Counterexamples to variational free energy compactness folk theorems

R.E. Turner, P. Berkes, M. Sahani, and D.J.C. Mackay

July 9, 2008

Abstract

Suppose we want to approximate some complicated distribution $p(X)$ by a simpler distribution $q(X)$. The approximating distribution can, for example, have a simpler parameteric form (e.g. Gaussian), or capture simpler dependencies (e.g. factored). One way of choosing the best approximation within some class, is to minimise the variational free-energy. It is then often the case that the approximating distribution, $q(X)$ is more compact than the true distribution. A natural question is whether there in fact a theorem which says the "optimized $q(X)$ is always more compact than $p(X)$". Or can we find a counterexample? In this note we provide several such couterexamples which indicate that the compactness folk theorem is a useful rule of thumb, rather than a law of the cosmos.

1 A first compactness conjecture and a counterexample

Consider a joint distribution over two binary latent variables x_1 and x_2. Let the joint distribution be parameterised via an energy, so that $P(x_1, x_2) = \frac{1}{Z}\exp(-E(x_1, x_2))$ and let the energy function be given by,

$$E(x_1, x_2) = \begin{cases} -\epsilon & x_1 = x_2 \\ 0 & x_1 \neq x_2. \end{cases} \tag{1}$$

Consider separable approximating distributions, $P(x_1, x_2) \approx Q(x_1, x_2) = Q_1(x_1)Q_2(x_2)$ where $Q_i(x_i) = \Delta_i^{x_i}(1-\Delta_i)^{1-x_i}$. In order to determine the "best" factored approximating distribution we minimise the Free-Energy (F) with respect to Δ_1 and Δ_2, where

$$F(Q) = \langle E \rangle_Q - S(Q). \tag{2}$$

The intuition behind this example is that if ϵ is a little larger than 0 then the true distribution is almost uniform. It then seems plausible that optimal approximating distribution will be precisely uniform ($\Delta_1 = \Delta_2 = \frac{1}{2}$), and therefore it will have a larger entropy than the true distribution.

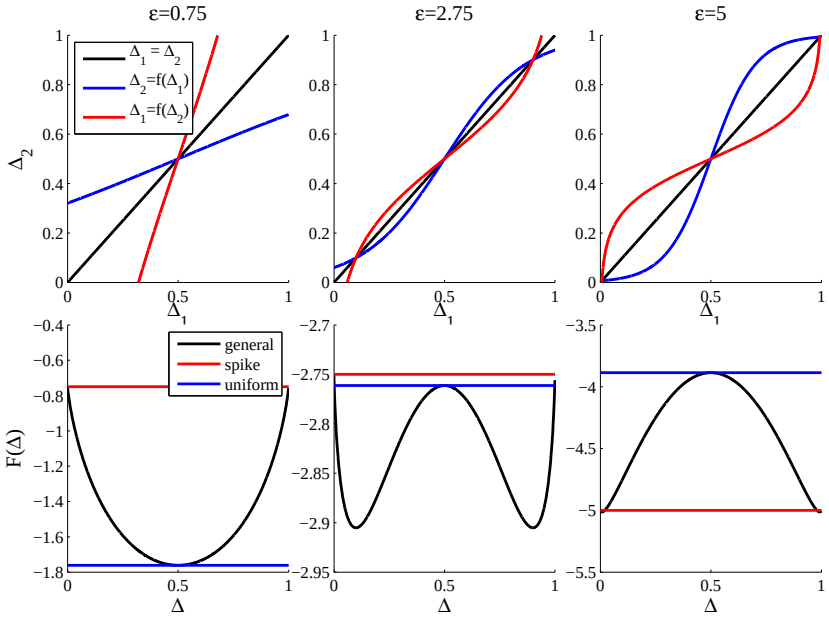

Figure 1: Top Row: Optima constraints. Bottom Row: Free-Energies

We now have to check whether this intuition is correct. To do this we need to calculate the free-energy which is composed of two terms. The first is the average energy,

$$\langle E \rangle_Q = -\epsilon \left(\Delta_1 \Delta_2 + (1 - \Delta_1)(1 - \Delta_2) \right), \tag{3}$$

and the second, the entropy,

$$S(Q) = S(Q_1) + S(Q_2), \tag{4}$$

$$S(Q_i) = -\Delta_i \log \Delta_i - (1 - \Delta_i) \log(1 - \Delta_i). \tag{5}$$

At the optima of the free-energy the derivative with respect to Δ_1 and Δ_2 are both zero which yields following pair of conditions,

$$\Delta_1 = f(\Delta_2), \quad \Delta_2 = f(\Delta_1) \tag{6}$$

$$f(z) = \frac{1}{1 + \exp(\epsilon(1 - 2z))}. \tag{7}$$

The two lines of points satisfying both of these constraints can be seen in the top row of Fig.1 for different settings of the energy ϵ. As f is a monotonically increasing function, these two conditions can only be satisfied on the line $\Delta_1 = \Delta_2$ and so the optimal Q must be symmetric.

