Approximate inference for the loss-calibrated Bayesian

Simon Lacoste-Julien
University of Cambridge

Ferenc Huszár
University of Cambridge

Zoubin Ghahramani
University of Cambridge

Abstract

We consider the problem of approximate inference in the context of Bayesian decision theory. Traditional approaches focus on approximating general properties of the posterior, ignoring the decision task – and associated losses – for which the posterior could be used. We argue that this can be suboptimal and propose instead to *loss-calibrate* the approximate inference methods with respect to the decision task at hand. We present a general framework rooted in Bayesian decision theory to analyze approximate inference from the perspective of losses, opening up several research directions. As a first loss-calibrated approximate inference attempt, we propose an EM-like algorithm on the Bayesian posterior risk and show how it can improve a standard approach to Gaussian process classification when losses are asymmetric.

1 INTRODUCTION

Bayesian methods have enjoyed a surge of popularity in machine learning over the last decade. Even though it is sometimes overlooked, the main theoretical motivations for the Bayesian paradigm are rooted in Bayesian decision theory (Berger, 1985), which provides a well-defined theoretical framework for rational decision making under uncertainty about a hidden parameter θ. The ingredients of Bayesian decision theory are an observation model $p(\mathcal{D}|\theta)$, a prior distribution $p(\theta)$, and a loss $L(\theta, a)$ for an action $a \in \mathcal{A}$. In this framework, the optimal action is chosen by minimizing its expected loss over the posterior $p(\theta|\mathcal{D})$. The independence of the posterior from the loss motivates the common practice of breaking decision making into two independent sub-problems: *inference*, whereby the posterior $p(\theta|\mathcal{D})$ is computed irrespectively of the loss; and then *decision*, whereby an action is chosen to minimize its expected loss over our posterior belief.

In practically interesting Bayesian models, however, the posterior is often computationally intractable and therefore one has to resort to approximate inference techniques, such as variational methods or Markov chain Monte Carlo. Most approaches to approximate inference ignore the decision theoretic loss and try to approximate the posterior based on its general features, such as matching its mode or higher order moments. While this is probably a reasonable approach for the simple losses usually considered or when the loss is unknown, they might fail to work well with asymmetric, non-trivial losses that appear in modern applications in machine learning.

The main message of the present paper is that when inference is carried out only approximately, treating (approximate) inference and decision making independently can lead to suboptimal decisions for a fixed loss under consideration. We thus investigate whether one can "calibrate" the approximate inference algorithm to a fixed loss, and propose an analysis framework to analyze this situation. We note that a related philosophy has already been applied in the frequentist discriminative machine learning literature, as for example with the use of *surrogate loss functions* (Bartlett et al., 2006; Steinwart and Christmann, 2008). In contrast, we focus in this paper on the pure subjectivist Bayesian viewpoint as we are not yet aware of the existence of such an investigation in this case. The contributions of the present paper can be summarized as follows:

1. In Sec. 2, we propose a general approximate inference framework based on Bayesian decision theory to guide our analysis. The framework naturally gives rise to a divergence between distributions that can be seen as a loss-calibrated generalization of the Kullback-Leibler divergence for general losses. We focus in this paper on the application of the framework to the predictive setting that is relevant to supervised machine learning.

2. In Sec. 3, we present an algorithmic template to derive loss-calibrated approximate inference algorithms for different losses by applying the varia-

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

tional Expectation-Maximization algorithm on the Bayesian posterior risk.

3. In Sec. 4, we investigate our approximation framework on the concrete setup of supervised learning. We apply the loss-calibrated EM algorithm to a Gaussian process classification model and analyze its performance in terms of the loss-calibrated framework. Our proof-of-concept experiments indicate that it improves over a loss-insensitive approximate inference alternative and that the advantage of loss-calibration is more prominent when misclassification losses are asymmetric.

2 BAYESIAN DECISION THEORY

We use Bayesian statistical decision theory as the basis of our analysis (see Ch. 2 of Robert (2001) or Ch. 1 of Berger (1985) for example). We review here its main ingredients:

- a (statistical) loss $L(\theta, a)$ which gives the cost of taking action $a \in \mathcal{A}$ when the world state is $\theta \in \Theta$;
- an observation model $p(\mathcal{D}|\theta)$ which gives the probability of observing $\mathcal{D} \in \mathcal{O}$ assuming that the world state is θ;
- a prior belief $p(\theta)$ over world states.

The loss L describes the decision task that we are interested in, whereas the observation model and the prior represent our beliefs about the world. Given these, the Bayesian evaluation metric for a possible action a after observing $\mathcal{D}$ is the *expected posterior loss* (also called the *posterior risk* (Schervish, 1995)): $\mathcal{R}_{p_\mathcal{D}}(a) \doteq \int_\Theta L(\theta, a)\, p(\theta|\mathcal{D}) d\theta$, and so the (Bayes) optimal action $a_{p_\mathcal{D}}$ is the one that minimizes $\mathcal{R}_{p_\mathcal{D}}$.

2.1 Supervised learning

We now relate this abstract decision theory setup to the typical supervised learning applications of machine learning. For a prediction task, the goal is to estimate a function $h : \mathcal{X} \to \mathcal{Y}$ where the output space $\mathcal{Y}$ can be discrete (classification) or continuous (regression). We suppose that we are given a fixed cost function $\ell(y, y')$ which gives the cost of predicting y' when the true output was y. We can cast this problem in the standard statistical decision theory setting by defining a suitable prediction loss for our action $a = h$, namely the standard generalization error from machine learning:

$$L(\theta, h) \doteq \mathbb{E}_{(x,y)\sim p(x,y|\theta)} \left[\ell\left(y, h(x)\right)\right]. \tag{1}$$

For the observation model, we will assume that we are given a training set $\mathcal{D} = \{(x_i, y_i)_{i=1}^N\}$ of labeled observations generated i.i.d. from $p(x, y|\theta)$. The goal of the learning algorithm is then to output a function h chosen from a set of (possibly non-parametric) hypotheses $\mathcal{H}$ after looking at the (training) data $\mathcal{D}$. From the pure Bayesian point of view, the best hypothesis $h_{p_\mathcal{D}}$ is clear: it is the one that minimizes the posterior risk, i.e. $h_{p_\mathcal{D}} \doteq \arg\min_{h\in\mathcal{H}} \mathcal{R}_{p_\mathcal{D}}(h)$.

