{E}_{\rho(\boldsymbol{\theta}|\lambda)}[\hat{L}(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D) + \frac{KL(\rho,\pi)}{n}. \quad \text{(B.6)}$$

Appendix H.2 shows a numerically stable algorithm to perform optimization over this objective function using modern black-box variational methods (Zhang et al., 2018). We refer to this learning method as PAC^2-Variational learning.

Tighter second-order Jensen bounds

One the key contributions of our work is the identification of the *Jensen inequality* as significant barrier when learning under model misspecification. Our assumption is that our learning strategy should further improve if we use tighter (second-order) Jensen bounds. (Liao & Berg, 2019) proposes an alternative second-order Jensen bound which is tighter than the one considered in Theorem 2. This new bound suggests a new learning algorithm, we called PAC^2_T-Variational learning, where the subscript T highlights that it relies on *tighter* Jensen bounds. The only difference with the presented approach in Equation (B.6) is the use of a different variance term, denoted $\hat{\mathbb{V}}_T(\rho, D)$,

$$\hat{\mathbb{V}}_T(\rho, D) = \frac{1}{n}\sum_{i=1}^{n} h(m_{\boldsymbol{x}_i}, \mu_{\boldsymbol{x}_i})\mathbb{E}_{\rho(\boldsymbol{\theta})}[(p(\boldsymbol{x}_i|\boldsymbol{\theta}) - p(\boldsymbol{x}_i))^2], \quad \text{(B.7)}$$

where $\mu_{\boldsymbol{x}_i} = \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}_i|\boldsymbol{\theta})]$, $m_{\boldsymbol{x}_i} = \max_{\boldsymbol{\theta}} p(\boldsymbol{x}_i|\boldsymbol{\theta})$ and $h(m,\mu) = \frac{\ln\mu - \ln m}{(m-\mu)^2} + \frac{1}{\mu(m-\mu)}$. Appendix H.1 further discusses this algorithm and Appendix H.2 shows a numerically stable form for learning with it.

Figures B.2 and B.3 compare the Bayesian/Variational pos-

[4] Appendix H.3 further discusses how to set this parameter.

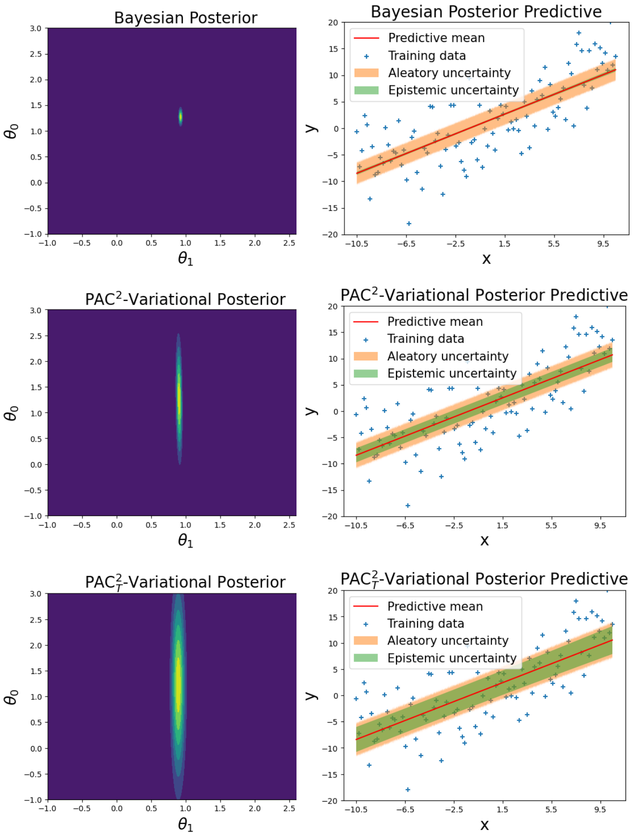

Figure B.2. **Linear Model**: The data generating model, $\nu(y|x)$, is $y \sim \mathcal{N}(\mu = 1 + x, \sigma^2 = 5)$. The probabilistic model, $p(y|x, \boldsymbol{\theta})$ is $y \sim \mathcal{N}(\mu = \theta_0 + \theta_1 x, \sigma^2 = 1)$. So, the probabilistic model is miss-specified, but note how model misspecification mainly affects to the parameter θ_0. The prior $\pi(\boldsymbol{\theta})$ is a standard Normal distribution. We learn from 100 samples. The Bayesian posterior $p(\boldsymbol{\theta}|D)$ is a bidimensional Normal distribution and can be exactly computed. The PAC^2_T-Variational posterior is computed using a bidimensional Normal distribution approximation family and the Adam optimizer. The uncertainty in the predictions is computed by random sampling first from $\boldsymbol{\theta} \sim \rho(\boldsymbol{\theta})$ and then from $y \sim p(y|x, \boldsymbol{\theta})$ for 100 times. We plot the predictive mean plus/minus two standard deviations. We distinguish between *epistemic uncertainty* which comes from the uncertainty in $\rho(\boldsymbol{\theta})$ and *aleatory uncertainty* which comes from the uncertainty in $p(y|x, \boldsymbol{\theta})$. The test log-likelihood of the posterior predictive distribution is -13.63, -12.99, -9.71 and -6.93 for the MAP, the Bayesian, the PAC^2-Variational and the PAC^2_T-Variational posterior predictive distributions, respectively.

terior and the Bayesian/Variational posterior predictive distribution and its PAC^2-Variational counter-parts for a simple misspecified linear model and for a neural network based model using sinusoidal data, respectively. In Appendix F, we also illustrate how PAC^2-Variational methods recover the Bayesian posterior and agrees with standard variational methods under perfect model specification in both models.

Finally, we highlight that the variational inference algorithm (see Equation (3)) can be similarly derived from the constraint minimization of the PAC-Bayes bound of Theorem 1, which misses the $\hat{\mathbb{V}}(\rho, D)$ term because it is based on a first-order Jensen bound, i.e., the expected log-loss $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$.

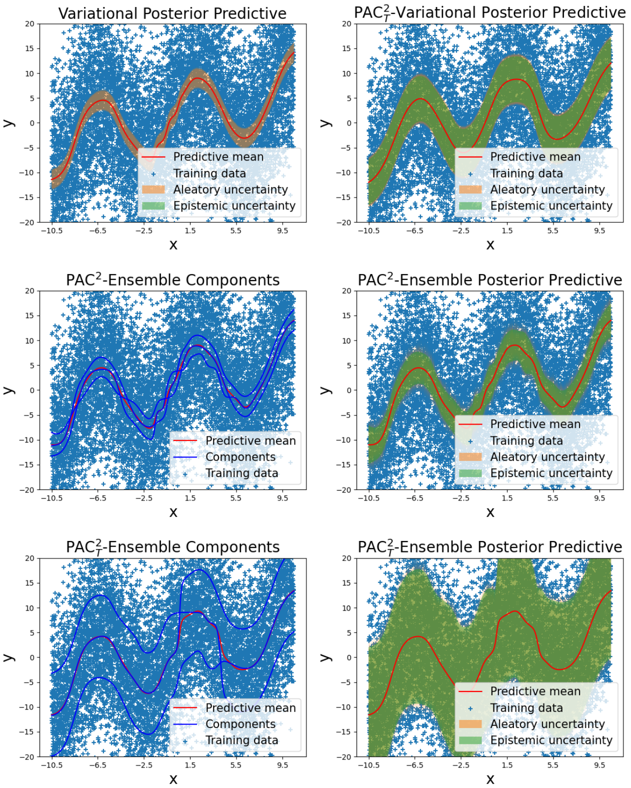

Figure B.3. **Neural Network Model:** The data generating model, $\nu(y|x)$, is a sinusoidal function plus Gaussian noise, $y = s(x) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2 = 10)$. The probabilistic model $p(y|x, \boldsymbol{\theta})$ is $y = f_{\boldsymbol{\theta}}(x) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2 = 1)$, where f is a MLP parametrized by $\boldsymbol{\theta}$, with one hidden layer with 20 units able to approximate $s(x)$. The prior $\pi(\boldsymbol{\theta})$ is a standard Normal distribution. 10000 training samples. The Variational and the PAC^2-Variational approximation family Q is a fully factorized mean field Normal distribution and it is optimized with the Adam optimizer. The test log-likelihood of the posterior predictive distribution is -50.44, -50.20, -36.03 and -24.55 for the MAP, the Variational, the PAC^2-Variational and the PAC^2_T-Variational models, respectively. While for the PAC^2-Ensemble and the PAC^2_T-Ensemble models is -39.16 and -15.88, respectively. Uncertainty estimations are computed as in Figure B.2.

