5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–9

Variational Prediction

Alexander A. Alemi ALEMI@GOOGLE.COM and **Ben Poole** POOLEB@GOOGLE.COM
Google Research

Abstract

Bayesian inference offers benefits over maximum likelihood, but it also comes with computational costs. Computing the posterior is typically intractable, as is marginalizing that posterior to form the posterior predictive distribution. In this paper, we present *variational prediction*, a technique for directly learning a variational approximation to the posterior predictive distribution using a variational bound. This approach can provide good predictive distributions without test time marginalization costs. We demonstrate Variational Prediction on an illustrative toy example.

1. Introduction

The promise of Bayesian inference is that it can provide accurate predictions by leveraging prior knowledge about the world and its mechanisms. Unfortunately, computing these predictions is costly, as it requires integrating over all possible parameter settings.

Given a parametric statistical *model* $p(x|\theta)$ and *prior* $p(\theta)$, the *posterior distribution* over the parameters of our model is given by: [1]

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \quad \text{with} \quad p(D) = \int d\theta\, p(\theta)p(D|\theta). \tag{1}$$

We marginalize out this posterior to compute the *posterior predictive* distribution:

$$p(x|D) = \int d\theta\, p(x|\theta)p(\theta|D), \tag{2}$$

which is the optimal predictive distribution, on average, in the well-specified case (Aitchison, 1975).

Unfortunately, both computing the posterior distribution $p(\theta|D)$ and marginalizing this posterior to form the posterior predictive $p(x|D)$ are often intractable.

Most work on scaling Bayesian inference has focused on tackling the first problem of posterior inference: how can we sample or approximate $p(\theta|D)$? The two major approaches are (1) Markov Chain Monte Carlo methods that aim to form a sampling chain that generates samples from the posterior, and (2) variational methods that search for a parametric distribution that is as close as possible to the true Bayesian posterior. Given infinite time to run Monte Carlo, or an infinitely flexible variational family, we could recover samples from the Bayesian posterior. Unfortunately, these approaches only address the first of our two problems. We still need to marginalize out this distribution to form the posterior predictive and inaccuracies in the approximate posterior could propagate to inaccuracies in the approximate posterior predictive.

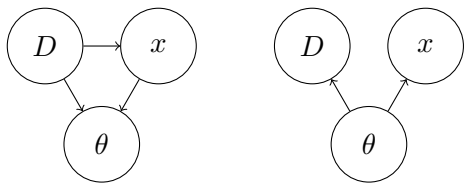
Why spend so much time, compute, and energy to estimate the posterior over parameters if our primary goal is to form accurate predictions? Can we shortcut the two-step process and target the predictive distribution directly? It turns out we can.

1. We use $p(D|\theta)$ as shorthand for the likelihood over a dataset, $D \equiv \{x_1, x_2, \ldots, x_N\}$, $p(D|\theta) \equiv \prod_i p(x_i|\theta)$.

© A.A. Alemi & B. Poole.

2. Variational Prediction

What we want is a direct variational predictive distribution that won't require marginalization. What we need is a principle to guide our search. What we'll do is start by considering two different ways to describe the world.



(a) World Q. (b) World P.

Figure 1: The graphical models under consideration. $q(D, x, \theta) = q(D)q(x|D)q(\theta|x, D)$ and $p(D, x, \theta) = p(D|\theta)p(x|\theta)p(\theta)$

In World P, the Bayesian world, we assume that we start by drawing a parameter value θ from some prior $p(\theta)$. This parameter value then generates not only the data we observe $p(D|\theta)$ but also generates any future data (our predictions) $p(x|\theta)$. Alternatively, in the real world, World Q, we start by observing data D drawn from some process outside of our control $q(D)$. We want to create a process by which we could use that data to directly make predictions, $q(x|D)$. Without any additional assumptions we allow for θ to depend on *both* of these: $q(\theta|x, D)$.

