The Variational Gaussian Approximation Revisited

Manfred Opper [*] Cédric Archambeau[†]

March 16, 2009

Abstract

The variational approximation of posterior distributions by multivariate Gaussians has been much less popular in the Machine Learning community compared to the corresponding approximation by factorising distributions. This is for a good reason: the Gaussian approximation is in general plagued by an $\mathcal{O}(N^2)$ number of variational parameters to be optimised, N being the number of random variables. In this work, we discuss the relationship between the Laplace and the variational approximation and we show that for models with Gaussian priors and factorising likelihoods, the number of variational parameters is actually $\mathcal{O}(N)$. The approach is applied to Gaussian process regression with non-Gaussian likelihoods.

1 Introduction

The variational approximation is among the most important techniques for treating intractable probabilistic models in the field of Machine Learning.

---

[*]Department of Computer, Technical University Berlin, Franklinstr. 28/29, D-10587 Berlin, Germany. `opperm@cs.tu-berlin.de`

[†]Centre for Computational Statistics and Machine Learning, University College London, Gower Street, London WC1E 6BT. `c.archambeau@cs.ucl.ac.uk`

An intractable probability distribution (usually the Bayesian posterior) is approximated by the closest distribution within a tractable family, where closeness is defined by the Kullback-Leibler divergence (Kullback & Leibler, 1951). In most applications, the tractable families contain distributions which factorize in all or in tractable subgroups of random variables (Beal, 2003; Winn, 2003). Hence, the method neglects correlations between variables which may be crucial in the learning of the hyperparameters.

If the random variables are continuous and unconstrained, the family of multivariate Gaussian densities suggests itself as a natural alternative to factorizing densities, allowing incorporation of correlations. Nevertheless, such a *variational Gaussian approximation* has been applied, to our knowledge, only a few times to problems in Machine Learning (e.g., Barber and Bishop (1998), Seeger (2000), Honkela and Valpola (2005)). One possible explanation is that the covariance matrix of the Gaussian requires a number of variational parameters to be optimised, which scales quadratically with the number of latent variables in the model.

However, we show that this prejudice against the variational Gaussian approximation is not always justified. We derive the exact fixed point conditions for the optimal setting of the variational parameters and find that for certain classes of probabilistic models, related to *Gaussian processes* (O'Hagan, 1978; Rasmussen & Williams, 2006), the number of free variational parameters will be only $2N$. While this fact seems to be known, at least in some parts of the Machine Learning community (Seeger, 1999, p. 119), several discussions have shown that many researchers were not aware of this result. We will demonstrate the method only on toy regression problems. Our results have also motivated the inclusion of the variational Gaussian approach within a larger study of different methods for classification with Gaussian processes (Nickisch & Rasmussen, 2008). In this complementary work a variety of comparisons can be found.

2 The variational Gaussian approximation

We consider probabilistic models for a set of observations $\mathbf{y} = (y_1, \ldots, y_M)^\top$ and a set of latent, unobserved random variables $\mathbf{x} = (x_1, \ldots, x_N)^\top$ defined by a joint probability distribution $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes a set of hyperparameters. We aim to approximate the posterior density $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{y}|\boldsymbol{\theta})}$ by a density $q(\mathbf{x})$, which belongs to a family of tractable densities. The op-

timal $q(\mathbf{x})$ is chosen to minimise the *variational free energy*

$$\mathcal{F}(q, \boldsymbol{\theta}) = -\ln p(\mathbf{y}|\boldsymbol{\theta}) + \text{KL}\left[q\|p\right], \tag{1}$$

where $\text{KL}\left[q\|p\right] = \int q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})} d\mathbf{x}$ is the *Kullback-Leibler* (KL) divergence. The free energy is an upper bound to the negative log-marginal probability of the observations and can be used to estimate hyperparameters by a variational EM algorithm (Dempster et al., 1977; Neal & Hinton, 1998).

