Variational objectives and KL Divergence

The Laplace approximation fitted a Gaussian distribution to a parameter posterior by matching a mode and the curvature of the log posterior at that mode. We saw that there are failure modes when the shape of the distribution at the mode is misleading about the over-all distribution.

Variational methods fit a target distribution, such as a parameter posterior, by defining an optimization problem. The ingredients are:

- A family of possible distributions $q(\mathbf{w};\alpha)$.
- A variational cost function, which describes the discrepancy between $q(\mathbf{w};\alpha)$ and the target distribution (for us, the parameter posterior).

The computational task is to optimize the *variational parameters* (here α).[1]

For this course, the variational family will always be Gaussian:

$$q(\mathbf{w};\ \alpha = \{\mathbf{m}, V\}) = \mathcal{N}(\mathbf{w};\ \mathbf{m},\ V).$$

So we fit the mean and covariance of the approximation to find the best match to the posterior according to our variational cost function. Although we won't consider other cases, the variational family doesn't have to be Gaussian. The variational distribution can be a discrete distribution if we have a posterior distribution over discrete variables.

Kullback–Leibler Divergence

The Kullback–Leibler divergence, usually just called the KL-divergence, is a common measure of the discrepancy between two distributions:

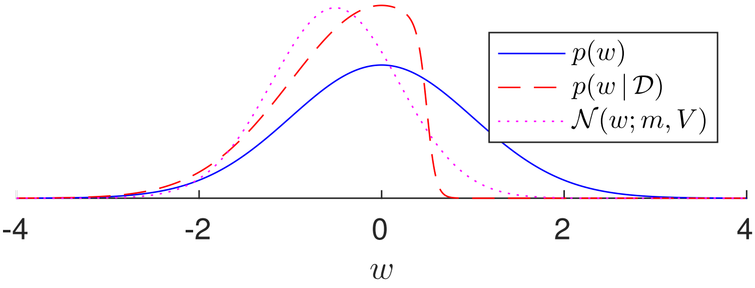
$$D_{\mathrm{KL}}(p\,||\,q) = \int p(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})}\, \mathrm{d}\mathbf{z}.$$

The KL-divergence is non-negative, $D_{\mathrm{KL}}(p\,||\,q) \geq 0$, and is only zero when the two distributions are identical.

The divergence doesn't satisfy the formal criteria to be a *distance*, for example, it isn't symmetric: $D_{\mathrm{KL}}(p\,||\,q) \neq D_{\mathrm{KL}}(q\,||\,p)$.

Minimizing $D_{\mathrm{KL}}(p\,||\,q)$

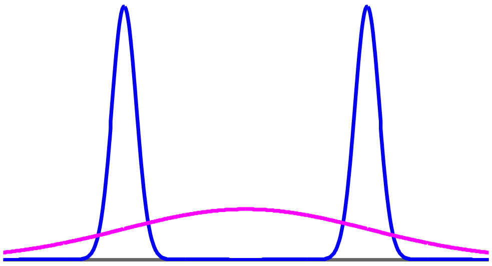
To minimize $D_{\mathrm{KL}}(p\,||\,q)$ we set the variational parameters $\mathbf{m}$ and V to match the mean and covariance of the target distribution p. The illustration below shows an example from the notes on Bayesian logistic regression. The Laplace approximation is poor on this example: the mode of the posterior is very close to the mode of the prior, and the curvature there is almost the same as well. The Laplace approximation sets the approximate posterior so close to the prior distribution (blue solid line below) that I haven't plotted it. A different Gaussian fit (magenta dotted line), with the same mean m and variance V as the posterior distribution, is a better summary of where the plausible parameters are than the Laplace approximation:



[1] The textbooks often avoid parameterizing q in their presentations of variational methods. Instead they describe the optimization problem as being on the distribution q itself, using the calculus of variations. We don't need such a general treatment in this course.

Optimizing $D_{\mathrm{KL}}(p(\mathbf{w}\,|\,\mathcal{D})\,||\,q(\mathbf{w};\alpha))$ tends to be difficult. The cost function is an expectation under the complicated posterior distribution that we are trying to approximate, and we usually can't evaluate it.

