PAC^m-Bayes

Narrowing the Empirical Risk Gap in the Misspecified Bayesian Regime

Warren R. Morningstar [1] **Alexander A. Alemi** [1] **Joshua V. Dillon** [1]

Abstract

While the decision-theoretic optimality of the Bayesian formalism under correct model specification is well-known (Berger, 2013), the Bayesian case becomes less clear under model misspecification (Grünwald et al., 2017; Ramamoorthi et al., 2015; Fushiki et al., 2005). To formally understand the consequences of Bayesian misspecification, this work examines the relationship between posterior predictive risk and its sensitivity to correct model assumptions, i.e., choice of likelihood and prior. We present the multi-sample PAC^m-Bayes risk. This risk is justified by theoretical analysis based on PAC-Bayes as well as empirical study on a number of toy problems. The PAC^m-Bayes risk is appealing in that it entails direct minimization of the Monte-Carlo approximated posterior predictive risk yet recovers both the Bayesian formalism as well as the MLE in its limits. Our work is heavily influenced by (Masegosa, 2019); our contributions are to align training and generalization risks while offering a tighter bound which empirically performs at least as well and sometimes much better.

1. Introduction

Pierre-Simon Laplace formulated one of the earliest Bayesian models (Laplace, 1781). Interested in the relative birth rates of boys and girls, he derived the Beta posterior for a Bernoulli likelihood with uniform prior. He then calculated the "posterior probability" that the girl birth rate exceeds the boy rate and found it to be about 10^{-42} from which he concluded that it is "as certain as any other moral truth" that humans give birth to more boys than girls (The presently accepted natural ratio is 105 males per 100 females (Ritchie, 2019)).

[1]Google Research, Mountain View, California, USA. Correspondence to: <wmorning@google.com>, <alemi@google.com>, <jvdillon@google.com>.

Under Review

Laplace's objective was to *infer* the parameter of his model. Broadly, science has followed suit. When a modern experiment such as the Large Hadron Collider at CERN processes terabytes of particle collision data (ATLAS Collaboration, 2012), or the Planck satallite maps the cosmic microwave background radiation (Planck Collaboration VI, 2019), they are in pursuit of the "moral truth" of some underlying parameter of the universe.

Contrast this with modern machine learning. The primary goal of modern machine learning is to build models that can form accurate *predictions*. We do not truly care about the value of the millionth weight in a deep neural network. We do not believe the parameters of the neural network are reflecting any "moral truths".

For well-specified models the goals of inference and prediction align. For misspecified models they might not. Optimizing for inference when you'll evaluate a model's predictive performance violates the golden rule of ML: *do unto training as done unto testing*. As made clear in the pivotal work of Masegosa (2019), both Bayesian inference and Maximum Likelihood target inferential rather than predictive risks, and can make poor predictions under model misspecification.

While this issue has been known for a while (Minka, 2000; Domingos, 1997), a general solution is elusive. In this work, we offer a tractable multi-sample bound on the true predictive risk that is asymptotically tight. This bound allows us to directly optimize predictive performance. We can smoothly interpolate between traditional Frequentist and Bayesian (inferential) approaches as well. We can provide the predictive performance of a finite mixture model without the philosophical or computational burden thereof.

2. Predictive and Inferential Risks

We begin with a *statistical model*: $p(X|\theta)$ defining a distribution of each observed datum X in terms of some parameters θ. After observing n data points drawn from some *true distribution* $X^n \overset{\text{def}}{=} \{X_i\}_i^n \overset{\text{iid}}{\sim} \nu(X)$, we form a distribution of parameters, $q(\Theta|\{x_i\}_i^n)$. In principle we can then compute the *predictive distribution*:

$$p(X|\{x_i\}_i^n) = \mathbb{E}_{q(\Theta|\{x_i\}_i^n)}\left[p(X|\Theta)\right]. \tag{1}$$

(For brevity we henceforth regard q's dependence on $\{x_i\}_i^n$ as implicit.) We judge the quality of this predictive distribution by measuring the relative entropy (Kullback-Leibler divergence) between it and the true distribution:

$$\begin{aligned}\mathsf{KL}[\nu(X); p(X|q)] &= \mathbb{E}_{\nu(X)}\left[\log \frac{\nu(X)}{p(X|q)}\right] \\ &= \mathbb{E}_{\nu(X)}[\log \nu(X)] - \mathbb{E}_{\nu(X)}\left[\log p(X|q)\right].\end{aligned}$$

Up to a constant outside our control (the continuous entropy of the true distribution) this defines the *true predictive risk*:

$$\mathcal{P}[q] \stackrel{\text{def}}{=} -\mathbb{E}_{\nu(X)}\left[\log \mathbb{E}_{q(\Theta)}[p(X|\Theta)]\right]. \tag{2}$$

In many cases the true predictive risk is ultimately what we care most about. It is what governs how much money our model will make, or how many lives it will save.

Not knowing the true distribution $\nu(X)$, we cannot directly minimize the true predictive risk. One thing we can compute is the *empirical predictive risk*:

$$\overline{\mathcal{P}}_n[q] \stackrel{\text{def}}{=} -\frac{1}{n}\sum_i^n \log \mathbb{E}_{q(\Theta)}[p(x_i|\Theta)]. \tag{3}$$

This is the observed average risk on the $\{x_i\}_i^n$ sample set. Akin to a training loss, the empirical predictive risk is a measure of how well we do at predicting the training data. If used as a target for optimization we can easily overfit. Training with this risk directly would amount to a type of *ensemble method* (Dietterich, 2000).

In contrast to the predictive risks, we'll also define the *inferential risks*. The *true inferential risk* (often just called the *true risk*):

$$\mathcal{R}[q] \stackrel{\text{def}}{=} -\mathbb{E}_{\nu(X)}\left[\mathbb{E}_{q(\Theta)}\left[\log p(X|\Theta)\right]\right], \tag{4}$$

and the corresponding *empirical inferential risk* (often called the *empirical risk*):

$$\overline{\mathcal{R}}_n[q] \stackrel{\text{def}}{=} -\frac{1}{n}\sum_i^n \mathbb{E}_{q(\Theta)}\left[\log p(x_i|\Theta)\right]. \tag{5}$$

For a variety of reasons, directly minimizing the inferential risk is fairly commonplace. It measures the average of the divergence between the true distribution $\nu(X)$ and the single-value parameter settings of the model $p(X|\theta)$. This is akin to doing variational optimization (Staines & Barber, 2012), and concentrates on a delta function corresponding to the best single-value parameter setting.

Jensen's inequality implies $-\log \mathbb{E}_{q(\Theta)}[p(x|\Theta)] \leq -\mathbb{E}_{q(\Theta)}[\log p(x|\Theta)]$ and so the inferential risks are upper bounds on the predictive risks:

$$\mathcal{P}[q] \leq \mathcal{R}[q] \qquad \overline{\mathcal{P}}_n[q] \leq \overline{\mathcal{R}}_n[q]. \tag{6}$$

In this way, minimizing the inferential risk is a *valid* strategy for achieving good predictions since it minimizes an upper bound on the predictive risk. When is it a *good* strategy? When is this bound tight? If our model is well-specified (Masegosa, 2019) (Proof replicated in appendix A.1). However, in cases of model misspecification this can break down severely. We would prefer to target the true predictive risk directly, but cannot since we do not know the true data distribution. What we need is a tractable bound on the true risks.

3. PAC-Bayes

While the empirical risks $(\overline{\mathcal{P}}, \overline{\mathcal{R}})$ provide unbiased estimates of the true risks $(\mathcal{P}, \mathcal{R})$, minimizing the empirical risks does not minimize the true risks:

$$\arg\min_q \mathcal{R}[q] = \arg\min_q \mathbb{E}\left[\overline{\mathcal{R}}_n[q]\right] \neq \mathbb{E}\left[\arg\min_q \overline{\mathcal{R}}_n[q]\right].$$

Said another way, the empirical risks do not provide a *bound* on the true risks.

Despite not being a valid bound, empirical risk minimization is quite popular. Minimizing the empirical (inferential) risk over the space of all possible distributions over parameters is the well known Maximum Likelihood method. This concentrates in a delta-function-like parameter distribution with all of its mass on the maximum likelihood parameter value.

We could similarly directly optimize the empirical predictive risk. In cases with bounded likelihoods this seems to perform decently well (e.g. the toy example of section 5) just as it does in the case of Maximum Likelihood. If our model is too expressive (e.g. unbounded likelihoods) minimizing the empirical risks will quickly start to concentrate on the *empirical* data distribution rather than the *true* distribution. Classic approaches limit model capacity by adding regularization or other tricks. If we instead had a valid bound on the true risks, we needn't worry. PAC-Bayes approaches provide such a bound.

We would really like to have some assurance that we won't overfit to our finite training data. We can formulate an upper bound on the true risks in terms of the empirical risks that nearly always hold. Such *probably approximately correct* (or PAC) bounds can be used to motivate Bayesian inference, demonstrating that the Bayesian posterior is the minimizer of a PAC-style upper bound on the true inferential risk $\mathcal{R}$ (Banerjee, 2006; Alquier et al., 2016) (Proof replicated in appendix A.2).

In light of these results we will define the following PAC-

inferential risk (or ELBO):

$$\widetilde{\mathcal{R}}_n[q;r,\beta] \stackrel{\text{def}}{=} \overline{\mathcal{R}}_n[q] + \frac{1}{\beta n}\mathsf{KL}\left[q(\Theta);r(\Theta)\right] \tag{7}$$

$$= \mathbb{E}_{q(\Theta)}\left[-\frac{1}{n}\sum_i \log p(x_i|\Theta) + \frac{1}{\beta n}\log\frac{q(\Theta)}{r(\Theta)}\right]. \tag{8}$$

Aside from constants independent of q, $\widetilde{\mathcal{R}}$ is a stochastic upper bound on $\mathcal{R}$. Intuitively, this is accomplished by ensuring that our parameter distribution $q(\Theta)$ can't stray too far from a *prior* $r(\Theta)$ we chose before looking at the data. Notice that ordinary Bayesian inference corresponds to minimizing this risk for $\beta = 1$ (Knoblauch et al., 2019; Bissiri et al., 2016). Furthermore, as $\beta \to \infty$ we recover the empirical risk $\overline{\mathcal{R}}$.

Because $\mathcal{P} \leq \mathcal{R}$, Bayesian inference is equivalent to (almost always) minimizing an upper bound on the *true predictive risk* $\mathcal{P}$. In the case of a well-specified model, $\min\mathcal{P} = \min\mathcal{R}$ and Bayesian inference targets not only optimal inferential power but also optimal predictive power. If you have the correct model, searching for the correct single parameter setting of the model is the right thing to do.

What ought we do if our model is misspecified?

4. PAC^m-Bayes

If our model is misspecified, there may be a large gap between the minimum of the predictive and inferential risks ($\min\mathcal{P} \ll \min\mathcal{R}$) as we'll demonstrate in our experiments below.

Our central contribution is to provide a new class of bounds, analogous to the PAC-style upper bounds on the inferential risk but targeting the predictive risk more directly.

The potential gap between the predictive and inferential risks came from invoking Jensen's inequality:

$$-\log\mathbb{E}_{q(\Theta)}[p(x|\Theta)] \leq -\mathbb{E}_{q(\Theta)}[\log p(x|\Theta)]. \tag{9}$$

The core insight is to explore a family of multisample stochastic bounds: (Burda et al., 2015; Mnih & Rezende, 2016)

$$-\log\mathbb{E}_{q(\Theta^m)}[p(x|\Theta^m)]$$

$$\leq -\mathbb{E}_{q(\Theta^m)}\left[\log\frac{1}{m}\sum_j^m p(x|\Theta_j)\right] \tag{10}$$

$$\leq -\mathbb{E}_{q(\Theta)}[\log p(x|\Theta)]. \tag{11}$$

Averaging a finite number of samples from our parameter distribution provides an unbiased estimate of the predictive likelihood. Taking the log of an unbiased estimator produces a *stochastic lower bound* (Burda et al., 2014; Grosse et al., 2016) that becomes tight asymptotically.

Our main result is in theorem 1 below. Despite looking rather complex, this PAC-Bound establishes that we are free to minimize the empirical predictive risk, without fear of overfitting, provided we simultaneously ensure that our parameter distribution remains close to some *prior* $r(\Theta)$ which we specified independent of the data. This is nearly always an upper bound on the true risk, with an offset determined by ψ (eq. (15)), a term which measures the gap (Δ, eq. (16)) in our true and empirical inferential risks if we drew parameter values from our prior. Most crucially, this ψ is independent of $q(\Theta)$ and so for optimization purposes can be dropped as a constant. We further show (Theorem 2, Appendix A.3) that under certain assumptions, ψ is finite and therefore Theorem 1 is nonvacuous. More specifically, if $\lambda = o(nm)$ then $\psi = o(m)$ and if $\lambda = o(n\log m)$ then $\psi = o(\log m)$.

This yields our proposed risk, $\widetilde{\mathcal{P}}_{n,m}$ (eq. (14)). Minimizing $\widetilde{\mathcal{P}}_{n,m}$ (eq. (14)) is equivalent to minimizing a stochastic upper bound on the true predictive risk $\mathcal{P}$, analogous to the relationship between $\widetilde{\mathcal{R}}$ and $\mathcal{R}$. See theorem 1 for a complete proof, though it follows directly from the traditional PAC-Bayes proof once we invoke the multisample bound. Furthermore, as we increase m we get tighter approximations to the true predictive risk.