B.2. PAC^2-Ensemble Learning

Ensemble models (Dietterich, 2000) are based on the combination of a group of models to obtain better predictions than the predictions of any single model of the group alone. Our work provides a novel explanation of why the so-called *diversity* of the ensemble (Kuncheva & Whitaker, 2003) is key to have powerful ensembles. Finally, we present a novel ensemble learning algorithm.

Let us denote ρ_E a mixture of Dirac-delta distributions centered around a set of E parameters $\{\boldsymbol{\theta}_j\}_{1\leq j\leq E}$,

$$\rho_E(\boldsymbol{\theta}) = \sum_{j=1}^{E} \frac{1}{E}\delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}). \tag{B.8}$$

The posterior predictive induced by ρ_E correspond to the averaging of the probabilities of E different models, $p_E(\boldsymbol{x}) = \mathbb{E}_{\rho_E(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})] = \frac{1}{E}\sum_{j=1}^{E} p(\boldsymbol{x}|\boldsymbol{\theta}_j)$.

Although we can directly adapt Theorem 3 to these settings, we would face the problem that the KL term would be equal to infinity because of the entropy of a Dirac-delta distribution is minus infinity. So, the bound would be completely vacuous. We can get rid of this problem by defining a prior $\pi_F(\boldsymbol{\theta})$ as a mixture of Dirac-delta distributions over finite-precision parameters. It means that the support of the distribution π_F is contained in $\boldsymbol{\Theta}_F$, which denotes the space of real number vectors of dimension M that can be represented under a finite-precision scheme using F bits to encode each element of the vector. So, we have that $supp(\pi_F) \subseteq \boldsymbol{\Theta}_F \subseteq \boldsymbol{\Theta} \subseteq \mathcal{R}^M$. Mathematically, this prior distribution π_F is expressed as,

$$\pi_F(\boldsymbol{\theta}) = \sum_{\boldsymbol{\theta}'\in\boldsymbol{\Theta}_F} w_{\boldsymbol{\theta}'}\delta_{\boldsymbol{\theta}'}(\boldsymbol{\theta}), \tag{B.9}$$

where $w_{\boldsymbol{\theta}'}$ are positive scalars values parametrizing this prior distribution. They satisfy that $w_{\boldsymbol{\theta}'} \geq 0$ and $\sum w_{\boldsymbol{\theta}'} = 1$. If the support of ρ_E is also contained in $\boldsymbol{\Theta}_F$, i.e., $supp(\rho_E) \subseteq \boldsymbol{\Theta}_F$, then the KL divergence between ρ_E and π_F is well defined,

$$KL(\rho_E, \pi_F) = \sum_{j=1}^{E} \frac{1}{E} \ln \frac{\frac{1}{E}\delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_j)}{w_{\boldsymbol{\theta}_j}\delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_j)} = \frac{1}{E}\sum_{j=1}^{E} \ln \frac{\frac{1}{E}}{w_{\boldsymbol{\theta}_j}}. \tag{B.10}$$

The following result provides a second-order PAC-Bayes bound for an ensemble of models.

Theorem B.4. *For any prior distribution π_F over $\boldsymbol{\Theta}_F$, as defined in Equation (B.9), and for any $\xi \in (0,1)$ and $c > 0$, with probability at least $1-\xi$ over draws of training data $D \sim \nu^n(\boldsymbol{x})$, for all densities ρ_E with $supp(\rho_E) \subset \boldsymbol{\Theta}_F$, as defined in Equation (B.8), simultaneously,*

$$CE(\rho_E) \leq \mathbb{E}_{\rho_E(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho_E) \leq$$
$$\mathbb{E}_{\rho_E(\boldsymbol{\theta})}[\hat{L}(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho_E, D) + \frac{KL(\rho_E, \pi_F) + \ln\frac{1}{\xi} + \psi'_{\pi_F,\nu}(c,n)}{c\,n},$$

where $\psi'_{\pi_F,\nu}(c,n)$ is the same term as in Theorem 3, and $\hat{\mathbb{V}}(\rho_E, D)$ is the empirical variance,

$$\hat{\mathbb{V}}(\rho_E, D) = \frac{1}{nE}\sum_{i=1}^{n}\sum_{j=1}^{E} \frac{(p(\boldsymbol{x}_i|\boldsymbol{\theta}_j) - p_E(\boldsymbol{x}_i))^2}{2\max_{\boldsymbol{\theta}} p(\boldsymbol{x}_i|\boldsymbol{\theta})^2}.$$

Proof. This result follows by considering Theorem 3 given a density ρ_E as defined in Equation (B.8). □

$\hat{\mathbb{V}}(\rho_E, D)$ can be interpreted as a measure of the *diversity* of the ensemble (Kuncheva & Whitaker, 2003). According to the above result, to learn optimal ensembles, we need to trade-off how well we fit the data (i.e., low values for $\mathbb{E}_{\rho_E(\boldsymbol{\theta})}[\hat{L}(\boldsymbol{\theta}, D)] = \frac{1}{E}\sum_{i=1}^{E}\hat{L}(\boldsymbol{\theta}_j, D)$) while also maintaining high *diversity* (i.e., high $\hat{\mathbb{V}}(\rho_E, D)$ values) to make the ensemble work (i.e., to reduce the generalization bound given in Theorem B.4). This is a clear explanation of why ensembles need diversity to generalize better, but it is out the scope of the this paper to further explore this claim.

Theorem B.4 shows how to design a learning algorithm for building ensembles by minimizing the generalization upper bound. The algorithm we propose is obtained by fixing $c = 1$ and discarding constant terms wrt ρ_E. We call it the *PAC^2-Ensemble learning algorithm*. Minimize wrt ρ_E reduces to a gradient-based optimization of the $\{\boldsymbol{\theta}_j\}_{1\leq j\leq E}$ parameters of the following objective function,

$$\arg\min_{\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_E\}} \frac{1}{E}\sum_{j=1}^{E}\hat{L}(\boldsymbol{\theta}_j, D) - \hat{\mathbb{V}}(\rho_E, D) - \frac{1}{E}\sum_{j=1}^{E}\frac{\ln\pi_F(\boldsymbol{\theta}_j)}{n} \tag{B.11}$$

We can also employ the tighter second-order Jensen bound mentioned in Section B.1 to derive a new objective function by replacing $\hat{\mathbb{V}}(\rho_E, D)$ with $\hat{\mathbb{V}}_T(\rho_E, D)$ (see Equation (B.7)). We call this algorithm *PAC^2_T-Ensemble learning*. Appendix H.2 discusses a numerically stable algorithm to perform gradient-based optimization over this objective function.