2.2 General approximation framework

The quantity central to the Bayesian methodology is the posterior $p_\mathcal{D}(\theta) \doteq p(\theta|\mathcal{D})$ which summarizes our uncertainty about the world. On the other hand, it is rarely computable in a tractable form, and so it is usually approximated with a tractable approximate distribution $q(\theta) \in \mathcal{Q}$. Popular approaches to this problem include sampling, variational inference – which minimizes $KL(q\|p_\mathcal{D})$, and expectation propagation – which minimizes $KL(p_\mathcal{D}\|q)$ (Minka, 2001). Most approximate inference approaches stop at q, though in the context of decision theory, we still need to *act*. In practice, one usually treats the approximate q as if it was the true posterior and chooses the action that minimizes what we will call the *q-risk*:

$$\mathcal{R}_q(h) \doteq \int_\Theta q(\theta) L(\theta, h) d\theta, \tag{2}$$

obtaining a *q-optimal* action h_q:

$$h_q \doteq \arg\min_{h\in\mathcal{H}} \mathcal{R}_q(h). \tag{3}$$

In this paper, we will assume that computing exactly the q-optimal action h_q for $q \in \mathcal{Q}$ is tractable, and focus on the problem of choosing a suitable q to approximate the posterior $p_\mathcal{D}$ *in order to yield a decision* h_q *with low posterior risk* $\mathcal{R}_{p_\mathcal{D}}(h_q)$, mimicking the standard methodology but crystallizing the decision theoretic goal. Given this approach, a (usually non-unique) optimal $q \in \mathcal{Q}$ is clearly:

$$q_{\text{opt}} = \arg\min_{q\in\mathcal{Q}} \mathcal{R}_{p_\mathcal{D}}(h_q), \tag{4}$$

though a practical algorithm might only be able to find an approximate minimizer to this quantity. In the case where $p_\mathcal{D} \in \mathcal{Q}$, $p_\mathcal{D}$ is obviously optimal according to this criterion.

We could interpret the above criterion as minimizing the following asymmetric non-negative discrepancy measure between distributions:

$$d_L(p\|q) \doteq \mathcal{R}_p(h_q) - \mathcal{R}_p(h_p). \tag{5}$$

Interestingly, the Kullback-Leibler divergence $KL(p\|q)$ can be interpreted as a special case of d_L for the task of posterior density estimation over Θ. In this task, an action h is a density over Θ and the standard density estimation statistical loss is $L(\theta, h) = -\log h(\theta)$. The q-risk $R_q(h)$ then becomes the cross-entropy $H(q, h) = -\int_\Theta q(\theta) \log(h(\theta)) d\theta$, and so $h_q = q$ assuming that

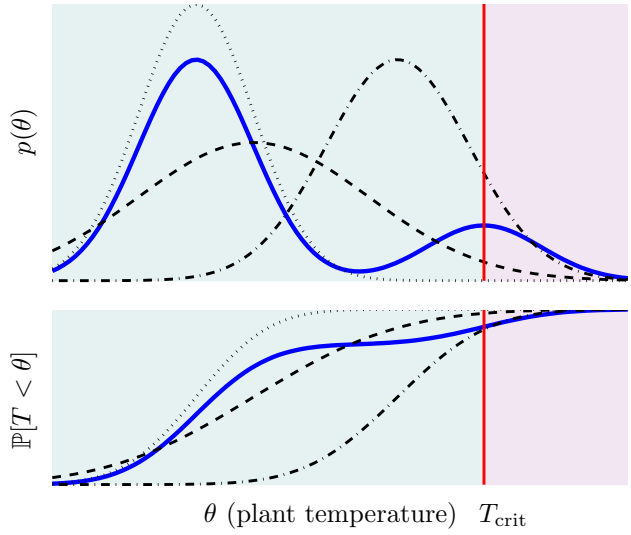

Figure 1: **Top:** Real bimodal posterior *(blue)* and three Gaussian approximations obtained by minimizing $KL(q\|p)$ *(q_1, dotted)*, $KL(p\|q)$ *(q_2, dashed)* or $d_L(p\|q)$ *(q_3, dash-dotted)* in the power plant example. **Bottom:** Cumulative distribution functions for the posterior and the three approximate distributions.

$q \in \mathcal{H}$. Under these assumptions, we obtain that $KL(p\|q) = d_L(p\|q)$ and so as was already known in statistics, $KL(p_{\mathcal{D}}\|\cdot)$ appears "loss-calibrated" for the task of posterior density estimation in our approximation framework. But this begs the natural question of whether minimizing d_L for a particular loss L provides optimal performance under other losses. We will show in Sec. 4.1 that even in the simple Gaussian linear regression setting, minimizing the KL divergence can be suboptimal in the squared loss sense, thus motivating us to seek loss-calibrated alternatives.