Dropping ψ (being constant in q), we can summarize the relationships between the risks as:

$$\mathcal{P} \lesssim \widetilde{\mathcal{P}}_{n,m} \leq \widetilde{\mathcal{P}}_{n,1} = \widetilde{\mathcal{R}}_n \gtrsim \mathcal{R} \geq \mathcal{P}. \tag{17}$$

The $\widetilde{\mathcal{R}}_n \gtrsim \mathcal{R}$ relationship is the classic PAC-Bayes result (Alquier et al., 2016), $\mathcal{R} \geq \mathcal{P}$ follows from Jensen's inequality (Masegosa, 2019), and the left hand side $\mathcal{P} \lesssim \widetilde{\mathcal{P}}_{n,m} \leq \widetilde{\mathcal{P}}_{n,1} = \widetilde{\mathcal{R}}_n$ is our contribution.

Masegosa (2019) identified the need for tighter bounds on predictive risks than $\mathcal{R}$, suggesting a family (PAC^2_T) of risks that utilize a second order Jensen bound. Unfortunately these are not asymptotically tight and so there exists an m for which PAC^m will be tighter. Additionally, while estimating the variance term in PAC^2_T requires care, PAC^m is minibatch friendly, having its expectation over the parameter distribution outermost in the objective. Later in our experiments we directly compare these approaches.

We now have two knobs we can use to adjust our risk: m, the number of samples we use to estimate the predictive distribution and β, a sort of inverse temperature used for adjusting the relative strength of the likelihood and prior terms. For $m = 1$ we recover the (inferential) risks we are used to, but for $m \geq 1$ we can form tighter bounds on the true predictive risk. With $\beta = 1$ we recover traditional Bayesian inference with an equal weighting of the likelihood and

Theorem 1. *For all $q(\Theta)$ absolutely continuous with respect to $r(\Theta)$, $X^n \overset{iid}{\sim} \nu(X)$, $\beta \in (0,\infty)$, $n,m \in \mathbb{N}$, $p(x|\theta) \in (0,\infty)$ for all $\{x \in \mathcal{X} : \nu(x) > 0\} \times \{\theta \in \mathcal{T} : r(\theta) > 0\}$, $\xi \in (0,1)$, and $\lambda_m \in [m,\infty)$, then with probability at least $1-\xi$:*

$$\mathcal{P}[q] \leq \widetilde{\mathcal{P}}_{n,m}[q;r,\beta] + \psi(\nu,n,m,\beta,r,\xi) \tag{12}$$

and furthermore (unconditionally):

$$\widetilde{\mathcal{P}}_{n,m}[q;r,\beta] \leq \widetilde{\mathcal{P}}_{n,m-1}[q;r,\beta] \leq \widetilde{\mathcal{P}}_{n,1}[q;r,\beta] = \widetilde{\mathcal{R}}_n[q,r,\beta] \tag{13}$$

where:

$$\widetilde{\mathcal{P}}_{n,m}[q;r,\beta] \overset{def}{=} -\frac{1}{n}\sum_i^n \mathbb{E}_{q(\Theta^m)}\left[\log\left(\frac{1}{m}\sum_j^m p(x_i|\Theta_j)\right)\right] + \frac{m}{\lambda_m}\mathsf{KL}\left[q(\Theta);r(\Theta)\right] \overset{def}{=} \mathsf{PAC}^m \tag{14}$$

$$\psi(\nu,n,m,\beta,r,\xi) \overset{def}{=} \frac{1}{\lambda_m}\log \mathbb{E}_{\nu(X^n)}\,\mathbb{E}_{r(\Theta^m)}\left[e^{\lambda_m \Delta_{n,m}}\right] - \frac{1}{\lambda_m}\log\xi \tag{15}$$

$$\Delta_{n,m} \overset{def}{=} \Delta(X^n,\Theta^m) \overset{def}{=} \frac{1}{n}\sum_i^n \log\left(\frac{1}{m}\sum_j^m p(X_i|\Theta_j)\right) - \mathbb{E}_{\nu(X)}\left[\log\left(\frac{1}{m}\sum_j^m p(X|\Theta_j)\right)\right] \tag{16}$$

Proof. Proof in appendix A.3. *Sketch*: form a multisample bound on the predictive risk and apply the traditional PAC-Bayes bound. □

prior terms, as $\beta \to \infty$ we recover purely empirical risks. For any β, including $\beta \leq 1$ we still maintain our stochastic bounds. Downweighting the KL term with respect to the prior, or *cold posteriors* has shown to be useful especially in the context of neural networks (Wenzel et al., 2020).

5. An Illustrative Toy Example

Consider trying to fit a Normal distribution to a set of observations with a fixed unit variance but unknown mean:

$$p(x|\theta) = \mathsf{Normal}(x;\theta,1) = (2\pi)^{-\frac{1}{2}} e^{-\frac{(x-\theta)^2}{2}}. \tag{18}$$

Imagine further that we are operating in a severe model misspecification regime. While our model is a unit variance Normal distribution, the true data distribution is a 30-70 mixture of two Normals with twice the standard deviation and separated by four times their standard deviation.

In fig. 1 (top) we show the predictive distributions that result from minimizing all of the risks discussed previously. In fig. 1 (bottom) we show the corresponding parameter distributions. The true data distribution is shown with the dark red curve in fig. 1 (top). The dark red tick marks on the axis show five ($n = 5$) samples which we took as our data. Table 1 shows the resulting KL divergences between the true data distribution and each of the found predictive distributions.

Minimizing $\overline{\mathcal{R}}$ (eq. (5)) is equivalent to Maximum Likelihood (grey curves), which concentrates its parameter dis-

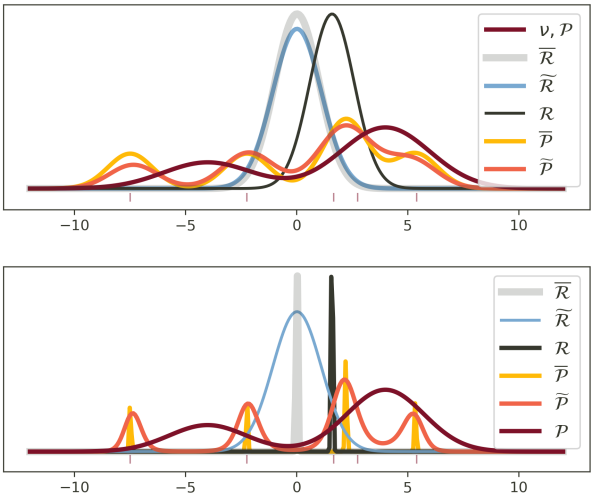

Figure 1. Toy Example: The top plot shows the resulting predictive distributions. The bottom plot shows the learned parameter distributions. Please see accompanying text for a full explanation.

Risk	$\overline{\mathcal{R}}$	$\widetilde{\mathcal{R}}$	$\mathcal{R}$	$\overline{\mathcal{P}}$	$\widetilde{\mathcal{P}}$	$\mathcal{P}$
KL	12.	9.6	10.	0.50	0.38	0

Table 1. KL Divergences between the true distribution and the predictive distributions for the various methods in fig. 1 (measured in bits).

tribution to a delta-function located at the empirical mean, and whose predictive distribution is simply a unit variance Normal distribution centered at that empirical mean.

Minimizing $\widetilde{\mathcal{R}}$ (eq. (7), aka ELBO) is equivalent to Bayesian inference (blue curves). In this case we used a weakly informative prior of a Normal centered at 0 with a variance of nine (light red curve in fig. 1 bottom). This risk prevents the parameter distribution from collapsing onto a delta function, but the resulting predictive distribution is quite similar to the one we found with Maximum Likelihood. It is still fundamentally unimodal as minimizing $\widetilde{\mathcal{R}}$ is still fundamentally looking for the best single parameter setting of our model.

Minimizing $\mathcal{R}$ (eq. (4)) is equivalent to Bayesian inference with infinite data (black curve). Here again our parameter distribution concentrates on a delta function, this time at the true distributions mean, but the resulting predictive distribution is the best single parameter setting we could achieve, a unimodal predictive distribution that doesn't match the true distribution all that well.

In contrast, minimizing the predictive risks (warm colors) do not look for single parameter settings of the model. Minimizing $\overline{\mathcal{P}}$ (eq. (3)) performs a sort of clustering of the data (yellow curve). While it might seem natural to allow each data point its own delta-like contribution in the parameter distribution, two of our samples are near enough that we achieve better empirical predictive risk by combining the two points into a single contribution to the parameter distribution with twice the weight but located at the two points' mean. The resulting predictive distribution remains multimodal and achieves a much lower divergence with respect to the true distribution (0.5 bits versus the ~ 10 bits for the traditional (inferential) risks).

Minimizing $\widetilde{\mathcal{P}}$ (eq. (14), aka PAC^m, here with $m \to \infty$ see appendix C.2) has a similar qualitative effect compared to the corresponding inferential case ($\widetilde{\mathcal{R}}$). The addition of the KL penalty with respect to some prior (here the same as used in the Bayesian case) prevents the parameter distribution from collapsing to a delta-comb.

Finally, in this case, even though we have rather gross model misspecification in the sense that our model $p(X|\Theta)$ is quite unlike the true distribution, our true distribution can be expressed as an infinite mixture of our model. Minimizing the true $\mathcal{P}$ (eq. (2)) can achieve perfect predictive performance (red curve). This is achieved with a bimodal Normal distribution in parameter space which when convolved with our Normal model gives the exact bimodal Normal data distribution we chose. This is also what we achieve asymptotically from $\widetilde{\mathcal{P}}$ in the limit of infinite data.

This toy example illustrates how and when we can hope to achieve better predictive performance from $\overline{\mathcal{P}}, \widetilde{\mathcal{P}}$ than from $\overline{\mathcal{R}}, \widetilde{\mathcal{R}}$. Namely, if some mixture of our model can get closer to the true distribution than the best single setting of the parameters, we expect approaches that target the predictive risks to outperform the inferential risks by a corresponding margin.

6. Related Work

By far the work most closely related to the PAC^m bound is the PAC^2_T bound presented in Masegosa (2019). PAC^2_T is based on a second order Jensen tightening of $\mathcal{P}$. While clearly instrumental to our work, PAC^2_T has a number of defects which PAC^m remedies. First, there always exists a m for which PAC^m is tighter than PAC^2_T (the proof is immediate from the asymptotic tightness of PAC^m in m). Second, the variance tightening term in PAC^2_T is non-degenerate only for bounded likelihoods; PAC^m has no such restriction. Third, the PAC^m risk, by directly targetting predictive risk satisfies the *golden rule*; the same cannot be said for PAC^2_T. Finally, in the experiments below we demonstrate that PAC^m generally matches or exceeds the test-set performance of PAC^2_T; as expected, both generally outperform ELBO.

The PAC^m proofs leverage multisample insights from the IWAE work (Burda et al., 2015). In response, Rainforth et al. (2018) question the utility of these tighter class of bounds and demonstrate that tighter bounds on the *marginal evidence* do not help learn useful posteriors. This valuable insight does not apply to PAC^m because our bound is not on the evidence marginal $p(\{x_i\}_i^n) = \mathbb{E}_{r(\Theta)}[\prod_i^n p(x_i|\Theta)]$ but rather on the *posterior predictive distribution*, $p(X|\{x_i\}_i^n) = \mathbb{E}_{q(\Theta|\{x_i\}_i^n)}[p(X|\Theta)]$. In fact, in the limit $m \to \infty$ we already explicitly encode the idea that the actual Bayesian posterior is not directly useful.

PAC^m offers real benefits in predictive performance in cases of model misspecification, in particular, when a mixture of our model family would be a better predictive model than the model itself. If so, why not simply fit mixture models? This certainly does work, as fig. 7 demonstrates. Mixtures have proven difficult to fit in general (Morningstar et al., 2020). The PAC^m family of risks subsume classic risks and offer theoretical generalization guarantees in cases of model misspecification, while remaining computationally tractable.

7. Experimental Results

Here we demonstrate that our new risk: PAC^m, can achieve better predictive performance on a suite of tasks. In all experiments, we compare the PAC^m-Bayes risk ($\widetilde{\mathcal{P}}$) to alternative objectives, including the PAC-Inferential Risk ($\widetilde{\mathcal{R}}$), also called the Evidence Lower Bound (ELBO), since it is a lower bound on the marginal likelihood. We also compare to alternative PAC-style bounds on the predictive risk, namely

the PAC^2_T objective proposed in (Masegosa, 2019).

7.1. Toy problems

7.1.1. Sinusoid

We start with a simple regression task. We follow (Masegosa, 2019) and draw data from a simple sinusoidal model:

$$\mu_x = 7 \sin\left(\frac{3x}{4}\right) + \frac{x}{2} \tag{19}$$

$$y \sim \text{Normal}(\mu_x, 10). \tag{20}$$

We use the predictive model presented in (Masegosa, 2019). We introduce model misspecification by underestimating the variance of the output data, and using $\text{Normal}(f(x, \omega), 1)$ as our likelihood function, where f is a neural network. By underestimating the variance, we require that the posterior capture the uncertainty present in the data. Full experimental details can be found in appendix C.3.