Figures B.3 illustrates this algorithm on a sinusoidal data sample. Appendix G also illustrates the advantage of using this approach wrt the variational approach introduced in Section B.1 in a multimodal data set.

We highlight again that we can also derive an ensemble learning algorithm from the PAC-Bayes bound of Theorem 1, which corresponds to the use of a first-order Jensen bound (see Section 3). In this case, the $\hat{\mathbb{V}}(\rho_E, D)$ term would disappear from Equation (B.11), and a global minimum would be achieved when all the $\{\boldsymbol{\theta}_j\}_{1\leq j\leq E}$ parameters collapse to the MAP parameter, $\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} \ln p(D|\boldsymbol{\theta}) + \ln\pi(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) - \frac{\ln\pi(\boldsymbol{\theta})}{n}$. So, the performance of an ensemble based on the minimization of the (first-order) PAC-Bayes bound of Theorem 1 is identical to the performance a single model ensemble.

C. Empirical Evaluation

We perform the empirical evaluation on two data sets, Fashion-Mnist (Xiao et al., 2017) and CIFAR-10 (Krizhevsky, 2009) and two prediction tasks. A supervised task, where the goal is to predict the class label of each image, and a self-supervised task, where the goal is to predict the pixels of the below half part of the image given the pixels of the upper half part of the same image. For the self-supervised task, we employ two data models: a Normal distribution for continuous value predictions and a Binomial one for binarized pixels. The underlying neural network is the same multi-layer perceptron with 20 hidden units used in the results of Figure B.3. Full details in Appendix I.

Figure C.4 shows the result of this evaluation. These results validate the main hypothesis of our work: the use of tighter Jensen bounds addressing the gap introduced when upper bounding the cross-entropy function $CE(\rho)$ leads to learning algorithms that generalize better. PAC^2-Variational and PAC^2-Ensemble methods, based on second-order Jensen bounds, have better predictive performance than standard variational methods and single model ensembles, based on first-order Jensen bounds (see the discussions at the end of Section B.1 and Section B.2, respectively). And, in turn, the PAC^2_T-Variational and PAC^2_T-Ensemble methods, based on tighter second-order Jensen bounds, generalize better than the PAC^2-Variational and PAC^2-Ensemble methods, respectively. The only exception is the classification task in CIFAR-10, where all the variational approaches do not perform better than the MAP model. We highlight that all the models of the ensembles are randomly initialized with the same parameters and are jointly optimized, so it is the variance term $\hat{\mathbb{V}}(\rho_E, D)$ the one which induces better generalization performance, because, otherwise, all the models of the ensemble would be identical.

D. Proofs

D.1. The Dirac-delta Function

The Dirac-delta function (Friedman, 1990), $\delta_{\boldsymbol{\theta}_0} : \boldsymbol{\Theta} \rightarrow \mathcal{R}^+$ with parameter $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, is a generalized function with the following property

$$\mathbb{E}_{\delta_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})}[f(\boldsymbol{\theta})] = \int \delta_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} = f(\boldsymbol{\theta}_0) \qquad \text{(D.12)}$$

for any continuous function $f : \boldsymbol{\Theta} \rightarrow \mathcal{R}$. The Dirac-delta function also defines the density function of a probability distribution because it is positive and, by the above property, $\int \delta_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$. This Dirac-delta distribution is a degenerated probability distribution which always samples the same value $\boldsymbol{\theta}_0$ because $\delta_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) = 0$ if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ (Friedman, 1990).

D.2. Proofs Section 3

Minimum of the expected log-loss

The following lemma defines that a Dirac-delta distribution center around $\boldsymbol{\theta}^\star_{ML}$ is a minimum of the expected log-loss,

Lemma D.4. *The distribution $\rho^\star_{ML}$ defined as a Dirac-delta distribution center around $\boldsymbol{\theta}^\star_{ML}$, $\rho^\star_{ML}(\boldsymbol{\theta}) = \delta_{\boldsymbol{\theta}^\star_{ML}}(\boldsymbol{\theta})$ is a minimum of the expected log-loss,*

$$\rho^\star_{ML} = \arg\min_{\rho} \mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})],$$

where $\boldsymbol{\theta}^\star_{ML} = \arg\min_\theta KL(\nu(\boldsymbol{x}), p(\boldsymbol{x}|\boldsymbol{\theta}))$.

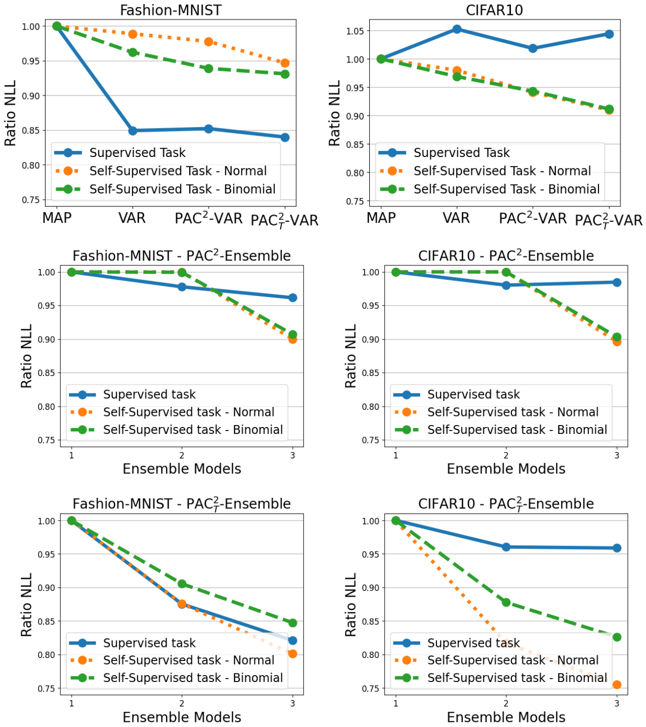

Figure C.4. **PAC^2-Variational and PAC^2-Ensemble Learning.** Top two figures shows the ratio of the test NLL wrt to a MAP model for the Variational (VAR), PAC^2-Variational (PAC^2-VAR) and PAC^2_T-Variational (PAC^2_T-VAR) methods. For the rest of the figures, we evaluate ensembles with two and three models and also report the ratio of the test NLL wrt a single model ensemble.

Proof. We first have that $KL(\nu(\boldsymbol{x}), p(\boldsymbol{x}|\boldsymbol{\theta})) = L(\boldsymbol{\theta}) - H(\nu)$, where $H(\nu)$ denotes the entropy of $\nu(\boldsymbol{x})$. As $H(\nu)$ is constant wrt $\boldsymbol{\theta}$, $\boldsymbol{\theta}^\star_{ML}$ is also a minimum of $L(\boldsymbol{\theta})$. We also have that $\int (L(\boldsymbol{\theta}) - H(\nu))\rho(\boldsymbol{\theta})d\boldsymbol{\theta} \geq \min_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}) - H(\nu)) \int \rho(\boldsymbol{\theta})d\boldsymbol{\theta}$, because $L(\boldsymbol{\theta}) - H(\nu) \geq 0$. Rearranging terms, we have that $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] \geq L(\boldsymbol{\theta}^\star_{ML}) = \mathbb{E}_{\delta_{\boldsymbol{\theta}^\star_{ML}}(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$. □

This result states that both the Bayesian and the Variational posterior converges to a minimum of the expected log-loss,

Lemma D.5. *Under the technical conditions established in (Kleijn et al., 2012; Wang & Blei, 2019), the Bayesian posterior $p(\boldsymbol{\theta}|D)$ and the Variational posterior $q^\star(\boldsymbol{\theta})$ converge, in the large sample limit and in probability, to a minimum of the expected log-loss, $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$.*

