Proceedings of Machine Learning Research 96:1–10, 2018. 1st Symposium on Advances in Approximate Bayesian Inference

An Exploration of Acquisition and Mean Functions in Variational Bayesian Monte Carlo

Luigi Acerbi[*] LUIGI.ACERBI@UNIGE.CH
University of Geneva, CMU, 1 rue Michel-Servet, 1206 Genève, Switzerland

Abstract

Variational Bayesian Monte Carlo (VBMC) is a novel framework for tackling approximate posterior and model inference in models with *black-box*, expensive likelihoods by means of a sample-efficient approach (Acerbi, 2018). VBMC combines variational inference with Gaussian-process (GP) based, active-sampling Bayesian quadrature, using the latter to efficiently approximate the intractable integral in the variational objective. VBMC has been shown to outperform state-of-the-art inference methods for expensive likelihoods on a benchmark consisting of meaningful synthetic densities and a real model-fitting problem from computational neuroscience. In this paper, we study the performance of VBMC under variations of two key components of the framework. First, we propose and evaluate a new general family of acquisition functions for active sampling, which includes as special cases the acquisition functions used in the original work. Second, we test different mean functions for the GP surrogate, including a novel *squared-exponential* GP mean function. From our empirical study, we derive insights about the stability of the current VBMC algorithm, which may help inform future theoretical and applied developments of the method.

Keywords: Bayesian quadrature; *black-box* inference; variational inference

1. Introduction

Many models in the computational sciences, in engineering, and machine learning are characterized by *black-box* expensive likelihoods. The research for active, sample-efficient methods to *optimize* such models by means of statistical surrogates — e.g., Gaussian processes (GPs; Rasmussen and Williams, 2006) — has been extremely succesful, spawning the field of Bayesian optimization (Jones et al., 1998; Brochu et al., 2010; Snoek et al., 2012; Shahriari et al., 2016; Acerbi and Ma, 2017). Despite the outstanding successes of GP-based surrogate modeling for optimization, a suprisingly few works have adopted a similar approach for the harder problem of full (approximate) *Bayesian inference*, which entails: (a) reconstructing the full posterior distribution (Kandasamy et al., 2015; Wang and Li, 2018); (b) computing the marginal likelihood, a key metric for model selection (Ghahramani and Rasmussen, 2002; Osborne et al., 2012; Gunter et al., 2014; Briol et al., 2015). To these ends, we recently proposed Variational Bayesian Monte Carlo (VBMC), an approximate inference framework that, by combining variational inference and Bayesian quadrature, efficiently computes both an approximate posterior and an estimate of the *evidence lower bound* (ELBO), a lower bound on the marginal likelihood (Acerbi, 2018). VBMC outperformed state-of-the-art inference algorithms for expensive likelihoods on a benchmark that

[*] Website: luigiacerbi.com. Alternative e-mail: luigi.acerbi@gmail.com.

includes synthetic likelihoods with realistic, challenging properties, and a real model-fitting problem from computational neuroscience (Acerbi, 2018).

The VBMC framework includes several algorithmic features which were mostly fixed in the original paper, and deserve further exploration. Key components of VBMC include:

1. The *acquisition function* $a(\boldsymbol{x})$ used in active sampling. The surrogate optimization of $a(\boldsymbol{x})$ decides which point $\boldsymbol{x} \in \mathcal{X}$ of the expensive likelihood is queried next, where $\mathcal{X} \subseteq \mathbb{R}^D$ is the domain of model parameters (we tested up to $D = 10$). The acquisition function embodies the crucial role of balancing exploration vs. exploitation.

2. The GP model. While GP covariance and likelihood functions are almost fully determined by the desire to have an analytical expression for the surrogate ELBO, there is some freedom in the design of the GP mean function under this constraint.