To illustrate the difference between traditional approaches to approximate inference and the loss-calibrated framework, consider the following simple problem. Suppose that we control a nuclear power-plant which has an unknown temperature θ that we model with Bayesian inference based on some measurements $\mathcal{D}$. The plant is in danger of over-heating, and as the operator, we can take two actions: either shut it down or keep it running. Keeping it running while the temperature is above a critical threshold T_{crit} will cause a nuclear meltdown, incurring a large loss $L(\theta > T_{\text{crit}}, \text{'on'})$. On the other hand, shutting down the power plant incurs a moderate loss $L(\text{'off'})$, irrespective of the temperature. Suppose that our current observations yielded a complicated multi-modal posterior $p_{\mathcal{D}}(\theta)$ (Fig. 1, solid curve) and that we thus chose to approximate it with a Gaussian. Now consider how various approaches would perform in terms of their Bayesian posterior risk. Minimizing $KL(q\|p_{\mathcal{D}})$ yields candidate q_1 which concentrates around the largest mode, ignoring entirely the second small mode around the critical temperature (Fig. 1, dotted curve). Minimizing $KL(p_{\mathcal{D}}\|q)$ gives a more global approximation: q_2 matches moments of the posterior, but still underestimates the probability of the temperature being above T_{crit}, thereby leading to a suboptimal decision (Fig. 1, dashed curve). q_3 is one of the minimizers of $d_L(p_{\mathcal{D}}\|q)$ in this setting, resulting in the same decision as $p_{\mathcal{D}}$ (Fig. 1, dash-dotted curve). Note that q_3 does not model all aspects of the posterior, but it estimates the Bayes-decision well. Because there are only two possible actions in this setup, the set $\mathcal{Q}$ is split in only two halves by the function $d_L(p_{\mathcal{D}}, q)$ and so there are infinitely many q_{opt}'s that are equivalent in terms of their risk. In contrast, in the predictive setting of section 2.1 where in addition we assume $\mathcal{X}$ and $p(x)$ to be continuous, we could obtain a finer resolution $d_L(p_{\mathcal{D}}\|q)$ which can potentially yield a unique optimizer.

3 LOSS-CALIBRATED EM

In the previous section, we argued that minimizing d_L should guide our choice of approximate posterior, though in practice this optimization also needs to be approximated. In this section, we propose a variational algorithm as a first general loss-calibrated alternative. In order to motivate it, recall that our general goal is to find an action $h_{p_{\mathcal{D}}}$ that minimizes the Bayesian posterior risk $\mathcal{R}_{p_{\mathcal{D}}}$:

$$h_{p_{\mathcal{D}}} = \arg\min_{h \in \mathcal{H}} \int_{\Theta} p(\theta|\mathcal{D}) L(\theta, h) d\theta. \tag{6}$$

This problem combines integration and optimization, which creates a chicken and egg problem of approximating the integration vs. the optimization. One way to solve this chicken and egg problem is to employ a strategy used by the well-known Expectation-Maximization (EM) algorithm (Dempster et al., 1977) which is normally applied to maximize the marginal likelihood, a similar integral over latent variables. EM can be derived from Jensen's inequality and doing coordinate ascent on a lower bound of the log-likelihood. In order to re-use this strategy here, we need to move from minimization to maximization to obtain inequalities in the correct direction. Assuming from now on that our loss function is bounded, we thus define the following *utility* function:

$$U_M(\theta, h) \doteq M - L(\theta, h), \tag{7}$$

where M is a fixed finite constant chosen so that $M > \sup_{\theta \in \Theta, h \in \mathcal{H}} L(\theta, h)$, hence $U_M(\theta, h) > 0$. In analogy with the q-risk $\mathcal{R}_q$, we define the *q-gain* $\mathcal{G}_q$:

$$\mathcal{G}_q(h) \doteq \int_{\Theta} q(\theta) U_M(\theta, h) d\theta. \tag{8}$$

$$\text{(E-step)} \quad q^{t+1} = \arg\min_{q \in \mathcal{Q}} \; KL\left(q \parallel \frac{p_{\mathcal{D}}(\cdot) U_M(\cdot, h^t)}{\mathcal{G}_{p_{\mathcal{D}}}(h^t)}\right)$$

$$\text{(M-step)} \quad h^{t+1} = \arg\max_{h \in \mathcal{H}} \int_{\Theta} q^{t+1}(\theta) \; \log U_M(\theta, h) d\theta$$

Table 1: Loss-EM updates

$$\begin{aligned} \text{(E-step)} \quad q^{t+1} &= \arg\min_{q \in \mathcal{Q}} \; KL\left(q \| p_{\mathcal{D}}\right) + \frac{\mathcal{R}_q(h^t)}{M} \\ \text{(M-step)} \quad h^{t+1} &= \arg\min_{h \in \mathcal{H}} \; \mathcal{R}_{q^{t+1}}(h) \end{aligned}$$

Table 2: Linearized loss-EM updates

Minimizing the q-risk is equivalent to maximizing the q-gain, as well as the log of the q-gain. So we have:

$$h_{p_{\mathcal{D}}} = \arg\max_{h \in \mathcal{H}} \; \log\left(\int_{\Theta} p_{\mathcal{D}}(\theta) U_M(\theta, h) d\theta\right), \qquad (9)$$

which is the optimization problem that we will approximate with (variational) EM.

3.1 Variational EM derivation

Assuming that $q(\theta) = 0 \Rightarrow p_{\mathcal{D}}(\theta) = 0$, we obtain the following lower bound from Jensen's inequality:

$$\begin{aligned} \log\left(\mathcal{G}_{p_{\mathcal{D}}}(h)\right) &= \log\left(\int_{\Theta} q(\theta) \frac{p_{\mathcal{D}}(\theta) U_M(\theta, h)}{q(\theta)} d\theta\right) \qquad (10) \\ &\geq \int_{\Theta} q(\theta) \log\left(\frac{p_{\mathcal{D}}(\theta) U_M(\theta, h)}{q(\theta)}\right) d\theta \doteq \mathcal{L}(q, h). \end{aligned}$$