Proof. This results follows from the convergence results given in (Kleijn et al., 2012; Wang & Blei, 2019), which state that $p(\boldsymbol{\theta}|D)$ and $q^\star(\boldsymbol{\theta})$ converge, in the large sample limit and in probability, to $\delta_{\boldsymbol{\theta}^\star_{ML}}(\boldsymbol{\theta})$. And, by Lemma D.4, $\delta_{\boldsymbol{\theta}^\star_{ML}}(\boldsymbol{\theta})$ is a minimum of the expected log-loss. □

We now characterize any distribution minimizing the expected log-loss,

Lemma D.6. *Let us denote $\Omega_{\boldsymbol{\theta}_0}$ the set of parameters which defines the same distribution than $\boldsymbol{\theta}_0$, i.e. if $\boldsymbol{\theta}' \in \Omega_{\boldsymbol{\theta}_0} \subseteq \boldsymbol{\Theta}$ then $\forall \boldsymbol{x} \in supp(\nu) \subseteq \mathcal{X}$ $p(\boldsymbol{x}|\boldsymbol{\theta}_0) = p(\boldsymbol{x}|\boldsymbol{\theta}')$, where $supp(\nu)$ denotes the support of $\nu(\boldsymbol{x})$. We have that any distribution $\rho^\star$ which is a minimum of the expected log-loss, $\rho^\star = \arg\min_\rho \mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$, satisfies that $supp(\rho^\star) \subseteq \Omega_{\boldsymbol{\theta}^\star_{ML}}$, where $supp(\rho^\star)$ denotes the support of $\rho^\star(\boldsymbol{\theta})$, and also that $\mathbb{V}(\rho^\star) = 0$.*

Proof. From Lemma D.4, we have that $\rho^\star_{ML}(\boldsymbol{\theta}) = \delta_{\boldsymbol{\theta}^\star_{ML}}(\boldsymbol{\theta})$ is a minimum of the expected log-loss. In consequence, if $\rho^\star$ is a minimum of the expected log-loss, then $\mathbb{E}_{\rho^\star}[L(\boldsymbol{\theta})] = \mathbb{E}_{\rho^\star_{ML}}[L(\boldsymbol{\theta})]$. From this equality, we arrive to the following equality $\mathbb{E}_{\rho^\star}[L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\star_{ML})] = 0$. And this last equality can only be satisfied if the conditions of the lemma hold because, as noted in the proof of Lemma D.4, $\boldsymbol{\theta}^\star_{ML}$ is a minimum of $L(\boldsymbol{\theta})$ and, as a consequence, $L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\star_{ML}) \geq 0$.

As $supp(\rho^\star) \subseteq \Omega_{\boldsymbol{\theta}^\star_{ML}}$, we have that, by definition, $\forall \boldsymbol{\theta}' \in supp(\rho^\star), \forall \boldsymbol{x} \in supp(\nu) \subseteq \mathcal{X}$ $p(\boldsymbol{x}|\boldsymbol{\theta}^\star_{ML}) = p(\boldsymbol{x}|\boldsymbol{\theta}')$. And we can deduce that $\mathbb{V}(\rho^\star) = 0$. □

D.3. Proof of Lemma 2

Lemma 2. *A distribution minimizing $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$, denoted $\rho^\star_{ML}$, is also a minimizer of the cross-entropy loss $CE(\rho)$ if and only if for any distribution ρ over $\boldsymbol{\Theta}$ we have that,*

$$KL(\nu(\boldsymbol{x}), p(\boldsymbol{x}|\boldsymbol{\theta}^\star_{ML})) \leq KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]).$$

And $\rho^\star_{ML}$ can always be characterized as a Dirac-delta distribution center around $\boldsymbol{\theta}^\star_{ML}$, i.e., $\rho^\star_{ML}(\boldsymbol{\theta}) = \delta_{\boldsymbol{\theta}^\star_{ML}}(\boldsymbol{\theta})$.

Proof. We first proof that if the inequality of the lemma holds, then $\rho^\star_{ML}$, is also a minimizer of the cross-entropy loss $CE(\rho)$. Due to the following equality,

$$KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]) = CE(\rho) - H(\nu). \quad \text{(D.13)}$$

Any ρ minimizing Equation (D.13) is also a minimum of the cross-entropy loss, $CE(\rho)$, because $H(\nu)$ is constant w.r.t. ρ. And, according to Lemma D.6, any density $\rho^\star_{ML}$ minimizing the expected log-loss satisfies that $\mathbb{E}_{\rho^\star_{ML}(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})] = p(\boldsymbol{x}|\boldsymbol{\theta}^\star_{ML})$.

So, if the inequality of Lemma 2 holds, $\rho^\star_{ML}$ is also a minimum of Equation (D.13) and, as a consequence, $\rho^\star_{ML}$ is also a minimum of the cross-entropy loss, $CE(\rho)$.

We now have to prove that if $\rho^\star_{ML}$ is also a minimizer of the cross-entropy loss $CE(\rho)$, then the inequality of the lemma also holds. Equivalently, we have that if the inequality of the lemma does not hold, then $\rho^\star_{ML}$ can not be a minimizer of Equation (D.13) and, as a consequence, is not a minimizer of $CE(\rho)$.

Finally, according to Lemma D.6 we always have that $\mathbb{E}_{\rho^\star_{ML}(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})] = p(\boldsymbol{x}|\boldsymbol{\theta}^\star_{ML})$, so $\rho^\star_{ML}$ can always be characterized as a Dirac-delta distribution center around $\boldsymbol{\theta}^\star_{ML}$, i.e., $\rho^\star_{ML}(\boldsymbol{\theta}) = \delta_{\boldsymbol{\theta}^\star_{ML}}(\boldsymbol{\theta})$. □

D.4. Proofs Section 4

Proof of Theorem 2

We start giving the proof of Theorem 2,

Theorem 2. *Any distributions ρ over $\boldsymbol{\Theta}$ satisfies that,*

$$CE(\rho) \leq \mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho),$$

where $\mathbb{V}(\rho)$ is the normalized variance of $p(\boldsymbol{x}|\boldsymbol{\theta})$ wrt $\rho(\boldsymbol{\theta})$,

$$\mathbb{V}(\rho) = \mathbb{E}_{\nu(\boldsymbol{x})}[\frac{1}{2\max_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})^2}\mathbb{E}_{\rho(\boldsymbol{\theta})}[(p(\boldsymbol{x}|\boldsymbol{\theta}) - p(\boldsymbol{x}))^2]].$$

Proof. Applying Taylor's theorem to $\ln y$ about μ with a mean-value form of the remainder gives,

$$\ln y = \ln \mu + \frac{1}{\mu}(y - \mu) - \frac{1}{2g(y)^2}(y - \mu)^2,$$

where $g(y)$ is a real value between y and μ. Therefore,

$$\begin{aligned}\mathbb{E}_{\rho(\boldsymbol{\theta})}[\ln p(\boldsymbol{x}|\boldsymbol{\theta})] &= \ln \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})] \\ &\quad - \mathbb{E}_{\rho(\boldsymbol{\theta})}[\frac{1}{2g(p(\boldsymbol{x}|\boldsymbol{\theta}))^2}(p(\boldsymbol{x}|\boldsymbol{\theta}) - p(\boldsymbol{x}))^2]\end{aligned}$$