To learn our variational predictive model, $q(x|D)$, we will align Q and P by minimizing the conditional KL:[2]

$$\left\langle \log \frac{q(x, \theta|D)}{p(x, \theta|D)} \right\rangle = \left\langle \log \frac{q(x|D)q(\theta|x, D)}{p(x|\theta)p(\theta|D)} \right\rangle = \left\langle \log \frac{q(x|D)q(\theta|x, D)}{p(x|\theta)p(D|\theta)p(\theta)} \right\rangle + \log p(D) \geq 0. \quad (3)$$

This establishes the Variational Prediction (VP) loss,

$$\mathcal{J} \equiv \left\langle \log \frac{q(x|D)q(\theta|x, D)}{p(x|\theta)p(D|\theta)p(\theta)} \right\rangle, \quad (4)$$

as a variational upper bound on the negative Bayesian marginal likelihood ($-\log p(D)$).

Furthermore, the joint conditional KL (eq. (3)) is also a variational upper bound on the KL divergence of our approximate predictive $q(x|D)$ to the Bayesian posterior predictive $p(x|D)$:

$$\left\langle \log \frac{q(x, \theta|D)}{p(x, \theta|D)} \right\rangle = \left\langle \log \frac{q(x|D)}{p(x|D)} \right\rangle + \left\langle \log \frac{q(\theta|x, D)}{p(\theta|x, D)} \right\rangle \geq \left\langle \log \frac{q(x|D)}{p(x|D)} \right\rangle \geq 0. \quad (5)$$

Since eq. (3) and eq. (4) differ only by the marginal likelihood, $\log p(D)$, if we fix the Bayesian model, P, this bound (**??**) helps ensure that our approximate-predictive $q(x|D)$ will approximate the true Bayesian posterior predictive. The tightness of this bound is controlled by how well $q(\theta|x, D)$ matches $p(\theta|x, D)$, an *augmented posterior*, i.e. the posterior you

2. All $\langle\cdot\rangle$ brackets in this work are expectations with respect to the full q distribution. $\langle\cdot\rangle \equiv \langle\cdot\rangle_q \equiv \mathbb{E}_q[\cdot]$

would obtain if you had observed an augmented dataset consisting of the original data and an additional observation at x. For some additional, independent insight, see appendix A.

Notice that eq. (5) ensures that our loss is a valid variational bound on the KL divergence between our variational predictive distribution, $q(x|D)$, and the true Bayesian predictive, $p(x|D)$, even if our variational augmented posterior is imperfect, our model or variational predictive is misspecified, or we have a finite dataset.

2.1. Conditioning

Extending the objective to conditional densities (e.g. regression rather than density estimation) is straightforward:

$$\mathcal{J}_{|} = \left\langle \log \frac{q(y|x,D)q(x|D)q(\theta|y,x,D)}{p(y|x,\theta)p(x)p(D|\theta)p(\theta)} \right\rangle_q + \log p(D) \geq \left\langle \log \frac{q(y|x,D)}{p(y|x,D)} \right\rangle_q . \tag{6}$$

This clearly bounds the KL divergence between a variational conditional posterior, $q(y|x,D)$, and the true Bayesian conditional posterior, $p(y|x,D)$. Interestingly, this KL divergence is minimized with respect to the q distribution, namely and importantly this includes a new $q(x|D)$, a distribution used to generate our synthetic input points and under our control. This means that we can choose which points to evaluate our predictive model at during training. In particular, for a classification or regression setting, if we have access to test-set inputs or unlabelled data, we could directly focus our efforts on learning a variational predictive that matched the true Bayesian predictive *on those points*.

2.2. Implicit Variational Augmented Posteriors

Operationally, we follow the procedure illustrated in appendix B. While minimizing the objective we (1) synthesize a new data-point x from our variational predictive distribution $q(x|D)$, (2) compute and sample a parameter value from some variational augmented posterior $\theta \sim q(\theta|x,D)$, and (3) score that synthetic data-point and parameter value according to their sources $q(x|D), q(\theta|x,D)$ and the Bayesian model $p(x|\theta)p(D|\theta)p(\theta)$. In contrast to most other forms of inference, our predictive models outputs are judged during the training process. This allows the VP method to scrutinize the predictive model off the data manifold.