EM amounts to maximizing the lower bound functional $\mathcal{L}(q, h)$ by coordinate ascent on q and h: the E-step computes $q^{t+1} = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q, h^t)$, while the M-step computes $h^{t+1} = \arg\max_{h \in \mathcal{H}} \mathcal{L}(q^{t+1}, h)$. Moreover, the difference between the quantity that we want to maximize and the lower bound is $\log\left(\mathcal{G}_{p_{\mathcal{D}}}(h)\right) - \mathcal{L}(q, h) = KL(q \| \tilde{p}_h)$, where

$$\tilde{p}_h(\theta) \doteq \frac{p_{\mathcal{D}}(\theta) U_M(\theta, h)}{\mathcal{G}_{p_{\mathcal{D}}}(h)}, \qquad (11)$$

and so the E-step is equivalently minimizing $KL(q \| \tilde{p}_h)$ as h is fixed. We summarize the obtained updates in Table 1 for what we will call the *loss-EM algorithm*. If $\tilde{p}_{h^t} \in \mathcal{Q}$, then $q^{t+1} = \tilde{p}_{h^t}$ and the E-step makes the lower bound tight, as in standard EM, guaranteeing that the original objective improves after each full iteration. On the other hand, we also allow $\mathcal{Q}$ to be a restricted family of tractable distributions, in which case we are using the variational version of EM which only optimizes a lower bound but which has still been applied successfully in the past (Ghahramani and Jordan, 1997; Jordan et al., 1999).

3.2 Linearized loss-EM

Although loss-EM produces a decision h that has good risk, this h is not guaranteed to minimize the q-risk for a particular q, and as such the algorithm does not directly provide us with a loss-calibrated approximate distribution q, as in Sec. 2.2. Also, the objective function in the M-step can be hard to compute and minimize. To address both of these issues, we suggest another approximation. In particular, using the fact that for $M \gg L$, $\log(1 - L/M) = -L/M + O(L^2/M^2)$, we can linearize the $\log U_M$ term in the loss-EM updates to obtain the linearized loss-EM updates given in Table 2. Recall that M was a constant chosen by us: it does not change the optimal action $h_{p_{\mathcal{D}}}$, still it influences the behavior of the loss-EM algorithm. As $M \to \infty$, the linearized and the loss-EM algorithms become basically equivalent as the linearization becomes perfect. On the other hand, we can also see that as $M \to \infty$, both algorithms reduce to the standard variational inference algorithm that minimizes $KL(q \| p_{\mathcal{D}})$, as the second term in the E-step of Table 2 vanishes. Thus, we can see the constant M as a parameter for the linearized loss-EM algorithm which allows us to interpolate between the standard KL approach for large M and a more principled coordinate ascent approach on the Bayesian posterior risk for medium M. It will usually be the case that linearized loss-EM has more tractable updates than loss-EM, but this is at the cost of not corresponding to a valid coordinate ascent algorithm on a lower bound of the posterior risk for medium M.

4 SUPERVISED LEARNING

In this section, we make our framework more concrete by investigating it in the predictive setting presented in Sec. 2.1. We recall that in order to apply our framework, we need to specify the loss, the action space, the Bayesian observation model and a tractable family $\mathcal{Q}$ of approximate distributions over the latent variable θ. In the predictive setting, an action is a prediction function $h : \mathcal{X} \to \mathcal{Y}$. We let the action space $\mathcal{H}$ be the set of all possible such functions here – we are thus in the non-parametric prediction regime where we are free to make arbitrary pointwise decision on $\mathcal{X}$. This gives us rich predictive possibilities as well as actually enables us to analytically compute h_q, as we will see in the next paragraph. For the observation model, we consider Bayesian non-parametric probabilistic models based on Gaussian processes (GPs), which have been

successfully applied to model a wide variety of phenomena in the past (Rasmussen and Williams, 2006). In Sec. 4.1, we first look at Gaussian process regression. In this case, we can obtain an analytic form for $p_{\mathcal{D}}$ and $\mathcal{R}_{p_{\mathcal{D}}}(h_q)$ which gives us some insights about the approximation framework as well as when minimizing the KL divergence can be suboptimal. Because the quadratic cost function is not bounded (and so $M = \infty$), we cannot directly apply our loss-EM algorithm for regression, but we can nevertheless get useful insights which suggest future research directions for regression with sparse GPs. In section 4.2, we consider Gaussian process classification (GPC) which will provide a test bed for the loss-EM algorithm. In both cases, we use a GP as our prior over parameters and let $\mathcal{Q}$ also be a family of GPs.

For both regression and classification, we will look at the discriminative regime inasmuch we are not modelling the marginal distribution of x: we assume that we are given a fixed test distribution $p(x)$ which enters in the generalization error $L(\theta, h)$ given by (1), but *not* for the generation of the training inputs x_i. In other words, we assume that $\mathcal{D} = \{(x_i, y_i)_{i=1}^N\}$ with y_i generated independently from $p(y|x_i, \theta)$ for each x_i, but we *do not* assume that x_i is generated from $p(x)$ – for example the training inputs could even be chosen deterministically or have different support than $p(x)$. We could think of the test input distribution $p(x)$ as coming from a large unlabeled corpus of examples or from the transductive setting which specifies where we want to make predictions. In this discriminative predictive setup, the loss (1) separates pointwise over $\mathcal{X}$:

$$L(\theta, h) = \int_{\mathcal{X}} p(x) \left(\int_{\mathcal{Y}} p(y|x, \theta)\, \ell\left(y, h(x)\right) dy \right) dx, \tag{12}$$

and the q-risk also takes the pointwise form (by pushing the marginalization over θ inside):

$$\mathcal{R}_q(h) = \mathbb{E}_{X \sim p(x)} \underbrace{\left[\int_{\mathcal{Y}} p_q(y|X)\, \ell\left(y, h(X)\right) dy \right]}_{\doteq \mathcal{R}_q(h(X)|X)}, \tag{13}$$

where the q-conditional-risk $\mathcal{R}_q(h(X)|X)$ was defined in terms of the q-marginalized predictive likelihood that we denote by $p_q(y|x)$:

$$p_q(y|x) \doteq \int_{\Theta} q(\theta) p(y|x, \theta) d\theta. \tag{14}$$

In the case of non-parametric h, the q-optimal action h_q can thus be analytically obtained as the pointwise minimum of the q-conditional-risk:

$$h_q(x) = \arg\min_{y \in \mathcal{Y}} \mathcal{R}_q(y|x). \tag{15}$$

4.1 Gaussian process regression

We now describe the Gaussian process regression setup, which actually requires a small redefinition from the standard approach in order to analyze our framework in a simple fashion. The standard approach to GP regression would be to use a Gaussian observation model $p(y|x, f) = \mathcal{N}(y|f(x), \sigma^2)$ with observation noise hyperparameter σ^2 and where the latent parameter for the observation model is actually a function $f : \mathcal{X} \to \mathbb{R}$. The prior over this parameter would be a Gaussian process (basically an infinite dimensional multivariate normal): $p(f) = GP(f|0, K)$, where $K(\cdot, \cdot)$ is the covariance kernel for the GP. In order to avoid the technical complications of looking at the KL divergence between infinite dimensional distributions[1], we make the following subtle but important observation about our framework: because our analysis is *conditioned* on the data (in terms of posterior risk optimization), it turns out that we can *equivalently* redefine our probabilistic observation model using a finite parameter vector θ of size N. We provide more details for this in Appendix 7.1. We stress that this is possible because we are only interested in the problem of finding an h that approximately minimizes the posterior risk; we are not considering for example the problem of updating the posterior with incoming observations. We are thus free to define a probabilistic model which actually depends on $\mathcal{D}$ for the purpose of analyzing the quantities arising in the framework of Sec. 2.2.

The equivalent probabilistic model that we can use is the following finite dimensional model:

$$p(\theta) = \mathcal{N}(\theta|0, K_{\mathcal{D}\mathcal{D}}^{-1}) \tag{16}$$

$$p(y|x, \theta) = \mathcal{N}(y|\mu_x(\theta), \sigma_x^2), \tag{17}$$

where $K_{\mathcal{D}\mathcal{D}}$ is the $N \times N$ matrix with (i, j) entry $K(x_i, x_j)$. We also define similarly $K_{x\mathcal{D}}$ as the $1 \times N$ row vector with i^{th} entry $K(x, x_i)$ as well as its transpose $K_{\mathcal{D}x}$ to write the conditional mean and variance of the observation model as follows:

$$\begin{aligned} \mu_x(\theta) &\doteq K_{x\mathcal{D}}\theta \\ \sigma_x^2 &\doteq \sigma^2 + K_{xx} - K_{x\mathcal{D}} K_{\mathcal{D}\mathcal{D}}^{-1} K_{\mathcal{D}x}. \end{aligned} \tag{18}$$

These expressions can be derived from the standard GP model by doing the change of variable $\theta = K_{\mathcal{D}\mathcal{D}}^{-1} f_{\mathcal{D}}$, where $f_{\mathcal{D}} \doteq (f(x_1), \ldots, f(x_N))^\top$. This change of variables has the advantage of yielding a h_q which does not require the expensive inversion of $K_{\mathcal{D}\mathcal{D}}$.

With our Bayesian observation model fully specified, we are now ready to analyze the q-risk for GP regression. Following the standard convention for regression, we

[1] See Csató (2002) for one way to define the KL divergence between GPs.

consider the quadratic cost function $\ell(y, y') = (y - y')^2$. The q-conditional-risk in (15) takes the simple form:

$$\mathcal{R}_q(y'|x) = \mathbb{V}\mathrm{ar}_q[Y|x] + (\mathbb{E}_q[Y|x] - y')^2, \tag{19}$$

where $\mathbb{E}_q[Y|x]$ and $\mathbb{V}\mathrm{ar}_q[Y|x]$ are the conditional mean and variance of $p_q(y|x)$ respectively. If we assume that q is a Gaussian with mean μ_q and covariance Σ_q, we get that the q-optimal action has the simple form $h_q(x) = \mathbb{E}_q[Y|x] = K_{x\mathcal{D}}\mu_q$. Note that in this case h_q does not depend on Σ_q and so we do not need to specify Σ_q for this application – the Bayesian posterior risk of h_q is agnostic to it. Because of our Gaussian observation model, the posterior $p_\mathcal{D}$ is also a Gaussian $\mathcal{N}(\mu_{p_\mathcal{D}}, \Sigma_{p_\mathcal{D}})$ which thus lies in $\mathcal{Q}$. We can now obtain an explicit expression for the excess posterior risk of h_q compared to the Bayes decision $h_{p_\mathcal{D}}$:

$$d_L(p_\mathcal{D}\|q) = (\mu_q - \mu_{p_\mathcal{D}})^\top \Lambda (\mu_q - \mu_{p_\mathcal{D}}), \tag{20}$$

where

$$\Lambda \doteq \int_\mathcal{X} p(x) K_{\mathcal{D}x} K_{x\mathcal{D}} dx \tag{21}$$

is a loss-sensitive term (i.e. is sensitive to where the test set distribution $p(x)$ lies). It is interesting to compare d_L with the KL divergence between two Gaussians:

$$KL(q\|p_\mathcal{D}) = c(\Sigma_q) + \frac{1}{2}(\mu_q - \mu_{p_\mathcal{D}})^\top \Sigma_{p_\mathcal{D}}^{-1} (\mu_q - \mu_{p_\mathcal{D}}) \tag{22}$$

where $c(\Sigma_q)$ is constant with respect to μ_q. Both are quadratic forms in $(\mu_q - \mu_{p_\mathcal{D}})$, but with different Hessians (we give an explicit formula for $\Sigma_{p_\mathcal{D}}^{-1}$ in Appendix 7.2). So the first interesting observation is that unless our family $\mathcal{Q}$ contains the true posterior mean (i.e. $\exists q \in \mathcal{Q}$ s.t. $\mu_q = \mu_{p_\mathcal{D}}$), the minimum KL solution will not necessarily minimize d_L – i.e. KL is not loss-calibrated.