Rearranging we have

$$\begin{aligned}&-\ln \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})] = \\ &-\mathbb{E}_{\rho(\boldsymbol{\theta})}[\ln p(\boldsymbol{x}|\boldsymbol{\theta})] - \mathbb{E}_{\rho(\boldsymbol{\theta})}[\frac{1}{2g(p(\boldsymbol{x}|\boldsymbol{\theta}))^2}(p(\boldsymbol{x}|\boldsymbol{\theta}) - p(\boldsymbol{x}))^2] \\ &\leq -\mathbb{E}_{\rho(\boldsymbol{\theta})}[\ln p(\boldsymbol{x}|\boldsymbol{\theta})] \\ &\qquad - \frac{1}{2\max_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})^2}\mathbb{E}_{\rho(\boldsymbol{\theta})}[(p(\boldsymbol{x}|\boldsymbol{\theta}) - p(\boldsymbol{x}))^2],\end{aligned}$$

where the inequality is derived from that fact $(p(\boldsymbol{x}|\boldsymbol{\theta}) - p(\boldsymbol{x}))^2$ is always positive and that for all $\theta \in supp(\rho)$ the term $g(p(\boldsymbol{x}|\boldsymbol{\theta}))$, which is a real number between $p(\boldsymbol{x}|\boldsymbol{\theta})$ and $\mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]$, is upper bounded by $\max_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})$. Finally, the result of the theorem is derived by taking expectation wrt $\nu(\boldsymbol{x})$ on both sides of the above inequality. □

Proof of Lemma 3

Before proving Lemma 3, we need to introduce the following preliminary result,

Lemma D.7. *If a density ρ' over $\boldsymbol{\Theta}$ has null variance, i.e. $\mathbb{V}(\rho')=0$, where $\mathbb{V}(\rho')$ is defined in Theorem 2, then we have the following equality,*

$$CE(\rho') = \mathbb{E}_{\rho'(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$$

Proof. We first have that if $\mathbb{V}(\rho')=0$, then all parameters in the support of ρ' induce the same probability distribution over $\boldsymbol{x}$. Because the variance term will be not null as soon as we have two parameters in the support of ρ' which induces two different probability distributions over $\boldsymbol{x}$. We then denote $\boldsymbol{\theta}'$ to a parameter in the support of ρ', i.e., $\boldsymbol{\theta}' \in supp(\rho')$.

In this case, we can deduce that if $\mathbb{V}(\rho')=0$, then $\mathbb{E}_{\rho'(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})] = \mathbb{E}_{\delta_{\boldsymbol{\theta}'}(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]$, because all the parameters in the support of ρ' induce the same distribution over $\boldsymbol{x}$. So, ρ' can be reparametrized as a Dirac-delta distribution. And, by Equation (D.12), we have that $\mathbb{E}_{\delta_{\boldsymbol{\theta}'}(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})] = p(\boldsymbol{x}|\boldsymbol{\theta}')$.

Finally, we have that,

$$\begin{aligned}\mathbb{E}_{\rho'(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] &= L(\boldsymbol{\theta}') \\ &= \mathbb{E}_{\nu(\boldsymbol{x})}[\ln \frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta}')}] \\ &= \mathbb{E}_{\nu(\boldsymbol{x})}[\ln \frac{1}{\mathbb{E}_{\rho'(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]}] \\ &= CE(\rho'),\end{aligned}$$

where the first equality follows because, as we have seen above, ρ' acts as a Dirac-delta distributions (see Equation (D.12)), the second equality follows from the definition of $L(\boldsymbol{\theta})$, the third equality follows again from the property of Dirac-delta distributions, and the last equality follows from the definition of the cross-entropy function CE. □

We now introduce the proof of Lemma 3.

Lemma 3. *Let us denote $\rho^\star_{J^2}$ and $\rho^\star_{ML}$ a distribution minimizing the second-order Jensen bound of Theorem 2 and $\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})]$, respectively. The following inequality holds*

$$KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho^\star_{J^2}}[p(\boldsymbol{x}|\boldsymbol{\theta})]) \leq KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho^\star_{ML}}[p(\boldsymbol{x}|\boldsymbol{\theta})]).$$

Under perfect model specification or in a degenerated model averaging setting (see Lemma 2) both KL terms are equal.

Proof. Let us define Δ the space of distributions ρ over $\boldsymbol{\Theta}$ whose variance is null, i.e., if $\rho \in \Delta$ then $\mathbb{V}(\rho')=0$, where $\mathbb{V}(\rho')$ is defined in Theorem 2. Then, we have that the minimum of the second-order Jensen bound for all the distributions $\rho \in \Delta$ can be written as,

$$\min_{\rho\in\Delta} \mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho) = \min_{\rho\in\Delta} \mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] = \mathbb{E}_{\rho^\star_{ML}(\boldsymbol{\theta})}[L(\boldsymbol{\theta})],$$

where the first inequality follows because if $\rho \in \Delta$ then $\mathbb{V}(\rho)=0$, and the second equality follows from Lemma D.6. We also have that

$$\mathbb{E}_{\rho^\star_{J^2}(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho^\star_{J^2}) \leq \mathbb{E}_{\rho^\star_{ML}(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] \qquad \text{(D.14)}$$

because, by definition, the left hand side of the inequality is the minimum of the second-order Jensen bound for all the distributions $\rho(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$, while the right hand side of the inequality is the minimum of the second-order Jensen bound but only for those distributions $\rho \in \Delta$.

By chaining the above inequality of Equation (D.14) with the second-order Jensen bound inequality of Theorem 2, we have

$$CE(\rho^\star_{J^2}) \leq \mathbb{E}_{\rho^\star_{ML}(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] \qquad \text{(D.15)}$$

Finally, we have that $\mathbb{E}_{\rho^\star_{ML}(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] = CE(\rho^\star_{ML})$ by Lemma D.7 because $\mathbb{V}(\rho^\star_{ML})=0$ due to Lemma D.6. So, combining $\mathbb{E}_{\rho^\star_{ML}(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] = CE(\rho^\star_{ML})$ with the inequality of Equation (D.15), we have that,

$$CE(\rho^\star_{J^2}) \leq CE(\rho^\star_{ML}) \qquad \text{(D.16)}$$

which proofs the inequality of the lemma, because $KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]) = CE(\rho) - H(\nu)$.

If we are under the conditions of Lemma 2, we have that for any distribution ρ over $\boldsymbol{\Theta}$, $KL(\nu(\boldsymbol{x}), p(\boldsymbol{x}|\boldsymbol{\theta}^\star_{ML})) \leq KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})])$. From this condition, we can deduce that for any distribution ρ over $\boldsymbol{\Theta}$,

$$C(\rho^\star_{ML}) \leq CE(\rho), \qquad \text{(D.17)}$$

because $KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]) = CE(\rho) - H(\nu)$ and $KL(\nu(\boldsymbol{x}), p(\boldsymbol{x}|\boldsymbol{\theta}^\star_{ML})) = KL(\nu(\boldsymbol{x}), \mathbb{E}_{\rho^\star_{ML}}[p(\boldsymbol{x}|\boldsymbol{\theta})]) = C(\rho^\star_{ML}) - H(\nu)$. Combining Equations (D.15) and (D.17), we have that $CE(\rho^\star_{J^2}) = C(\rho^\star_{ML})$ under the conditions of Lemma 2, proving that the inequality of the lemma becomes an equality. □

Characterizing under which conditions the inequality of Lemma 3 becomes strict is an open problem and a subject of future research.