In general, specifying an augmented posterior means building a *conditional* posterior approximation. For large-scale problems this naively creates an new intractable task. Inspired by MAML (Finn et al., 2017), in the toy experiments below, we've defined the augmented posterior implicitly as a single gradient update of an approximate (unconditional) posterior:

$$q(\theta|y,x,D) = q(\theta'|D) \tag{7}$$

$$\theta' = \theta - \lambda \nabla \left\langle -\beta \log p(y|x,\theta) + \log \frac{q(\theta|D)}{p(\theta)} \right\rangle_{q(\theta|D)} . \tag{8}$$

This requires specifying an approximate posterior $q(\theta|D)$ (at the same cost as in variational Bayes), plus two additional parameters: λ, a learning rate for the update, and β, an effective inverse temperature for the ELBO used to update the approximate posterior. This β sets the effective number of observations the synthetic point is considered equivalent to during the posterior update.

3. Toy Example

Let's demonstrate the method on a simple toy example. Consider a two-parameter sinusoidal curve-fitting problem:

$$x \sim \mathcal{U}(0,1) \tag{9}$$

$$y \sim \mathcal{N}(\mu(x), 1) \tag{10}$$

$$\mu(x) = \sin(2\pi f x + \phi)\,. \tag{11}$$

The x values are drawn uniformly on the unit interval. The y values follow a sinsusoidal curve with frequency, f, and phase, ϕ. The observational model is Gaussian with unit variance. For our target distribution we'll set $f = 1, \phi = 1$ and sample 8 data-points to serve as our dataset as shown in blue in fig. 2 below. The predictive distributions are shown in orange.