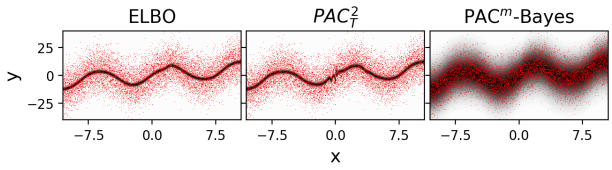

Figure 2. Red points show data from a simple sinusoidal model. Black contours show the predictive distribution formed using 10^4 samples from the learned posterior distribution. The left panel shows a model trained using ELBO, the middle panel shows a model trained using PAC^2_T, and the right panel a model trained using PAC^m. ELBO and PAC^2_T underestimate the observed variance, while PAC^m captures the uncertainty better.

We show the predictive models alongside the observed data in Figure 2. The first noteworthy observation is that while ELBO appears to do a reasonable job of predicting the mean, it underestimates the variance substantially. This is expected behavior. Here, ELBO attempts to learn the posterior over the mean of the data. Because the likelihood is misspecified, the posterior concentrates on the (true) mean of the data. Note that it concentrates *more* than it would if the model were well specified; by getting the model wrong, ELBO becomes overconfident in its prediction. This can be hazardous when predicting future data.

Our second observation is that while PAC^2_T does a marginally better job of accounting for the observed uncertainty (its predictive distribution is slightly wider than ELBO), it still underpredicts the variance of the data. In contrast, on the right we show the predictive model learned by optimizing our PAC^m objective, which appears to recover a much better approximation to the predictive distribution. To quantify this, we use the KL-Divergence between the posterior predictive distribution, and the true generative distribution. We find that ELBO gets a test set negative log-posterior-predictive $\text{nlpp} = 50.3$, while PAC^2_T gets $\text{nlpp} = 5.9$, and PAC^m gets $\text{nlpp} = 3.9$. This corresponds to $\text{KL} = 46.2$ for ELBO, $\text{KL} = 2.2$ for PAC^2_T, and $\text{KL} = 0.2$ for PAC^m. Note that our observed performance of PAC^2_T is substantially better than that reported in (Masegosa, 2019), which reports $\text{nlpp} = 25.2$ on the test set in their paper. We attribute this improvement in performance to a modification of their training objective to utilize m samples, rather than just two. Note also that the computational and memory cost of the PAC^2_T training objective scales quadratically in the number of samples.

7.1.2. Mixture

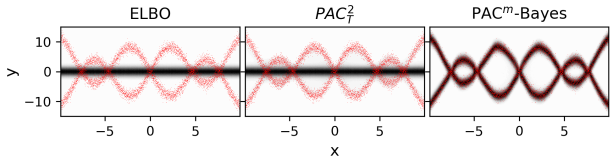

Figure 3. Similar to Figure 2, but where the data was generated from a mixture. All models were fit using only a single mode in the posterior. Here, ELBO offers little predictive power, while PAC^2_T does better but still places much of its probability away from the observed data. PAC^m offers a much better predictive model, even though it only uses a single posterior mode.

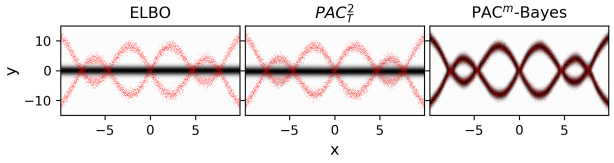

Figure 4. Similar to Figure 3, but where the surrogate posterior is multimodal. We find that the results are unchanged, despite the increased flexibility afforded to ELBO and PAC^2_T through the use of multiple modes in the posterior.

We also studied the mixture experiment from (Masegosa, 2019). In this experiment, we draw data from an equally weighted mixture with two components with opposite means:

$$\mu_x = 7 \sin\left(\frac{3x}{4}\right) + \frac{x}{2} \tag{21}$$

$$Z \sim \text{Rademacher} \tag{22}$$

$$Y \sim \text{Normal}(Z\mu_x, 1) \tag{23}$$

Full experimental details can be found in appendix C.4, though it is similar to that of the sinusoid experiment above. The predictive distributions for these models are shown in Figure 3. Similar to the sinusoid experiment, ELBO predicts the mean of $p(y|x)$. However, because the data is generated from a mixture, many samples from the true generative distribution do not fall near the mean. This leads to an extremely poor predictive model when ELBO is used in training. In contrast, both PAC_T^2 and PAC^m place probability mass near where samples of the data are found, though PAC_T^2 still places most of its mass near the mean. Interestingly, PAC^m appears to recover a close approximation to the true generative distribution, despite having only a unimodal surrogate posterior.

In order to assess whether the improvements we're seeing from the modified approaches is due to what is effectively a richer posterior approximation, we repeat the experiment above but with an explicitly multimodal posterior: a mixture of independent normal distributions. We fix the categorical probabilities of the mixture here, and use *stratified* sampling to facilitate gradient propagation through the mixture, following the procedures outlined in (Morningstar et al., 2020). We show the results from this in Figure 4. Similar to Figure 3, we find that ELBO learns the mean of the predictive distribution, but fails to produce a reasonable predictive model. Similarly, PAC_T^2 appears to consolidate much of its probability on the mean as well, though it includes some samples which track the observed data. As a result, it has a lower but still fairly large KL Divergence from the generative distribution (9.09, compared to 14.19 for ELBO). However, PAC^m does better still, closely tracking both modes in the posterior and hitting a KL divergence substantially lower than alternatives (0.63).

The observed failure of ELBO may seem confusing at first, since the model has multiple modes in the posterior. Its failure is, again, a consequence of model misspecification, this time brought about by the use of a unimodal likelihood which cannot offer a good explanation for all of the observed data. In fact, our analysis here along with prior analysis from (Morningstar et al., 2020) shows that multimodality in the posterior is extremely difficult to recover in models trained with ELBO. Intuitively, we understand this as follows: Traditionally, the posterior is observed to be multimodal when two distinct parameter settings can offer reasonable competing explanations for a single outcome (i.e. they have the same likelihood). Here the situation is different; we actually observe multiple distinct outcomes for a single input. If we knew that multiple distinct outcomes could exist we would want to explicitly accommodate this by using a mixture likelihood, but we generally do not know if this is a possibility. Assuming a unimodal likelihood, if we optimize the inferential risk, we find that the best explanation for the observed data is just 0 since it predicts all of the data equally poorly. Of course, we are uninterested in offering a reasonable explanation for all of the observed data, and are instead interested in ensuring that the entire posterior predictive distribution accurately predicts new examples. If we therefore optimize bounds on the predictive risk (e.g. PAC^m, PAC_T^2), we observe much better predictive performance.

Bayesian inference performs well if the model is well specified, in this case this would mean having a two-component likelihood. We show the results of this model in Figure 7 (in appendix B.2). As expected, all models recover a similar predictive distribution, which has low KL from the generating distribution. Even here, we do find that optimizing the Predictive Risk appears to offer marginally better returns than optimizing alternatives ($\text{KL} = 0.007$ for PAC^m, vs $\text{KL} = 0.017$ for ELBO and $\text{KL} = 0.07$ for PAC_T^2).

8. Image Experiments

8.1. Structured Prediction

We also test our objective on structured prediction tasks (e.g. Sohn et al., 2015). For this, we train a Bayesian neural network to predict the bottom half of an image, using only the top half as an input. We test this on 3 different image datasets: MNIST (LeCun, 1998), FashionMNIST (Xiao et al., 2017), and CIFAR-10 (Krizhevsky et al., 2009). For our likelihood, we use a Normal distribution where each pixel is considered independent. Following (Masegosa, 2019), we further fix the scale of the likelihood distribution to 1/255. These choices are interesting for two reasons. First, this setup also replicates the training setup which is often employed in training naive Variational Autoencoders (see e.g., Kingma & Welling, 2013; Tomczak & Welling, 2018), where the output variance is either fixed or shared between pixels (for non binarized images) and where all output pixels are assumed to be independent. Second, this setup is a misspecified model since we know that the pixels in the data are not independent, at least not at the granularity which we are able to capture in most models. We therefore hypothesize that PAC^m should be able to offer improvements in predictive performance.

We train models and measure performance as a function of m and the loss function used in training. For each value of m and each loss, we conduct 5 trials with different initializations to estimate the uncertainty in our final test set negative log-posterior-predictive probability. We show the results in Figure 5. We find that the performance of ELBO is roughly static in m, with much of the observed variation consistent with noise. This is consistent with our expectation. In contrast, PAC_T^2 and PAC^m exhibit rapid improvement in performance with m, showing that the model is, in fact, misspecified and that these models are therefore able to offer

meaningful improvement in predictive performance. Interestingly, we also appear to observe a saturation in m when the number of samples approaches the size of the batch. This could occur for two reasons. First, as we show in theorem 2, the model may ultimately cease to improve in m because the increasing value of ψ may overcome the tightening of $\widetilde{\mathcal{P}}$. Alternatively, this could be due to empirical variance in the gradients introduced by minibatch training. This is similar to the findings from (Rainforth et al., 2018) who showed that variance in the gradients results in an impedance to effective learning which eventually overcomes tightness. Alternative gradient estimators such as that from (Tucker et al., 2018) may help to solve this issue.

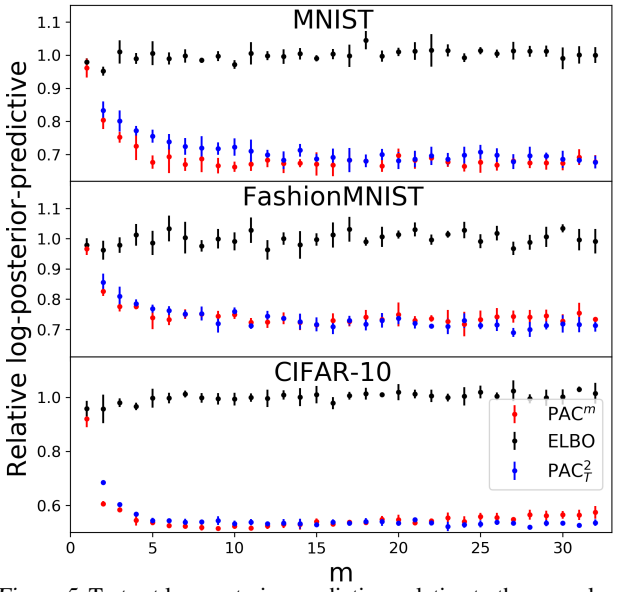

Figure 5. Test set log-posterior-predictive, relative to the mean log-posterior-predictive recovered using ELBO. Black points show models trained using ELBO, while red points show models trained using PAC^m and blue points show models trained with PAC^2_T. We see that PAC^m appears to offer the best log-posterior-predictive, but that performance either saturates or degrades as the sample number becomes to large, indicating issues with the optimization procedure.

8.2. Classification

So far, we have experimented with models where the likelihood was either purposefully misspecified in order to highlight the generalization gap introduced by minimizing $\widetilde{\mathcal{R}}$ compared to PAC^m or where we expect that it is misspecified because the assumptions we make about the output data are likely to be incorrect (e.g. pixels are likely non-independent in most images). It is unclear the degree to which this is an issue for many real-world applications where we use highly expressive deep neural network models, but in many cases we are still forced to make incorrect modeling assumptions for the sake of convenience.

It is equally interesting to consider the performance of PAC^m when the likelihood is well specified, but when other parts of the Bayesian model (the prior) are not. A good example of such a scenario is image classification, where we expect that a categorical distribution is a reasonable choice of likelihood. To test this scenario, in appendix B we present additional experiments where we use Bayesian convolutional neural networks to classify images from the datasets used in section 8.1. We consider two cases: (1) being Bayesian over the weights of the model (the "global" variables), or (2) being Bayesian over the activations of the model (the "local" variables). This latter case has been explored in works such as (Alemi et al., 2016). Here we consider the same approach, except where we minimize a PAC-Bayesian bound on the predictive likelihood.

As we expect, for classification problems, we find that though the model appears to be mostly well-specified, PAC^m learns models that make better predictions at the same cost, measured in terms of the KL divergence between the posterior and the prior.

9. Conclusion

Bayesian inference minimizes a stochastic upper bound on the predictive risk but the tightness of this bound is limited by model misspecification. In this work we proposed PAC^m, a new bound on predictive risk which is asymptotically tight yet can also recover the traditional Bayesian posterior (but not simultaneously both). Moreover, minimizing this bound respects the *golden rule* by minimizing the same predictive loss we use to evaluate our model, while providing generalization gaurentees. We demonstrated that PAC^m outperforms ELBO and PAC^2_T (Masegosa, 2019) on misspecified Bayesian models.

Although optimizing our bound empirically leads to models which make better predictions, theorem 2 shows that the bound loosens at best like $o(log(m))$ thus increasing the generalization gap. We appear to observe this empirically in our experiments as m approaches the order of magnitude of n. However we note that in these experiments we chose $\lambda = o(nm)$ meaning the generalization bound loosened at $o(m)$; additional experimentation with $\lambda = o(n\sqrt{\log m})$ is warranted, since that may enable use of larger m. We note however that large m is typically not computational practical anyway. Additionally we note that in practice one would tune λ based on held-out performance; for our experiments we made no such tuning (See the end of appendix A.3 for additional discussion).