Proof of Theorem 3

Before proving Theorem 3, we need to introduce the following result,

Lemma D.8. *For any prior distribution π over $\boldsymbol{\Theta}$ and for any distribution ρ over $\boldsymbol{\Theta}$, the second-order Jensen bound of Theorem 2 bound can be expressed as follows,*

$$\mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho) = \mathbb{E}_{\rho(\boldsymbol{\theta},\boldsymbol{\theta}')}[L(\boldsymbol{\theta},\boldsymbol{\theta}')],$$

where $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, $\rho(\boldsymbol{\theta}, \boldsymbol{\theta}') = \rho(\boldsymbol{\theta})\rho(\boldsymbol{\theta}')$, and $L(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is defined as

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\theta}') =& \mathbb{E}_{\nu(\boldsymbol{x})}[\ln \frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta})} \\ & - \frac{1}{2 \max_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})^2} \left(p(\boldsymbol{x}|\boldsymbol{\theta})^2 - p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}')\right)], \end{aligned}$$

Proof. The proof is straightforward by applying first this equality,

$$\mathbb{E}_{\rho(\boldsymbol{\theta})}[(p(\boldsymbol{x}|\boldsymbol{\theta}) - p(\boldsymbol{x}))^2] = \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})^2] - \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]^2,$$

and, after that, the following equality,

$$\begin{aligned} \mathbb{E}_{\rho(\boldsymbol{\theta}, \boldsymbol{\theta}')}[p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}')] =& \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})\mathbb{E}_{\rho(\boldsymbol{\theta}')}[p(\boldsymbol{x}|\boldsymbol{\theta}')]] \\ =& \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]^2 \end{aligned}$$

□

We now proceed to the proof of Theorem 3.

Theorem 3. *For any prior distribution π over $\boldsymbol{\Theta}$ independent of D and for any $\xi \in (0, 1)$ and $c > 0$, with probability at least $1 - \xi$ over draws of training data $D \sim \nu^n(\boldsymbol{x})$, for all distribution ρ over $\boldsymbol{\Theta}$, simultaneously,*

$$CE(\rho) \leq \mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho) \leq$$

$$\mathbb{E}_{\rho(\boldsymbol{\theta})}[\hat{L}(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D) + \frac{KL(\rho, \pi) + \frac{1}{2}\ln\frac{1}{\xi} + \frac{1}{2}\psi'_{\pi,\nu}(c, n)}{cn},$$

where $\psi'_{\pi,\nu}(c, n)$ is the same term as in Theorem 1 adapted to this setting and $\hat{\mathbb{V}}(\rho, D)$ is the empirical version of $\mathbb{V}(\rho)$.

Proof. By Lemma D.8, we can express the problem using an *extended log-loss function* $L(\boldsymbol{\theta}, \boldsymbol{\theta}')$. Then, we apply (Alquier et al., 2016, Theorem 4.1) to this loss using as prior $\pi(\boldsymbol{\theta}, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{\theta}')$ and have

$$\begin{aligned} \mathbb{E}_{\rho(\boldsymbol{\theta}, \boldsymbol{\theta}')}[L(\boldsymbol{\theta}, \boldsymbol{\theta}')] \leq& \mathbb{E}_{\rho(\boldsymbol{\theta}, \boldsymbol{\theta}')}[\hat{L}(\boldsymbol{\theta}, \boldsymbol{\theta}', D)] \\ & + \frac{1}{\lambda}(KL(\rho(\boldsymbol{\theta}, \boldsymbol{\theta}'), \pi(\boldsymbol{\theta}, \boldsymbol{\theta}')) + \ln\frac{1}{\xi} + \psi'_{\pi,\nu}(\lambda, n)), \end{aligned}$$

where $\psi'_{\pi,\nu}(\lambda, n)$ is defined as

$$\psi'_{\pi,\nu}(\lambda, n) = \ln \mathbb{E}_{\pi(\boldsymbol{\theta}, \boldsymbol{\theta}')}\mathbb{E}_{D \sim \nu^n(\boldsymbol{x})}[e^{\lambda(L(\boldsymbol{\theta}, \boldsymbol{\theta}') - \hat{L}(\boldsymbol{\theta}, \boldsymbol{\theta}', D))}].$$

The PAC-Bayes bound of the theorem follows by rewriting $\mathbb{E}_{\rho(\boldsymbol{\theta}, \boldsymbol{\theta}')}[L(\boldsymbol{\theta}, \boldsymbol{\theta}')] = \mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] - \mathbb{V}(\rho)$ and $\mathbb{E}_{\rho(\boldsymbol{\theta}, \boldsymbol{\theta}')}[\hat{L}(\boldsymbol{\theta}, \boldsymbol{\theta}', D)] = \mathbb{E}_{\rho(\boldsymbol{\theta})}[\hat{L}(\boldsymbol{\theta}, D)] - \hat{\mathbb{V}}(\rho, D)$, and noting that $KL(\rho(\boldsymbol{\theta}, \boldsymbol{\theta}'), \pi(\boldsymbol{\theta}, \boldsymbol{\theta}')) = 2KL(\rho(\boldsymbol{\theta}), \pi(\boldsymbol{\theta}))$. Finally, we reparametrized λ as $\lambda = 2cn$. □

E. First-order vs second-order Jensen bounds

Figure E.5 illustrates the behavior of first-order and second-order Jensen bounds under perfect model specification and model-misspecification. In this case, the model space is composed of two Binomial models. The ρ distribution can be defined with a single parameter between 0 and 1. Extremes ($\rho = 0$ or $\rho = 1$) values choose a single model. Non-extremes values induce an averaging of the models. Left figure shows the situation of model misspecification because there exists an optimal ρ distribution minimizing the cross-entropy function (i.e., achieving optimal prediction performance), but the expected log-loss is minimized with a Dirac-delta distribution (i.e., $\rho = 1$), choosing the best single model. While the minimum of the second-order Jensen bound achieves a better result. Right figure shows the case of perfect model specification. In this case, the data generating distribution corresponds to one of the models, $\rho = 1$ is the distribution minimizing the cross-entropy loss (i.e., achieving optimal prediction performance). In this case, both the expected log-loss and the second-order Jensen bound are minimized with a Dirac-delta distribution (i.e., $\rho = 1$), choosing the best single model and achieving optimal results.

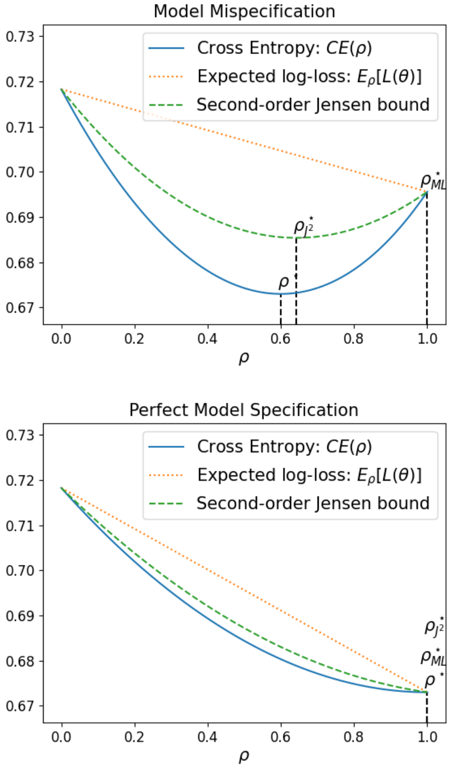

Figure E.5. First-Order vs Second-Order Jensen Bounds.

F. Learning under Perfect Model Specification

Figure F.6 also illustrates the behavior of the PAC^2-Variational learning methods under perfect model specification. In this case, we can observe that the posterior predictive of the PAC^2-Variational algorithm matches the posterior predictive distribution of standard Bayesian/Variational methods.

G. Multimodal Data

Figure G.7 illustrates the performance of the Variational, PAC^2-Variational and PAC^2-Ensemble approaches over multimodal data. The figure shows how variational approaches based on a mean-field Gaussian approximation family are not able to properly capture the multimodality nature of the data. However, ensemble approaches defines a more flexible approximate family (i.e., a mixture of Dirac-delta distributions) for the posterior distribution and can perfectly capture this multimodality. Again, we also see like the approach based on tighter second-order Jensen bounds performs much better.