We also make the following high-level observations for which we provide more details in Appendix 7.2. For GP regression, $\mu_{p_\mathcal{D}}$ has an explicit formula but takes $O(N^3)$ to compute due to the inversion of the kernel matrix. For computational efficiency, some proposals have been made in the GP literature to use a *sparse* μ_q instead (Quińonero-Candela and Rasmussen, 2005). We can thus consider $\mathcal{Q}$ to be a set of Gaussians with sparse mean with support on only a fixed subset of $\mathcal{D}$ of size k. It actually turns out that we can compute the sparse mean $\mu_{q_{\mathrm{sp}}^{\mathrm{KL}}}$ that minimizes the KL (22) over $\mathcal{Q}$ in $O(k^3)$ due to fortuitous cancellations[2]. Unfortunately, the minimizer $\mu_{q_{\mathrm{sp}}^{\mathrm{opt}}}$ of d_L (20) with sparse constraints does not yield similar cancellations and still requires $O(N^3)$ time to compute. It thus leaves open how to obtain *efficiently* an approximate sparse solution with lower Bayesian risk than $\mu_{q_{\mathrm{sp}}^{\mathrm{KL}}}$. Equations (20) and (21) make it clear though that the sparse approximations to the GP should take the test distribution $p(x)$ in consideration, especially if $p(x)$ is quite different of the training input distribution in $\mathcal{D}$. We see this question as an interesting open problem.

[2]See also section 2.3.6 in Snelson (2007) for the interpretation of sparse GPs as KL minimizers.

4.2 Gaussian process classification

After having looked at an example for which we could compute the posterior analytically, we now consider one where the posterior is intractable and on which we can apply the loss-EM algorithm. We look at Gaussian process binary classification ($\mathcal{Y} = \{-1, +1\}$). We allow for an asymmetric binary cost function: the cost $\ell(y, y')$ is zero for $y = y'$ and has false positive value $\ell(-1, +1) = c_+$ and false negative value $\ell(+1, -1) = c_-$. We use the probit likelihood model $p(y|x, f) = \Phi(yf(x)) = \int_{z \le yf(x)} N(z|0, 1)dz$, i.e. Φ is the cumulative distribution function of a univariate normal, and we use a GP prior on f. Using the same trick as mentioned at the beginning of Sec. 4.1, we use a finite parametrization $\theta = K_{\mathcal{D}\mathcal{D}}^{-1} f_\mathcal{D}$ and redefine the equivalent (in terms of posterior risk) probabilistic model:

$$p(\theta) = N(\theta|0, K_{\mathcal{D}\mathcal{D}}^{-1}) \tag{23}$$

$$p(y|x, \theta) = \Phi\left(y \frac{K_{x\mathcal{D}}\theta}{\sigma_x}\right), \tag{24}$$

where σ_x^2 is as in (18), but with $\sigma^2 = 1$. We also assume the transductive scenario where we are given a test set $\mathcal{S}$ of S points $\{x_s\}_{s=1}^S$, i.e. $p(x) = \frac{1}{S}\sum_s \delta_{x_s}$.

We use again a Gaussian approximate posterior $q = \mathcal{N}(\mu_q, \Sigma_q)$ which enable us to get a closed form for the marginalized predictive likelihood (14):

$$p_q(y|x) = \Phi\left(y \frac{K_{x\mathcal{D}}\mu_q}{\sigma_q(x)}\right), \tag{25}$$

where $\sigma_q^2(x) \doteq \sigma_x^2 + K_{x\mathcal{D}}\Sigma_q K_{\mathcal{D}x}$ (and so unlike in the regression case, we see here that Σ_q can influence the decision boundary in the case of asymmetric cost function). The q-optimal action with general formula (15) has then the following analytic form:

$$h_q(x) = \mathrm{sign}\{K_{x\mathcal{D}}\mu_q - \sigma_q(x) b_c\}, \tag{26}$$

where b_c is a threshold depending on the amount of cost asymmetry $b_c \doteq \Phi^{-1}(c_+/(c_- + c_+))$ (see Appendix 7.3 for details). In the E-step of loss-EM, we need to minimize $-\int_\Theta q(\theta) \log \tilde{p}_{h^t}(\theta) d\theta - H(q)$ with respect to q, where $\tilde{p}_{h^t}$ is defined in (11) and corresponds to a loss-sensitive weighting of the posterior distribution. By analogy to a standard methodology for GP classification, we use a Laplace approximation of the intractable $\tilde{p}_{h^t}$ (which corresponds to a second order Taylor expansion of $\log \tilde{p}_{h^t}(\theta)$ around the mode $\hat{\theta}$ of $\tilde{p}_{h^t}$). This yields

1: Initialize h^0 to a random function.
2: **for** $t = 0$ to T **do**
3: (Laplace E-step) Maximize $\log \tilde{p}_{h^t}$ using conjugate gradient to get $\hat{\theta}$.
4: Set $\mu_{q^{t+1}} = \hat{\theta}$ and $\Sigma^{-1}_{q^{t+1}} = -\nabla\nabla \log \tilde{p}_{h^t}(\hat{\theta})$.
5: (Linearized M-step)
Set $h^{t+1}(x_s) = h_{q^{t+1}}(x_s)$ as per (26) $\forall x_s \in \mathcal{S}$.
6: **if** $h^{t+1} = h^t$ **then return** h^{t+1}.
7: **end for**