ACKNOWLEDGEMENTS

We'd like to thank Ben Poole, Sergey Ioffe, and Rif A. Saurous for comments on the draft.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283, 2016.

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken elbo. In *International Conference on Machine Learning*, pp. 159–168. PMLR, 2018.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

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.

Asadi, K. and Littman, M. L. An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning*, pp. 243–252, 2017.

ATLAS Collaboration. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1 – 29, 2012. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2012.08.020. URL http://www.sciencedirect.com/science/article/pii/S037026931200857X.

Banerjee, A. On bayesian bounds. In *Proceedings of the 23rd international conference on Machine learning*, pp. 81–88, 2006.

Berger, J. O. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

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, 2016.

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.

Burda, Y., Grosse, R. B., and Salakhutdinov, R. Accurate and conservative estimates of mrf log-likelihood using reverse annealing, 2014.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Catoni, O. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

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

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M., and Saurous, R. A. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*, 2017.

Domingos, P. M. Why does bagging work? a bayesian account and its implications. In *KDD*, pp. 155–158. Citeseer, 1997.

Fushiki, T. et al. Bootstrap prediction and bayesian prediction under misspecified models. *Bernoulli*, 11(4):747–758, 2005.

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.

Grosse, R. B., Ancha, S., and Roy, D. M. Measuring the reliability of mcmc inference with bidirectional monte carlo. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2451–2459. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6290-measuring-the-reliability-of-mcmc-inference-with-bidirectional-monte-carlo.pdf.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Knoblauch, J., Jewson, J., and Damoulas, T. Generalized variational inference: Three arguments for deriving new posteriors, 2019.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Laplace, P.-S. Mémoire sur les probabilités. *Mémoires de l'Académie Royale des sciences de Paris*, 1778:227–332, 1781.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Masegosa, A. R. Learning under model misspecification: Applications to variational and ensemble methods. *arXiv preprint arXiv:1912.08335v3*, 2019.

Minka, T. P. Bayesian model averaging is not model combination. *Available electronically at http://www. stat. cmu. edu/minka/papers/bma. html*, pp. 1–2, 2000.

Mnih, A. and Rezende, D. J. Variational inference for monte carlo objectives, 2016.

Morningstar, W. R., Vikram, S. M., Ham, C., Gallagher, A., and Dillon, J. V. Automatic differentiation variational inference with mixtures. *arXiv preprint arXiv:2003.01687*, 2020.

Murphy, K. P. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16, 2007.

Piponi, D., Moore, D., and Dillon, J. V. Joint distributions for tensorflow probability. *arXiv preprint arXiv:2001.11819*, 2020.

Planck Collaboration VI. *Planck* 2018 results. VI. Cosmological parameters. *AAP, in press*, 2019.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. Tighter variational bounds are not necessarily better. *arXiv preprint arXiv:1802.04537*, 2018.

Ramamoorthi, R., Sriram, K., Martin, R., et al. On posterior concentration in misspecified models. *Bayesian Analysis*, 10(4):759–789, 2015.

Ritchie, H. Gender ratio. *Our World in Data*, 2019. https://ourworldindata.org/gender-ratio.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pp. 3483–3491, 2015.

Staines, J. and Barber, D. Variational optimization, 2012.

Tomczak, J. and Welling, M. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pp. 1214–1223. PMLR, 2018.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. Doubly reparameterized gradient estimators for monte carlo objectives. *arXiv preprint arXiv:1810.04152*, 2018.

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

A. Proofs

This section proves our main theoretical result (theorem 1) as well as presents additional theory relevant to PAC^m.

A.1. Relationship Between Predictive and Inferential Risks In the Presence of Model Misspecification

The following two results are adapted from (Masegosa, 2019) to our notation and given here for the reader's convenience. These results examine conditions under which solutions to the inferential risk, $\min_{q(\Theta)} \mathcal{R}[q]$, are equivalent to solutions to the predictive risk, $\min_{q(\Theta)} \mathcal{P}[q]$. That is, these lemmas show that model misspecification introduces a gap between *predictive risk* ($\mathcal{P}$) and *inferential risk* ($\mathcal{R}$). This gap is potentially problematic because machine learning practitioners care about $\mathcal{P}$ but minimize (an approximation of) $\mathcal{R}$.

Lemma 1. $\arg\min_{q(\Theta)} \mathcal{R}[q] \equiv \arg\min_{q(\Theta)} \mathcal{P}[q]$ *only if for any distribution* ρ *over* Θ, $\mathsf{KL}[\nu(X); p(X|\theta^{(\mathsf{ml})})] \leq \mathsf{KL}[\nu(X); \mathbb{E}_{\rho(\Theta)}[p(X|\Theta)]]$, *and* $q^{(\mathsf{ml})}(\Theta) = \arg\min_{q(\Theta)} \mathcal{R}[q] \equiv \delta(\Theta - \theta^{(\mathsf{ml})})$ *where* δ *is the Dirac-delta distribution.*

Proof. (Sketch.) Note that,

$$\begin{aligned}\mathcal{P}[q] &= \mathsf{KL}[\nu(X), \mathbb{E}_{q(\Theta)}\, p(X|\Theta)] + \mathsf{H}[\nu(X)] \\ &\leq \mathbb{E}_{q(\Theta)}\, \mathsf{KL}[\nu(X), p(X|\Theta)] + \mathsf{H}[\nu(X)] \\ &= \mathcal{R}[q]\end{aligned}$$

where the inequality is Jensen's. Since the theorem condition implies $\mathcal{R}[q^*] \leq \mathcal{P}[q]$ then $\mathcal{R}[q^*] \leq \min_q \mathcal{P}[q] \leq \mathcal{R}[q^*]$ and the claim follows. (See Lemma 2 of (Masegosa, 2019) for original proof; our sketch is based on a sandwich argument.) □

Lemma 2. *If there exists a density* ρ *over* Θ *such that* $\mathsf{KL}[\nu(X); \mathbb{E}_{\rho(\Theta)}[p(X|\Theta)] < \mathsf{KL}[\nu(X); p(X|\theta^{(\mathsf{ml})}]$, *then a minimizer of* $\mathcal{R}$ *is not a minimizer of* $\mathcal{P}$ *where* $q^{(\mathsf{ml})}(\Theta) \stackrel{def}{=} \arg\min_{q(\Theta)} \mathcal{R}[q] \equiv \delta(\Theta - \theta^{(\mathsf{ml})})$ *where* δ *is the Dirac-delta distribution.*

Proof. (Sketch.) The condition of this lemma implies that $q^{(\mathsf{ml})}(\Theta)$ cannot be a minimizer of $\mathcal{P}$ however it is the minimizer of $\mathcal{R}$. (See Lemma 3 of (Masegosa, 2019) for original proof.) □

A.2. PAC-Bayes Relationships

This section presents two well-known PAC-Bayes results as special cases of theorem 1.

Corollary 1. *Under the conditions of theorem 1, then with probability at least* $1 - \xi$, $\mathcal{P}[q] \leq \widetilde{\mathcal{R}}_n[q; r, \beta] + \psi_1$.

Proof. Immediate from theorem 1 when $m = 1$. □

Corollary 2. *The Bayesian posterior* $p(\Theta|\{x_i\}_i^n) \propto r(\Theta) \prod_i^n p(x_i|\Theta)$ *minimizes* PAC^m *when* $m = \beta = 1$.

Proof. PAC^m is equivalently PAC when $m = \beta = 1$ for which the claim is proven by (Germain et al., 2016). □

A.3. PAC^m-Bayes Theory

Theorem 1. *For all* $q(\Theta)$ *absolutely continuous with respect to* $r(\Theta)$, $X^n \stackrel{iid}{\sim} \nu(X)$, $\beta \in (0, \infty)$, $n, m \in \mathbb{N}$, $p(x|\theta) \in (0, \infty)$ *for all* $\{x \in \mathcal{X} : \nu(x) > 0\} \times \{\theta \in \mathcal{T} : r(\theta) > 0\}$, $\xi \in (0, 1)$, *and* $\lambda_m \in [m, \infty)$, *then with probability at least* $1 - \xi$*:*

$$\mathcal{P}[q] \leq \widetilde{\mathcal{P}}_{n,m}[q; r, \beta] + \psi(\nu, n, m, \beta, r, \xi) \tag{12}$$

and furthermore (unconditionally):

$$\widetilde{\mathcal{P}}_{n,m}[q; r, \beta] \leq \widetilde{\mathcal{P}}_{n,m-1}[q; r, \beta] \leq \widetilde{\mathcal{P}}_{n,1}[q; r, \beta] = \widetilde{\mathcal{R}}_n[q, r, \beta] \tag{13}$$

where:

$$\widetilde{\mathcal{P}}_{n,m}[q;r,\beta] \stackrel{def}{=} -\frac{1}{n}\sum_{i}^{n}\mathbb{E}_{q(\Theta^m)}\left[\log\left(\frac{1}{m}\sum_{j}^{m}p(x_i|\Theta_j)\right)\right] + \frac{m}{\lambda_m}\mathsf{KL}\left[q(\Theta);r(\Theta)\right] \stackrel{def}{=} \mathsf{PAC}^m \tag{14}$$

$$\psi(\nu,n,m,\beta,r,\xi) \stackrel{def}{=} \frac{1}{\lambda_m}\log\mathbb{E}_{\nu(X^n)}\mathbb{E}_{r(\Theta^m)}\left[e^{\lambda_m\Delta_{n,m}}\right] - \frac{1}{\lambda_m}\log\xi \tag{15}$$

$$\Delta_{n,m} \stackrel{def}{=} \Delta(X^n,\Theta^m) \stackrel{def}{=} \frac{1}{n}\sum_{i}^{n}\log\left(\frac{1}{m}\sum_{j}^{m}p(X_i|\Theta_j)\right) - \mathbb{E}_{\nu(X)}\left[\log\left(\frac{1}{m}\sum_{j}^{m}p(X|\Theta_j)\right)\right] \tag{16}$$

Proof. Write:

$$g(\Theta^m;X) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{j}^{m}p(X|\Theta_j)$$

$$\overline{\mathcal{G}}_{n,m}[q] \stackrel{\text{def}}{=} -\frac{1}{n}\sum_{i}^{n}\mathbb{E}_{q(\Theta^m)}\left[\log g(\Theta^m;x_i)\right]$$

$$\mathcal{G}_m[q] \stackrel{\text{def}}{=} -\mathbb{E}_{\nu(X)}\mathbb{E}_{q(\Theta^m)}\left[\log g(\Theta^m;X)\right]$$

For the first claim:

Jensen's inequality implies $-\log\mathbb{E}_{q(\Theta^m)}\left[g(\Theta^m;X)\right] \leq \mathbb{E}_{q(\Theta^m)}\left[-\log g(\Theta^m;X)\right]$. Applying $\mathbb{E}_{\nu(X)}$ to both sides implies $\mathcal{P}[q] \leq \mathcal{G}_m[q]$.

To complete the proof of the first claim, we now show $p(\mathcal{G}_m[q] \leq \widetilde{\mathcal{P}}_{n,m}[q;r,\beta] + \psi_{n,m}) \geq 1-\xi$. Make the substitution, $f(\Theta^m;\{x_i\}_i^n) \stackrel{\text{def}}{=} \lambda_m\Delta(\{x_i\}_i^n,\Theta^m)$ (for some non-stochastic λ_m) to Lemma 3 ("Compression Lemma") and rearrange:

$$\begin{aligned}-\mathbb{E}_{q(\Theta^m)}\mathbb{E}_{\nu(X)}[\log g(\Theta^m;X)] \leq &-\mathbb{E}_{q(\Theta^m)}\mathbb{E}_{\nu(X|\{x_i\}_i^n)}[\log g(\Theta^m;X)]\\ &+\tfrac{1}{\lambda_m}\mathsf{KL}[q(\Theta^m),r(\Theta^m)] + \tfrac{1}{\lambda_m}\log\mathbb{E}_{r(\Theta^m)}\left[e^{\lambda_m\Delta(\{x_i\}_i^n,\Theta^m)}\right].\end{aligned}$$

For the KL term, note that Lemma 5 ("KL-divergence iid") implies $\mathsf{KL}[q(\Theta^m),r(\Theta^m)] = m\,\mathsf{KL}[q(\Theta),r(\Theta)]$.

For the rightmost term (a log moment generating function conditioned on $\{x_i\}_i^n$), make substitutions $Z \stackrel{\text{def}}{=} \mathbb{E}_{r(\Theta^m)}[e^{\lambda_m\Delta(\{x_i\}_i^n,\Theta^m)}]$ and $p \stackrel{\text{def}}{=} \nu(X^n)$ to Lemma 4 ("Log Markov Inequality") to conclude:

$$\nu_{X^n}\left(\log\mathbb{E}_{r(\Theta^m)}\left[e^{\lambda_m\Delta(X^n,\Theta^m)}\middle|X^n\right] \leq \log\mathbb{E}_{\nu(X^n)}\mathbb{E}_{r(\Theta^m)}\left[e^{\lambda_m\Delta(X^n,\Theta^m)}\right] - \log\xi\right) \geq 1-\xi.$$

Scale the inner inequality by $\frac{1}{\lambda_m}$ (which doesn't change the probability) and combine this result with the previous two to prove the first claim.