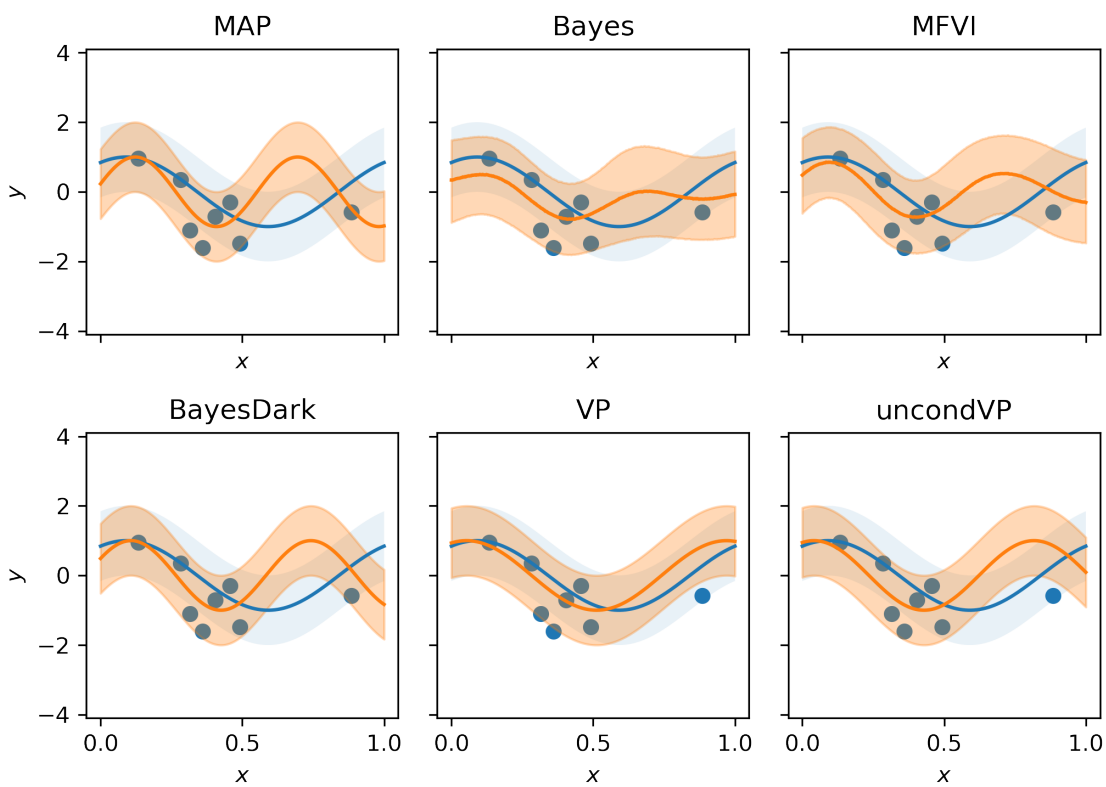


Figure 2: Six different predictive distributions learned for the toy example. In each panel, the true distribution and data are shown in blue, the fit predictive distribution is shown in orange.

We compared six different inference techniques: (1) Maximum a posteriori (MAP) estimation of the prediction distribution, (2) the exact Bayesian posterior predictive (Bayes),

with a $\mathcal{N}(0, 16)$ prior on both $\log f$ and ϕ, (3) the predictive distribution obtained by marginalizing out a Mean Field Variational Inference (MFVI) approximate posterior, using a factorized Gaussian, (4) the Bayesian Dark Knowledge (BayesDark) (Balan et al., 2015) distilled predictive distribution, (5) the proposed Variational Predictive (VP) method, and (6) a baseline version of Variational Prediction that uses an unconditional augmented posterior (uncondVP), i.e. the same approximate posterior used for the other methods.

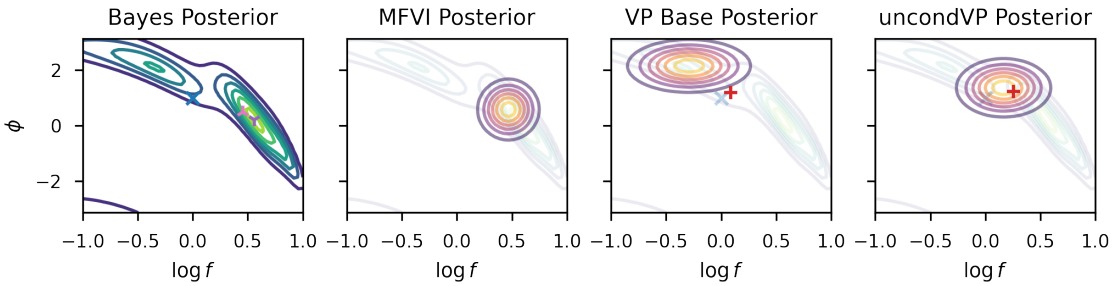


Figure 3: The exact Bayesian posterior for the data shown in blue in fig. 2 is shown on the left. The × marks the true parameter values. The ⋎ marks the MAP parameters, ⋏ marks the BayesDark parameters. The second and last column show the MFVI and uncondVP learned approximate posteriors, respectively. For the VP example, we learned a conditional posterior which is hard to visualize, shown is the base unconditional posterior that is modified with MAML. For the VP and uncondVP solutions the corresponding learned predictive distribution parameters are indicated by +.

In fig. 3 we visualize the solutions in parameter space. Notice that the true Bayesian posterior (left panel) is bimodal and neither mode is particularly well aligned with the true parameters (×). It's corresponding prediction distribution cannot be represented by a single frequency and phase. The MAP predictive distribution (top-left of fig. 2) uses a single point-estimate, the global maximum of the true Bayesian posterior (⋎ in fig. 3). The learned MFVI posterior is shown in the second column of fig. 3. As is usually the case, variational Bayesian inference learns an approximate posterior that aims to minimize the KL divergence from the approximate posterior to the true, which leads to a mode-seeking behavior and the approximate posterior, here, concentrates at the higher frequency mode. This means that the posterior predictive formed when marginalizing out this approximate posterior has predominately the wrong frequency, shown in the top-right of fig. 2.

