6.882: Bayesian Modeling and Inference **Lecture Date: February 15, 2018**

Two Problems with Variational Expectation Maximization for Time Series

Lecturers: Phyllis Ju and Helian Feng *Scribes: Christina Ji and Kevin Kainan Li*

Outline

Part 1. Variational Inference

- Evaluating the variational Bayes approximation.
- Functionals
- Properties of posterior distributions
- Examples: univariate Gaussian and correlated Gaussian

Part 2. Turner et al 2011 paper

- Variational EM
- Two problems
- Class discussions

1 Functionals used for evaluating VB approximation

Definition 1 (Functional)**.** (Bishop, 2006, p. 462) A functional is a function mapping a distribution to a scalar.

Examples of functionals are the expectation and variance operators, e.g., for a distribution $\mathbf{P}$

$$\mathbf{E}(\mathbf{P}) = \mathbf{E}_{\mathbf{P}}\,\Theta \quad \text{and} \quad \mathbf{Var}(\mathbf{P}) = \mathbf{Var}_{\mathbf{P}}\,\Theta \tag{1}$$

where Θ has distribution $\mathbf{P}$.

Consider a Bayesian application. Suppose we observe heights $x_1, x_2, \ldots, x_N$, and we want to figure out the population mean height and population height variation. E.g., we have a generative model consisting of a likelihood:

$$x_n|\mu,\tau \overset{iid}{\sim} \mathcal{N}(\mu,\tau^{-1}) \tag{2}$$

and priors

$$\tau \sim \mathsf{Gamma}(a_0,b_0) \quad \text{and} \quad (\mu|\tau) \sim \mathsf{Normal}(\mu_0,(\lambda_0\tau)^{-1}). \tag{3}$$

for some fixed hyperparameters $a_0, b_0, \mu_0, \lambda_0$. Note that τ here is the "precision", a.k.a. the reciprocal of the likelihood variance parameter.

The posterior mean of μ gives us an estimate of the mean population height and the posterior variance of μ gives us a sense of how well we know the mean of μ. If the variance is high, we don't have enough data to really pin down the value of μ. Likewise, the posterior mean of $\tau^{-1/2}$ gives us an estimate of how much heights vary in the population (in particular, an estimate of the standard deviation of the height distribution). And the posterior variance of $\tau^{-1/2}$ tells us how well we know this quantity. If the posterior variance is large, we have a lot of uncertainty about the value of $\tau^{-1/2}$.

Let D represent all of our data $\{x_1,\ldots,x_N\}$. If we approximate the exact posterior $p(\cdot|D)$ with q^*, then the quality of the approximation might be gauged by how well we approximate the two functionals above—especially if this is what we report in our analysis.

$\mathbf{E}_{q^*}\,\Theta \overset{?}{=} \mathbf{E}_{p(\cdot|D)}\,\Theta$.

$\mathbf{Var}_{q^*}\,\Theta \overset{?}{=} \mathbf{Var}_{p(\cdot|D)}\,\Theta$.

2 Properties of posterior distributions

Suppose that, conditional on $\theta_0 \in \Theta$, the data $X_1, \ldots, X_n \in \mathbf{R}$ is a random sample from a distribution $\pi_0(\cdot \mid \theta_0)$. Given a distribution p_0 on $\mathbf{R} \times \Theta$, let $p(\theta) = p_0(\theta \mid X_1, \ldots, X_n)$ be the posterior. Assuming sufficiently nice regularity conditions, the posterior has the following properties.

Consistent. As $n \to \infty$, $p(U) \to 1$ for every neighborhood of $\theta_0 \in U \subseteq \Theta$. In other words, the posterior concentrates around the truth.

Asymptotic Normality. For large n, the distribution p is normal

$$p(\theta) \approx \mathcal{N}(\theta_0, I_n(\theta_0)^{-1}) \tag{4}$$

where $I_n(\theta_0)$ is the Fisher information matrix. (In some sense, this motivates why we study Gaussian examples below.) This result is the Bernstein von-Mises Theorem; see van der Vaart (1998) for more details.