There are a number of ways to see this. My current favourite is to assume that there is a point where the above conditions are satisfied for which $\Delta_1 \neq \Delta_2$. Call this point $\mathbf{\Delta}^{(1)} = [\Delta_1^{(1)}, \Delta_2^{(1)}]$. Due to symmetry there must also then be a solution at $\mathbf{\Delta}^{(2)} = [\Delta_1^{(2)}, \Delta_2^{(2)}] = [\Delta_2^{(1)}, \Delta_1^{(1)}]$. Assume (wlg) that $\Delta_1^{(1)} < \Delta_1^{(2)}$, then this means $\Delta_2^{(1)} > \Delta_2^{(2)}$. Alternatively, we can reason that both solutions have to lie on the same monotonically increasing function (f) and therefore if $\Delta_1^{(1)} < \Delta_1^{(2)}$ then $\Delta_2^{(1)} < \Delta_1^{(2)}$. This is a contradiction and so $\Delta_1^{(1)} = \Delta_2^{(1)}$.

Now we have reduced things to a one-dimensional problem and so it is easy to visualise. In particular the free-energy is,

$$F(\Delta) = -\epsilon(1 + 2\Delta^2 - 2\Delta) + 2\Delta \log \Delta + 2(1 - \Delta)\log(1 - \Delta), \tag{8}$$

and this has been plotted for the same three energies ϵ as before in the lower row of Fig.1. For $0 < \epsilon < 2$ the minimum corresponds to the unform distribution $\Delta = \frac{1}{2}$ [1]. This is the counterexample we were looking for and so we are home and dry. For $\epsilon > 2$ the approximating distribution starts to put more mass on one or other of the settings, so that as $\epsilon \to \infty$ the approximating distribution tends to a spike $Q(x_1) = Q(x_2) \to [1, 0]$ or $Q(x_1) = Q(x_2) \to [0, 1]$. In this case the approximation is more compact than the posterior ($H(Q) = 0$ versus $H(P) = \log 2$).

2 A counterexample to an alternative compactness folk theorem

To recap, so far we know that when using variational approximations to make factored approximations to a) correlated Gaussians, the approximation is always compact, but that for b) correlated discrete variables, the approximation is sometimes compact and sometimes not.

Here we turn to different example when we factorise between two sets of latent variables, where each set has a different form. We consider a simple linear model with non-Gaussain (Student-t) latent variables (ICA). To aid computation we can replace the Student-t prior with an equivalent hierarchical prior in which we first draw a precision s_i from a Gamma distribution and then draw the source x_i from a Gaussian of that precision. To keep things simple to visualise, the model will contain two sources and the observations will be formed

[1] The optima in the new free energy obey $\Delta = \frac{1}{1+\exp(\epsilon(1-2\Delta))}$. This always is true at $\Delta = \frac{1}{2}$ and so there is always a maximum or minimum here. For $\epsilon < 2$ there are no other optima. This can be seen by computing the derivative, $\frac{df}{d\Delta} = \frac{\Delta\epsilon}{2\cosh(\epsilon(\Delta - \frac{1}{2}))}$, which is always less than one if $\epsilon < 2$. This means there are no other solutions in this regime. From the free-energy we can see that the point $\Delta = \frac{1}{2}$ is a maximum. For $\epsilon > 2$ there are two maxima, and a minimum at $\Delta = \frac{1}{2}$. For a closer look at the pitch-fork bifurcation as ϵ moves from below to above 2 see Fig.2.

