Understanding Covariance Estimates in Expectation Propagation

William Stephenson
Department of EECS
Massachusetts Institute of Technology
Cambridge, MA 02139
wtstephe@csail.mit.edu

Tamara Broderick
Department of EECS
Massachusetts Institute of Technology
Cambridge, MA 02139
tbroderick@csail.mit.edu

Abstract

Bayesian inference for most models of modern interest requires approximation of the posterior distribution. Variational inference (VI) methods formulate posterior approximation as a particular optimization problem and have grown in popularity due to their fast runtime on large datasets. There is recent work that suggests it might be beneficial to instead consider an alternative optimization objective to that used by VI, such as the one employed by expectation propagation (EP). EP has recently been shown to scale to large datasets and complex models, and existing theory suggests we should expect EP to overestimate—rather than underestimate—the variance of parameters. We show through two examples that there are actually two regimes, one in which EP overestimates covariances and one in which it underestimates. We prove that under some conditions, the objective function behind EP switches between these regimes depending on the log-concavity / convexity of the approximate distribution used.

1 Introduction

Let $x = (x_1, \ldots, x_N)$ represent N observed data points and $\theta \in \mathbb{R}^K$ a K-dimensional parameter vector. We assume some generative model for the data described by a likelihood $p(x|\theta)$ and a prior $p(\theta)$. A common task in Bayesian inference is find the posterior distribution $p_x(\theta) := p(\theta|x)$. For any mildly complex choice of generative model, the posterior cannot be computed exactly and must be approximated. In practice, however, the end product of a Bayesian analysis is often some functional of the posterior; for instance, we might report the posterior mean and variance for each parameter. In these cases, a full description of the posterior is unnecessary, and computational savings might be achieved by focusing on a faster method that is accurate for the desired functional.

Markov Chain Monte Carlo methods have been a traditional mainstay of approximate Bayesian inference and enjoy a number of nice theoretical properties but are often too slow in practice. Variational inference (VI), especially the variant known as *mean-field variational inference* (MFVI), has grown in popularity recently due to its speed on large datasets [6, 3]. But MFVI is known to have a number of practical failings—such as underestimating the uncertainty of model parameters, sometimes dramatically [2, 15, 8]. A number of recent papers aim to correct this issue either via a post-hoc correction to MFVI or within the VI framework [14, 4]. But these new methods often have poorer scaling properties—either in the number of data points or dimension of the parameter or both—than MFVI.

At first glance, it seems that a promising alternative would be found in *expectation propagation* (EP) [9]. The motivation behind EP is to minimize an alternative optimization objective to that used in VI. VI aims to minimize the *reverse Kullback-Leibler divergence* $KL(q||p_x)$ over q in some class of tractable distributions. By contrast, EP is motivated by minimization of the *forward* KL-

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

divergence $KL(p_x||q)$ over q. The same conventional wisdom that states VI tends to underestimate uncertainties also states that minimizing the forward KL-divergence yields overestimates of parameter uncertainties [2, 10, 5]. In some ways, this is a safer direction; overestimating uncertainty should lead to more conservative decisions. Additionally, overestimates might be more amenable to post-hoc corrections than underestimates. While there has been work on correcting the entirety of the approximate distribution found by EP [12, 13], there has not yet been a correction that focuses directly on the covariance—which might yield a more computationally efficient approach. Another reason to consider EP is that recent work demonstrates it can be scaled to large datasets and suggests it may even have better locality properties than VI in complex models [7, 3].

However, conventional wisdom and practical experiments seem not to agree in the literature. As one example of uncertainty overestimation, [2, 10, 5] state or show that we expect EP to, e.g., mode-average by stretching across modes of a multimodal posterior. Conversely, [13, 7] show practical examples of EP fitting to a single posterior mode. Before investigating EP's covariance estimates, or indeed before using EP in general, we must understand exactly how EP operates in practice. In the following, we review the forward KL-divergence in Section 2. We examine the proposition that minimizing this objective encourages mode-averaging, and we give an example of mode-averaging behavior in Section 3, whereas we demonstrate mode-seeking behavior in Section 4. In Section 5, we prove that for a somewhat restrictive class of posteriors, the mode-averaging / mode-seeking behavior of this objective depends on the log-concavity / convexity of the approximate distribution used. We emphasize that our results mainly pertain to the objective *motivating* EP, which is not necessarily what EP actually minimizes. We leave further discussion of this point for future work.

