Natural Gradient for Variational Bayesian Learning

Antti Honkela, Matti Tornio, Tapani Raiko, and Juha Karhunen

Helsinki University of Technology, Adaptive Informatics Research Centre,
P.O.Box 5400, FI-02015 TKK, Finland.
Email: {Antti Honkela, Matti Tornio, Tapani.Raiko, Juha.Karhunen}@tkk.fi
URL: http://www.cis.hut.fi/projects/bayes/

We introduce an efficient natural conjugate gradient algorithm for variational Bayesian learning. The algorithm is based on the geometry of the approximating distribution which is assumed to have a sufficiently simple form for enabling efficient computations. The new algorithm is most useful for models outside the conjugate exponential family, and it can provide superior convergence compared with previous methods.

Variational Bayesian learning [5, 2] and related variational methods [4] have become popular techniques in machine learning and graphical models. They often provide good approximations and robustness against overfitting at a reasonable computational cost compared to sampling-based methods.

Variational Bayesian learning is based on approximating the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X}, \mathcal{H})$ with a tractable approximation $q(\boldsymbol{\theta}|\boldsymbol{\xi})$, where $\boldsymbol{X}$ is the data, $\boldsymbol{\theta}$ are the parameters of the model $\mathcal{H}$, and $\boldsymbol{\xi}$ are the (variational) parameters of the approximation. The approximation is fitted by maximizing a lower bound on marginal log-likelihood

$$\mathcal{B}(q(\boldsymbol{\theta}|\boldsymbol{\xi})) = \left\langle \log \frac{p(\boldsymbol{X}, \boldsymbol{\theta}|\mathcal{H})}{q(\boldsymbol{\theta}|\boldsymbol{\xi})} \right\rangle = \log p(\boldsymbol{X}|\mathcal{H}) - D_{\mathrm{KL}}(q(\boldsymbol{\theta}|\boldsymbol{\xi})||p(\boldsymbol{\theta}|\boldsymbol{X}, \mathcal{H})), \tag{1}$$

where $\langle\cdot\rangle$ denotes expectation over q. This is equivalent to minimizing the Kullback–Leibler $D_{\mathrm{KL}}(q||p)$ divergence between q and p [2].

Finding the optimal approximation can be seen as an optimization problem, where the lower bound $\mathcal{B}(q(\boldsymbol{\theta}|\boldsymbol{\xi}))$ is maximized with respect to the variational parameters $\boldsymbol{\xi}$. This is often solved using a variational EM algorithm by updating sets of parameters alternatively while keeping the others fixed. Both VE and VM steps can individually implicitly optimally utilize the Riemannian structure of $q(\boldsymbol{\theta}|\boldsymbol{\xi})$ for conjugate exponential family models [7]. Nevertheless, the EM based methods are prone to slow convergence, especially under low noise.

The formulation of variational Bayesian learning as an optimization problem allows applying generic optimization algorithms to maximize $\mathcal{B}(q(\boldsymbol{\theta}|\boldsymbol{\xi}))$, but this is rarely done in practice because the problems are quite high dimensional. Additionally many of the parameters are in different roles and the lack of this specific knowledge of the geometry of the problem can seriously hinder generic optimization tools.

In this work, we have applied natural gradient for optimizing the bound $\mathcal{B}(q(\boldsymbol{\theta}|\boldsymbol{\xi}))$. Natural gradient is in this variational Bayesian learning problem superior to conventional gradient, because the space $S = \{\boldsymbol{\xi} \in \mathbf{R}^n\}$ is a curved Riemannian manifold. For a scalar function $\mathcal{F}(\boldsymbol{\xi})$ defined on a Riemannian manifold S, the direction of steepest ascent is given by the natural gradient [3]

$$\tilde{\nabla}\mathcal{F}(\boldsymbol{\xi}) = \mathbf{G}^{-1}(\boldsymbol{\xi})\nabla\mathcal{F}(\boldsymbol{\xi}). \tag{2}$$

where the matrix $\mathbf{G}(\boldsymbol{\xi}) = (g_{ij}(\boldsymbol{\xi}))$ is called the Riemannian metric tensor.

For the space of probability distributions $q(\boldsymbol{\theta}|\boldsymbol{\xi})$, the most common Riemannian metric tensor is given by the Fisher information matrix [1]

$$I_{ij}(\boldsymbol{\xi}) = g_{ij}(\boldsymbol{\xi}) = E\left\{ \frac{\partial \ln q(\boldsymbol{\theta}|\boldsymbol{\xi})}{\partial \xi_i} \frac{\partial \ln q(\boldsymbol{\theta}|\boldsymbol{\xi})}{\partial \xi_j} \right\} = E\left\{ -\frac{\partial^2 \ln q(\boldsymbol{\theta}|\boldsymbol{\xi})}{\partial \xi_i \partial \xi_j} \right\}, \tag{3}$$

where the last equality is valid given certain regularity conditions [6].

In approximate inference, the approximation $q(\boldsymbol{\theta}|\boldsymbol{\xi})$ is often chosen such that disjoint groups of variables are independent: $q(\boldsymbol{\theta}|\boldsymbol{\xi}) = \prod_i q_i(\boldsymbol{\theta}_i|\boldsymbol{\xi}_i)$. This simplifies the computation of the natural gradient, as the Fisher information matrix becomes block-diagonal, and the required matrix inversion can be performed very efficiently. This is a key point in our approach.

For getting an even more efficient algorithm for high-dimensional problems, we have combined natural gradient learning with the conjugate gradient method [8]. We have compared the resulting Riemannian conjugate gradient algorithm with the conjugate gradient algorithm and with the heuristic algorithm for optimizing $q(\boldsymbol{\theta}|\boldsymbol{\xi})$ introduced in [9] for the nonlinear state-space model studied in [9]. For synthetic data, the Riemannian conjugate gradient algorithm converged much faster than the standard conjugate gradient algorithm, while the heuristic algorithm performed poorly. For real-world speech data, the Riemannian conjugate gradient algorithm outperformed the heuristic algorithm by a factor more than 10 while the conjugate gradient algorithm had problems in converging at all in a reasonable time.

References

[1] S. Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Lecture Notes in Statistics.* Springer-Verlag, 1985.

[2] C. Bishop. *Pattern Recognition and Machince Learning.* Springer, Cambridge, 2006.

[3] S. Douglas and S. Amari. Natural-gradient adaptation. In S. Haykin, editor, *Unsupervised Adaptive Filtering, Vol. I: Blind Source Separation*, pages 13–61. Wiley, New York, 2000.

[4] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In M. Jordan, editor, *Learning in Graphical Models*, pages 105–161. The MIT Press, Cambridge, MA, USA, 1999.

[5] H. Lappalainen and J. Miskin. Ensemble learning. In M. Girolami, editor, *Advances in Independent Component Analysis*, pages 75–92. Springer-Verlag, Berlin, 2000.

[6] M. Murray and J. Rice. *Differential Geometry and Statistics.* Chapman & Hall, 1993.

[7] M. Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13(7):1649–1681, 2001.

[8] S. T. Smith. *Geometric Optimization Methods for Adaptive Filtering.* PhD thesis, Harvard University, Cambridge, Massachusetts, 1993.

[9] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.

Category: Graphical models
Preference: Poster or oral
Presenter: Prof. Juha Karhunen