In this paper, we perform an empirical evaluation of variants of these two main features of the VBMC algorithm. First, we recap in Section 2 the main formulation of VBMC. In Section 3.1, we introduce a novel family of acquisition functions, which includes as specific cases the two acquisition functions described in the original paper. In Section 3.2, we introduce a novel GP mean function. We then report in Section 4 results of our experiments with different acquisition and mean functions. Finally, we discuss in Section 5 our findings and further extensions of the framework.

Code for the VBMC algorithm is available at: `https://github.com/lacerbi/vbmc`.

2. Variational Bayesian Monte Carlo (VBMC)

We summarize here the main features of VBMC; see Acerbi (2018) for details. Let $f = p(\mathcal{D}|\boldsymbol{x})p(\boldsymbol{x})$ be the expensive *target* log joint probability (unnormalized posterior), where $p(\mathcal{D}|\boldsymbol{x})$ is the model likelihood for dataset $\mathcal{D}$ and parameter vector $\boldsymbol{x}$, and $p(\boldsymbol{x})$ the prior.

In each iteration t, the algorithm: (1) actively samples sequentially a batch of n_{active} 'promising' new points that maximize a given acquisition function, and for each selected point $\boldsymbol{x}^*$ evaluates the target $\boldsymbol{y}^* \equiv f(\boldsymbol{x}^*)$;[1] (2) trains a GP surrogate model of the log joint f, given the training set $\boldsymbol{\Xi}_t = \{\mathbf{X}_t, \boldsymbol{y}_t\}$ of points and their associated observed values so far; (3) updates the variational posterior approximation, indexed by $\boldsymbol{\phi}_t$, by optimizing the surrogate ELBO. This loop repeats until reaching a termination criterion (e.g., budget of function evaluations). We use $n_{\text{active}} = 5$, as in Acerbi (2018). VBMC includes an initial *warm-up* stage to converge faster to regions of high posterior probability (see Acerbi, 2018).

Variational Posterior The variational posterior is a flexible mixture of K Gaussians, $q(\boldsymbol{x}) \equiv q_{\boldsymbol{\phi}}(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{\Sigma}\right)$, where w_k, $\boldsymbol{\mu}_k$, and σ_k are, respectively, the mixture weight, mean, and scale of the k-th component; $\boldsymbol{\Sigma}$ is a common diagonal covariance matrix $\boldsymbol{\Sigma} \equiv \text{diag}[{\lambda^{(1)}}^2, \ldots, {\lambda^{(D)}}^2]$; and the number of components K is set adaptively. The vector $\boldsymbol{\phi}$ summarizes all variational parameters.

Gaussian Process Approximation In VBMC, the log joint f is approximated by a GP with a squared exponential (rescaled Gaussian) kernel, a Gaussian likelihood with small observation noise (for numerical stability), and a *negative quadratic* mean function (see

[1] When possible, we apply a rank-1 update of the current GP posterior after each new evaluation.

Section 3.2). Initially, the GP hyperparameters are estimated via MCMC sampling (Neal, 2003); marginalization over the GP hyperparameter posterior is crucial to properly represent model uncertainty, a key element of active sampling. Training of the GP model switches to gradient-based optimization when the contribution of the variance of the ELBO due to sampling decreases below a given threshold, suggesting that the posterior over hyperparameters is reasonably summarized by a point estimate (see Acerbi, 2018 for details).

The Evidence Lower Bound (ELBO) Using the GP surrogate f, and for a given variational posterior $q_{\boldsymbol{\phi}}$, we can estimate the posterior mean of the surrogate ELBO as

$$\mathbb{E}_{f|\boldsymbol{\Xi}}\left[\text{ELBO}(\boldsymbol{\phi})\right] = \mathbb{E}_{f|\boldsymbol{\Xi}}\left[\mathbb{E}_{\boldsymbol{\phi}}\left[f\right]\right] + \mathcal{H}[q_{\boldsymbol{\phi}}], \tag{1}$$