3 Univariate Gaussian example

Since the Gaussian distribution is widely used, we give an explicit derivation of variational inference in the Gaussian case[1]. Specifically, the goal is to infer the posterior distribution for the mean μ and precision τ of a normal distribution $\mathcal{N}(\mu, \tau^{-1})$ given conditionally i.i.d. observations $\mathcal{D} = \{x_1, \ldots, x_N\}$. The likelihood function is

$$p(\mathcal{D}|\mu, \tau) = \big(\frac{\tau}{2\pi}\big)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \tag{5}$$

Suppose that the prior of τ and μ are given by

$$\tau \sim \mathsf{Gamma}(a_0, b_0) \quad \text{and} \quad (\mu|\tau) \sim \mathsf{Normal}(\mu_0, (\lambda_0\tau)^{-1}). \tag{6}$$

The variational approximation we will use is

$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau). \tag{7}$$

The optimal choice q_μ^* and q_τ^* are given by

$$\begin{aligned}
\ln q_\mu^*(\mu) &= \mathbf{E}_\tau[\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau)] + \text{const} \\
&= -\frac{\mathbf{E}\,\tau}{2}\left\{\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right\} + \text{const} \\
&= -\frac{\mathbf{E}\,\tau}{2}\{(\lambda_0 + N)\mu^2 - 2\mu_0\mu\lambda_0 - 2\bar{x}N\mu\} + \text{const} \\
&= -\frac{\mathbf{E}\,\tau}{2}(\lambda_0 + N)\left\{\mu - \frac{\lambda_0\mu_0 + \bar{x}N}{\lambda_0 + N}\right\}^2 + \text{const}
\end{aligned} \tag{8}$$

where we completed the square in moving from line three to four; see this source[2]. We now can read off

$$q_\mu^* \sim \mathsf{Normal}(\mu_N, \lambda_N^{-1}), \quad \text{where} \quad \mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N} \text{ and } \lambda_N = (\lambda_0 + N)\,\mathbf{E}\,\tau. \tag{9}$$

[1] As we did not have time for the full derivation in class, the material in this section is supplemented from section 10.1.3 in (Bishop, 2006, pg.470-472).

[2] `https://learnbayes.org/index.php?option=com_content\&view=article\&id=77:completesquare` — from Morey (2012).

Similarly, we find

$$\begin{aligned}\ln q_\tau^*(\tau) &= \mathbf{E}_\mu[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)] + \ln p(\tau) + \text{const} \\ &= -\frac{\tau}{2}\mathbf{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \frac{N}{2}\ln\tau + (a_0-1)\ln\tau - b_0\tau + \text{const}\end{aligned} \tag{10}$$

Once again, pattern matching with the log likelihoood for the Gamma distribution gives

$$q_\tau^* \sim \mathsf{Gamma}(a_N, b_N), \quad \text{where} \quad a_N = a_0 + \frac{N}{2} \text{ and } b_N = b_0 + \frac{1}{2}\mathbf{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] \tag{11}$$

An illustration of coordinate ascent in this case is illustrated in the following figure.