(This proof was inspired by (Masegosa, 2019).)

For the second claim:

Note that $m/\lambda_m \leq 1$ for all m hence the KL terms of $\widetilde{\mathcal{P}}_{n,m}$ and $\widetilde{\mathcal{R}}_n$–which are not otherwise a function of m–cannot grow in m and may be safely ignored. The equality $\widetilde{\mathcal{P}}_{n,1}[q;r,\beta] = \widetilde{\mathcal{R}}_n[q,r,\beta]$ is true by definition;

$g(\Theta^1; X) = p(X|\Theta)$. To complete the proof it is sufficient to show $\overline{\mathcal{G}}_{n,m} \leq \overline{\mathcal{G}}_{n,m-1}$. I.e.,

$$\begin{aligned}
\overline{\mathcal{G}}_{n,m}[q] &= -\frac{1}{n}\sum_{i}^{n} \mathbb{E}_{q(\Theta^m)}\left[\log \frac{1}{m}\sum_{j}^{m} p(x_i|\Theta_j)\right] \\
&= -\frac{1}{n}\sum_{i}^{n} \mathbb{E}_{q(\Theta^m)}\left[\log \frac{1}{m}\sum_{j}^{m} \frac{1}{m-1}\sum_{k\neq j}^{m} p(x_i|\Theta_k)\right] \\
&\leq -\frac{1}{m}\sum_{j}^{m} \frac{1}{n}\sum_{i}^{n} \mathbb{E}_{q(\Theta^m)}\left[\log \frac{1}{m-1}\sum_{k\neq j}^{m} p(x_i|\Theta_k)\right] \\
&= \frac{1}{m}\sum_{j}^{m} \overline{\mathcal{G}}_{n,m-1}[q] \\
&= \overline{\mathcal{G}}_{n,m-1}[q].
\end{aligned}$$

The inequality is Jensen's and the second-to-last equality follows from Θ^m being independent.

(This proof is inspired by (Burda et al., 2015).)

□

While Theorem 1 is technically true, additional assumptions are needed to ensure it is nonvacuous, i.e., $\psi(\nu, n, m, \beta, r, \xi) < \infty$. Theorem 2 (below) affirms this is the case when $\Delta(X, \theta)$ is *everywhere sub-gaussian* for all $\{\theta \in \mathcal{T} : r(\theta) > 0\}$ and furthermore suggests that the optimal λ is given by $\lambda^* = n\beta\sqrt{\log(\max(2, m))}$.

We emphasize that sub-gaussianity is only assumed for $n = m = 1$, yet our proof holds for $n, m \geq 1$. This assumption is similar to that made by Germain et al. (2016), however we assume $\Delta(X, \theta)$ is everywhere sub-gaussian whereas they assume $\Delta(X, \Theta)$ is *jointly sub-gaussian*. We note that their Corollaries 4 and 5 (the relevant claims) have incorrect proofs which do not obviously follow from joint sub-gaussianity; our Theorem 2 with $m = 1$ serves as a correction. As also indicated in Germain et al. (2016), our ψ's finiteness is also provable by the stronger assumption that $p(x|\theta) \in [a, b]$ where $a, b \in \mathbb{R}$ and for all $\{x \in \mathcal{X} : \nu(x) > 0\} \times \{\theta \in \mathcal{T} : r(\theta) > 0\}$. (Catoni, 2007; Alquier et al., 2016) However, we refain from making such claims, preferring the arguably more general assumptions of Theorem 2.

Theorem 2. *Making the assumptions of Theorem 1 except that $\lambda \in (0, \infty)$ and additionally that for all $\{\theta \in \mathcal{T} : r(\theta) > 0\}$, $\Delta(X, \theta)$ is sub-gaussian with standard deviation $s_\theta \in (0, s]$ for $s \in (0, \infty)$, i.e., $\log \mathbb{E}_{\nu(X)}\left[e^{\lambda\Delta(X,\theta)}\right] \leq \frac{1}{2}\lambda^2 s_\theta^2 \leq \frac{1}{2}\lambda^2 s^2$, then:*

$$\psi_{n,m} = \frac{1}{\lambda}\log \mathbb{E}_{\nu(X^n)}\,\mathbb{E}_{r(\Theta^m)}\left[e^{\lambda\Delta(X^n,\Theta^m)}\right] - \frac{1}{\lambda}\log\xi \tag{24}$$

$$\leq \left(\frac{\lambda s^2}{2n} + \frac{n\log m}{\lambda} + \log m\right) - \frac{1}{\lambda}\log\xi. \tag{25}$$

Additionally, writing $\beta = s^{-1}\sqrt{2}$ and assuming $\xi = 1$ then,

$$\lambda^* = n\beta\sqrt{\log\max(2, m)}, \tag{26}$$

minimizes Equation (25) for $m > 1$ and is a constant when $m = 1$.

Proof. Begin by noting that,

$$\Delta(x, \{\theta_j\}_j^m) \stackrel{\text{def}}{=} \log \frac{1}{m} \sum_j^m p(x|\theta_j) - \mathbb{E}_{\nu(X)} \left[\log \frac{1}{m} \sum_j^m p(X|\theta_j) \right] \tag{27}$$

$$\leq \max \left\{ \log p(x|\theta_j) \right\}_j^m - \mathbb{E}_{\nu(X)} \left[\log \frac{1}{m} \sum_j^m p(X|\theta_j) \right] \tag{28}$$

$$= \max \left\{ \log p(x|\theta_j) - \mathbb{E}_{\nu(X)} \left[\log \frac{1}{m} \sum_k^m p(X|\theta_k) \right] \right\}_j^m \tag{29}$$

$$\leq \max \left\{ \log p(x|\theta_j) - \mathbb{E}_{\nu(X)} \left[\log p(X|\theta_j) \right] \right\}_j^m + \log m \tag{30}$$

$$= \max \left\{ \Delta(x, \theta_j) \right\}_j^m + \log m. \tag{31}$$

The first inequality follows from the upper bound in Lemma 9. The second inequality follows from the negative of the lower bound in Lemma 9, i.e.,

$$-\log \frac{1}{m} \sum_k^m p(X|\theta_k) \leq -\max\{\log p(x|\theta_j)\}_j^m + \log m \leq -\log p(x|\theta_k) + \log m \text{ for all } k \in \{1, \ldots, m\}.$$

Combining this fact and the fact that $e^{\max\{a_j\}_j^m} = \max\{e^{a_j}\}_j^m \leq \sum_j^m e^{a_j}$, implies:

$$e^{\frac{\lambda}{n}\Delta(x, \{\theta_j\}_j^m)} \leq e^{\frac{\lambda}{n}\left(\max \left\{ \Delta(x, \theta_j) \right\}_j^m + \log m \right)} \tag{32}$$

$$= m^{\frac{\lambda}{n}} \max \left\{ e^{\frac{\lambda}{n}\Delta(x, \theta_j)} \right\}_j^m \tag{33}$$

$$\leq m^{\frac{\lambda}{n}} \sum_j^m e^{\frac{\lambda}{n}\Delta(x, \theta_j)}. \tag{34}$$

Combining this fact with the everywhere sub-gaussianity of $\Delta(X, \theta)$ implies:

$$\log \mathbb{E}_{\nu(X^n)} \mathbb{E}_{r(\Theta^m)} \left[e^{\lambda \frac{1}{n} \sum_i^n \Delta(X_i, \Theta^m)} \right] \tag{35}$$

$$= \log \mathbb{E}_{r(\Theta^m)} \left[\prod_i^n \mathbb{E}_{\nu(X)} \left[e^{\frac{\lambda}{n} \Delta(X, \Theta^m)} \right] \right] \tag{36}$$

$$= \log \mathbb{E}_{r(\Theta^m)} \left[\left(\mathbb{E}_{\nu(X)} \left[e^{\frac{\lambda}{n} \Delta(X, \Theta^m)} \right] \right)^n \right] \tag{37}$$

$$\leq \log \mathbb{E}_{r(\Theta^m)} \left[\left(\mathbb{E}_{\nu(X)} \left[\sum_j^m e^{\frac{\lambda}{n} \Delta(X, \Theta_j)} \right] \right)^n \right] + \lambda \log m \tag{38}$$

$$= \log \mathbb{E}_{r(\Theta^m)} \left[\left(\sum_j^m \mathbb{E}_{\nu(X)} \left[e^{\frac{\lambda}{n} \Delta(X, \Theta_j)} \right] \right)^n \right] + \lambda \log m \tag{39}$$

$$\leq \log \mathbb{E}_{r(\Theta^m)} \left[\left(\sum_j^m e^{\frac{\lambda^2 s^2}{2n^2}} \right)^n \right] + \lambda \log m \tag{40}$$

$$= \log \mathbb{E}_{r(\Theta^m)} \left[\left(m e^{\frac{\lambda^2 s^2}{2n^2}} \right)^n \right] + \lambda \log m \tag{41}$$

$$= \frac{\lambda^2 s^2}{2n} + (\lambda + n) \log m \tag{42}$$

Adding $-\log \xi$ and scaling by $\frac{1}{\lambda}$ completes the first part of the proof.

It now remains to find the optimal λ for $\xi = 1$. For $m = 1$ we resign ourselves to finding a constant, e.g., $\lambda = n\beta\sqrt{\log 2}$ where $\beta = s^{-1}\sqrt{2}$. For $m > 1$ note that $\frac{\lambda s^2}{2n} + \frac{n \log m}{\lambda} + \log m$ is convex in $\lambda > 0$ since $m, n > 0$. Solving for the root of the gradient we find $\lambda^* = n\beta\sqrt{\log \max(2, m)}$. □

Theorem 2 indicates that $\lambda = o(n)$ is sufficient to ensure nonvacuousness of Theorem 1 for all n and a fixed, finite m. Unfortunately Theorem 2 also indicates that no choice of λ ensures $\psi_{n,m} \to 0$ as $n, m \to \infty$. That is, even for $\lambda = o(n\sqrt{\log m})$ the Theorem 1 bound loosens at rate $o(\log m)$. Worse, Theorem 1 actually assumes $\lambda = o(nm)$ to ensure monotonic tightening to the predictive risk (see λ_m assumption in Theorem 1); in this regime the Theorem 1 bound looses at rate $o(m)$. Despite these concerns we note the following facts:

1. Regardless of ψ growing at best like $o(\log m)$ or nominally like $o(m)$, Theorem 1 remains non-vacuous for any $\lambda = o(n)$ and finite m.

2. In practice we recommend choosing λ by cross-validation and for each n, m regime. This implies the n, m-parameterization is merely a theoretical consideration (especially in light of point 1 above).

3. Theorem 2 is an upper bound and may or may not be made tighter. Theorem 2 assumptions are arguably fairly weak and stronger assumptions might help, e.g., bounded likelihood or $\Delta(X, \{\theta\}_j^m)$ being everywhere sub-gaussian (as opposed to $\Delta(X, \theta)$ being everywhere sub-gaussian).

4. The practitioner would not typically use large m. Given that computational complexity also grows in m, we expect the vast majority of cases to use $m \leq 50$ and to see improvements over $m = 1$.

A.4. Lemmas

In this section we present several Lemmas used to simplify this paper's proofs. Most of the Lemmas are well-known and are given here for the reader's convenience.

Lemma 3 (Compression). *If $p(\Theta)$ is absolutely semicontinuous wrt $r(\Theta)$ and $\mathbb{E}_{r(\Theta)}[e^{f(\Theta)}] < \infty$, then $\mathbb{E}_{p(\Theta)}[f(\Theta)] \leq \mathsf{KL}\left[p(\Theta), r(\Theta)\right] + \log \mathbb{E}_{r(\Theta)}[e^{f(\Theta)}]$.*

Proof. Write $q(\Theta) \stackrel{\text{def}}{=} \frac{r(\Theta)e^{f(\Theta)}}{\mathbb{E}_{r(\Theta)}[e^{f(\Theta)}]}$ and note that Lemma 6 implies, $0 \leq \mathsf{KL}\left[p(\Theta), q(\Theta)\right] = \mathsf{KL}\left[p(\Theta), r(\Theta)\right] - \mathbb{E}_{p(\Theta)}[f(\Theta)] + \log \mathbb{E}_{r(\Theta)}[e^{f(\Theta)}]$. □

Proof due to (Banerjee, 2006; Zhang, 2006).

