1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–6

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.

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

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

The VBMC framework includes several algorithmic features which we mostly fixed in our 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 obtain analytical expressions for the ELBO (see Section 2), 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, and we discuss their meaning for the framework.

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) sequentially samples a batch of 'promising' new points that maximize a given acquisition function, and evaluates the target f at each of them; (2) trains a GP 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 ELBO. This loop repeats until reaching a termination criterion (e.g., budget of function evaluations).

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 $\boldsymbol{\Sigma} \equiv \mathrm{diag}[{\lambda^{(1)}}^2, \ldots, {\lambda^{(D)}}^2]$; and the number of components K is set adaptively. The vector $\boldsymbol{\phi}$ summarizes all variational parameters.

The Evidence Lower Bound (ELBO) In VBMC, the log joint f is approximated by a GP with a squared exponential (rescaled Gaussian) kernel, a Gaussian likelihood, and a *negative quadratic* mean function (see below). The ELBO is then estimated as

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

where $\mathcal{H}[q_{\boldsymbol{\phi}}]$ is the entropy of the variational posterior $q_{\boldsymbol{\phi}}$. 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). 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

The ideal acquisition function for VBMC has to balance exploration of uncertain regions and exploitation of regions with high probability mass to ensure a fast convergence of the variational posterior as closely as possible to the ground truth. 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. 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 Section A in the Appendix).

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 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 the *squared exponential* GP mean function,

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

where m_0 is a constant offset, $\boldsymbol{x}_{\text{m}}$ is the location of the maximum, h the height of the 'bump', and $\boldsymbol{\omega}$ is a vector of length scales. 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*, $m_{\text{CN}}(\boldsymbol{x}) = m_0$. Crucially, all these mean functions afford analytical expressions for the expected log joint in Equation (1), by means of Bayesian quadrature.

4. Results

Procedure We tested variants of VBMC to perform inference on (1) three families of synthetic target likelihoods, for $D \in \{2, 6, 10\}$ (*lumpy*: mildly multimodal distributions;

Student: heavy-tailed distributions; *cigar*: distributions with highly-correlated parameters); and (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 one V1 and one V2 cell ($D = 7$; Goris et al., 2015). We evaluated performance by tracking the absolute error between the ELBO and the true log marginal likelihood (LML), and the "Gaussianized" symmetrized KL divergence (gsKL) between approximate posterior and ground truth.[1] For both metrics, a usable solution is expected to be (much) less than 1. For each problem, we allow a budget of $50 \times (D + 2)$ likelihood evaluations. For more details on the benchmark procedure, see Acerbi (2018). Plots with results are reported in Section B of the Appendix.

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}); $\alpha(n) = \ln n, \beta = 1, \gamma = 1$ ($a_{\ln}$); $\alpha(n) = \sqrt{n}, \beta = 1, \gamma = 1$ (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 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, 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). 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.

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).

By combining variational inference and Bayesian quadrature, the VBMC framework yields great promise for both theoretical and applied advances in approximate inference.

1. The 'Gaussianized' KL-divergence only considers differences in mean and covariance matrix.
2. a_{us} and a_{pro} were introduced and tested in Acerbi (2018); we report them here for comparison.

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.

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.

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.

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.

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. 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

$$\log a_{\text{gus}}(\boldsymbol{x}) \propto \log V_{\boldsymbol{\Xi}}(\boldsymbol{x}) + \widetilde{\beta} \log q_{\boldsymbol{\phi}}(\boldsymbol{x}) + \widetilde{\gamma} \overline{f}_{\boldsymbol{\Xi}}(\boldsymbol{x}), \qquad \text{with } \widetilde{\beta} = \frac{\beta}{\alpha}, \widetilde{\gamma} = \frac{\gamma}{\alpha}. \tag{4}$$

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

Appendix B. Supplementary Results

By default, VBMC uses the a_{pro} acquisition function and m_{NQ} GP mean function. We show here the results for several 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}).

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 (obtained via bootstrap). 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. For a comparison between VBMC and several other inference algorithms, see Acerbi (2018).

