Notes on some basic probability stuff

David Meyer

dmm@{1-4-5.net,uoregon.edu,brocade.com,...}

March 22, 2016

1 Introduction

Note well: There are likely to be many mistakes in this document. That said...

Much of what is described here follows from two simple rules:

$$\text{Sum Rule:} \qquad P(\mathcal{X}) = \sum_{y} P(\mathcal{X}, \mathcal{Y}) \tag{1}$$

$$\text{Product Rule:} \qquad P(\mathcal{X}, \mathcal{Y}) = P(\mathcal{X}|\mathcal{Y})P(\mathcal{Y}) \tag{2}$$

- The Sum Rule is sometimes called marginalization
- The Product Rule is part of the proof of the Hammersley-Clifford Theorem

Remember also that $\mathcal{X} = \{x_i\}_{i=1}^{|\mathcal{X}|}$, where each x_i is a realization of the random variable x[1]. Lets also say that the set Θ of probability distribution parameters can be used to explain the *evidence* $\mathcal{X}$. Then we say that the "manner in which the evidence $\mathcal{X}$ depends on the parameters Θ" is the *observation model*. The analytic form of the observation model is the likelihood $P(\mathcal{X}|\Theta)$.

[1] Each observation x_i is, in general, a data point in a multidimensional space.

2 Estimating the parameters Θ with Bayes' Theorem

Note that

$$P(\Theta, \mathcal{X}) = P(\mathcal{X}, \Theta) \tag{3}$$
$$P(\Theta, \mathcal{X}) = P(\Theta|\mathcal{X})P(\mathcal{X}) \tag{4}$$
$$P(\mathcal{X}, \Theta) = P(\mathcal{X}|\Theta)P(\Theta) \tag{5}$$
$$P(\Theta|\mathcal{X})P(\mathcal{X}) = P(\mathcal{X}|\Theta)P(\Theta) \tag{6}$$

Solving for $P(\Theta|\mathcal{X})$ we get Bayes' Theorem

$$P(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)P(\Theta)}{P(\mathcal{X})} \tag{7}$$
$$\tag{8}$$

Said another way

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}} \tag{9}$$

You might also see Bayes' Theorem written using the *Law of Total Probability*[2] which is sometimes written as follows:

$$P(A) = \sum_n P(A \cap B_n) \qquad \text{\# by the } \textit{Sum Rule} \text{ (Equation 1)} \tag{10}$$
$$= \sum_n P(A, B_n) \qquad \text{\# in the notation used in Equation 1} \tag{11}$$
$$= \sum_n P(A|B_n)P(B_n) \qquad \text{\# by the } \textit{Product Rule} \text{ (Equation 2)} \tag{12}$$

so that the posterior distribution $P(\mathcal{C}_1|\mathbf{x})$ for two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ given input vector $\mathbf{x}$ would look like

$$P(\mathcal{C}_1|\mathbf{x}) = \frac{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) + P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} \tag{13}$$

[2]The Law of Total Probability is a combination of the Sum and Product Rules

Interestingly, the posterior distribution is related to logistic regression as follows: First recall that the posterior $P(\mathcal{C}_1|\mathbf{x})$ is

$$P(\mathcal{C}_1|\mathbf{x}) = \frac{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1) + P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} \tag{14}$$

Now, if we set

$$a = \ln \frac{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)} \tag{15}$$

we can see that

$$P(\mathcal{C}_1|\mathbf{x}) = \frac{1}{1 + e^{-a}} = \sigma(a) \tag{16}$$

that is, the sigmoid function.

2.1 Maximum Likelihood Estimation (MLE)

Given all of that, for the MLE we seek the value of Θ that maximizes the likelihood $P(\mathcal{X}|\Theta)$ for our observations $\mathcal{X}$. Remembering that $\mathcal{X} = \{x_1, x_2, \ldots\}$ and that the x_i are iid, the value of Θ we seek maximizes

$$\prod_{x_i \in \mathcal{X}} P(x_i|\Theta) \tag{17}$$

Because of the product it is easier to use the log[3], we use the log likelihood $\mathcal{L}$:

$$\mathcal{L} = \sum_{x_i \in \mathcal{X}} \log P(x_i|\Theta) \tag{18}$$

and define $\hat{\Theta}_{ML}$ as follows:

$$\hat{\Theta}_{ML} = \underset{\Theta}{\operatorname{argmax}}\, \mathcal{L} \tag{19}$$

[3]and since log(x) is monotonically increasing it doesn't effect the argmax

The maximization is obtained by (calculus tricks):

$$\frac{\partial \mathcal{L}}{\partial \theta_i} = 0 \quad \forall \theta_i \in \Theta \tag{20}$$