If we restrict the approximate posterior $q(\mathbf{x})$ to be a multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, i.e.

$$q(\mathbf{x}) = (2\pi)^{-N/2}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \tag{2}$$

we get

$$\mathcal{F}(q, \boldsymbol{\theta}) = -\left\langle \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \right\rangle_q - \frac{N}{2}\ln 2\pi e - \frac{1}{2}\ln|\boldsymbol{\Sigma}|. \tag{3}$$

Hence, setting the derivatives of $\mathcal{F}(q, \boldsymbol{\theta})$ with respect to the variational parameters equal to zero leads to

$$0 = \nabla_{\boldsymbol{\mu}} \left\langle \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \right\rangle_q \qquad \text{and} \qquad \boldsymbol{\Sigma}^{-1} = -2\nabla_{\boldsymbol{\Sigma}} \left\langle \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \right\rangle_q. \tag{4}$$

In general, this will require the computation of $N(N+3)/2$ free variational parameters, which is much larger than the typically $\mathcal{O}(N)$ number of parameters often required for factorising variational distributions.

General properties

The variational Gaussian approach can be compared to the well-known Laplace approximation, where the mean of a Gaussian density is fitted *locally* at a point $\mathbf{x}$ which maximises the posterior $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$. The covariance is then built from the curvature of the log-posterior at the maximum. Hence, we have

$$0 = \nabla_{\mathbf{x}} \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \qquad \text{and} \qquad \boldsymbol{\Sigma}^{-1} = -\nabla_{\mathbf{x}}\nabla_{\mathbf{x}} \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}). \tag{5}$$

By contrast, (4) can be rewritten in two different ways using (18) and (19)

$$0 = \nabla_{\boldsymbol{\mu}} \left\langle \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \right\rangle_q = \left\langle \nabla_{\mathbf{x}} \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \right\rangle_q, \tag{6}$$

$$\boldsymbol{\Sigma}^{-1} = -\nabla_{\boldsymbol{\mu}}\nabla_{\boldsymbol{\mu}} \left\langle \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \right\rangle_q = -\left\langle \nabla_{\mathbf{x}}\nabla_{\mathbf{x}} \ln p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \right\rangle_q. \tag{7}$$

The second set of equalities on both lines shows that we have a global approximation: the conditions of the Laplace approximation hold on *average*. Another interpretation comes from the first set of equalities. Here we see that the variational Gaussian method is equivalent to applying Laplace's method to a new (implicitly defined) probability density $\tilde{q}(\boldsymbol{\mu}) \propto e^{\langle \ln p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \rangle_q}$, which is defined over the space of parameters $\boldsymbol{\mu}$.

3 Gaussian prior models

We will now specialise to a class of models which are of the following form:

$$p(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) = \frac{1}{Z_0} e^{-\sum_n V_n - \frac{1}{2} \mathbf{x}^\top \mathbf{K}^{-1} \mathbf{x}}, \tag{8}$$

where $\mathbf{K}$ is a positive definite matrix, V_n is a shorthand notation for $V(y_n, x_n)$ and Z_0 is the normalization constant (including the factor $|\mathbf{K}|^{1/2}$). A typical application is inference with Gaussian process ($\mathcal{GP}$) models (O'Hagan, 1978; Rasmussen & Williams, 2006), where $\mathbf{x} = (x(\mathbf{s}_1), \ldots, x(\mathbf{s}_N))^\top$ denotes the values of a latent function $x(\mathbf{s})$ at inputs $\mathbf{s}_1, \ldots, \mathbf{s}_N$, $\mathbf{K}$ is the kernel matrix and $V_n = -\ln p(y_n | x_n)$ denotes the negative log-likelihood.

From (8), we get

$$\mathcal{F}(q, \boldsymbol{\theta}) = \sum_n \langle V_n \rangle_{q_n} + \frac{1}{2} \text{tr} \{ \mathbf{K}^{-1} \boldsymbol{\Sigma} \} + \frac{1}{2} \boldsymbol{\mu}^\top \mathbf{K}^{-1} \boldsymbol{\mu} - \frac{1}{2} \ln |\boldsymbol{\Sigma}|$$
$$+ \ln Z_0 - \frac{N}{2} \ln 2\pi e, \tag{9}$$

where $\langle \cdot \rangle_{q_n}$ indicates that the expectation is taken with respect to the marginal $q(x_n)$, the univariate Gaussian with mean μ_n and variance Σ_{nn}.