Table 3: Laplace Linearized Loss-EM for GPC

a Gaussian approximation $\tilde{p}_{h^t}(\theta) \simeq \mathcal{N}(\theta|\mu_{q^{t+1}}, \Sigma_{q^{t+1}})$. Hence minimizing the KL with this approximation will yield back the same Gaussian for q assuming it is unrestricted. We present the full algorithm in Table 3. We use the conjugate gradient algorithm to find a local maximum of $\log \tilde{p}_{h^t}(\theta)$. We present its gradient here as it provides interesting insights on the loss-sensitivity of the algorithm:

$$\frac{\partial}{\partial\theta}\log\tilde{p}_{h^t}(\theta) = -K_{\mathcal{D}\mathcal{D}}\theta + \sum_{x_i\in\mathcal{D}} a_{x_i}\frac{y_i}{p(y_i|x_i,\theta)}K_{\mathcal{D}x_i} + \frac{1}{S}\sum_{x_s\in\mathcal{S}} a_{x_s}\frac{h^t(x_s)\ell\left(-h^t(x_s), h^t(x_s)\right)}{U_M(\theta, h^t)}K_{\mathcal{D}x_s}, \quad (27)$$

where $a_x \doteq \sigma_x^{-1}N(K_{x\mathcal{D}}\theta/\sigma_x|0,1)$. The first term of (27) comes from the prior; the second from the likelihood and the third from the loss. By comparing the third term with the second, we see that the effect of the loss term on the gradient is to push the gradient in the directions of the previous decision $h^t(x_s)$ and proportional to the cost of a false prediction. Unsurprisingly, if the cost is symmetric, we expect the effect to be smaller, as we will see in our synthetic experiments.

5 EXPERIMENTS

As a proof-of-concept, we conducted the following synthetic experiments testing the performance of our linearized loss-EM algorithm for GP classification (Table 3). We generated 100 synthetic datasets, each with 15 univariate training inputs sampled from a uniform distribution on $[0, 1]$, denoted by $\mathcal{U}(0, 1)$. For each dataset, a fixed random function was drawn from the GPC prior and used to generate at random the binary labels y_i according to the GPC observation model.

To investigate the effect of the test distribution $p(x)$ on our method, we generated three different transductive test sets of size 1000, with inputs sampled from $\mathcal{U}(0, 1)$, $\mathcal{U}(0.2, 1.2)$ and $\mathcal{U}(0.5, 1.5)$ respectively (columns of Table 4), and repeated these experiments 10 times to get significance results. We used five different loss matrices: the loss for false negatives was constant at $c_- = 1$, the loss for false positives c_+ was varied so that the decision threshold $p_{thresh} = \frac{c_+}{c_-+c_+}$ changed linearly between 0.5 and 0.05 (rows of Table 4).

For each dataset, we compared three methods for approximate inference: Laplace approximation, expectation propagation (EP) and loss-EM (run separately for each loss and test set combination). Both Laplace and EP are standard approaches to GP classification (Rasmussen and Williams, 2006). To evaluate the performance of the methods, we used the following criterion based on the posterior risk:

$$\tilde{R}(q) = \frac{R_{p_\mathcal{D}}(h_q) - R_{p_\mathcal{D}}(h_{p_\mathcal{D}})}{R_{p_\mathcal{D}}(-h_{p_\mathcal{D}}) - R_{p_\mathcal{D}}(h_{p_\mathcal{D}})}. \quad (28)$$

where $-h_{p_\mathcal{D}}$ is the classifier that always makes the opposite prediction to the optimal classifier – thus $R_{p_\mathcal{D}}(-h_{p_\mathcal{D}})$ provides an upper bound on the posterior risk of any classifier. $\tilde{R}(q)$ is thus normalized to take values between 0 (posterior-optimal) and 1 (maximum risk), enabling us to aggregate performance measures over trials of different difficulty. We estimated $R_{p_\mathcal{D}}(h_q)$ by sampling a large number of $\theta^{(i)} \sim p_\mathcal{D}(\theta)$ with hybrid Monte Carlo sampling (Neal, 2010), and averaging the corresponding values of $L(\theta^{(i)}, h_q)$ (12). The numbers reported in Table 4 are the mean $\tilde{R}$ values, excluding the "easy" scenarios for which $\tilde{R}(q)$ were zero for all methods. We note that EM usually converged in less than 5 iterations for M set to the maximum loss.

We observed that loss-EM provided some improvement over the direct Laplace approximation of the posterior when the loss is asymmetric. This is in line with our expectation that loss-calibration is more critical when the loss is asymmetric. Another observation is that EP dominates the other approaches on these simple 1D synthetic examples. This could be because EP is particularly effective at approximating the posterior in GP classification as was already known (Nickisch and Rasmussen, 2008) and definitively superior to Laplace approximation. We also note that EP aims at minimizing $KL(p_\mathcal{D}\|q)$, whereas our particular EM algorithm is closer to optimizing $KL(q\|p_\mathcal{D})$. These findings motivate future research into algorithms that minimize d_L more directly – one possibility could be to use EP to approximate $\tilde{p}_{h^t}$ in step 3 of Table 3.

6 DISCUSSION

Related work. As mentioned in the introduction, the discriminative machine learning community has already produced several inherently "loss-calibrated" algorithms. A common learning approach is to optimize a regularized upper bound (called *surrogate loss*) of the empirical generalization error that directly depends on the cost function, such as in modern versions of large margin approaches (Steinwart and Christmann, 2008). See also the concurrently submitted work of Stoyanov et al. (2011) which estimates the parameters in graphical models using empirical risk minimization and taking

c_+	p_{tresh}	$\mathbf{p_{test}} = \mathcal{U}(\mathbf{0},\mathbf{1}) = \mathbf{p_{train}}$			$\mathbf{p_{test}} = \mathcal{U}(\mathbf{0.2},\mathbf{1.2})$			$\mathbf{p_{test}} = \mathcal{U}(\mathbf{0.5},\mathbf{1.5})$		
		Lapl	L-EM	EP	Lapl	L-EM	EP	Lapl	L-EM	EP
1.00	0.5000	.0009	**.0009**	.0005	**.0027**	.0035	.0023	.0157	.0187	.0158
0.63	0.3875	.0008	.0008	.0005	.0031	**.0026**	.0024	.0400	**.0371**	.0348
0.38	0.2750	.0025	**.0022**	.0020	.0088	**.0065**	.0035	.0382	.0387	.0249
0.19	0.1625	.0099	**.0084**	.0011	.0207	**.0196**	.0031	**.0360**	.0370	.0098
0.05	0.0500	.1891	**.1890**	.0033	.1184	**.1183**	.0024	.0414	**.0413**	.0011