where $\mathbb{E}_{f|\boldsymbol{\Xi}}\left[\mathbb{E}_{\boldsymbol{\phi}}\left[f\right]\right]$ is the (expected) expected log joint, and $\mathcal{H}[q_{\boldsymbol{\phi}}]$ is the entropy of the variational posterior. Crucially, our choice of variational family and of GP representation affords an analytical computation of the posterior mean and variance of the expected log joint (and of their gradients) by means of Bayesian quadrature (BQ; O'Hagan, 1991; Ghahramani and Rasmussen, 2002; see also Appendix A). Entropy and its gradient are estimated via simple Monte Carlo and the reparameterization trick (Kingma and Welling, 2013; Miller et al., 2017), such that Equation (1) is amenable to stochastic optimization (Kingma and Ba, 2014).

3. Exploring the Components of VBMC

3.1. Acquisition Functions

In principle, VBMC needs to solve a complex sequential decision-making problem which consists of evaluating the expensive log joint f at a sequence of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ such that the approximate posterior $q_{\boldsymbol{\phi}}$ converges as closely as possible to the ground truth, for a given budget of function evaluations. In practice, such problem is intractable and we instead adopt a heuristic — the acquisition function — which, based on our current model of the log joint, grades which points are more advantageous to evaluate next. The ideal acquisition function for VBMC should balance exploitation of known regions of high probability mass (so as to refine our approximation) and exploration of uncertain regions (which might contain yet undiscovered amounts of probability mass). While we generally want to find probability *mass*, for convenience we will use probability *density* as a proxy.

We introduce here a novel family of *generalized uncertainty sampling* (GUS) acquisition functions,

$$a_{\text{gus}}(\boldsymbol{x}) = V_{\boldsymbol{\Xi}}^{\alpha}(\boldsymbol{x}) q_{\boldsymbol{\phi}}^{\beta}(\boldsymbol{x}) \exp\left(\gamma \overline{f}_{\boldsymbol{\Xi}}(\boldsymbol{x})\right), \qquad \alpha, \beta, \gamma \geq 0, \tag{2}$$

where $\overline{f}_{\boldsymbol{\Xi}}(\boldsymbol{x})$ and $V_{\boldsymbol{\Xi}}(\boldsymbol{x})$ are, respectively, the GP posterior predictive mean and variance at $\boldsymbol{x}$ given the current training set $\boldsymbol{\Xi}$, and $q_{\boldsymbol{\phi}}$ is the variational posterior. Equation (2) is a generalization of the expression of the variance of the *integrand* involving the log joint in Equation (1).

For $\alpha = 1$, Equation (2) with $\beta = 2, \gamma = 0$ is equivalent to *vanilla uncertainty sampling*, whereas with $\beta = 1, \gamma = 1$ we obtain *prospective uncertainty sampling*, as described in Acerbi (2018). Here, we also consider the case $\beta = 0, \gamma = 2$, which ignores the current variational posterior and performs full *GP-uncertainty sampling*. By increasing α, we increase the focus on exploration (regions of high uncertainty) vs. exploitation (regions of high posterior probability). A particularly interesting option is to make α iteration-dependent, motivated by acquisition functions such as UCB (Srinivas et al., 2010). Here, we consider $\beta, \gamma = 1$, with $\alpha(n) = \max(1, \log n)$ (*logarithmic*) and $\alpha(n) = \sqrt{n}$ (*square root*), where n is the number of points in the training set. Note that Equation (2) can be reduced from 3 to 2 parameters with virtually no loss of generality (see Appendix B).

3.2. GP Mean Functions

In VBMC, the GP (prior) mean function implicitly affects exploration vs. exploitation by setting the value of the GP posterior mean far away from points in the current training set.