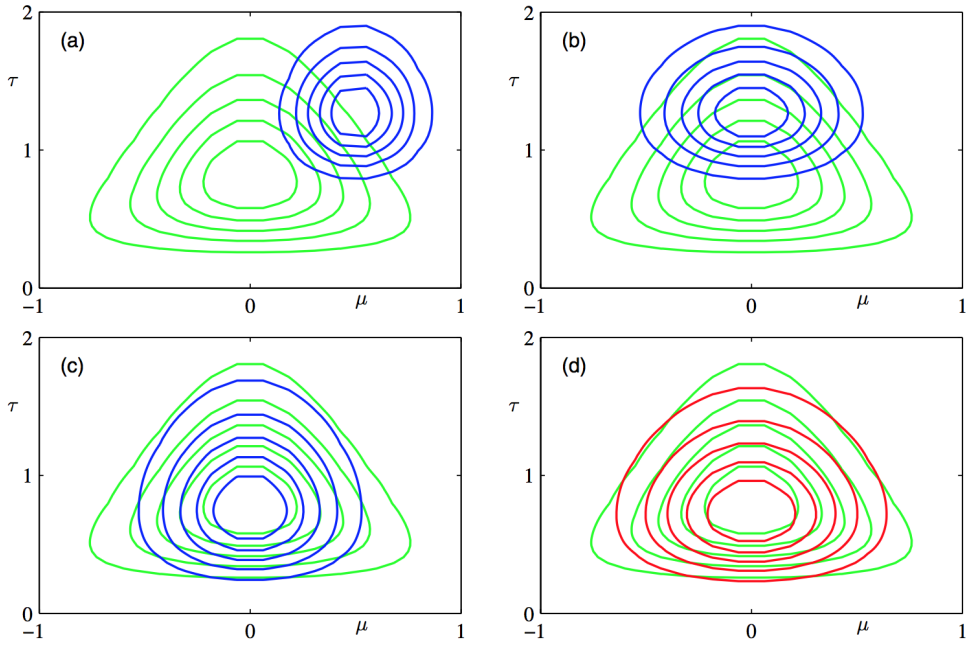

Figure 10.4 Illustration of variational inference for the mean μ and precision τ of a univariate Gaussian distribution. Contours of the true posterior distribution $p(\mu, \tau|D)$ are shown in green. (a) Contours of the initial factorized approximation $q_\mu(\mu)q_\tau(\tau)$ are shown in blue. (b) After re-estimating the factor $q_\mu(\mu)$. (c) After re-estimating the factor $q_\tau(\tau)$. (d) Contours of the optimal factorized approximation, to which the iterative scheme converges, are shown in red.

Figure 1: The x and y-axes are μ and τ, respectively. From (a) to (b), μ is updated, so the factorized approximation only changes along the x-axis. From (b) to (c), τ is updated, so the factorized approximation only changes along the y-axis. (d) shows the factorized approximation when the iterative scheme has converged. The approximation cannot perfectly fit the true distribution because μ and τ are not actually independent as assumed in the factorization. (Bishop, 2006, pg.472)

4 Correlated Gaussians example and Multimodality

This part was covered separately by the two presenters. We've chosen to present the complete example here. (Bishop, 2006, pg.466-468) covers the example in detail.

Consider a Gaussian distribution $\mathbf{z} = (z_1, z_2) \sim \mathsf{Normal}(\mu, \Gamma)$ with and precision have elements

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}. \tag{12}$$

Using the factorized approximation gives

$$\begin{aligned} \ln q_1^*(z_1) &= \mathbf{E}_{z_2}[\ln p(z)] + \text{const} \\ &= -\frac{1}{2} z_1^2 \Lambda_{11} + z_1 \mu_1 \Lambda_{11} - z_1 \Lambda_{12} (\mathbf{E}[z_2] - \mu_2) + \text{const} \\ &= -\frac{1}{2\Lambda_{11}^{-1}} \left(z_1^2 - 2 z_1 \left(\mu_1 - \Lambda_{12} (\mathbf{E}[z_2] - \mu_2) \Lambda_{11}^{-1} \right) \right) + \text{const} \\ &= -\frac{1}{2\Lambda_{11}^{-1}} \left(z_1 - \left(\mu_1 - \Lambda_{12} (\mathbf{E}[z_2] - \mu_2) \Lambda_{11}^{-1} \right) \right)^2 + \text{const} \end{aligned} \tag{13}$$

In line 3, any terms not including z_1 are included in the constant. Completing the square in lines 4 and 5 matches the log likelihood form for a Gaussian distribution. We can read off the distribution:

$$q^*(z_1) = \mathsf{Normal}(z_1 | m_1, \Lambda_{11}^{-1}) \quad \text{where} \quad m_1 = \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (\mathbf{E}\, z_2 - \mu_2) \tag{14}$$

By symmetry,

$$q^*(z_2) = \mathsf{Normal}\left(z_2 | m_2, \Lambda_{22}^{-1}\right) \quad \text{where} \quad m_2 = \mu_2 - \Lambda_{22}^{-1} \Lambda_{21} (\mathbf{E}\, z_1 - \mu_1) \tag{15}$$

Here, the expectations $\mathbf{E}\, z_1$ and $\mathbf{E}\, z_2$ are taken with respect to $q^*(z_1)$ and $q^*(z_2)$, respectively. In the general case (e.g., non-Gaussian), solving for $q^*(z_1)$ and $q^*(z_2)$ involves cycling through the variables—keeping q_i fixed while updating q_j—until some convergence criterion is met. In our present example, however, it is clear that the equations above have solutions with $\mathbf{E}\, z_1 = \mu_1$ and $\mathbf{E}\, z_2 = \mu_2$. In other words, when approximating a two-dimensional correlated Gaussian with independent Gaussians, the means are matched exactly. This phenomenon is illustrated in the Figure below.