2 Kullback-Leibler Divergence and Expectation Propagation

The motivating optimization problem for EP is to choose an approximating distribution q^* for the posterior. In particular, we aim to choose q^* from some class of distributions $\mathcal{Q}$ such that q^* minimizes the forward Kullback-Leibler (KL) divergence, $KL(p_x||q)$. Typically the distributions in $\mathcal{Q}$ are more computationally convenient than the exact posterior; e.g., one can more easily calculate desirable functionals under $q \in \mathcal{Q}$. So, when the optimal q^* exists, we wish to choose

$$q^*(\theta) = \arg\min_{q\in\mathcal{Q}} KL(p_x||q) = \arg\max_{q\in\mathcal{Q}} \int p(\theta|x) \log q(\theta) d\theta = \arg\max_{q\in\mathcal{Q}} E(q), \tag{1}$$

where we have defined $E(q) := \int p(\theta|x) \log q(\theta) d\theta$.

Recall that the full KL integrand is $p(\theta|x) \log(p(\theta|x)/q(\theta))$. So the traditional intuition [2] is that we expect that q should not have very little mass where p_x has a lot of mass, or the $\log$ factor will be large. If this were always true, we would expect q^* to span all high-probability modes of p_x. But we see from Eq. 1 that the $p(\theta|x) \log p(\theta|x)$ term is constant in q^* and effectively disappears from the objective, so it is perhaps not so simple. In fact, we see that q^* does not always stretch across the posterior modes below.

3 An Analytic Example of Fitting a Multimodal Posterior

We first give a case in which we can analytically maximize Eq. 1 and the conventional wisdom for Eq. 1 always applies; that is, the q^* minimizing $KL(p_x||q)$ stretches across multiple modes of the posterior. Suppose the exact posterior is a K-component mixture of D-dimensional diagonal-covariance Gaussians: $p(\theta|x) = \sum_{k=1}^K \pi_k N(\theta; \mu_k, \Gamma_k^{-1})$, where π_k are weights with $\sum_{k=1}^K \pi_k = 1$.

As is common in practice [2], we assume the approximating family $\mathcal{Q}$ follows the *mean-field* assumption; that is, any $q \in \mathcal{Q}$ factorizes as $q(\theta) = \prod_i q(\theta_i)$. Specifically, here we suppose q is a single Gaussian, and the mean-field assumption implies q has diagonal covariance Λ^{-1}: $q(\theta) = N(\theta; \eta, \Lambda^{-1})$. For any fixed η, we apply Eq. 1 to find that $E(q)$ is maximized by setting the dth diagonal entry of Λ to:

$$\Lambda_{dd} = \left[\sum_{k=1}^K \pi_k (\Gamma_{kd} + (\mu_{kd} - \eta_d)^2) \right]^{-1} \tag{2}$$

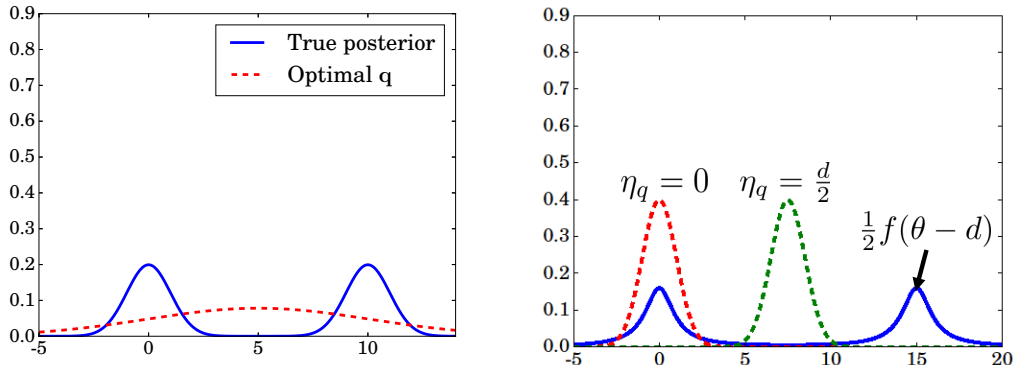

Figure 1: *Left*: Illustration of Section 3, where the true posterior is a 1-D mixture of Gaussians. The optimal Gaussian q^* fits across both modes of p_x with a wide variance, regardless of the separation between them. *Right*: Illustration of Section 5. For a symmetric bimodal posterior $p_x(\theta) = \frac{1}{2}f(\theta) + \frac{1}{2}f(\theta - d)$, a convex $\log q(\theta)$ yields mode-seeking behavior ($\eta_q = 0$) whereas a concave $\log q(\theta)$ yields mode-averaging behavior ($\eta_q = d/2$).