In our prior work, we argued that a *negative quadratic* function is theoretically preferable to *constant* because it ensures that the posterior GP predictive mean $\overline{f}$ is a proper log probability distribution (that is, it is integrable when exponentiated; Acerbi, 2018). On the other hand, a mean function that decreases too quickly may curb exploration outside the training set. As an intermediate alternative, we introduce here the *squared exponential* GP mean function,

$$m_{\text{SE}}(\boldsymbol{x}) = m_0 + h \exp\left[-\frac{1}{2} Q(\boldsymbol{x})\right], \qquad \text{with} \;\; Q(\boldsymbol{x}) \equiv \sum_{i=1}^{D} \frac{\left(x^{(i)} - x_{\text{m}}^{(i)}\right)^2}{{\omega^{(i)}}^2}, \tag{3}$$

where m_0 is a constant offset, h the height of the 'bump', and $\boldsymbol{x}_{\text{m}}$ and $\boldsymbol{\omega}$ are vectors of, respectively, location and scale parameters. For comparison, the standard *negative quadratic* GP mean function for VBMC is $m_{\text{NQ}}(\boldsymbol{x}) = m_0 - \frac{1}{2}Q(\boldsymbol{x})$, and a typical mean function for GP regression is *constant*, $m_{\text{CN}}(\boldsymbol{x}) = m_0$.

The interpretation of m_{NQ}, once exponentiated, is that of a global multivariate normal approximation with diagonal covariance (e.g., similar to an axis-aligned Laplace approximation), whereas m_{SE} can be thought of as a *locally* Gaussian approximation (near the maximum), which becomes constant asymptotically. In any case, note that the (prior) mean function does not constrain the shape of the GP — that is, the *posterior* GP mean may well be multimodal and non-axis aligned. The role of the GP mean function is mostly in dictating the GP behavior far from observed points.

Crucially, all the considered mean functions afford analytical expressions for the expected log joint in Equation (1), by means of Bayesian quadrature (see Appendix A). Note that, of these functions, only m_{NQ} leads to a proper posterior distribution; but whether this property matters in practice for the algorithm remains an empirical question.

4. Experiments

Procedure We tested variants of VBMC to perform inference of the posterior distribution and model evidence on the following families of problems:

1. Three families of synthetic target likelihoods, for $D \in \{2, 6, 10\}$. *Lumpy*: mildly multimodal distributions obtained as clumped mixtures of twelve multivariate Gaussians; *Student*: heavy-tailed, multivariate Student's t distributions; *Cigar*: single multivariate Gaussians with highly correlated covariance matrix.

2. A real model-fitting problem from computational neuroscience, with two posterior densities computed from a complex model of neuronal orientation selectivity in visual cortex, applied to neural recordings of, respectively, one V1 and one V2 cell ($D = 7$; data and model from Goris et al., 2015).

See Acerbi (2018) for an extended description of the tested target distributions.

We evaluated inference performance by tracking (a) the absolute error between the ELBO and the true log marginal likelihood (LML), and (b) the "Gaussianized" symmetrized Kullback-Leibler divergence (gsKL) between approximate posterior and ground truth.

a. We measure the model evidence approximation in terms of absolute error from ground truth, since differences of LML are used for model comparison. For reference, differences of LML of 10 points or more are often presented as *decisive* evidence in favor of one model (Kass and Raftery, 1995), while errors $\ll 1$ can be considered negligible.

b. The gsKL, introduced in Acerbi (2018), is defined as the symmetrized KL divergence between two multivariate normal distributions with mean and covariances equal, respectively, to the moments of the approximate posterior and the moments of the true posterior. For reference, two Gaussians with unit variance and whose means differ by $\sqrt{2}$ (resp., $\frac{1}{2}$) have a gsKL of 1 (resp., $\frac{1}{8}$).

As a rule of thumb, for both metrics we consider a solution "usable" if it is at least less than 1 (ideally, much lower than that). For synthetic problems, we evaluated ground-truth posteriors and model evidence analytically or via simple numerical integration; for real model-fitting problems, we employed extensive MCMC sampling (Acerbi, 2018).