Table 4: Performance of Laplace approximation *(Lapl)*, Loss-EM *(L-EM)* and expectation propagation *(EP)* applied to GP classification on synthetic datasets as a function of the shift between the test and training distributions *(columns)* and the asymmetry of loss *(rows)*. Smaller numbers mean better performance (see text). Numbers in bold indicate a significant difference according to the Wilcoxon signed rank test at $p = 0.01$ level between *Lapl* and *L-EM* over the 10 repetitions. *EP* is consistently better.

approximate inference in consideration. Their objective is somewhat different inasmuch as these approaches are aimed at minimizing the frequentist risk – an average over possible training sets, whereas the Bayesian approach tries to make the most of the *given set of observations* by conditioning on it. We see these two approaches as complementary, rather than conflicting, and hope that our framework will attract more interest in analyzing the decision theoretic basis of Bayesian methods used in machine learning.

A closely related approach at midpoint between the Bayesian methodology and the frequentist one is Maximum Entropy Discrimination (MED) by Jaakkola et al. (1999). Following the more modern treatment of Jebara (2011), MED aims at solving the following optimization problem (using our notation):

$$q^{MED} = \arg\min_{q\in\mathcal{Q}} KL(q(\theta)||p(\theta)) + C\sum_i \xi_i \qquad (29)$$
$$\text{s.t. } \xi_i + p_q(y_i|x_i) \geq p_q(y|x_i) + \ell(y_i, y)\ \forall i, y \in \mathcal{Y},$$

though in practice they use $\int_\Theta q(\theta)\log p(y|x,\theta)d\theta$ rather than $p_q(y|x)$ for computational reasons. The MED optimization problem can be contrasted to our linearized E-step of Table 2. MED uses the data through a hinge upper bound (Joachims et al., 2009) on the empirical error (the ξ_i part), whereas we use the data $\mathcal{D}$ through the likelihood term of $p_{\mathcal{D}}$. The term $\mathcal{R}_q(h^t)$ can be contrasted to the ξ_i part as being a Bayesian loss on data labeled by h^t (our previous best guess) instead of the empirical error on $\mathcal{D}$ as it is for MED.

Finally, we note that Dawid (1994) has provided an extensive analysis of the discrepancy d_L that we defined in (5). He analyzed its relationship to losses and 'scoring rules', and studied the question of which losses would yield a unique minimizer.

Summary and future directions. Our main goal with this paper was to emphasize that, when faced with a particular decision task with a fixed loss, an approximate inference method should take the loss into consideration. We took initial steps into what we believe will become a rich field of interesting research questions. We proposed a general decision theoretic framework in which we identified minimization of the loss divergence d_L as an objective of loss-calibrated approximate inference. We designed a variational EM algorithm and applied it in the context of non-parametric Bayesian classification. Our synthetic experiments indicated that our loss-calibrated method improved over its loss-insensitive counterpart, i.e. Laplace approximation, but was outperformed by EP, motivating as a line of future research the loss-calibration of EP. Moreover, the loss-calibrated framework highlights which key ingredients need to be considered when calibrating approximate inference to a task. Considering these ingredients, we see the following as promising applications for our framework:

1. non-trivial ℓ: Our experiments suggest that the loss-calibration is more pronounced in the case of asymmetric losses, which suggests that the approach has most benefits for applications where complex, structured losses are used, such as in structured prediction (Bakir et al., 2007).

2. parametric decision boundary: restricting $\mathcal{H}$ to a parametric family – e. g. in consideration of computational efficiency – induces tradeoffs in the performance that different approximate q's can achieve. Therefore, the approximate inference algorithm needs to be calibrated to those tradeoffs.

3. semi-supervised learning and covariate shift: information can enter our framework through the test distribution $p(x)$ which can be arbitrarily different than the empirical distribution of training inputs. We could thus handle the covariate shift problem (Sugiyama et al., 2007) with a set of unlabelled examples from the test distribution.

Acknowledgments This work was supported by the EPSRC grants EP/F026641/1 and EP/F028628/1.

References

G. H. Bakir, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan. *Predicting Structured Data*. The MIT Press, 2007.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 1985.

L. Csató. *Gaussian Processes - Iterative Sparse Approximations*. PhD thesis, Aston University, 2002.

A. P. Dawid. Proper measures of discrepancy, uncertainty and dependence with applications to predictive experimental designs. Technical Report 139, Department of Statistical Science at University College London, 1994. (revised in 1998).

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Z. Ghahramani and M. I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–275, 1997.

T. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. In *Advances in Neural Information Processing Systems 12*. MIT Press, Cambridge, MA, 1999.

T. Jebara. Multitask sparsity via Maximum Entropy Discrimination. *Journal of Machine Learning Research*, 12:75–110, 2011.

T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77 (1):27–59, 2009.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, 1999.

T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

R. M. Neal. MCMC using Hamiltonian dynamics. In G. J. S. Brooks, A. Gelman and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press, 2010.

H. Nickisch and C. E. Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 2008.

J. Quińonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1935–1959, 2005.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, USA, 2006.

C. P. Robert. *The Bayesian Choice*. Springer, New York, 2001.

M. J. Schervish. *Theory of Statistics*. Springer, New York, 1995.

E. Snelson. *Flexible and efficient Gaussian process models for machine learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2007.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

V. Stoyanov, J. Eisner, and A. Ropson. Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In G. Gordon and D. Dunson, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, Fort Lauderdale, FL, USA, April 2011. Journal of Machine Learning Research.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.