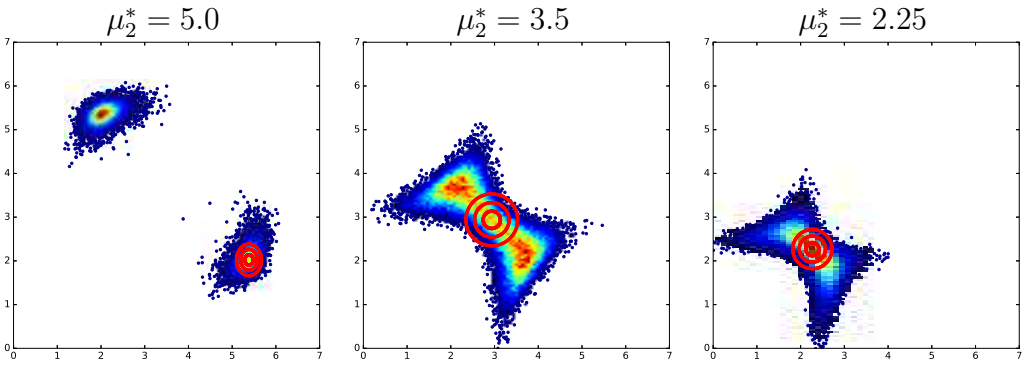

Figure 2: Density plots of samples from posterior $p(\mu_1, \mu_2|x)$ of a 1-D Gaussian mixture model, as described in Section 4, with mean and covariance found by EP (red) plotted on top. In each case, the true $\mu_1^* = 2$, while μ_2^* varies from left to right as $5, 3.5, 2.25$. In the case of a strongly bimodal posterior (*left*), EP fits to only one mode of the posterior.

and the mean $\eta = \sum_k \pi_k \mu_k$. A brief derivation can be found in Appendix A. In the case of $\pi_k = 1/K$, we see that q places significant mass on every mode of $p(\theta|x)$, as expected from the conventional wisdom; see Fig. 1 (left) for an illustration.

4 An Empirical Example: Gaussian Mixture Model

We now give an empirical case that contrasts the results of Section 3; that is, we find EP fitting q to a single mode of the posterior. We take the case of Gaussian mixture models, where the *likelihood* distribution $p(x|\theta)$ is a mixture of Gaussians. We then expect the posterior to be multimodal, but not a finite mixture of Gaussians as in Section 3. We note that capturing all multimodality in the current example would typically be undesirable as symmetry turns each "real" mode into $K!$ modes. Still, this provides a simple illustration of a real multimodal posterior.

We sampled 100 1-dimensional points from a GMM with $K = 2$ components and true parameters $\pi_1^* = \pi_2^* = 0.5$, $\sigma_1^* = \sigma_2^* = 1.0$, and variable means. We ran three experiments, with fixed $\mu_1^* = 2$, and varied $\mu_2^* = \{5, 3.5, 2.25\}$. Our choice of approximate distribution follows [13] and factorizes as:

$$q(\pi, \mu, \Lambda) = q(\pi) \prod_{k=1}^{K} q\left(\mu_k, \sigma_k^{-2}\right),$$

where each $q(\mu_k, \sigma_k^{-2})$ is a normal-Wishart distribution and $q(\pi)$ a Dirichlet distribution. In this case, it is not possible to analytically minimize the KL-divergence in Eq. 1, so we run the actual EP algorithm [9], which we emphasize does not necessarily minimize $KL(p_x||q)$. Our chosen q has a Student's t-distribution as its marginal $q(\mu)$, which quickly approaches the Gaussian q used in Section 3, so we might expect similar behavior as in Section 3.

In Fig. 2, we show MCMC samples from the true marginal posterior $p(\mu_1, \mu_2|x)$ along with the mean and covariance of the EP approximate distribution $q(\mu)$. When the modes of p_x are very distant, EP centers itself at a single mode of p_x. However, as the modes come closer, EP chooses to place q in-between them. Though this is similar behavior to that observed experimentally by [7, 13], it is surprisingly distinct from that in Section 3 and the typically cited behavior of minimizing the KL-divergence $KL(p_x||q)$.