For each VBMC variant we performed at least 20 runs per inference problem, with randomized starting points, and for each performance metric we report the median and 95% CI of the median (bootstrapped). For each problem, we allow a budget of $50 \times (D+2)$ likelihood evaluations. For more details on the benchmark procedure, see Acerbi (2018).

Algorithms In this paper, we focus on comparing different versions of the VBMC algorithm (see Acerbi, 2018 for a comparison between VBMC and several other inference algorithms). By default, VBMC uses the a_{pro} acquisition function and m_{NQ} GP mean function. We show here results for the following variants of the VBMC algorithm:

1. Different acquisition functions: vanilla uncertainty sampling (a_{us}); GP-uncertainty sampling (a_{gpus}); iteration-dependent logarithmic uncertainty sampling ($a_{\ln}$) and square-root uncertainty sampling (a_{sqrt}).

2. Different GP mean functions: constant (m_{CN}); squared exponential (m_{SE}).

Results for synthetic likelihoods are shown in Figure 1, for the neuronal model in Figure 2. In Figure 1, a_{us} performs almost identically to a_{pro}, such that the plots for these two acquisition functions are overlapping almost everywhere.

Acquisition Functions For the GUS acquisition function, described in Equation (2), we consider the following parameter settings: $\alpha = 1, \beta = 2, \gamma = 0$ (a_{us}); $\alpha = 1, \beta = 1, \gamma = 1$ (a_{pro}); $\alpha = 1, \beta = 0, \gamma = 2$ (a_{gpus}); $\alpha(n) = \ln n, \beta = 1, \gamma = 1$ ($a_{\ln}$); $\alpha(n) = \sqrt{n}, \beta = 1, \gamma = 1$

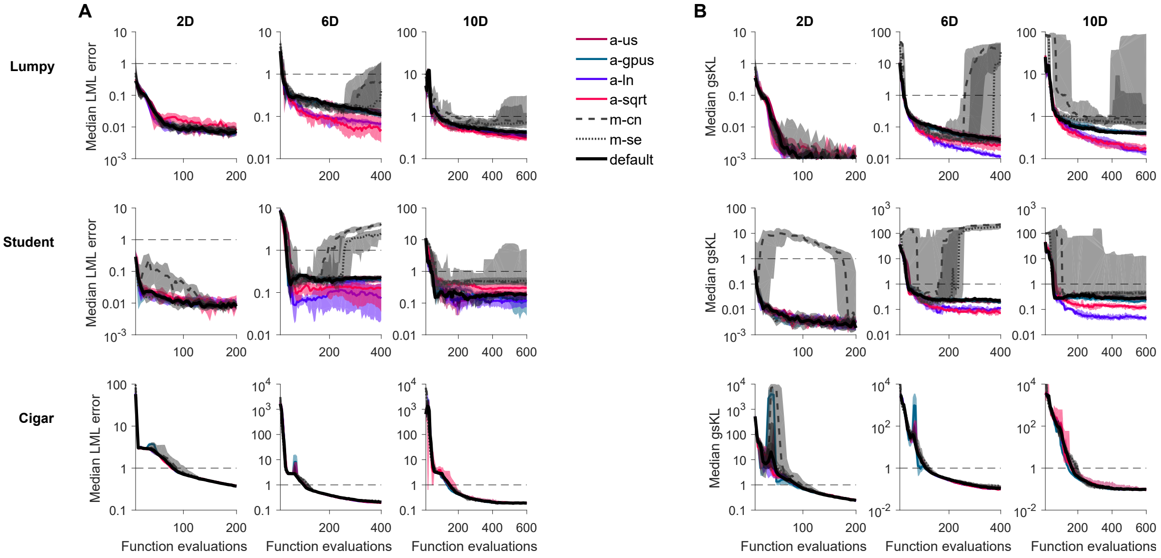

Figure 1: **Synthetic likelihoods. A.** Median absolute difference between the ELBO and true log marginal likelihood (LML), as a function of likelihood evaluations, on the *lumpy* (top), *Student* (middle), and *cigar* (bottom) problems, for $D \in \{2, 6, 10\}$ (columns). **B.** Median "Gaussianized" symmetrized KL divergence between the variational posterior and ground truth. For both metrics, shaded areas are 95 % CI of the median, and we consider a desirable threshold to be < 1 (dashed line).