To learn direct (marginalization-free) variational predictive models, we first replicate the Bayesian Dark Knowledge technique of Balan et al. (2015), which attempts to *distill* the true Bayesian posterior predictive distribution into a parametric $q(y|x, D)$, minimizing $KL[p(y|x, D)|q(y|x, D)]$, the opposite KL compared with the VP objective (eq. (6)). This method requires exact samples from the true Bayesian posterior predictive. For the learned predictive model, $q(y|x, D)$, we simply used the true likelihood $q(y|x, D) = p(y|x, \theta)$ with

learned $\log f$ and ϕ parameters. The resulting predictive model is shown on the bottom left of fig. 2 with the learned parameters denoted with $\curlywedge$ in fig. 3. Notice that these parameters are at the center of the MFVI approximate posterior.

For the conditional VP method (eq. (6)), we used a uniform $\mathcal{U}(0,1)$ distribution for the synthetic x distribution $q(x|D)$. We used a parametric copy of the likelihood for the predictive model, $q(y|x,D)$, as with BayesDark, and used the MAML style augmented posterior (eq. (8)) acting on the factorized Gaussian posterior of the MFVI method.

With these choices in place, the learned predictive model in the bottom-left panel of fig. 2 is a nice fit to the true distribution. We emphasize that this predictive model doesn't require any marginalization and is defined by a point estimate for the parameters of the likelihood. The specific parameter values found are shown by the $+$ in fig. 3, along with the learned base augmented posterior, i.e. the learned approximate posterior that gets updated by the synthetic point. The learned effective-learning-rate was $\lambda = 0.004$ and the learned effective-inverse-temperature was $\beta = 12.8$ in this example.

In all of the other inferential methods, the learned predictive distribution was only ever scored on the data itself, it was never tasked with generating predictions. Meanwhile, in the VP method we are using the predictive model at training time to generate new synthetic data that also must be explained by the Bayesian model. We suspect that it's this enforced sense of in internal consistency that improves the VP predictive model.

To isolate the effect and utility of the MAML-style conditioning in the augmented posterior, lastly we reran the Variational Prediction method but used an unconditional augmented posterior (uncondVP), i.e. the same mean-field approximate posterior that was used for the MFVI experiment. Now, the learned approximate posterior is tasked with explaining not only the original data, but also a single synthetic draw from the predictive model. This isn't equivalent to the MFVI method because the approximate posterior sees the additional point; and it's not equivalent to the VP method because the posterior no longer differs under distinct draws from the predictive model. Removing the conditioning leads to a worse predictive distribution as shown in the bottom right of fig. 2.

4. Conclusion

The toy model illustrates that Variational prediction can learn cheap but good predictive models at little additional cost beyond variational Bayes. However, we've encountered issues trying to scale the VP objective up to larger problems, mainly due to variance in the loss which makes convergence difficult. We hope to resolve these issues in future work, perhaps with techniques such as in Roeder et al. (2017); Grathwohl et al. (2018); Antoniou et al. (2019).

In this paper, we've introduced the VP method. VP is a new inferential technique for learning variational approximations to Bayesian posterior predictive distributions that doesn't require (1) the posterior predictive distribution itself, (2) the exact posterior distribution, (3) exact samples from the posterior, (4) or any test time marginalization. We are excited to see if this method can be shown to be workable on larger-scale problems.

References

Felix V Agakov and David Barber. An auxiliary variational method. In *Neural Information Processing: 11th International Conference, ICONIP 2004, Calcutta, India, November 22-25, 2004. Proceedings 11*, pages 561–566. Springer, 2004.

James Aitchison. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975.

Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml, 2019.

Anoop Korattikara Balan, Vivek Rathod, Kevin P Murphy, and Max Welling. Bayesian dark knowledge. In *Advances in Neural Information Processing Systems*, pages 3438–3446, 2015.

Julian Besag. A candidate's formula: A curious result in bayesian prediction. *Biometrika*, 76(1):183–183, 1989.

Aurélie Boisbunon and Yuzo Maruyama. Inadmissibility of the best equivariant predictive density in the unknown variance case. *Biometrika*, 101(3):733–740, 2014.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

Edward I George, Feng Liang, and Xinyi Xu. Improved minimax predictive densities under kullback-leibler loss. *The Annals of Statistics*, pages 78–91, 2006.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation, 2018.