Each term $\langle V_n \rangle_{q_n}$ depends only on the mean μ_n and the diagonal element Σ_{nn} of the full covariance $\boldsymbol{\Sigma}$. As a result, the second equation in (4) shows that the nondiagonal elements of $\boldsymbol{\Sigma}^{-1}$ are simply equal to those of $\mathbf{K}^{-1}$ and that the optimal covariance will be of the form

$$\boldsymbol{\Sigma} = \left(\mathbf{K}^{-1} + \boldsymbol{\Lambda} \right)^{-1}, \tag{10}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with $\boldsymbol{\lambda} \equiv (\ldots \ \lambda_n \ \ldots)^\top$ on its diagonal. Hence, we can use the N elements λ_n as new parameters. We found it also useful to represent the mean parameters in the form $\boldsymbol{\mu} = \mathbf{K} \boldsymbol{\nu}$ with a vector

4

$\boldsymbol{\nu}$ of N new effective parameters. Inserting these definitions into the free energy, a short calculation shows that the gradients of the free energy with respect to the new $2N$ parameters are given by

$$\mathbf{g_{\nu}} \equiv \nabla_{\boldsymbol{\nu}}\mathcal{F}(q,\boldsymbol{\theta}) = \mathbf{K}(\boldsymbol{\nu} - \bar{\boldsymbol{\nu}}), \qquad (11)$$

$$\mathbf{g_{\lambda}} \equiv \nabla_{\boldsymbol{\lambda}}\mathcal{F}(q,\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\Sigma} \circ \boldsymbol{\Sigma})(\boldsymbol{\lambda} - \bar{\boldsymbol{\lambda}}), \qquad (12)$$

where $\circ$ denotes the Hadamard product. $\bar{\boldsymbol{\nu}} \equiv (\ \ldots \ -\partial \langle V_n \rangle_{q_n} / \partial \mu_n \ \ldots \ )^{\top}$ and where $\bar{\boldsymbol{\lambda}} \equiv (\ \ldots \ 2\partial \langle V_n \rangle_{q_n} / \partial \Sigma_{nn} \ \ldots \ )^{\top}$. We use these gradients within a nonlinear conjugate gradient method with back-tracking (Hager & Zhang, 2005) to optimise the parameters.

One could generalize this approach to models where only a few of the likelihood terms depend on more than a single variable. In this case as well, a relatively small number of variational parameters would have to be be optimised.

Derivatives of the Gaussian expectations

The computation of the gradients requires explicit expressions for $\langle V_n \rangle_{q_n}$ for which there is often no analytical solution. However, one can circumvent this problem by using the Gaussian identities (18) and (19), along with $\left\langle \frac{\partial V_n}{\partial x_n} \right\rangle \Sigma_{nn} = \langle (x_n - \mu_n)V_n \rangle$:

$$-\bar{\nu}_n = \frac{\partial \langle V_n \rangle_{q_n}}{\partial \mu_n} = \left\langle \frac{\partial V_n}{\partial x_n} \right\rangle_{q_n} = \frac{\langle (x_n - \mu_n)V_n \rangle_{q_n}}{\Sigma_{nn}}, \qquad (13)$$

$$\frac{\bar{\lambda}_n}{2} = \frac{\partial \langle V_n \rangle_{q_n}}{\partial \Sigma_{nn}} = \frac{1}{2}\left\langle \frac{\partial^2 V_n}{\partial x_n^2} \right\rangle_{q_n} = \frac{\langle (x_n - \mu_n)^2 V_n \rangle_{q_n} - \Sigma_{nn} \langle V_n \rangle_{q_n}}{2\Sigma_{nn}^2}. \qquad (14)$$

As a consequence, the evaluation of these expectations does not require to compute the first and second order derivatives of V_n explicitly. They can either be naively estimated by sample averages, the samples being generated from the univariate Gaussian marginals q_n, or by more elaborate techniques such as Gauss-Hermite quadrature (Liu & Pierce, 1994), provided V_n satisfies some smoothness properties.