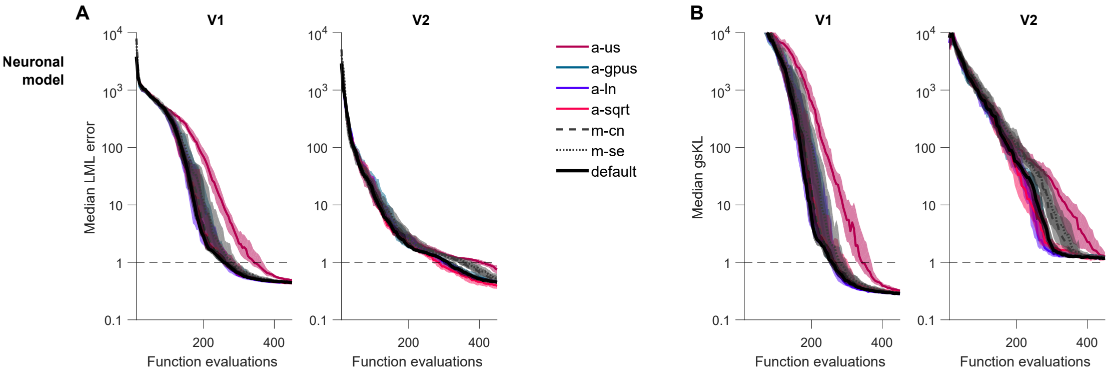

Figure 2: **Real neuronal model likelihoods. A.** Median absolute difference between the ELBO and true LML, as a function of likelihood evaluations, for two distinct neurons ($D = 7$). **B.** Median "Gaussianized" symmetrized KL divergence between the variational posterior and ground truth. See also Figure 1.

(a_{sqrt}).[2] Somewhat surprisingly, the performance of VBMC in our benchmark is quite robust across parameters of the generalized uncertainty sampling acquisition function. The only notable results are that: on real data (but not on synthetic functions) a_{us} performs

[2] a_{us} and a_{pro} were introduced and tested in Acerbi (2018); we report them here for comparison.

substantially worse than the other choices; on some synthetic functions (but not on real data), a_{ln} and less so a_{sqrt} perform marginally better than the rest. More challenging benchmark densities may be able to reveal larger differences in the performance of various acquisition functions, but for now our original recommendation of using a_{pro} still holds.

GP Mean Functions Having fixed the acquisition function to a_{pro}, we tested two additional GP mean functions, m_{CN} and m_{SE}. On several problems, both variants perform worse than the originally proposed m_{NQ}. In particular, we observe that the variational posterior becomes unstable when it finds a solution with an "infinitely flat" mixture component — the reason being that the GP posterior mean tends to a small nonzero value far away from the current training set (that is, the exponentiated GP is not a proper, integrable probability density). Heuristic solutions, such as bounding the scaling factor of each variational component (preventing them from "exploding" to infinity), allow the algorithm to run, but the presence of these runaway components still negatively affect performance. Thus, the negative quadratic GP mean function introduced (somewhat understatedly) in Acerbi (2018) is a crucial component for the success and stability of the algorithm.

5. Discussion

We investigated the performance of VBMC under different acquisition functions belonging to the generalized uncertainty sampling (GUS) family, and different GP mean functions compatible with Bayesian quadrature.

On the one hand, our findings could appear as a 'null result' in that for none of the investigated features we obtained a systematic improvement over our original choices for the VBMC algorithm (except perhaps for sporadic improvements with the iteration-dependent a_{ln}). On the other hand, this work provides empirical validation for seemingly arbitrary choices in the original paper, now justified by showing that either (1) the algorithm is fairly robust to changes in the details of the feature (i.e., parameters of GUS), or (2) the original choice is best among a few reasonable alternatives for both empirical and theoretical reasons (i.e., only the negative quadratic GP mean function, m_{NQ}, realizes a *proper* posterior distribution, required for stability).