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

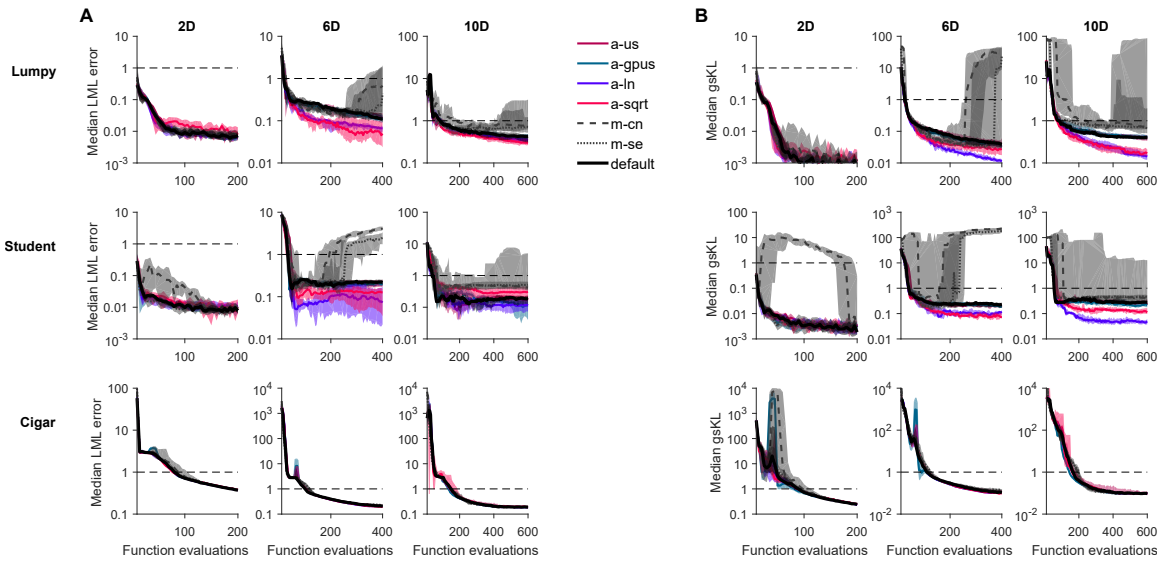

Figure 1: **Synthetic likelihoods. A.** Median absolute difference between the ELBO and true log marginal likelihood (LML), as a function of likelihood evaluations, on the *lumpy* (top), *Student* (middle), and *cigar* (bottom) problems, for $D \in \{2, 6, 10\}$ (columns). **B.** Median "Gaussianized" 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 $\lesssim 1$ (dashed line).

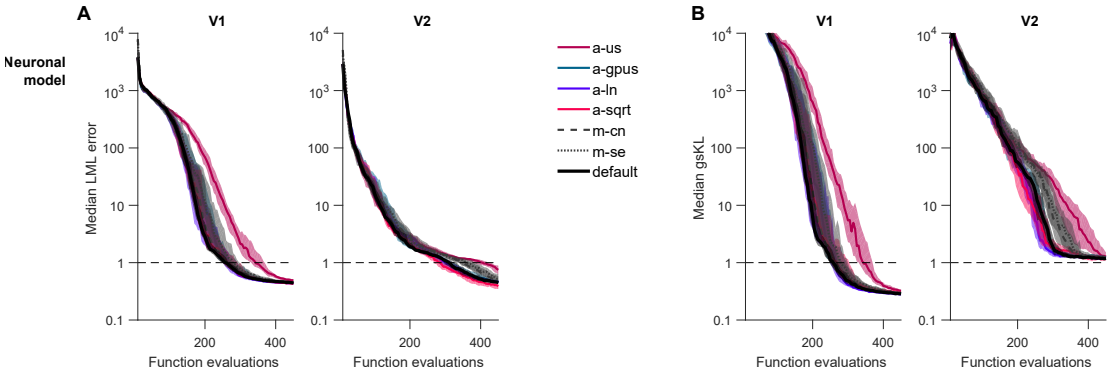

Figure 2: **Real neuronal model likelihoods. A.** Median absolute difference between the ELBO and true LML, as a function of likelihood evaluations, for two distinct neurons ($D = 7$). **B.** Median "Gaussianized" symmetrized KL divergence between the variational posterior and ground truth. See also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