Lemma 4 (Log Markov Inequality). *For any $\xi \in (0, 1]$ and random variable $Z \sim p$ with $p(Z \leq 0) = 0$ then $p(\log Z \leq \log \mathbb{E}_p[Z] - \log \xi) \geq 1 - \xi$.*

Proof. Markov's inequality states that $p(Z > t) \leq \frac{\mathbb{E}_p[Z]}{t}$ for non-negative random variable $Z \sim p$ and $t > 0$. Substituting $t = \frac{\mathbb{E}_p[Z]}{\xi}$ implies $p(Z > \frac{\mathbb{E}_p[Z]}{\xi}) \leq \xi$. Combining this with the fact that $\log$ is a non-decreasing bijection implies $p(\log Z > \log \mathbb{E}_p[Z] - \log \xi) \leq \xi$. Examining the complement interval completes the proof. □

Lemma 5 (KL-divergence iid). *If $p(\Theta^m) \stackrel{def}{=} \prod_j^m p(\Theta_j)$ and $r(\Theta^m) \stackrel{def}{=} \prod_j^m r(\Theta_j)$, then $\mathsf{KL}[p(\Theta^m), r(\Theta^m)] = m\,\mathsf{KL}[p(\Theta), r(\Theta)]$.*

Proof. $\mathsf{KL}[p(\Theta^m), r(\Theta^m)] = \mathbb{E}_{\prod_j^m p(\Theta_j)}\left[\log \frac{\prod_j^m p(\Theta_j)}{\prod_j^m r(\Theta_j)}\right] = m\,\mathsf{KL}\left[p(\Theta), r(\Theta)\right]$. □

Lemma 6 (Gibb's Inequality). *If $p(\Theta)$ is absolutely semicontinuous wrt $r(\Theta)$, then $\mathsf{KL}[p, q] \geq 0$.*

Proof. $\mathsf{KL}[p, q] = -\mathbb{E}_{p(x)}\left[\log \frac{q(x)}{p(x)}\right] \geq -\log \mathbb{E}_{p(x)}\left[\frac{q(x)}{p(x)}\right] = -\log 1 = 0$ where the inequality is Jensen's. □

Lemma 7 (ψ non-negative). *Under the conditions of theorem 1 and if $\mathcal{G}_m[r], \overline{\mathcal{G}}_{n,m}[r] < \infty$, then $\psi(\nu, n, m, \beta, r, \xi) \geq 0$.*

Proof. Jensen's inequality implies $e^{\mathbb{E}\,Z} \leq \mathbb{E}\,e^Z$. Applying $\log$ to both sides (a monotonically increasing function), implies $\mathbb{E}\,Z \leq \log \mathbb{E}\,e^Z$. Substitute $Z \stackrel{\text{def}}{=} \beta n m \Delta_{n,m}$ (see eq. (16)) and note $\mathbb{E}_{\nu(X^n)r(\Theta^m)}\,Z = 0$ by definition. Finally, note $\log z \leq z - 1$ for $z > 0$ implies $-\log \xi \geq 0$ for $\xi \in (0, 1]$. □

Lemma 8 (Log-Average-Exp Bound – Parametric).

$$-\log \frac{1}{n}\sum_i^n e^{x_i} \leq \begin{cases} -\frac{1}{\phi}\log \frac{1}{n}\sum_i^n e^{\phi x_i} & 0 < \phi \leq 1, \\ -\frac{1}{n}\sum_i^n x_i & \phi = 0. \end{cases}$$

Proof. Write $\mathsf{lse}(x) = \log \sum_i^n e^{x_i}$ and $\mathsf{softmax}_i(a) = \exp\left(a_i - \mathsf{lse}(a)\right)$.

For the $\phi \in (0, 1]$ case, note that lse convexity and Jensen's inequality imply $\mathsf{lse}\left(\phi a + (1-\phi)b\right) \leq \phi\,\mathsf{lse}(a) + (1-\phi)\,\mathsf{lse}(b)$ for $a, b \in \mathbb{R}^n$. Multiplying by $-\frac{1}{\phi}$ and rearranging yields $-\mathsf{lse}(a) \leq -\frac{1}{\phi}\,\mathsf{lse}\left(\phi a + (1-c)b\right) + \frac{1}{\phi}(1-\phi)\,\mathsf{lse}(b)$. Substituting $a = x - \log n$ and $b = -\log n$ proves the first case.

For the $\phi = 0$ case, note that L'Hopital's rule implies:

$$\lim_{\phi\to 0}\frac{1}{\phi}\,\mathsf{lse}\left(\phi x - \log n\right) = \lim_{\phi\to 0}\frac{\frac{\partial}{\partial\phi}\,\mathsf{lse}\left(\phi x - \log n\right)}{\frac{\partial}{\partial\phi}\phi} = \lim_{\phi\to 0}\frac{\sum_i^n \mathsf{softmax}_i\left(\phi x - \log n\right) x_i}{1} = \frac{1}{n}\sum_i^n x_i$$

since $\lim_{\phi\to 0}\mathsf{lse}(\phi x - \log n) = \lim_{\phi\to 0}\phi = 0$. The bound follows from this fact, the convexity of $-\log z$, and Jensen's inequality: $-\log \frac{1}{n}\sum_i^n e^{x_i} \leq -\frac{1}{n}\sum_i^n \log e^{x_i}$. □

Lemma 8 is potentially useful because it shows that minimizing $-\frac{1}{\phi}\log\sum_j^m p(X|\Theta_j)^\phi$ is still consistent with minimizing PAC^m (i.e., $\phi = 1$) in the sense that $\phi \in [0, 1]$ implies an upper bound. This result might be useful for mitigating some of the gradient variance observed in the Monte Carlo approximation of PAC^m for large m; this conjecture is left for future work. (We note that all experiments reported in this paper use $\phi = 1$.)

Lemma 8 similarly exists in (Asadi & Littman, 2017) though our proof differs slightly.

Lemma 9 (Log-Average-Exp Bound – Simple)**.**

$$\max\left(\frac{1}{n}\sum_i^n x_i, \max\{x_i\}_i^n - \log n\right) \leq \log\frac{1}{n}\sum_i^n e^{x_i} \leq \max\{x_i\}_i^n \tag{43}$$

Proof. For the upper bound, note that:

$$\begin{aligned}
\log\frac{1}{n}\sum_i^n e^{x_i} &= \max\{x_j\}_j^n + \log\frac{1}{m}\sum_i^n e^{x_i - \max\{x_j\}_j^n} \\
&\leq \max\{x_j\}_j^n + \log\frac{1}{n}\sum_i^n e^0 \\
&= \max\{x_j\}_j^n
\end{aligned}$$

For the lower bound, note that:

$$\log\frac{1}{n}\sum_i^n e^{x_i} \geq \log e^{\max\{x_j\}_j^n} - \log n = \max\{x_j\}_j^n - \log n$$

and by Jensen's inequality,

$$-\log\frac{1}{n}\sum_i^n e^{x_i} \leq -\frac{1}{n}\sum_i^n \log e^{x_i}.$$

□

B. Additional Experimental Results

In this section we present additional experimental results which were omitted from the main text due to space constraints.

B.1. Mixture

In Figure 3 and Figure 4, we showed histograms of samples from the predictive model alongside the observed data. Though we find that PAC^2_T does a much better job of predicting the data than ELBO (measured by the KL Divergence between the predictive model and the true generative model), this result is not obvious from looking at the figures presented in the main paper. Here we provide additional figures to show that models trained with PAC^2_T do appear to assign some samples from the predictive model to the observed data.

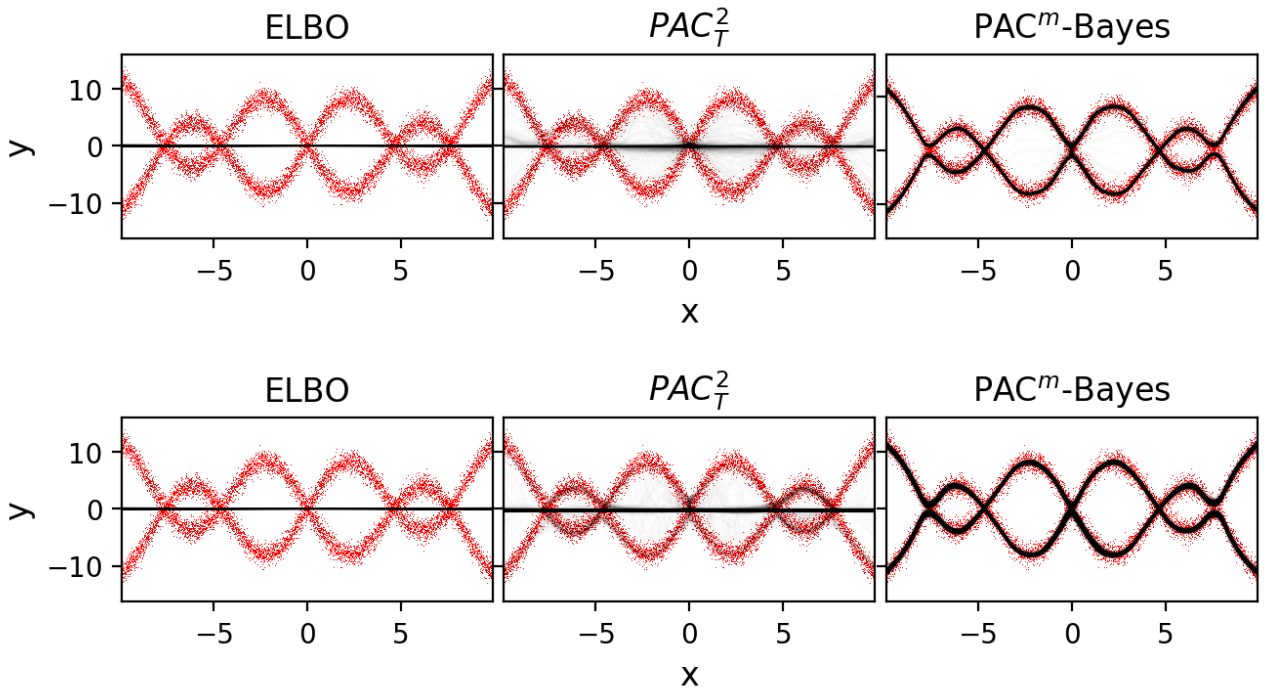

Figure 6. Visualization of the means of the likelihood predicted using samples from the posterior distribution. The top row shows the results for a unimodal posterior, and the bottom row shows the results for a multimodal posterior. We find that elbo assigns all predictions to y=0, while PAC^2_T appears to have occasional samples that track the observed data. PAC^m only has samples which track the observed data.

Instead of visualizing the histogram of samples from the predictive model, we instead draw 1000 samples from the posterior. For each posterior sample, we show the predicted mean of the output distribution as a function of x. Effectively, we want to see two curves tracing each mode in the output data. We show the results in Figure 6. For a 1 component model, we find that PAC^2_T places most of the probability mass on the mean, with tails that can be seen reaching toward the modes. However, we find that there are relatively few samples from the model which track any mode in the data. For a 2 component model, we find that there are proportionally more samples which track each of the modes, but these are still much less frequent than samples which merely follow the mean in the data. For reference, the peak probability density for samples near the data is roughly 30 times less than the density at the mean. This is better than ELBO which places all of its probability near the mean. It is also worse than PAC^m, for which the 1 component model only has very few samples which fall away from either of the modes, and for which the 2 component model only has samples at each mode.

B.2. Well Specified Mixture

To show that all losses perform equivalently when the loss is well specified, we show here an experiment we ran where the likelihood is assumed to be multimodal. Here we used a mixture of normal distributions, with fixed categorical distribution and component variance. This left us to predict only the mean, similarly to the other mixture experiments. Additional details are presented in subsection C.4.

We show the results in Figure 7. As expected, since ELBO is tight for well specified models, it does a reasonable job of recovering the true predictive distribution. However, we should also note that the models which optimize bounds on the predictive risk also perform comparably. In fact, PAC^m observes a marginally lower KL Divergence from the true generative distribution. We measure $\text{KL} = 0.007$ for PAC^m, $\text{KL} = 0.017$ for ELBO, and $\text{KL} = 0.07$ for PAC^2_T. We did not evaluate if the discrepancy is simply due to variance in the optimization or if the lower KL observed from PAC^m is a result of optimizing a tighter bound.

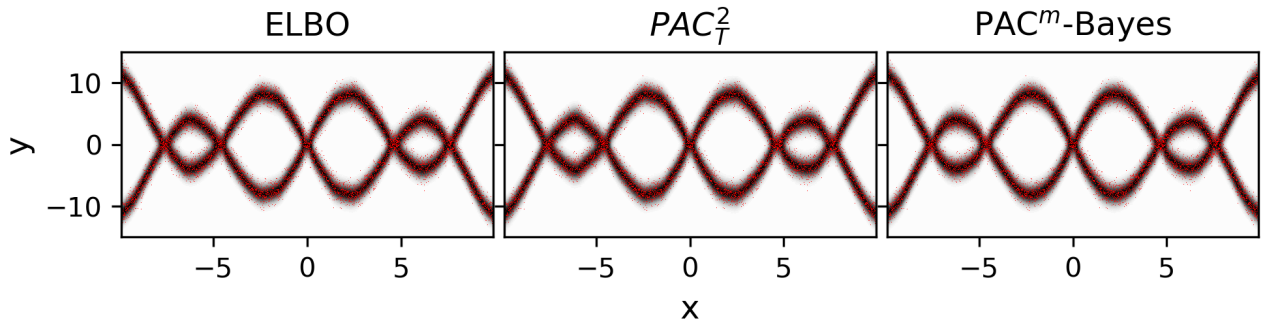

Figure 7. Similar to Figure 3, but where the likelihood is multimodal. This demonstrates performance for a well-specified model. We see that in this scenario, all three losses recover a good predictive model.