Alternative GP Representations Specifically with respect to properties of the GP surrogate used for VBMC, two main questions remain open.

First, we can ask if there are more complex covariance or mean functions of interest that are still amenable to closed-form Bayesian quadrature (see Appendix A). For example, we could include a quadratic form directly in the covariance function, as opposed to the mean function. Another interesting direction is to consider covariance or mean functions that are linear combinations and products of location-dependent radial basis functions, which could be used to introduce non-stationary behavior in the GP (Martinez-Cantin, 2018).

Second, we note that the Gaussian kernel, due to its smoothness, may be inadequate to model some posterior distributions. More in general, there may be problems for which we want to use covariance or mean functions which do not support a closed-form expression for the surrogate ELBO. In these cases, it might be worth dropping the analytical requirement, and approximate the expected log joint and its gradient numerically. Investigating VBMC under a more general class of representations remains avenue for future work.

References

Luigi Acerbi. Variational Bayesian Monte Carlo. *Advances in Neural Information Processing Systems*, 31:8222–8232, 2018.

Luigi Acerbi and Wei Ji Ma. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Advances in Neural Information Processing Systems*, 30:1834–1844, 2017.

François-Xavier Briol, Chris Oates, Mark Girolami, and Michael A Osborne. Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. *Advances in Neural Information Processing Systems*, 28:1162–1170, 2015.

Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

Zoubin Ghahramani and Carl E Rasmussen. Bayesian Monte Carlo. *Advances in Neural Information Processing Systems*, 15:505–512, 2002.

Robbe LT Goris, Eero P Simoncelli, and J Anthony Movshon. Origin and function of tuning diversity in macaque visual cortex. *Neuron*, 88(4):819–831, 2015.

Tom Gunter, Michael A Osborne, Roman Garnett, Philipp Hennig, and Stephen J Roberts. Sampling for inference in probabilistic models with fast Bayesian quadrature. *Advances in Neural Information Processing Systems*, 27:2789–2797, 2014.

Nikolaus Hansen, Sibylle D Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18, 2003.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

Kirthevasan Kandasamy, Jeff Schneider, and Barnabás Póczos. Bayesian active learning for posterior estimation. *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations*, 2013.

Ruben Martinez-Cantin. Funneled Bayesian optimization for design, tuning and control of autonomous systems. *IEEE Transactions on Cybernetics*, (99):1–12, 2018.

Andrew C Miller, Nicholas Foti, and Ryan P Adams. Variational boosting: Iteratively refining posterior approximations. *Proceedings of the 34th International Conference on Machine Learning*, 70:2420–2429, 2017.

Radford M Neal. Slice sampling. *Annals of Statistics*, 31(3):705–741, 2003.

Anthony O'Hagan. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29 (3):245–260, 1991.

Michael Osborne, David K Duvenaud, Roman Garnett, Carl E Rasmussen, Stephen J Roberts, and Zoubin Ghahramani. Active learning of model evidence using Bayesian quadrature. *Advances in Neural Information Processing Systems*, 25:46–54, 2012.

C. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2016.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25:2951–2959, 2012.

Niranjan Srinivas, Andreas Krause, Matthias Seeger, and Sham M Kakade. Gaussian process optimization in the bandit setting: No regret and experimental design. *ICML-10*, pages 1015–1022, 2010.

Hongqiao Wang and Jinglai Li. Adaptive Gaussian process approximation for Bayesian inference with expensive likelihood functions. *Neural Computation*, pages 1–23, 2018.