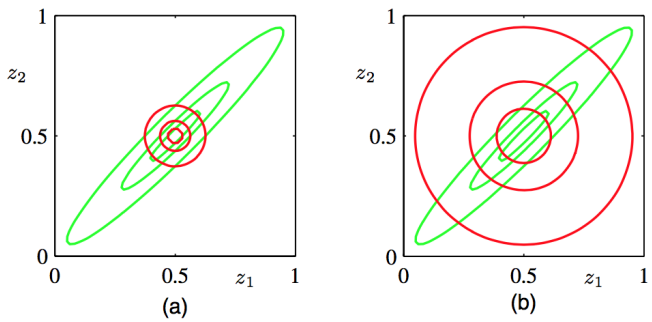

Figure 10.2 Comparison of the two alternative forms for the Kullback-Leibler divergence. The green contours corresponding to 1, 2, and 3 standard deviations for a correlated Gaussian distribution $p(\mathbf{z})$ over two variables z_1 and z_2, and the red contours represent the corresponding levels for an approximating distribution $q(\mathbf{z})$ over the same variables given by the product of two independent univariate Gaussian distributions whose parameters are obtained by minimization of (a) the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$, and (b) the reverse Kullback-Leibler divergence $\mathrm{KL}(p\|q)$.

Figure 2: (Bishop, 2006, pg.468)

The left figure is estimated using forward $\mathrm{KL}(q||p)$, while the right figure is estimated using reverse $\mathrm{KL}(p||q)$. The two methods are explained in detail here[3].

[3] `https://wiseodd.github.io/techblog/2016/12/21/forward-reverse-kl/` — from Kristiadi (2018)

Forward KL weights the difference between $p(z)$ and $q(z)$ by $p(z)$. As a result, $q(z) > 0$ anywhere $p(z) > 0$. This is called zero avoiding. In general, it's unusual to have strictly zero mass in a posterior, so typically only an approximate version of this issue arises. When forward KL is applied to the multivariate normal example, the variance is controlled by the direction of smallest variance of $p(z)$.

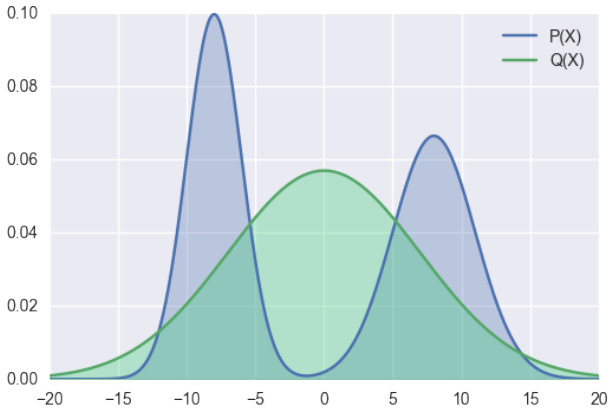

Figure 3: Forward KL in Kristiadi (2018)

Reverse KL is essentially the opposite. The difference is weighted by $q(z)$, resulting in some portion of $p(z)$ not being approximated by $q(z)$. This is called zero forcing. Again, the exact zero case rarely occurs in practice. In the multivariate normal example, the variance of $q(z)$ is controlled by the direction of largest variance of $p(z)$.

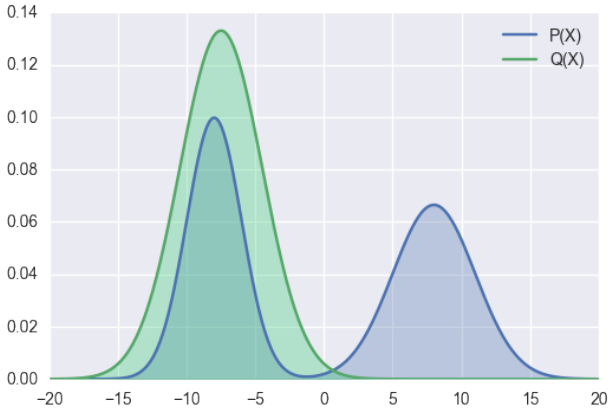

Figure 4: Reverse KL in Kristiadi (2018)

(Bishop, 2006, pg.469) shows another view of multimodality using contours of an overlapping mixture of two Gaussians. Forward KL($p \parallel q$) averages over the two Gaussians in a. Reverse KL($q \parallel p$) finds one of the modes, as shown in b and c.