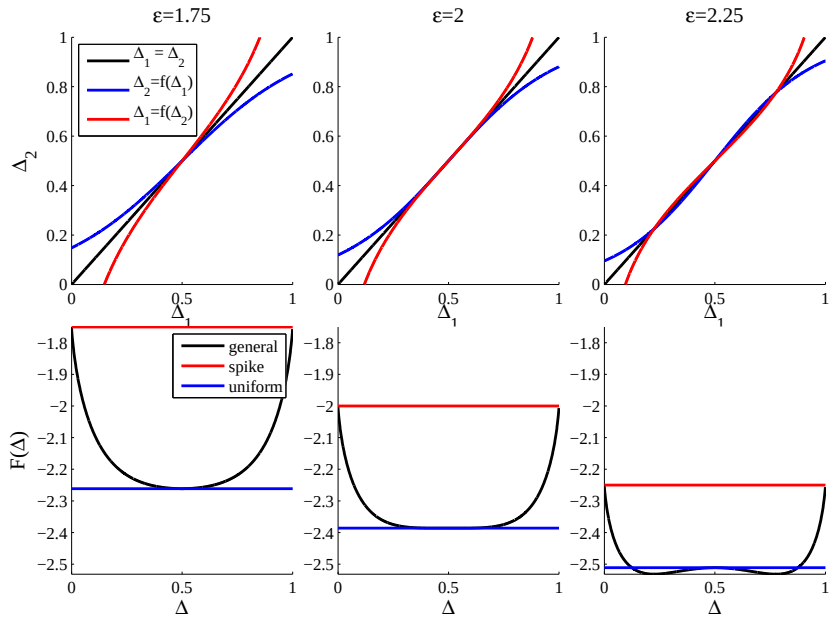

Figure 2: The pitch-fork bifurcation. Top Row: Optima constraints. Bottom Row: Free-Energies

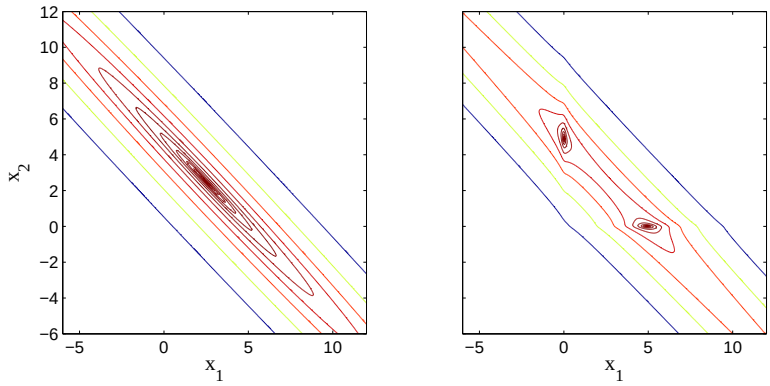

Figure 3: Contour plots of a) the true posterior distribution over sources ($p(\boldsymbol{x}|y)$)(left), and b) the variational approximation ($q(\boldsymbol{x})$) (right). Parameter settings: $y = 5$, $\sigma_y^2 = 0.1$, $a = 1.01$.

by adding the two sources and adding Gaussian noise,

$$p(s_i|a,b) = \text{Gamma}(s_i; a, b) = \frac{1}{Z(a,b)} s_i^{a-1} \exp(-b s_i), \tag{9}$$

$$p(x_i|s_i) = \text{Norm}(x_i; 0, s_i), \tag{10}$$

$$p(y|x_1, x_2, s_y) = \text{Norm}(y; x_1 + x_2, s_y). \tag{11}$$

$b = a - 1$ ensures the marginal variance of the latents under the prior is unity.

The posterior over precisions and sources ($p(\boldsymbol{x}, \boldsymbol{s}|y)$) is complex (e.g. see Fig.3, left) and we approximate it using a factored distribution ($q(\boldsymbol{x}, \boldsymbol{s}) = q(\boldsymbol{x})q(\boldsymbol{s})$). The average log-joint is,

$$\begin{aligned}\log p(y, x_1, x_2, s_1, s_2) = &-\frac{s_y}{2}(y - x_1 - x_2)^2 \\ &+ \frac{1}{2}\left(\log s_1 + \log s_2\right) - \frac{s_1}{2}x_1^2 + \frac{s_2}{2}x_2^2 \\ &+ (a-1)\left(\log s_1 + \log s_2\right) - b(s_1 + s_2),\end{aligned} \tag{12}$$

and so the optimal updates are,

$$q(\boldsymbol{x}) = \text{Norm}(\boldsymbol{x}; \mu, \Sigma), \quad \Sigma^{-1} = \begin{bmatrix} s_y + \langle s_1 \rangle & s_y \\ s_y & s_y + \langle s_2 \rangle \end{bmatrix}, \; \mu = y s_y \Sigma \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{13}$$

$$q(s_i) = \text{Gamma}(s_i; a_i, b_i), \quad a_i = a - \frac{1}{2}, \quad b_i = b + \frac{1}{2}\langle x_i^2 \rangle. \tag{14}$$

By iterating these updates until the Free-energy converges we can find the (locally) optimal variational approximation (e.g. see Fig.3, right).