5 Understanding the Different Behavior

Although it is at first surprising that the seemingly similar examples of Section 3 and Section 4 yield different behavior, we now show that this difference is due to the concavity / convexity of $\ell(\theta - \eta_q) := \log q(\theta)$, with η_q being the mean of q. Notably, we show this holds for any symmetric bimodal true posterior p_x:

Theorem 1. *Suppose $\ell(\theta - \eta_q) := \log q(\theta)$ is symmetric around η_q and concave in θ. Then for any bimodal symmetric true posterior of the form $p_x(\theta) = \frac{1}{2}f(\theta) + \frac{1}{2}f(\theta - d)$, the choice of $\eta_q = \frac{d}{2}$ yields a KL-divergence $KL(p_x||q)$ at least as small as the choice of $\eta_q = 0$. Conversely, if ℓ is convex, $\eta_q = 0$ yields a smaller KL-divergence.*

Proof. A short proof can be found in Appendix B. □

This theorem explains the above behavior. In Section 3, q was a Gaussian so that $\log q(\theta) = -\theta^2/2$ is concave, and we indeed saw that the optimum of $KL(p_x||q)$ placed q in-between the modes of p_x (i.e. $\eta_q = d/2$). In Section 4, our choice of $q(\mu)$ was a Student's t-distribution, which has ℓ neither convex nor concave so that Theorem 1 does not apply. Still, we can give a heuristic argument about its behavior in Fig. 2. A Student's t-distribution with ν degrees of freedom is log-convex outside the interval $[-\sqrt{\nu}, \sqrt{\nu}]$ and log-concave inside this interval [1]. If $\sqrt{\nu}$ is large enough to cover both modes of p_x, we effectively have a log-concave distribution, which Theorem 1 implies will prefer mode-averaging. On the other hand, if $\sqrt{\nu}$ is small compared to d, we have a "mostly" log-convex distribution, in which case Theorem 1 implies it will be mode-seeking. We see this behavior Fig. 2; the ν recovered by EP is roughly the same across all three cases, and EP moves to mode-averaging when the posterior modes are sufficiently close.

6 Conclusion

We have demonstrated the tendency for EP to fit to either one or many modes of the posterior and given a theoretical explanation as to why and when this occurs. A number of questions remain: 1) Can we derive similar properties for other divergences used in practice [5]? 2) Can we derive the exact separation d between posterior modes at which EP changes between fitting one mode to both? 3) How does this mode-fitting behavior affect practical applications of EP, and can we see it in previous applications of EP in models such as LDA [11]? Additionally, we have not given much attention to the fact that EP does not necessarily maximize the objective given by Eq. 1, and instead approximates it. How this approximation affects the above three questions is another avenue for future work.

Acknowledgments. WS and TB were supported in part by DARPA award FA8750-17-2-0019.

References

[1] M. Bagnoli and T. Bergstrom. Log-concave probability and its applications. *Economic Theory*, 2004.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*, chapter 10. Springer, 2009.

[3] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming variational bayes. In *Advances in Neural Information Processing Systems 26*. 2013.

[4] R. J. Giordano, T. Broderick, and M. I. Jordan. Linear response methods for accurate covariance estimates from mean field variational bayes. In *Advances in Neural Information Processing Systems 28*. 2015.

[5] J. M. Hernandez-Lobato, Y. Li, M. Rowland, D. Hernandez-Lobato, T. Bui, and R. E. Turner. Black-box α-divergence minimization. In *Proceedings of the 33rd International Conference on Machine Learning*. 2016.

[6] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

[7] Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic expectation propagation. In *Advances in Neural Information Processing Systems 28*. 2015.

[8] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*, chapter 33. Cambridge University Press, 2003.

[9] T. Minka. Expectation propagation for approximate bayesian inference. In *Uncertainty in Artificial Intelligence*. 2001.

[10] T. Minka. Divergence measures and message passing. Technical report, 2005.

[11] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*. 2002.

[12] M. Opper, U. Paquet, and O. Winther. Perturbative corrections for approximate inference in Gaussian latent variable models. *Journal of Machine Learning Research*, 14:2857–2898, 2013.

[13] U. Paquet, O. Winther, and M. Opper. Perturbation corrections in approximate inference: Mixture modelling applications. *Journal of Machine Learning Research*, 10:1263–1304, 2009.

[14] D. Tran, D. M. Blei, and E. M. Airoldi. Copula variational inference. In *Advances in Neural Information Processing Systems 28*. 2015.

[15] R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In A. Green, A. T. Cemgil, and S. Chiappa, editors, *Bayesian Time Series Models*. Cambridge University Press, 2011.