Even if we could find the mean and covariance of our distribution (approximating them would be possible) the answer may not be sensible. Matching the mean of a bimodal distribution will find an approximation centred on implausible parameters:



Our predictions are not likely to be sensible if we mainly use parameters that are not plausible given the data.

Minimizing $D_{\mathrm{KL}}(q\,||\,p)$

Most variational methods in Machine Learning minimize $D_{\mathrm{KL}}(q(\mathbf{w};\alpha)\,||\,p(\mathbf{w}\,|\,\mathcal{D}))$, partly because we are better at optimizing this cost function. (There are also other sensible variational principles, but we won't cover them in this course.) Minimizing the KL-divergence this way around encourages the fit to concentrate on plausible parameters:

$$\begin{aligned} D_{\mathrm{KL}}(q(\mathbf{w};\alpha)\,||\,p(\mathbf{w}\,|\,\mathcal{D})) &= \int q(\mathbf{w};\alpha)\log\frac{q(\mathbf{w};\alpha)}{p(\mathbf{w}\,|\,\mathcal{D})}\,\mathrm{d}\mathbf{w}\\ &= -\int q(\mathbf{w};\alpha)\log p(\mathbf{w}\,|\,\mathcal{D})\,\mathrm{d}\mathbf{w} + \underbrace{\int q(\mathbf{w};\alpha)\log q(\mathbf{w};\alpha)\,\mathrm{d}\mathbf{w}}_{\text{negative entropy, } -H(q)} \end{aligned}$$

To make the first term small, we avoid putting probability mass on implausible parameters. As an extreme example, there is an infinite penalty for putting probability mass of q on a region of parameters that are impossible given the data. The second term is the negative *entropy* of the distribution q.[2] To make the second term small we want a high entropy distribution, one that is as spread out as possible.

Minimizing this KL-divergence usually results in a Gaussian approximation that finds one mode of the distribution, and spreads out to cover as much mass in that mode as possible. However, the distribution can't spread out to cover low probability regions, or the first term would grow large. See Murphy Figure 21.1 for an illustration.

If we substitute the expression for the posterior from Bayes' rule,

$$p(\mathbf{w}\,|\,\mathcal{D}) = \frac{p(\mathcal{D}\,|\,\mathbf{w})\,p(\mathbf{w})}{p(\mathcal{D})},$$

into the KL-divergence, we get a spray of terms:

$$D_{\mathrm{KL}}(q\,||\,p) = \underbrace{\mathbb{E}_q[\log q(\mathbf{w})] - \mathbb{E}_q[\log p(\mathcal{D}\,|\,\mathbf{w})] - \mathbb{E}_q[\log p(\mathbf{w})]}_{J(q)} + \log p(\mathcal{D}).$$

The first three terms, equal to $J(q)$ in Murphy, depend on the variational distribution (or its parameters), so we minimize these terms. The final term, $\log p(\mathcal{D})$ is the log-marginal

[2] H is the standard symbol for entropy, and has nothing to do with a Hessian (also H; sorry!).

likelihood (also known as the "model evidence"). Knowing that the KL-divergence is non-negative gives us a bound on the marginal likelihood:

$$D_{\mathrm{KL}}(q\,||\,p) \geq 0 \;\Rightarrow\; \log p(\mathcal{D}) \geq -J(q).$$

Thus, fitting the variational objective is optimizing a lower bound on the log marginal likelihood. Recently "the ELBO" or "Evidence Lower Bound" has become a popular name for $-J(q)$.[3]

Optimization methods for $D_{\mathrm{KL}}(q\,||\,p)$

The literature is full of clever (non-examinable) iterative ways to optimize $D_{\mathrm{KL}}(q\,||\,p)$ for different models.

Could we use standard optimizers? The hardest term to evaluate is:

$$\mathbb{E}_q[\log P(\mathcal{D}\,|\,\mathbf{w})] = \sum_{n=1}^{N} \mathbb{E}_q[\log P(y^{(n)}\,|\,\mathbf{x}^{(n)}, \mathbf{w})],$$

which is a sum of (possibly simple) integrals. In the last few years variational inference has become dominated by stochastic gradient descent, which updates the variational parameters using unbiased approximations of the variational cost function and its gradients.