We now consider the following version of the compactness theorem:

$$H\left(\int p(\boldsymbol{x},\boldsymbol{s})d\boldsymbol{s}\right) \geq H\left(\int q(\boldsymbol{x},\boldsymbol{s})d\boldsymbol{s}\right) = H(q(\boldsymbol{x})). \tag{15}$$

i.e. the entropy of the marginal of the posterior distribution is greater than the entropy of the corresponding factor in the variational approximation. (The first example did *not* violate this theorem as the marginals of the true distribution were always uniform in that case.) In the current case the entropy of the variational factor is easy to compute as it is just the entropy of a bivarate Gaussian,

$$H(q(\boldsymbol{x})) = 1 + \frac{1}{2}\log\det 2\pi\Sigma. \tag{16}$$

The entropy of the posterior is harder to compute and so we have to turn to approximation methods. Ideally we would like to upper-bound this entropy and then show that this upper bound is smaller than that of the variational approximation. However, we will do the easy thing and approximate the entropy by gridding up the space. As a check, we approximate the entropy of the variational approximation and compare it to the 'true' value computed analytically. By doing this for a number of different grid-sizes we get an idea of the error-bars too. From Fig.4 it is clear that the approximating distribution is less compact using the new compactness-criteria.

3 Fitting one dimensional distributions using the variational free-energy

In this section we consider approximating a mixture of Gaussians ($p(x|\theta) = \sum_{k=1}^{K}\pi_k \mathrm{Norm}(x;\mu_k,\sigma_k^2)$) and a mixture of student-t distributions ($p(x|\theta) = \sum_{k=1}^{K}\pi_k \mathrm{student}(x;\mu_k,\alpha_k,\beta_k^2)$) with a Gaussian ($q(x) = \mathrm{Norm}(x;\mu_q,\sigma_q^2)$) by minimising the variational KL $\arg\min_{\mu_q,\sigma_q^2}\mathrm{KL}(q(x)||p(x))$.

As figs.5 and 6 show, the approximation can have less entropy or greater entropy than the true distribution depending on the true distribution.

Acknowledgments

Supported by the Gatsby Charitable Foundation.

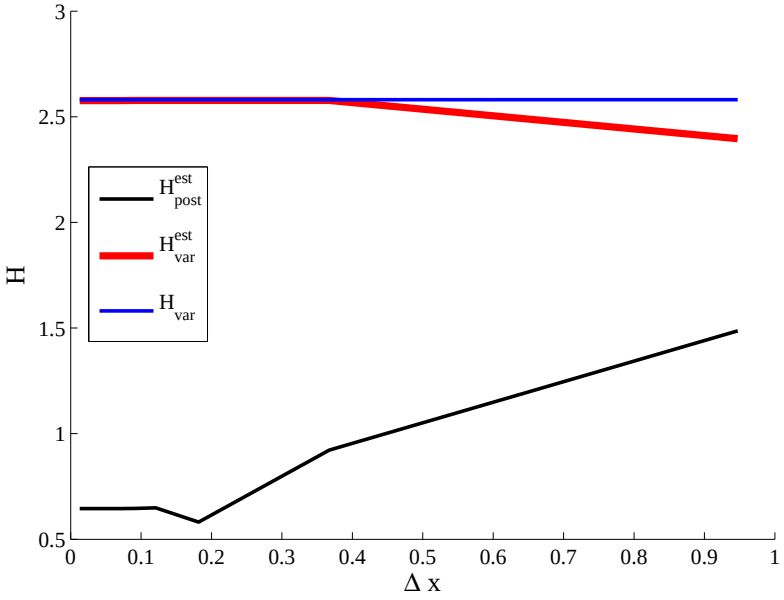

Figure 4: Entropies of the posterior distributions in Fig.3 as a function of the grid resolution Δx. The differences between the estimated variational entropy and the true entropy are small for the fine grid size and much smaller than the difference between the variation and posterior entropies.