4 Application to robust Gaussian process regression

We test our approach on the Boston housing regression data. The aim is to predict the median value of a home. The input data is 13-dimensional. More information on this data set can be found at `http://lib.stat.cmu.edu/datasets/`. We investigate two noise models, *Laplace* and *Cauchy* noise, which have heavier tails compared to a simple Gaussian and are thus expected to be less sensitive to outliers. The likelihoods are respectively given by

$$p(y|x,\eta) = \frac{\eta}{2}\, e^{-\eta|y-x|} \qquad \text{and} \qquad p(y|f,\gamma) = \frac{1}{\pi\gamma}\left\{1 + \frac{(y-x)^2}{\gamma^2}\right\}^{-1}, \quad (15)$$

where $\eta > 0$ and $\gamma > 0$ are the noise parameters. In order to estimate the kernel parameters, which we denote by $\boldsymbol{\theta} = \{\theta_i\}_i$, and the noise parameters, we resort to gradient descent algorithms (Nocedal & Wright, 2000). The gradient of the variational free energy w.r.t. θ_i is given by

$$g_{\theta_i} = -\frac{1}{2}\text{tr}\left\{\left(\bar{\boldsymbol{\nu}}\bar{\boldsymbol{\nu}}^\top - \bar{\mathbf{B}}^{-1}\right)\frac{\partial \mathbf{K}}{\partial \theta_i}\right\}, \qquad \bar{\mathbf{B}} = \mathbf{K} + \bar{\boldsymbol{\Lambda}}^{-1}, \qquad (16)$$

where $\bar{\boldsymbol{\Lambda}}$ is a diagonal matrix with $\bar{\boldsymbol{\lambda}}$ on its diagonal. When computing this gradient, we have kept the variational parameters fixed, that is $\boldsymbol{\nu} = \bar{\boldsymbol{\nu}}$ and $\boldsymbol{\Lambda} = \bar{\boldsymbol{\Lambda}}$. The reason is that the implicit derivatives vanish at $\bar{\boldsymbol{\nu}}$ and $\bar{\boldsymbol{\Lambda}}$, which are stationary points of the free energy $\mathcal{F}(q, \boldsymbol{\theta})$. The overall training algorithm requires thus to perform an inner and an outer optimization loop. After each gradient step (16), one needs to recompute $\bar{\boldsymbol{\nu}}$ and $\bar{\boldsymbol{\Lambda}}$ using (11) and (12). To compute approximate predictions for $x(\mathbf{s}_*)$ at inputs $\mathbf{s}_*$ which are not in the training set using the approximate Gaussian on $\mathbf{x}$, we follow Rasmussen and Williams (2006, p. 44).

Table 1 shows the average test mean squared error (MSE) and the 1-standard deviation of the MSE for the standard $\mathcal{GP}$, the variational Gaussian approximation assuming Laplace noise and the variational Gaussian approximation assuming Cauchy noise. All models use a squared exponential kernel function with common length scale and common multiplicative constant. The standard $\mathcal{GP}$ assumes additive Gaussian noise with variance σ^2. All hyperparameters are optimised by gradient techniques. We use 5-fold cross-validation to estimate the MSE. It can be observed from Table 1 that

Table 1: Average test mean squared error (MSE) and 1-standard deviation of the MSE for the Boston housing data. See text for details.

LIKELIHOOD	CAUCHY	LAPLACE	GAUSSIAN
BOSTON	47.92 ± 16.13	42.35 ± 13.65	53.75 ± 22.02

both variational Gaussian approximations outperform the standard $\mathcal{GP}$. Assuming Laplace distributed noise seems to be advantageous over the Cauchy noise, suggesting that there are no strong outliers in the data.