H. Learning algorithms

H.1. Tighter Jensen Bounds

The next result shows how we can define a tighter second-order Jensen bound using the Jensen inequality presented in (Liao & Berg, 2019).

Theorem H.5. *Any distribution ρ over $\boldsymbol{\Theta}$ satisfies the following inequality,*

$$CE(\rho) \leq \mathbb{E}_{\rho(\boldsymbol{\theta})}[L(\boldsymbol{\theta})] - \mathbb{V}_T(\rho),$$

where $\mathbb{V}_T(\rho)$ is the normalized variance of $p(\boldsymbol{x}|\boldsymbol{\theta})$ wrt $\rho(\boldsymbol{\theta})$,

$$\mathbb{V}_T(\rho) = \mathbb{E}_{\nu(\boldsymbol{x})}[h(m_{\boldsymbol{x}}, \mu_{\boldsymbol{x}})\mathbb{E}_{\rho(\boldsymbol{\theta})}[(p(\boldsymbol{x}|\boldsymbol{\theta}) - p(\boldsymbol{x}))^2]].$$

and $\mu_{\boldsymbol{x}} = \mathbb{E}_{\rho(\boldsymbol{\theta})}[p(\boldsymbol{x}|\boldsymbol{\theta})]$, $m_{\boldsymbol{x}} = \max_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})$ and $h(m,\mu) = \frac{\ln\mu - \ln m}{(m-\mu)^2} + \frac{1}{\mu(m-\mu)}$

Proof sketch. Apply (Liao & Berg, 2019)'s result to the random variable $p(\boldsymbol{x}|\boldsymbol{\theta})$, following the same strategy used in the proof of Theorem 2. □

As shown in (Liao & Berg, 2019), we have that $\forall \boldsymbol{x} \in \mathcal{X} \quad h(m_{\boldsymbol{x}}, \mu_{\boldsymbol{x}}) \geq \frac{1}{2\max_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})^2}$. In consequence, the above bound is tighter because $\mathbb{V}_T(\rho) \geq \mathbb{V}(\rho)$. Similarly, we can show that the same inequality for the empirical versions of both variance terms, i.e., $\hat{\mathbb{V}}_T(\rho, D) \geq \hat{\mathbb{V}}(\rho, D)$.

The issue with the introduction of the $\mathbb{V}_T(\rho)$ term is that we can not derive the corresponding second-order PAC-Bayes bound using the same approach of Theorem 3. This is therefore an open and future research problem.

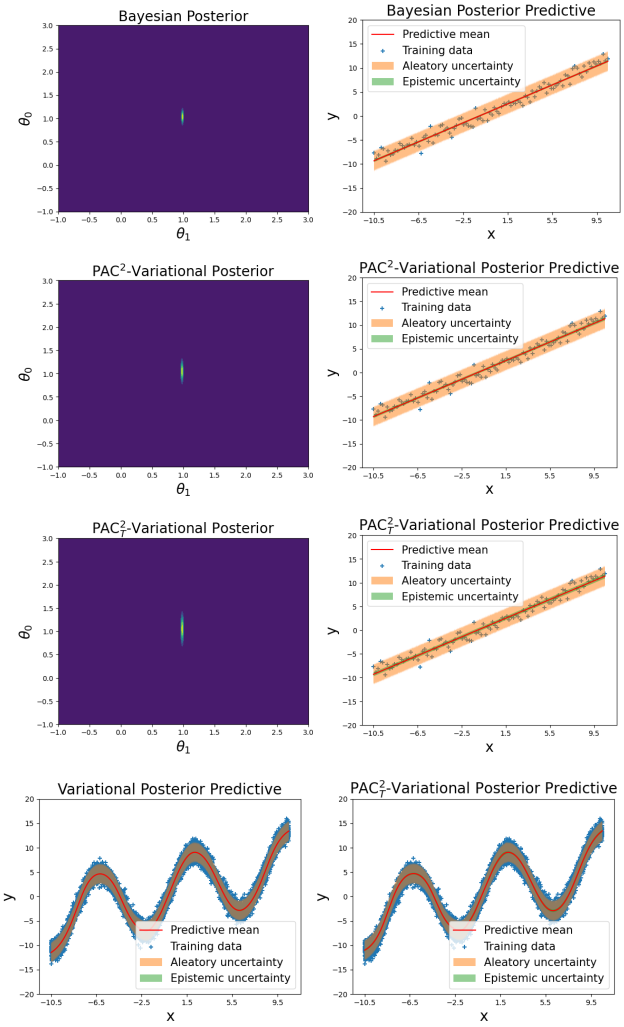

Figure F.6. **Perfect Model Specification**: For the top six figures, we have the same settings that in Figure F.6, but in this case we are under perfect model specification i.e., $\nu(y|x) = \mathcal{N}(\mu = 1+x, \sigma^2 = 1)$. For the two figures at the bottom, we have the same settings than in Figure B.3, but, again, under perfect model specification i.e., $\nu(y|x) = \mathcal{N}(s(x), \sigma^2 = 1)$. For the linear model, the test log-likelihood of the posterior predictive distribution is -1.43, -1.41, -1.41 and -1.42 for the MAP, the Bayesian, the PAC^2-Variational and the PAC^2_T-Variational posterior predictive distributions, respectively. For the sinusoidal data, the test log-likelihood of the posterior predictive distribution is -1.42 for the MAP, the Variational, the PAC^2-Variational and the PAC^2_T -Variational posterior predictive distributions. The test log-likelihood is computed from an independent test set of 10000 samples.

H.2. Numerically stable PAC^2-Variational Learning

PAC^2-Variational Inference is based on the optimization of Equation (B.6). This optimization is not feasible in this current formulation. We can rewrite Equation (B.6) by employing the expression provided in Lemma D.8, by multiplying and dividing the term $\mathbb{V}_T$ by $2\max_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})^2$ and, also, by

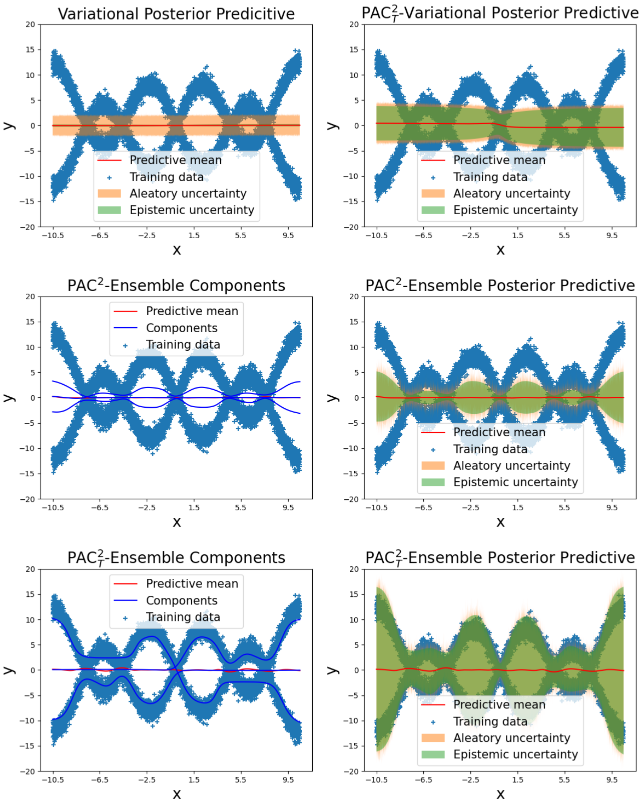

Figure G.7. **Multimodal data:** Same settings than in previous Figures B.3. But now the data generative function is a mixture of two sinusoidal functions (i.e. multimodal data) with $\sigma^2 = 1$. The test log-likelihood of the posteriors predictive distribution is -18.82, -12.72, -8.55, -13.30 and -4.00 for the Variational, PAC^2-Variational, PAC^2_T-Variational, PAC^2-Ensemble and PAC^2_T-Ensemble models, respectively.