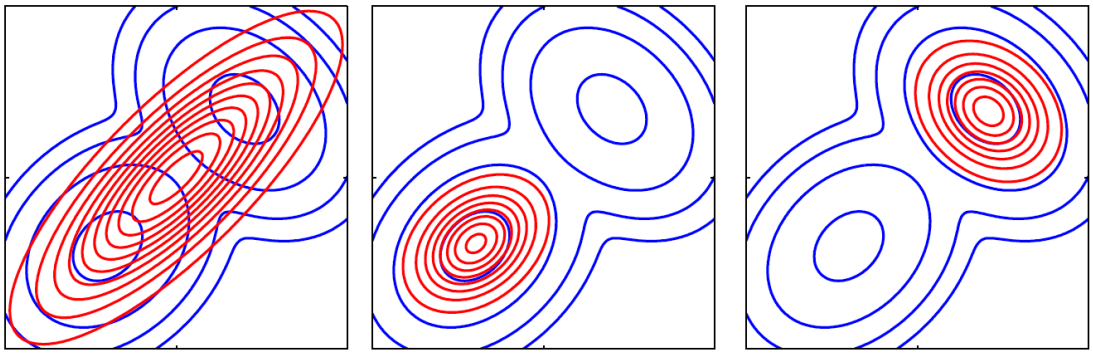

Figure 10.3 Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$. (b) As in (a) but now the red contours correspond to a Gaussian distribution $q(\mathbf{Z})$ found by numerical minimization of the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$. (c) As in (b) but showing a different local minimum of the Kullback-Leibler divergence.

Figure 5: (Bishop, 2006, pg.469)

We refer the reader to Bishop (2006) section 10.2 for an in-depth derivation of variational mixtures of Gaussians. But note that in Turner and Sahani we see that this compactness isn't always the case. See figure below.

Also note that we are seeing examples in class of cases where we can at least locally minimize forward KL. Reverse KL is more challenging due to the integration over the unknown posterior.

5 The vEM Algorithm

The EM algorithm can be understood as coordinate descent on the free energy, given by

$$F(q(z), \theta) = \log p(x \mid \theta) - \mathrm{KL}(q(z) \,\|\, p(z \mid x, \theta) \tag{16}$$

Alternate maximizations between q and θ will eventually find the parameters that maximizes the likelihood. However, maximizing over q is often intractable: for example, the analytic form of $p(z \mid x)$ is not available; or the number latent variables may be too large.

The varational EM (vEM) approach instead optimizes $q \mapsto F(q, \theta)$ over a restricted class Q of posteriors. For example, Q may be a parameterized class of posteriors

$$Q = \{q_\theta : \theta \in \Theta\}. \tag{17}$$

Alternatively, Q be the class of distributions that factor across disjoint sets of latent variables

$$Q = \left\{ q : q = \prod_{1 \le i \le I} q_i(z_{C_i}) \right\} \tag{18}$$

where $C_1, \ldots, C_I$ is a fixed partition of the latent variables. In this "mean field" approach, maximizing the free energy subject to

$$\int q_i(z_{C_i}) dz_{C_i} = 1, \quad \text{for each } i = 1, \ldots, I \tag{19}$$

leads to the optimality conditions

$$q_i(z_{C_i}) \propto \exp(\mathbf{E}_{-i} \log p(Z, X \mid \theta)) \tag{20}$$

where the expectation $\mathbf{E}_{-i}$ is taken over the product distribution $\prod_{j \neq i} q_j(z_{C_{-i}})$.

To gain intuition about how this works (see Bishop, pg. 468 figure above), let's consider approximating a correlated Gaussian with independent Gaussians. Suppose

$$Z \sim \mathsf{Normal}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \right) \tag{21}$$

The mean field approximation to (the distribution of) Z is a factored distribution to (Z_1, Z_2). Referencing the previous section, we see that

$$Z_1 \sim \mathsf{Normal}(m_1, \Lambda_{11}^{-1}), \quad \text{where} \quad m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbf{E}\, Z_2 - \mu_2). \tag{22}$$

Likewise, $Z_2 \sim \mathsf{Normal}(m_2, \Lambda_{22}^{-1})$, where m_2 is defined similarly.

6 Two Problems with vEM

The two problems with vEM pointed out by Turner and Sahani (2011) are compactness and bias. In particular, variational EM fails to propagate uncertainty across time in time series models (the paper demonstrates this with a Gaussian factor model).