5 Conclusion

In this paper, we have reconsidered the variational Gaussian approximation. We have clarified its relation to the Laplace approximation. We have shown that it is a tractable approach at least for models with Gaussian priors and factorising likelihoods, which naturally occur within the Gaussian process framework. We have also discussed several ways to compute the Gaussian expectations. The variational Gaussian approximation may also be naturally combined with variational sparse Gaussian approximations in order to speed up the inference for large datasets. We will give such extensions and comparisons of the method with other techniques in a forthcoming paper.

A Appendix

Derivatives of multivariate Gaussian expectations with respect to the mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$ are most conveniently computed using the characteristic function $G(\mathbf{k}) = \langle e^{i\mathbf{k}^\top \mathbf{x}} \rangle_q = e^{-\frac{1}{2}\mathbf{k}^\top \boldsymbol{\Sigma} \mathbf{k} + i\mathbf{k}^\top \boldsymbol{\mu}}$ of the Gaussian measure q. Using standard Fourier analysis, we can express expectations of any function $V(\mathbf{x})$ as

$$
\begin{aligned}
\langle V(\mathbf{x}) \rangle_q &= \frac{1}{(2\pi)^n} \int G(\mathbf{k}) e^{-i\mathbf{k}^\top \mathbf{y}} V(\mathbf{y}) \, d\mathbf{y} \, d\mathbf{k} \\
&= \frac{1}{(2\pi)^n} \int e^{-\frac{1}{2}\mathbf{k}^\top \boldsymbol{\Sigma} \mathbf{k} + i\mathbf{k}^\top (\boldsymbol{\mu} - \mathbf{y})} V(\mathbf{y}) \, d\mathbf{y} \, d\mathbf{k}.
\end{aligned} \tag{17}
$$

This shows that any derivative with respect to $\boldsymbol{\mu}$ is equivalent to $(-)$ the derivative of the exponential under the integral with respect to $\mathbf{y}$. This in turn, using integrations by parts with respect to $\mathbf{y}$, yields

$$\nabla_{\boldsymbol{\mu}} \langle V(\mathbf{x}) \rangle_q = \langle \nabla_{\mathbf{x}} V(\mathbf{x}) \rangle_q \tag{18}$$

$$\nabla_{\boldsymbol{\Sigma}} \langle V(\mathbf{x}) \rangle_q = \frac{1}{2} \langle \nabla_{\mathbf{x}} \nabla_{\mathbf{x}} V(\mathbf{x}) \rangle_q = \frac{1}{2} \nabla_{\boldsymbol{\mu}} \nabla_{\boldsymbol{\mu}} \langle V(\mathbf{x}) \rangle_q. \tag{19}$$

where the second equality in the last line follows from the first line.

References

Barber, D., & Bishop, C. M. (1998). Ensemble learning for Multi-Layer Networks. *Advances in Neural Information Processing Systems 10 (NIPS)*. The MIT Press.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. Doctoral dissertation, Gatsby Computational Neuroscience Unit, University College London, United Kingdom.

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

Hager, W. W., & Zhang, H. (2005). A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM Journal on Optimization*, *16*, 170–192.

Honkela, A., & Valpola, H. (2005). Unsupervised variational Bayesian learning of nonlinear models. *Advances in Neural Information Processing Systems 17 (NIPS)* (pp. 593 – 600). The MIT Press.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Liu, Q., & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, *81*, 624–629.

Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.), *Learning in graphical models*, 355–368. The MIT press.

Nickisch, H., & Rasmussen, C. E. (2008). Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*. Submitted.

Nocedal, J., & Wright, S. J. (2000). *Numerical optimization*. Springer.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society B*, *40*, 1–42.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, Massachusetts: The MIT Press.

Seeger, M. (1999). Bayesian methods for support vector machines and Gaussian processes. Master's thesis, University of Karlsruhe, Germany.

Seeger, M. (2000). Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. *Advances in Neural Information Processing Systems 12 (NIPS)* (pp. 603–609). The MIT Press.

Winn, J. (2003). *Variational message passing and its applications*. Doctoral dissertation, Department of Physics, University of Cambridge, United Kingdom.