B.3. Bayesian Neural Network - Stochastic Weights

For classification experiments using stochastic weights, we give the model the full images from each dataset and attempt to predict the output class. For this, we assume the following graphical model:

$$\Theta \sim r(\Theta) \tag{44}$$

$$\text{for } i = 1 \dots n: \quad y_i \sim p(Y_i | z_T(x_i, \Theta)) \tag{45}$$

where z_T is the output of a T-layer neural network where each layer's parameters are specified by a partitioning of the random vector θ. For example, if z_T is a multilayer perceptron, it might be defined by the recurrence $z_t(x, \theta) = a_{t-1}(z_{t-1}(x, \theta))w_t + b_t$. where $\{(w_t, b_t)\}_t^T$ is partition of vector θ and with appropriately reshaped members and where $a(\cdot)w$ is a (row-) vector-matrix product.

We experimented with classification using a Bayesian Neural Network on several popular benchmarking datasets. Experimental details can be seen in subsection C.5. Similar to (Alemi et al., 2018), we evaluate our models as a function of the constant β by producing the relationship between predictive negative log-likelihood (distortion) and KL divergence in the model (rate), a measure of compression of the model. For global experiments, we show the results in Figure 8. At the end of the day, all models achieve a comparable accuracy (within the experimental uncertainty). However, we find that $\text{PAC}^m$and PAC^2_T do so at lower rate than ELBO, and also have the dominant Pareto-frontier in the information, indicating that it may be doing a better job of distilling useful information from the data.

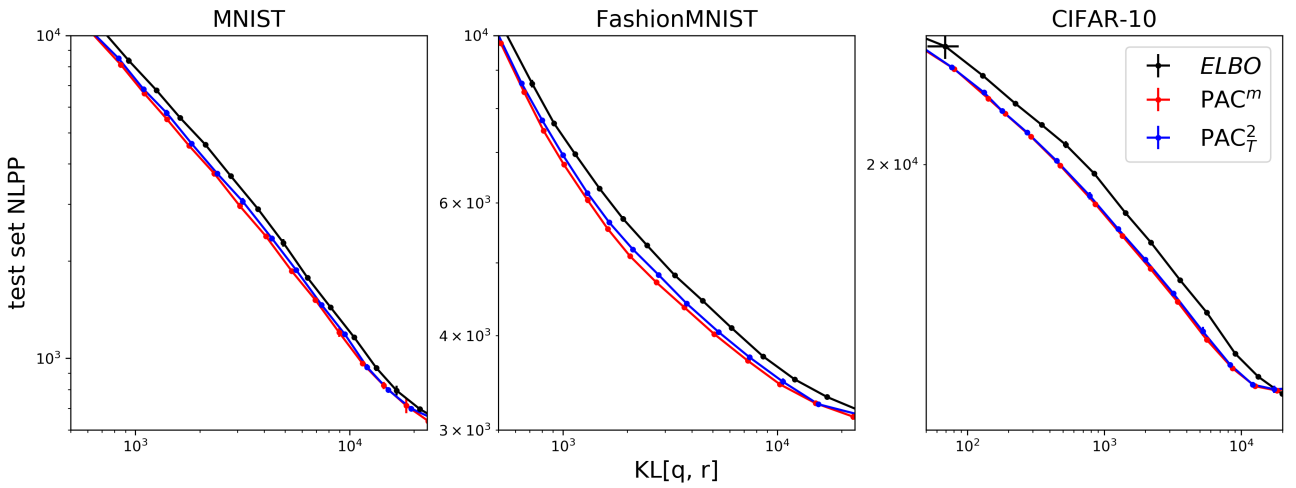

Figure 8. Predictive negative log-likelihood as a function of the KL divergence between the learned posterior and the prior (a measurement of the information content contained in the posterior distribution). The Pareto frontier for each model is shown as the solid line, as measured by the indicated points. Lower and to the left is "better." While all models have comparable performance, we find that models which optimize PAC-Bayesian bounds on the predictive likelihood do a better job of distilling information from the dataset, and therefore require fewer bits to produce equivalent accuracies. PAC^m performs best.

B.4. Bayesian Neural Network - Stochastic Activations

We also consider classification using a different, and non-traditional, type of Bayesian Neural Network wherein we treat the activations of an intermediate layer in the model as the random variables. This corresponds to the following graphical model.

$$\begin{aligned} &\text{for } i = 1 \dots n: \\ &\quad Z_i \sim r(Z) && (46) \\ &\quad y_i \sim p(Y_i|Z_i) && (47) \end{aligned}$$

In this formulation neither evidence x_i nor deep neural network are directly present in the assumed generative process. Rather, these ideas appear only in the construction of the surrogate posterior, i.e., $Z_i \sim q(Z|x_i, \theta)$. For example, one might assume $q(Z_i|x_i, \theta) \equiv \mathsf{Normal}(\mu_x, \sigma_x)$ where μ_x, σ_x are computed from two outputs of a DNN evaluated on x_i and using parameters θ (both of which are regarded as being non-stochastic). This type of setup is familiarized by Variational Autoencoders (Kingma & Welling, 2013), and in deep variational information bottleneck (Alemi et al., 2016) models which use this graphical model to optimize for the log-evidence (or a bound on mutual informations) to set up either an unsupervised generative model (VAE) or a supervised predictive model (VIB). For these experiments, we follow this previous work and use a deep neural network as an "encoder" which predicts the parameters of the posterior distribution, and a "decoder" which uses the latent variable to define $p(Y_i|Z_i)$. As in (Kingma & Welling, 2013), we use the *reparameterization trick* to differentiate through the posterior sampling, which facilitates the optimization of the encoder parameters.

Experimental details are presented in subsection C.5. To evaluate performance, we show the Negative log-posterior-predictive probability as a function of the KL divergence between the posterior and the prior. The results are shown in Figure 9. Notably, we find that PAC^m has a Pareto frontier which advances noticeably beyond the other alternatives. This means that the model needs to learn less information in the posterior in order to make reasonable predictions on the data. We further show the classification accuracy as a function of KL in Figure 10. We find that both PAC^m and PAC^2_T appear to require significantly less information in the latent representation in order to make useful predictions. PAC^m still appears to have the dominant Pareto frontier in this space, though it is often ambiguous that it performs "better" than PAC^2_T in this space. However, it still appears that both outperform ELBO which appears to undergo posterior collapse at relative high rates, as indicated by the sudden sharp decrease in accuracy.

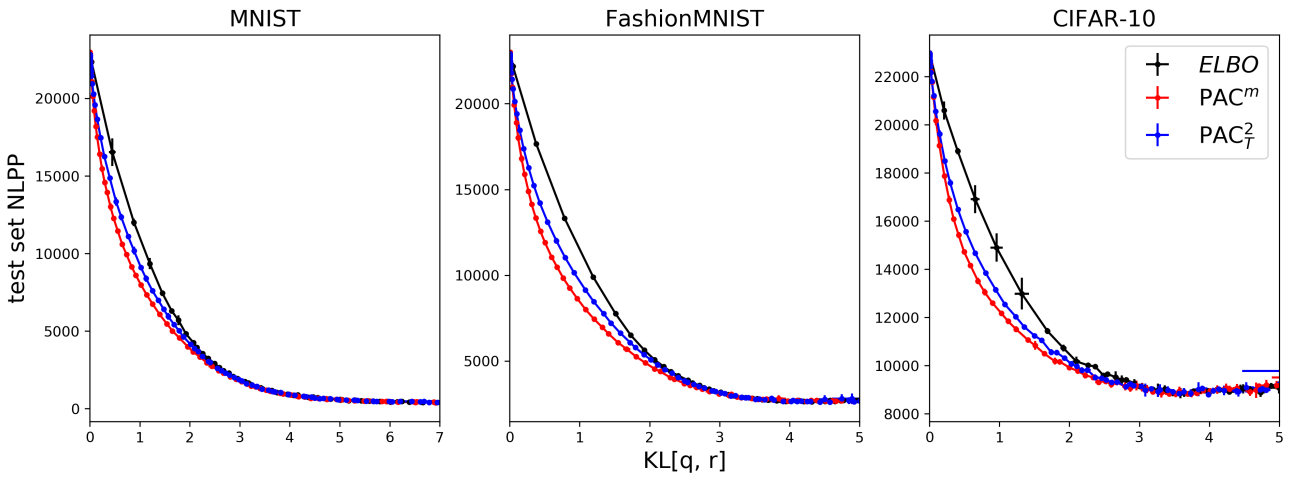

Figure 9. Similar to Figure 8, but where the posterior is defined over activations of an intermediate layer of the network, rather than all of the weights. Similar to before, lower and to the left is "better." We find that in this context, PAC^m clearly has the dominant pareto frontier.

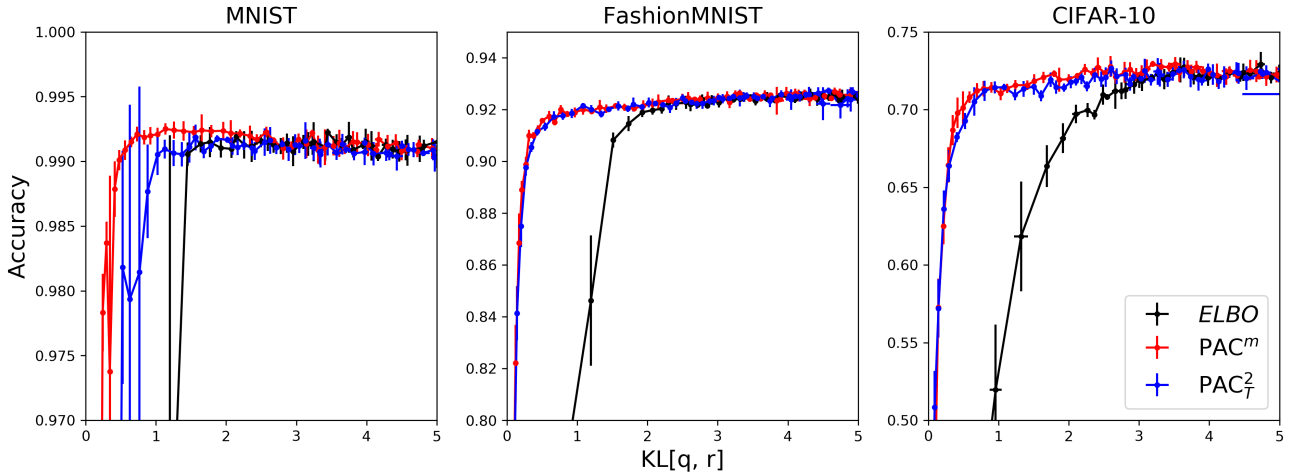

Figure 10. Classification accuracy as a function of the KL Divergence between the posterior and the prior. PAC^m and PAC^2_T consistently offer higher accuracy as a function of KL (i.e. for more *compressed* posteriors), with PAC^m appearing to do slightly better on MNIST. Sharp decreases in classification accuracy, along with corresponding large uncertainty in final accuracy correspond to a sudden collapse of the posterior which occurs for sufficiently large β.

C. Experimental Details

C.1. Example Code

Here we provide example code for computing each loss. In all cases, we assume that one has a posterior, prior, and likelihood, where the posterior and prior are over the weights of the model, and the likelihood is a function which takes in weights and inputs and returns a probability distribution over outputs. All of the following use tensorflow and tensorflow probability (Abadi et al., 2016; Dillon et al., 2017). Additional arguments are xy and y; the inputs to the model and outputs from the model, m; the number of samples to draw from the posterior, β; the weight to place on the KL penalty, and n; the number of examples in the dataset.

```
def elbo(prior, likelihood, posterior, x, y, m, beta, n):
  w = posterior.sample(m)
  ll = likelihood(x, w).log_prob(y)
  kl = tf.reduce_mean(
      posterior.log_prob(w) - prior.log_prob(w),
      axis=0)
  nll = -tf.reduce_mean(ll, axis=(0, 1))
  return nll + kl / (beta * n)
```

Figure 11. TF Probability (Dillon et al., 2017) implementation of ELBO loss for a unimodal global latent variable models (e.g., BNN).

```
def pacm(prior, likelihood, posterior, x, y, m, beta, n):
  w = posterior.sample(m)
  ll = likelihood(x, w).log_prob(y)
  kl = tf.reduce_mean(
      posterior.log_prob(w) - prior.log_prob(w),
      axis=0)
  nlpp = -tf.reduce_mean(
      tfp.math.reduce_logmeanexp(ll, axis=0),
      axis=0)
  return nlpp + kl / (beta * n)
```

Figure 12. TF Probability (Dillon et al., 2017) implementation of PAC^m loss for a unimodal global latent variable models (e.g., BNN). Note that this is identical to ELBO, with the exception of the use of the negative log-posterior-predictive rather than the negative log-likelihood.