multiplying by n which does not affect the minimization. Equation (B.6) can be expressed as follows,

$$\mathbb{E}_{\rho(\boldsymbol{\theta},\boldsymbol{\theta}'|\lambda)}[\sum_{i=1}^{n} -\ln p(\boldsymbol{x}_i|\boldsymbol{\theta}) - h(\alpha_{\boldsymbol{x}_i})\hat{\mathbb{V}}(\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{\theta}')] + KL(\rho,\pi) \tag{H.18}$$

where $\rho(\boldsymbol{\theta},\boldsymbol{\theta}'|\boldsymbol{\lambda}) = \rho(\boldsymbol{\theta}|\boldsymbol{\lambda})\rho(\boldsymbol{\theta}'|\boldsymbol{\lambda})$ and

$$\begin{aligned}
m_{\boldsymbol{x}_i} &= \max_{\boldsymbol{\theta}} \ln p(\boldsymbol{x}_i|\boldsymbol{\theta})\\
\alpha_{\boldsymbol{x}_i} &= \ln(\exp(\ln p(\boldsymbol{x}_i|\boldsymbol{\theta}) - m_{\boldsymbol{x}_i})\\
&\quad + \exp(\ln p(\boldsymbol{x}_i|\boldsymbol{\theta}') - m_{\boldsymbol{x}_i})) - \ln 2\\
h(\alpha_{\boldsymbol{x}}) &= \frac{\alpha_{\boldsymbol{x}}}{(1-\exp(\alpha_{\boldsymbol{x}}))^2} + \frac{1}{\exp(\alpha_{\boldsymbol{x}})(1-\exp(\alpha_{\boldsymbol{x}}))}\\
\mathbb{V}(\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{\theta}') &= \exp(2\ln p(\boldsymbol{x}_i|\boldsymbol{\theta}) - 2m_{\boldsymbol{x}_i})\\
&\quad - \exp(\ln p(\boldsymbol{x}_i|\boldsymbol{\theta}) + \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}') - 2m_{\boldsymbol{x}_i}).
\end{aligned}$$

For supervised classification problems, we fix $m_{\boldsymbol{x}_i} = 0$, assuming it is possible to make always a perfect classification. For regression tasks, we sample two parameters[5] $\boldsymbol{\theta}', \boldsymbol{\theta}'' \sim \rho(\boldsymbol{\theta}|\boldsymbol{\lambda})$ and take $m_{\boldsymbol{x}_i} = \max(\ln p(\boldsymbol{x}_i|\boldsymbol{\theta}), \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}')) + \epsilon$, with $\epsilon = 0.1$ to avoid numerically stability problems when computing $h(\alpha_{\boldsymbol{x}})$. Even though, better strategies can be defined to compute $m_{\boldsymbol{x}_i}$.

We can minimize Equation (H.18) using any gradient-based optimizing algorithm. Unbiased estimates of the gradient of Equation (H.18) can be computed using appropriate Monte-Carlo gradient estimation methods (Mohamed et al., 2019). We apply *stop-gradient* operation over $m_{\boldsymbol{x}_i}$ and $h(\alpha_{\boldsymbol{x}_i})$ to avoid problems deriving a *max* or a *log-sum-exp* operation.

H.3. Numerically stable PAC^2-Ensemble Learning

PAC^2-Ensemble Learning is done by minimizing Equation (B.11). We now show how to express this function in a numerically stable way, using the same strategy employed in Appendix H.2,

$$\begin{aligned}
\sum_{j=1}^{E}\sum_{k=1}^{E}\sum_{i=1}^{n} &-\ln p(\boldsymbol{x}_i|\boldsymbol{\theta}_j)\\
&- h(\alpha_{\boldsymbol{x}_i})\exp(2\ln p(\boldsymbol{x}_i|\boldsymbol{\theta}_j) - 2m_{\boldsymbol{x}})\\
&+ h(\alpha_{\boldsymbol{x}_i})\exp(\ln p(\boldsymbol{x}_i|\boldsymbol{\theta}_j) + \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}_k) - 2m_{\boldsymbol{x}})\\
&- \ln \pi(\boldsymbol{\theta}_j)
\end{aligned}$$

where we also need the following definitions

$$\begin{aligned}
m_{\boldsymbol{x}_i} &= \max_{\boldsymbol{\theta}} \ln p(\boldsymbol{x}_i|\boldsymbol{\theta})\\
\alpha_{\boldsymbol{x}_i} &= \ln \sum_{j=1}^{E} \exp(\ln p(\boldsymbol{x}_i|\boldsymbol{\theta}_j) - m_{\boldsymbol{x}_i}) - \ln E\\
h(\alpha_{\boldsymbol{x}}) &= \frac{\alpha_{\boldsymbol{x}}}{(1-\exp(\alpha_{\boldsymbol{x}}))^2} + \frac{1}{\exp(\alpha_{\boldsymbol{x}})(1-\exp(\alpha_{\boldsymbol{x}}))}.
\end{aligned}$$

Again, for supervised classification problems, we fix $m_{\boldsymbol{x}_i} = 0$, assuming it is possible to make always a perfect classification. For the rest of the cases, we take $m_{\boldsymbol{x}_i} = \max_j \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}_j)$.

Again, we apply *stop-gradient* operation over $m_{\boldsymbol{x}}$ and $h(\alpha_{\boldsymbol{x}_i})$ to avoid problems deriving a *max* or a *log-sum-exp* operation.

H.4. Setting the constant c

The presence of an arbitrary constant is a commonplace in PAC-Bayes bounds. According to PAC-Bayesian principles, this parameter can also be minimized. There are some specific approaches for doing that (Dziugaite & Roy, 2017) as this bound does not simultaneously hold for all $c > 0$ values. In this work, we set $c = 1$ to establish the previously mentioned link between PAC-Bayesian theory and Bayesian statistics (Germain et al., 2016). By optimizing this constant c, we may get some further increase in performance. However, it would require to bound the $\psi'_{\pi,\nu}(c,n)$ term, because,

[5] We employ same samples used for the gradient estimation.

in this case, this term would be involved in the optimization. We could apply the approaches presented in (Alquier & Guedj, 2018; Germain et al., 2016) to provide computable upper bounds over $\psi'_{\pi,\nu}(c,n)$, but it would require strong assumptions and would only apply to very simple models.

Another alternative would be to employ an independent validation data set. In this case, we could perform a grid search and choose the c value which leads to the model with better performance.

I. Empirical Evaluation Setup

For the artificial data sets, we employed the following experimental settings: a multilayer perceptron (MLP) with 20 hidden units and a hyperbolic tangent activation function, the Adam optimizer is used with learning rate 0.01, full-batch training and number of epochs equal to 5000. We also use 100 Monte-Carlo samples to compute the posterior predictive distribution, defined in Equation (1), for variational methods.

For the experiments with real data sets, we employ a MLP with 20 hidden units and a relu activation function, the Adam optimizer is used with learning rate 0.001, mini-batches with 100 samples and 100 epochs. We use 20 Monte-Carlo samples to compute the posterior predictive distribution for variational methods. We employ default train and test datasets.

The images of the CIFAR-10 are transformed to gray scale using *Tensorflow* method, *tf.image.rgb_to_grayscale*. Fashion-MNIST's and CIFAR-10's pixels values are normalized to the range 0-1. For this reason, for the self-supervised task with a Normal data model, we set the scale of the Normal distribution to 1/255.