Appendix A. Expected Log Joint via Bayesian Quadrature

An interesting question is which covariance and mean functions afford an analytical computation of the expected log joint in Equation (1). For a given variational posterior $q_{\boldsymbol{\phi}}$ represented by a Gaussian mixture model, as per Section 2, the expected log joint is

$$\mathbb{E}_{\boldsymbol{\phi}}\left[f(\boldsymbol{x})\right] = \sum_{k=1}^{K} w_k \int \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{\Sigma}\right) f(\boldsymbol{x}) d\boldsymbol{x} \equiv \sum_{k=1}^{K} w_k \mathcal{I}_k. \tag{S1}$$

We recall that the posterior predictive mean of a GP f, given training data $\boldsymbol{\Xi} = \{\mathbf{X}, \boldsymbol{y}\}$, where $\mathbf{X}$ are n training inputs with observed values $\boldsymbol{y}$, is (Rasmussen and Williams, 2006)

$$\overline{f}(\boldsymbol{x}) = \kappa(\boldsymbol{x}, \mathbf{X}) \left[\kappa(\mathbf{X}, \mathbf{X}) + \sigma_{\text{obs}}^2 \mathbf{I}_n\right]^{-1} (\boldsymbol{y} - m(\mathbf{X})) + m(\boldsymbol{x}), \tag{S2}$$

where $\kappa(\cdot, \cdot)$ and $m(\cdot)$ are, respectively, the GP covariance and mean functions. Thus, for each integral in Eq. S1, we have in expectation over the GP posterior (Acerbi, 2018)

$$\begin{aligned}\mathbb{E}_{f|\boldsymbol{\Xi}}\left[\mathcal{I}_k\right] &= \int \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{\Sigma}\right) \overline{f}(\boldsymbol{x}) d\boldsymbol{x} \\ &= \left[\sigma_f^2 \int \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{\Sigma}\right) \kappa\left(\boldsymbol{x}, \mathbf{X}\right) d\boldsymbol{x}\right] \left[\kappa(\mathbf{X}, \mathbf{X}) + \sigma_{\text{obs}}^2 \mathbf{I}\right]^{-1} (\boldsymbol{y} - m(\mathbf{X})) \\ &\quad + \sigma_f^2 \int \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{\Sigma}\right) m\left(\boldsymbol{x}\right) d\boldsymbol{x}.\end{aligned} \tag{S3}$$

From Equation (S3), we see that functional forms for the covariance and mean that would afford an analytical calculation of the integrals are Gaussian, exponential, polynomial, and products and linear combinations of such elementary forms (Ghahramani and Rasmussen, 2002). We are not aware of other general functional forms that could be meaningfully used in this context.

Appendix B. Reduced Formulation of Generalized Acquisition Function

We show here that the generalized acquisition function described by Equation (2) can be reduced from three to two parameters with virtually no loss of generality.

First, the location of the optimum of a function is invariant to monotonic[3] transformations of the output, and moreover in VBMC we optimize the acquisition function using CMA-ES (Hansen et al., 2003), which only uses the ranking of the objective function — making it invariant to monotonic transformation of the objective. Thus, we can apply a monotonic transformation to the acquisition function with absolutely no change to the entire optimization process. Second, we assume that for any "uncertainty sampling" acquisition function we want to keep dependence on the GP posterior predictive variance, that is $\alpha > 0$.

With these considerations, we can rewrite Equation (2) as

$$\log a_{\text{gus}}(\boldsymbol{x}) \propto \log V_{\boldsymbol{\Xi}}(\boldsymbol{x}) + \widetilde{\beta} \log q_{\boldsymbol{\phi}}(\boldsymbol{x}) + \widetilde{\gamma} \overline{f}_{\boldsymbol{\Xi}}(\boldsymbol{x}), \qquad \text{with } \widetilde{\beta} = \frac{\beta}{\alpha}, \widetilde{\gamma} = \frac{\gamma}{\alpha}. \tag{S4}$$

which only depends on two parameters, and the logarithmic form is numerically convenient to avoid overflows.

[3] In all this paragraph, we mean monotonic with positive derivative.