Note that PAC^m and ELBO are almost identical. The only difference between the two is that PAC^m uses a log-mean-exp over the sample dimensions to get the negative log-posterior-predictive probability rather than the expected negative log-likelihood. Note also that this is not the case with PAC^2_T, which relies on the additional computation of a complicated variance term. This term has memory and compute cost which scales in the number of samples, though this will likely be sub-dominant to the memory cost of the forward pass in the network itself. It also relies on several tricks to encourage stability, and for the likelihood to be bounded in order for it to not converge to $-\infty$.

```
def pac2t(prior, likelihood, posterior, x, y, m, beta, n, smoothing_constant=0.1):
  w = posterior.sample(m)
  ll = likelihood(x, w).log_prob(y)
  nll = -tf.reduce_mean(ll, axis=(0, 1))
  kl = tf.reduce_mean(posterior.log_prob(w) - prior.log_prob(w), axis=0)
  # We now compute the Masegosa "variance."
  lmx = tf.stop_gradient(
      tf.reduce_max(ll, axis=0, keepdims=True) + smoothing_constant)
  ll_max_centered = ll - lmx
  al = tfp.math.reduce_logmeanexp(ll_max_centered, axis=0)
  h = 2. * tf.stop_gradient(al / (1 - tf.math.exp(al))**2 +
      1. / (tf.math.exp(al) * (1 - tf.math.exp(al))))
  var1 = h * tf.math.exp(2 * ll_mac_centered)
  var2 = tf.math.reduce_mean(
      h * tf.math.exp(
          ll_max_centered[tf.newaxis] +
          ll_max_centered[:, tf.newaxis]),
      axis=0)
  variance = tf.math.reduce_mean(var1 - var2, axis=(0, 1))
  return nll - variance + kl / (beta * n)
```

Figure 13. TF Probability (Dillon et al., 2017) implementation of PAC^2_T loss for a unimodal global latent variable models (e.g., BNN).

C.2. Toy Model

The toy problem in fig. 1 was as described in section 5. The true data distribution came from a 30-70 mixture of two Normal distributions, with a variance of 1 and means at -2 and 2. The model was a standard Normal with fixed unit variance, the only learned parameter being the mean. Five datapoints were drawn, as indicated by the hash marks near the axis in the figures. The inferential risks were determined analytically, as the solution takes on the closed form (Murphy, 2007).

For the PAC-predictive risk, the posterior was found numerically with an iterative procedure. The sought after parameter distribution was represented by the values the density took on a grid with 500 points from -30 to 30. If we take $m \to \infty$ in $\widetilde{\mathcal{P}}_{n,m}$ in eq. (14), we have:

$$\widetilde{\mathcal{P}}_{n,\infty}[q; r, \beta] = -\frac{1}{n}\sum_{i}^{n} \log\left(\int d\theta\, q(\theta) p(x_i|\theta)\right) + \frac{1}{\beta n}\mathsf{KL}\left[q(\Theta); r(\Theta)\right] \tag{48}$$

Trying to minimize this functional with respect to $q(\Theta)$ using calculus of variations (along with the constraint that $q(\Theta)$ integrates to 1) suggests an iterative procedure to find the optimal parameter distribution:

$$q^{n+1}(\Theta) \propto r(\Theta) \exp\left(\beta \sum_i \frac{p(x_i|\Theta)}{p^{(n)}(x_i)}\right) \tag{49}$$

$$p^{(n+1)}(x_i) = \alpha p^{(n)}(x_i) + (1-\alpha)\int d\theta\, q^{(n+1)}(\theta) p(x_i|\theta). \tag{50}$$

We iterated these equations numerically, representing the parameter distribution as the values it took on a grid of 500 points between -30, and 30. Equation (49) sets the new estimate for the parameter distribution in terms of the current estimate for the data point marginal likelihoods. Notice the proportionality here, as we then numerically normalized the density after setting it to the right hand side of eq. (49). Then in eq. (50) we update our estimates of the data point marginal likelihoods, which act as sort of weights for the generalized Boltzmann distribution that is our parameter distribution. For the figure in the paper the mixing fraction α was set to 0.9.

The empirical predictive risk was minimized numerically. For the empirical predictive risk, an explicit mixture was fit, in this instance a 300 component Normal distribution, all with fixed unit variance. This is akin to searching for a 300 component atomic posterior distribution ($q(\Theta) = \sum_i \lambda_i \delta(\Theta - \theta_i)$). This was minimized with `adagrad` trained until it reached a fixed point to within a tolerance of 10^{-5}. Repeated runs all gave the same result. Though this was assuming the parameter distribution was itself atomic, experiments with a setup as was done for the PAC-predictive risk verified that the paramter distribution quickly does collect to an delta-comb.

C.3. Sinusoid

As mentioned in Section 7.1.1, for our first experiment we tried to predict data drawn from the following sinusoid model:

$$\text{for } i = 1 \dots n:$$

$$\mu_{x_i} = 7 \sin\left(\frac{3x_i}{4}\right) + \frac{x_i}{2} \tag{51}$$

$$Y_i \sim \mathsf{Normal}(\mu_{x_i}, 10). \tag{52}$$

We generate data for 10^4 evenly spaced values of $x \in [-10.5, 10.5]$.

For our neural network, we use a two layer MLP with 20 hidden units, and a hyperbolic tangent activation function. We use a Normal distribution for the posterior, with both mean and variance as trainable variables. The initial values of the means were set to 0, and the initial variances were set to 1. The variances were constrained to be positive using the $\exp$ bijector available in tensorflow probability. We similarly use $\mathsf{Normal}(0, 1)$ for the prior over each weight and bias. For the likelihood, we use the MLP to predict the mean of a normal distribution, whose variance is fixed to 1.

We train all models using Adam (Kingma & Ba, 2014) with a learning rate of 0.01 and no learning rate decay. We use full batch training, for 10^5 steps. For Figure 2, we used $m = 100$ samples from the posterior during training, and $\beta = 1$ for all models. For all models, we evaluate performance using the log-posterior predictive, constructed using 10^3 samples from the posterior.

C.4. Mixture Experiments

For our second experiment, we use data generated from a two component mixture distribution:

$$\begin{aligned}
&\text{for } i = 1 \dots n: \\
&\quad \mu_{x_i} = 7 \sin\left(\frac{3x_i}{4}\right) + \frac{x_i}{2} && (53) \\
&\quad Z_i \sim \mathsf{Rademacher} && (54) \\
&\quad Y_i \sim \mathsf{Normal}(Z_i \mu_{x_i}, 1) && (55)
\end{aligned}$$

The model setup was largely similar to the Sinusoid experiment described in subsection C.3, but with one major difference: For these experiments we added an additional hidden layer to the networks to aid in expressiveness. We also used Exponential Linear Unit (ELU) activations instead of tanh to facilitate gradient propagation more easily. For these experiments we used both a unimodal posterior, as well as a mixture posterior, but the underlying distribution was implemented similarly to the sinusoid. When considering a multimodal posterior, we fixed the component probabilities to 0.5 and used stratified sampling to integrate over the discrete categorical random variable. For unimodal likelihoods, we used a normal distribution whose mean was predicted by the model, and which had a fixed variance of 1. When considering a mixture likelihood, we compared situations with both 1 and 2 components in the posterior. The MLP was set up to predict the means of a two component mixture of gaussian distributions, whose component probabilities were fixed to 0.5, and whose component variances were fixed to 1.

All models were again trained using Adam with an initial learning rate of 0.01, but this time we added a small amount of learning rate decay with a decay rate of 0.5 and a decay timescale of 10^5 steps. Because the model was only trained for 10^5 steps, the learning rate only undergoes one half-life. We did not study if holding the learning rate fixed changed the results at all, though it is doubtful that it did. We employed full batch training, and used $m = 100$ samples from the posterior. To evaluate the models qualitatively, as in Figure 3, we used 10^5 samples from the posterior to construct the predictive distribution. For each sample, we computed a forward pass for 10^3 evenly spaced values of x and drew a single sample from the resulting likelihood. This gave us 10^5 samples from the predictive distribution for each x. We then computed the 1-d histogram of the predictive distribution at each x, which we used to display the predictive models as in Figure 3. To quantitatively evaluate models, we used 10^4 samples from the posterior to construct the predictive distribution, and then computed the KL divergence from the known generative distribution for each of the (x, y) pairs in an independently generated test set.

C.5. Image Experiments

C.5.1. Structured Prediction

For the structured prediction experiments, we attempt to solve the following problem: Given the top half of an image, predict the bottom half of the image. We follow the setup in (Masegosa, 2019), for this experiment which we now describe. We use the experimental TFP Neural Networking toolbox (`tfp.experimental.nn`, (Dillon et al., 2017)) and TFP Joint Distributions (Piponi et al., 2020) to compactly specify BNNs. For the posterior and prior, we used Normal distributions. The posterior had learnable variables to represent the mean and variance, while the mean and variance in the prior are assumed to be constant. The model predicts the mean of a Normal distribution with fixed variance of $1/255$ which we use as our likelihood. For the network architecture, we used a 3-layer MLP with 50 hidden units, and ELU activations. We therefore fed our input images to the model as flattened vectors. For CIFAR-10, we converted the image to grayscale to reduce the number of pixels and simplify the model.

All models were trained using Adam with an initial learning rate of 0.001, decayed by 0.5 every 10^5 steps. We train models for 500 epochs. We used a batch size of 128 during training. We tested performance on the heldout evaluation set as a function of m, ranging from $m = 1$ to $m = 32$, where each m corresponds to the number of samples used during training. Reconstruction performance was quantified using the log posterior-predictive (nlpp), which we measure using 100 samples from the posterior. We train 5 different models independently, with different random initializations and different shufflings of the training set in order to obtain the uncertainties in the final performance of the model.

C.5.2. Classification - Stochastic Weights

For our likelihood, we use a categorical distribution with 10 possible outcomes (all image datasets we consider have 10 output classes). Similar to previous experiments, we used Normal distributions for both the Posterior and the Prior, where the posterior uses variables to represent the location and scale of the normal distribution, and where the prior uses fixed values, both of which were initialized or fixed to 0 for the location and 1 for the scale respectively. We used the same architecture as the structured prediction experiments: A 3-layer MLP with 50 hidden units and ELU activations. All input images were normalized to the [-1, 1] interval before being passed to the first layer of the network.

Similar to the structured prediction experiments, we trained with a batch size of 128. We optimized our model for 100 epochs using Adam, with an initial learning rate of 10^{-4}, which we decayed by a factor of 0.5 every 10^5 steps. We study performance as a function of β, and fix the number of samples used during training to $m = 4$. We consider 33 values of c spaced logarithmically and ranging from $\left[10^{-3} - 10^{3}\right]$. We evaluate performance by computing both the log-posterior-predictive nlpp, on the heldout evaluation set. We use 100 samples to construct the posterior predictive distribution.

C.5.3. Classification - Stochastic Activations

For this model, we use the latent embedding z_i to codify each example in a 16 dimensional latent space. To increase the capacity of the model and since the memory cost is much lower, we consider a convolutional neural network for the encoder. This layer uses the following architecture[1]: 4 convolutional layers, followed by 2 dense layers. For each layer, we use LeakyReLU activations. Alternating convolutional layers use a stride of 2. The last dense layer predicts the parameters of the posterior. All images were normalized to the range [-1, 1] prior to being passed to the network.

For the posterior, we used a Multivariate Normal Distribution. The encoder predicts the location and the cholesky decomposition of the precision matrix. For the prior, we used a Multivariate Normal distribution with mean 0, and an identity covariance. For the likelihood, we used a categorical distribution. During training, we used a batch size of 128. We optimized our model for 100 epochs, using Adam with an initial learning rate of 10^{-4}, which was decayed by a factor of 0.5 every 10^5 training steps. Similar to our stochastic weights experiments, we evaluated performance as a function of β using 100 log-spaced bins between 10^{-3} and 10. We evaluate performance by computing both the nlpp and the Accuracy on the evaluation set. For both, we used 100 samples from the posterior to construct the predictive distribution.

D. Quick Reference: Comparing Different Losses

Here we depict several losses closely related to PAC^m and highlight their structural similarities and differences. The likelihood is $p(y|Z)$ and the prior/posterior discrepancy term is $\frac{r(Z)}{q(Z)}$.

$$\text{ELBO} \stackrel{\text{def}}{=} -\mathbb{E}_{q(Z^m)}\left[\frac{1}{m}\sum_j^m \log\left(p(y|Z_j)\right) + \frac{1}{m}\sum_j^m \log\left(\frac{r(Z_j)}{q(Z_j)}\right)\right]$$

$$\text{PAC}^m \stackrel{\text{def}}{=} -\mathbb{E}_{q(Z^m)}\left[\log\left(\frac{1}{m}\sum_j^m p(y|Z_j)\right) + \frac{1}{\beta}\frac{1}{m}\sum_j^m \log\left(\frac{r(Z_j)}{q(Z_j)}\right)\right]$$

$$\text{IWAE} \stackrel{\text{def}}{=} -\mathbb{E}_{q(Z^m)}\left[\log\left(\frac{1}{m}\sum_j^m p(y|Z_j)\frac{r(Z_j)}{q(Z_j)}\right)\right]$$

$$\text{PAC}^2_T \stackrel{\text{def}}{=} \text{ELBO} - \mathbb{E}_{q(Z^m)}\left[\text{SampleVariance}(y, Z^m)\right]$$

Where we colored the average-log, log-average terms, re-framed all losses in terms of multiple samples, and where:

- ELBO is the evidence lower bound (as a loss). (Blei et al., 2017)

[1]This architecture follows the encoder from the VAE example at https://www.tensorflow.org/probability/examples/Probabilistic_Layers_VAE

- IWAE is the importance weighted autoencoder loss of Burda et al. (2015).
- PAC_T^2 is the loss of Masegosa (2019).