A Derivation for Mixture of Gaussians Posterior

Here, we derive the results from Section 3.1. We start with the assumption that the posterior and approximate distributions over $\theta \in \mathbb{R}^D$ are of the form:

$$p(\theta|x) = \sum_{k=1}^{K} \pi_k N(\theta; \mu_k, \Gamma_k^{-1}) \ , \quad q(\theta) = N(\theta; \eta, \Lambda^{-1}).$$

The integral in Eq. 1 that we want to maximize is just the expectation of $\log q(\theta)$ under p:

$$\mathbb{E}_p[\log q(\theta)] = \mathbb{E}\left[-\frac{1}{2}(\theta^T\Lambda\theta - 2\theta^T\Lambda\eta + \eta^T\Lambda\eta) + \frac{1}{2}\log|\Lambda| - \frac{D}{2}\log 2\pi\right]$$

$$= -\frac{1}{2}\left(\sum_{k=1}^{K}\pi_k\sum_{d=1}^{D}\left(\Lambda_{dd}(\mu_{kd}^2 - 2\mu_{kd}\eta_d + \eta_d^2 + \Gamma_{kd})\right)\right) + \sum_{d=1}^{D}\log\Lambda_{dd} - \frac{D}{2}\log 2\pi.$$

Taking the derivative with respect to Λ_{dd} and setting it equal to zero gives Eq. 2:

$$\Lambda_{dd} = \left[\sum_{k=1}^{K}\pi_k\left((\mu_{kd} - \eta_d)^2 + \Gamma_{kd}\right)\right]^{-1}.$$

To find the optimal η, we can appeal to the moment-matching properties of $KL(p_x||q)$. That is, if $\mathcal{Q}$ represents exponential family distributions—a popular class of distributions that allows easy normalization and moment calculation, then Eq. 1 is maximized by choosing the parameters of q so that its expected sufficient statistics are equal to their expectation under p_x [2]; that is, if $q(\theta) \propto \exp \lambda^T \theta$, we can minimize Eq. 1 by picking λ such that:

$$\mathbb{E}_{p_x}[\theta] = \mathbb{E}_q[\theta].$$

In particular, since q here is a Gaussian, we must set η as:

$$\eta = \mathbb{E}_p[\theta] = \sum_{k=1}^{K} \pi_k \mu_k.$$

B Proof of Theorem 1

We restate and prove Theorem 1 from Section 5:

Theorem 1. *Suppose $\ell(\theta - \eta_q) := \log q(\theta)$ is symmetric around η_q and concave in θ. Then for any bimodal symmetric true posterior of the form $p_x(\theta) = \frac{1}{2}f(\theta) + \frac{1}{2}f(\theta - d)$, the choice of $\eta_q = \frac{d}{2}$ yields a KL-divergence $KL(p_x||q)$ at least as small as the choice of $\eta_q = 0$. Conversely, if ℓ is convex, $\eta_q = 0$ yields a smaller KL-divergence.*

Proof. Consider ℓ concave and symmetric around 0 and p_x of the given form. Recalling that minimizing $KL(p_x||q)$ is equivalent to maximizing Eq. 1, we evaluate Eq. 1 as a function of η_q:

$$\begin{aligned} E(\eta_q) &= \int \left(\frac{1}{2} f(\theta) \ell(\theta - \eta_q) + \frac{1}{2} f(\theta - d) \ell(\theta - \eta_q) \right) \mathrm{d}\theta \\ &= \int f(\theta) \left(\frac{1}{2} \ell(\theta - \eta_q) + \frac{1}{2} \ell(\theta - \eta_q + d) \right) \mathrm{d}\theta, \end{aligned} \tag{3}$$

which holds by a change of variable $f(\theta - d)\ell(\theta - \eta_q) = f(\theta')\ell(\theta' - \eta_q + d)$. Now, we can consider setting $\eta_q = 0$ (placing q on top of one mode of p_x) or $\eta_q = \frac{d}{2}$ (placing q in-between the modes of p_x):

$$E(0) = \int f(\theta) \left(\frac{1}{2} \ell(0) + \frac{1}{2} \ell(\theta + d) \right) \mathrm{d}\theta \ , \quad E\left(\frac{d}{2}\right) = \int f(\theta) \ell\left(\theta + \frac{d}{2}\right) \mathrm{d}\theta,$$

where the evaluation of $E(\frac{d}{2})$ uses the symmetry of ℓ around 0. Since $f \geq 0$, we see that if ℓ is concave, the integrand on the right is strictly larger, giving $E\left(\frac{d}{2}\right) \geq E(0)$. Conversely, if ℓ is convex, we get $E(0) \geq E\left(\frac{d}{2}\right)$. □