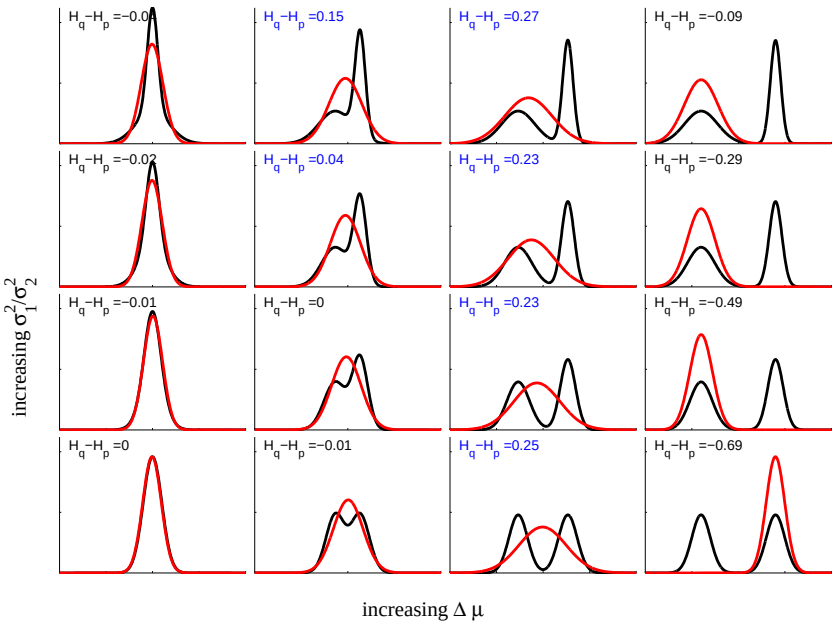

Figure 5: Approximating a Mixture of two Gaussians (black line) with a Gaussian (red line). The parameters of the mixture were set so that each component has equal weight ($\pi_1 = \pi_2 = \frac{1}{2}$). The difference between the means ($\mu_1 - \mu_2$) increases left to right, and the ratio of the variances (σ_1^2/σ_2^2) increases from bottom to top. The bottom left is therefore a mixture of two Gaussians with the same mean and variance and this is another Gaussian. The approximation is therefore exact and the entropy difference is therefore zero. In general the entropy of the approximation can be less than (black font) or greater than (blue font) the true entropy.

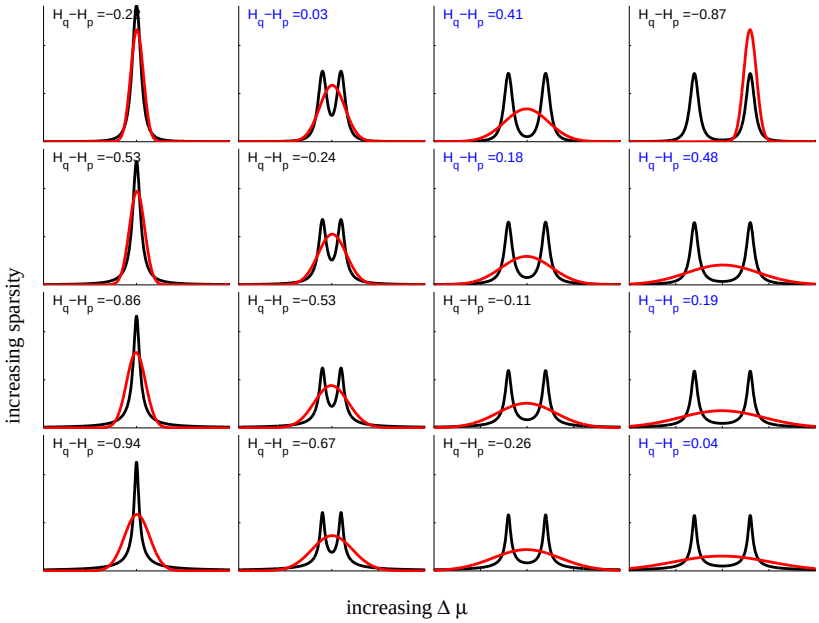

Figure 6: Approximating a Mixture of two student-t distributions (black line) with a Gaussian (red line). The parameters of the mixture were set so that each component has equal weight ($\pi_1 = \pi_2 = \frac{1}{2}$). The sparseness for all examples is very high and it increases from top to bottom. The difference in the means increases left to right. The entropy difference is shown above each plot and in general the entropy of the approximation can be less than (black font) or greater than (blue font) the true entropy. For less sparse mixtures of student t distributions, something like the top right plot is typical in that the approximation matches a single mode and it is compact.