6.1 Bias

The following figure from Turner and Sahani (2011) shows the behavior of EM and vEM, highlighting bias in particular. To provide a more concise summary of the long caption:

In the panels in (A), the left side is exact EM; the right variational EM. "Each update consists of an E-Step, which moves vertically to the optimal setting of μ_q, and an M-Step, which moves horizontally to the optimal setting of σ_y^2". In the bottom right, we see that variational EM (black) doesn't quite reach the lower bound on optimal free energy (grey).

In the panels in (B), the black line is the log-likelihood at the current parameters. The grey line is the free energy, which becomes tight to the log-likelihood after the E-step. The grey vertical line is the optimal q found in the M-step. The light grey is from the previous step. At iteration 10, the optimal free energy is clearly to the right of the optimal log-likelihood. Therefore, variational EM is biased to where the variational approximation is tightest.

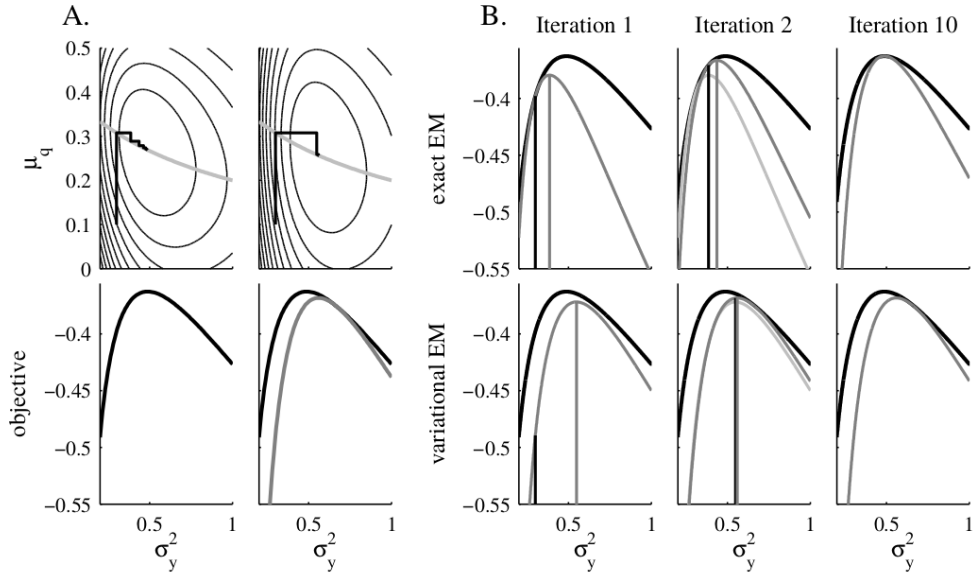

Figure 6: Figure 1.1 in Turner and Sahani (2011).

6.2 Compactness

The authors argue that while natural intuition dictates that "the free-energy should be as tight to the log-likelihood as possible" (Turner and Sahani, 2011, pg.9)),"from the perspective of learning it is more important to be equally tight everywhere. In other words, it is more important for the KL-term to be as parameter-independent as possible." (Turner and Sahani, 2011, pg.10)

They show this by comparing mean field approximation (MF), which has the loosest bound, with factorization across chains (FC) and factorization over time (FT). Because of explaining away or temporal correlation, FC and FT can over-estimate or under-estimate the variance and weights. MF provides an average of the two, so it is closest to the actual distribution.

However, tightness no longer gives the best approximation for larger dimensions or longer sequences in time. The authors' results depend on the particular model being estimated, and few general results or rules of thumb are given.

They explore compactness and entropy when estimating a mixture of two Gaussians in the figure below:

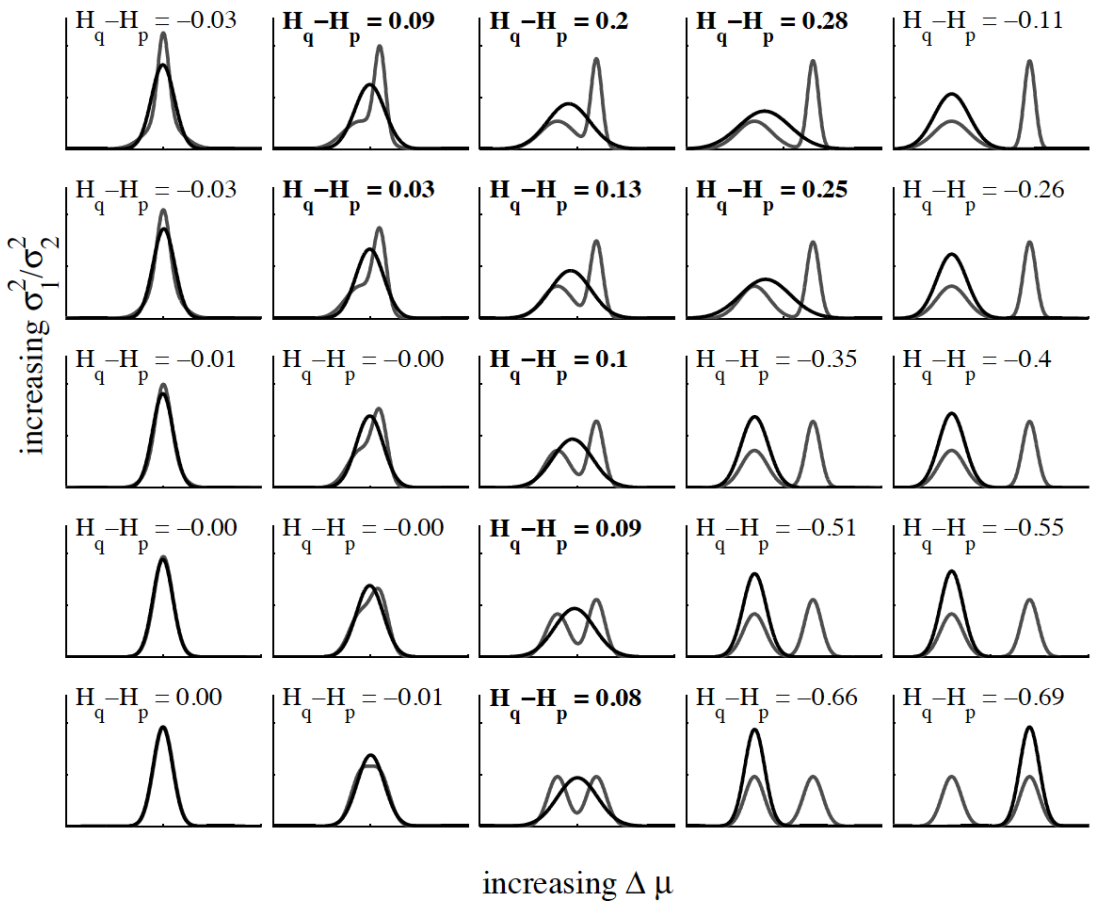

Figure 7: Figure 1.2 in Turner and Sahani (2011).

The x-axis is the difference between the two means. Going right, the two peaks become farther apart. As the difference increases, the approximation is more likely to match only one of the distributions. The y-axis is the ratio between the two standard deviations. Going up, the left peak becomes more spread out compared to the right peak. The unimodal estimations is more likely to match the original Gaussian with larger variance. The authors bolded the positive entropy because in those subplots, the estimated unimodal

distribution has higher entropy than the two individual peaks combined. However, because the variational approximation tends to be more compact, the mean field approximation is over-confident precisely when its estimation is poorest.

7 Class Discussions

One discussion that came up in class was the trade-off between compactness and more efficient computation. For example, even though some regions actually have zero probability, it is better to choose a posterior without zero mass. The computation would be more efficient.

The lecturers also compared variational approximation with maximum a posteriori (MAP) and Markov Chain Monte Carlo (MCMC). Variational approximation is unable to propagate uncertainty through time, while and MAP does not even try to given uncertainty estimates. MCMC can be slower at computation than variational inference.

Lastly, we can go full circle back to the reflection question Professor Broderick gave us. Quoting her note on Piazza:

> This paper shows an example where a larger class of nice distributions doesn't necessarily give "better" results, where here "better" refers to better estimates of the exact posterior mean. But conversely, of course there can be cases where a larger class of nice distributions does give "better" results. (E.g., an extreme case of this is choosing Q to be all distributions.) The real question I hope everyone is asking themselves is: so how do I know what to do in practice? Which Q class should I use, and how would I know going into a problem what to do?

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Kristiadi, A. (2018). KL divergence: Forward vs reverse?

Morey, R. D. (2012). Completing the square.

Turner, R. E. and Sahani, M. (2011). *Two problems with variational expectation maximisation for time series models*, pages 104–124. Cambridge University Press.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.