Reading

Reading for variational inference in Murphy: Sections 21.1–21.2.

Overview of Gaussian approximations

Laplace approximation:

- Straightforward to apply
- 2nd derivatives ⇒ certainty of parameters
- Incremental improvement on MAP estimate
- Approximation of marginal/model likelihood

Variational methods:

- Optimization: fit variational parameters of q (not $\mathbf{w}$!)
- Usually $D_{\mathrm{KL}}(q\,||\,p)$, not $D_{\mathrm{KL}}(p\,||\,q)$
- Bound of marginal/model likelihood
- Optimization: traditionally harder to apply. Now becoming automatic as well.

The KL divergence appears in other contexts

In 2016 I was asked: *"In MLP, we use the KL divergence cost function to train neural nets. More specifically, we use the divergence between the 1-0 labels and the model output… Does it count as using a variational method?"*

The answer is no. One principle for fitting a single setting of the weights is to estimate the KL between the true label distribution and the model's distribution. A Monte Carlo estimate of this KL uses samples from the true distribution found in the training data. Optimizing this objective is the same as maximum likelihood training. Variational methods (in the context of Bayesian methods) fit a posterior distribution over the weights with a distribution.

[3] Variational inference was originally inspired by work in statistical physics, and with that analogy, $+J(q)$ is also called the variational free energy.

Non-examinable: Information theory

The KL-divergence gives the average storage wasted by a compression system that encodes a file based on model q instead of the optimal distribution p. MacKay's book is the place to read about the links between machine learning and compression.

Non-examinable: $D_{\mathrm{KL}}(p||q)$ and moment matching

One question that you may have is: "Why does minimizing $D_{\mathrm{KL}}(p||q)$ with Gaussian q lead to matching the mean and covariance of p?" If you substitute in a Gaussian q into the formula, differentiate and set to zero, this result will drop out (eventually). However, the working for a multivariate Gaussian will be messy. It's easier to show a more general result. I give the (non-examinable) details here.

I'm going to match two distributions over $\mathbf{w}$. My variational parameters, defining q will be θ. We choose q to be the member of an *exponential family*:

$$q(\mathbf{w}) = \frac{1}{Z(\theta)} \exp(\theta^\top \Phi(\mathbf{w})),$$

where $Z = \int \exp(\theta^\top \Phi(\mathbf{w}))\,\mathrm{d}\mathbf{w}$, and $\Phi(\mathbf{w})$ is a vector of statistics, defining the approximating family. If we choose Φ to contain each variable w_d, and the product of each pair of variables, $w_c w_d$, then we are fitting a Gaussian distribution. Substituting our approximating family into the variational objective:

$$D_{\mathrm{KL}}(p\,||\,q) = \int p(\mathbf{w}) \log \frac{p(\mathbf{w})}{q(\mathbf{w})}\,\mathrm{d}\mathbf{w}$$

which up to a constant with respect to θ is[4]:

$$-\int p(\mathbf{w})\theta^\top \Phi(\mathbf{w})\,\mathrm{d}\mathbf{w} + \log Z(\theta).$$

We differentiate the KL wrt θ, to get

$$-\int p(\mathbf{w})\Phi(\mathbf{w})\,\mathrm{d}\mathbf{w} + \frac{1}{Z(\theta)} \int \Phi(\mathbf{w}) \exp(\theta^\top \Phi(\mathbf{w}))\,\mathrm{d}\mathbf{w}.$$

The gradient is a difference of expectations, one under p and the other under q. The gradient is zero when the expectations are equal:

$$\mathbb{E}_p[\Phi(\mathbf{w})] = \mathbb{E}_q[\Phi(\mathbf{w})].$$

That is, the minimum of the objective function (ok, I'd need to do slightly more work to prove this turning point is a minimum) is when the expected statistics defined by Φ all match in the two distributions. For a Gaussian, that means the mean and covariance match.

[4] A common mistake is to omit $Z(\theta)$, which is not a constant wrt θ.