Feng Liang and Andrew Barron. Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50 (11):2708–2726, 2004.

Geoffrey Roeder, Yuhuai Wu, and David Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference, 2017.

Edward Snelson and Zoubin Ghahramani. Compact approximations to bayesian predictive distributions. In *Proceedings of the 22nd international conference on Machine learning*, pages 840–847, 2005.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Appendix A. Candidate's Formula

In a single-page paper, Besag (1989) shares a curious result that appeared on a candidate's final exam:

$$p(x|D) = \frac{p(x|\theta)p(\theta|D)}{p(\theta|x, D)}. \tag{12}$$

Remarkably the right hand side holds for any θ if the p's are all taken to be consistent with some Bayesian model $p(x|\theta)p(\theta)$. The Bayesian posterior predictive distribution can be found by relating the likelihood, posterior and an *augmented posterior*, namely the posterior you would compute if you observed not only the data (D), but also the prediction x as an additional datapoint.

The proof is straightforward: simply factorize the joint distribution $p(x, \theta, D)$ out two ways: $p(\theta, x, D) = p(\theta|x, D)p(x|D) = p(x|\theta)p(\theta|D)$ and notice that because of the Markov chain $D \to \theta \to x$ we have that $p(x|\theta, D) = p(x|\theta)$. ■

As this demonstrates, if we knew the exact augmented posterior, we would be able to determine the exact posterior predictive without needing to explicitly marginalize. Our variational approach instead steers a variational approximation to the posterior predictive by using a variational approximation to the augmented posterior.

Appendix B. Pseudo-code

```python
def variational_prediction(predictive_model, aug_posterior, likelihood, prior, data):
    # first we sample from the predictive model
    x = predictive_model.sample()
    # then we compute the approximate posterior induced by that sample
    approx_posterior = aug_posterior(x)
    # now we proceed as we would normally with and ELBO by sampling a parameter value
    theta = approx_posterior.sample()
    # and then computing all of the relevant log probabilities.
    return (predictive_model.log_prob(x) + approx_posterior.log_prob(theta)
            - likelihood.log_prob(x) - likelihood.log_prob(data) - prior.log_prob(theta))
```

Listing 1: A (pseudo-)python implementation of the variational prediction loss.

Appendix C. Related Work

Aitchison (1975) introduced the distinction between *estimative* procedures that use point estimates of parameters and *predictive* approaches that involved marginalizing over parameter distributions. In that work, Aitchison showed that the Bayesian posterior predictive distribution is optimal in terms of the average Kullback-Leibler divergence from the true distribution as well as how some predictive distributions can always outperform standard estimative approaches like Maximum Likelihood by this same metric.

Follow-up work (George et al., 2006; Liang and Barron, 2004; Boisbunon and Maruyama, 2014) have extended and further analyzed these types of scenarios. However, in these cases the predictive density generated requires the form of marginalization we aimed to avoid in this work. Here we are interested in defining a type of *estimative* procedure that can claim

to more directly target predictive performance. We did so by use of an auxillary variational method to generate our bounds, akin to Agakov and Barber (2004).

Both Snelson and Ghahramani (2005) and Balan et al. (2015) attempt to learn compact *estimative* representations of the Bayesian posterior predictive, though targeting the opposite KL divergence as done here. They aim to learn a model for the predictive $q(x|D)$ that is as close as possible to the true posterior predictive $p(x|D)$ as measured by $\left\langle \log \frac{p(x|D)}{q(x|D)} \right\rangle_{p(x|D)}$. This requires being able to generate samples from the true Bayesian posterior predictive. In Balan et al. (2015), Stochastic Gradient Langevin Dynamics (Welling and Teh, 2011) was used to generate these samples. These samples were then distilled into a compact distribution. In our work we don't need to presuppose we can generate exact samples from the true Bayesian posterior and have more freedom to leverage tractable variational approximations to the augmented posterior while maintaining our bounds.