Note finally that in a Generalized Linear Regression setting, we have

$$\eta = \mathbf{w}^T \mathbf{x} + b \tag{21}$$

$$p(y|\mathbf{x}) = p(y|g(\eta); \theta) \tag{22}$$

where $g(.)$ is an *inverse link function*, also referred to as an activation function. For example, if the link function is the logistic function, then the inverse link function $g(\eta) = \frac{1}{1+e^{-\eta}}$ and the negative log-likelihood $\mathcal{L}$ is

$$\mathcal{L} = -\log p(y|g(\eta); \theta) \tag{23}$$

2.2 Maximum a Posteriori (MAP) Estimation of Θ

Recall that

$$P(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)P(\Theta)}{P(\mathcal{X})} \tag{24}$$

We are seeking the value of Θ that maximizes $P(\Theta|\mathcal{X})$, so the solution can be stated as

$$\hat{\Theta}_{MAP} = \underset{\Theta}{\operatorname{argmax}}\, P(\Theta|\mathcal{X}) \tag{25}$$

$$= \underset{\Theta}{\operatorname{argmax}}\, \frac{P(\mathcal{X}|\Theta) \cdot P(\Theta)}{P(\mathcal{X})} \tag{26}$$

However, since $P(\mathcal{X})$ does not depend on Θ, we can write

$$\hat{\Theta}_{MAP} = \underset{\Theta}{\operatorname{argmax}}\, P(\mathcal{X}|\Theta) \cdot P(\Theta) \tag{27}$$

$$= \prod_{x_i \in \mathcal{X}} P(x_i|\Theta) \cdot P(\Theta) \tag{28}$$

If we again take the log, we get

$$\hat{\Theta}_{MAP} = \underset{\Theta}{\operatorname{argmax}} \left(\sum_{x_i \in \mathcal{X}} \log P(x_i|\Theta) + \log P(\Theta) \right) \tag{29}$$

2.3 Notes

- Both MLE and MAP are point estimates for Θ (contrast probability distributions)
- MLE notoriously overfits
- MAP allows us to take into account knowledge about the prior (which is a sort of a regularizer)
- Bayesian estimation, by contrast, calculates the full posterior distribution $P(\Theta|\mathcal{X})$

2.4 Bayesian Estimation

Recall that Bayesian estimation calculates the full posterior distribution $P(\Theta|\mathcal{X})$, where

$$P(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)\ P(\Theta)}{P(\mathcal{X})} \tag{30}$$

In this case, however, the denominator $P(\mathcal{X})$ cannot be ignored, and we know from the *sum* and *product* rules that

$$P(\mathcal{X}) = \int_{\Theta} P(\mathcal{X}, \Theta)\ d\Theta \tag{31}$$

$$= \int_{\Theta} P(\mathcal{X}|\Theta)\ P(\Theta)\ d\Theta \tag{32}$$

putting it all together we get

$$P(\Theta|\mathcal{X}) = \frac{P(\mathcal{X}|\Theta)\ P(\Theta)}{\int_{\Theta} P(\mathcal{X}|\Theta)\ P(\Theta)\ d\Theta} \tag{33}$$

If we want to be able to derive an algebraic form for the posterior $P(\Theta|\mathcal{X})$, the most challenging part will be finding the integral in the denominator. This is where the idea of *conjugate priors* and appoximate inference approaches (*Monte Carlo Integration* and *Variational Bayesian methods*[4]) are useful. Need to futher expand this...

[4]Variational Bayesian methods are a family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning. They are typically used in complex statistical models consisting of observed variables (usually termed "data") as well as unknown parameters and latent variables, with various sorts of relationships among the three types of random variables, as might be described by a graphical model.

3 Monte Carlo Integration

Suppose we have a distribution $p(\theta)$ (perhaps a posterior) the we want to sample quantities of interest from. To do this analytically, we need to take an integral of the form

$$I = \int_{\Theta} g(\theta)\, p(\theta)\, d\theta \tag{34}$$

where $g(\theta)$ is some function of θ (typically $g(\theta) = \theta$ (the mean), etc). Need a deeper analysis here (note to self), but the punchline is that you can estimate I using *Monte Carlo Integration* as follows: Sample M values (θ^i) from $p(\theta)$ and calculate

$$\hat{I}_M = \frac{1}{M} \sum_{i=1}^{M} g(\theta^i) \tag{35}$$

Note that this works fine if the samples from $p(\theta)$ are iid[5] but if not, we can use a Markov Chain to draw "slightly dependent" samples and depend on the *Ergodic Theorem* (see Section 8.2).

4 Variational Inference: Basic Setup

Consider a joint distribution of latent variables $\mathbf{z} = z_{1:M} = \{z_1, \ldots, z_M\}$ and observations $\mathbf{x} = x_{1:N}$ and

$$p(z, x) = p(z)p(x|z) \tag{36}$$

In Bayesian models, the latent variables help govern the distribution of the data. A Bayesian model draws the latent variables from the prior distribution $p(z)$ and then relates them to the observations through the likelihood $p(x|z)$. Inference in a Bayesian model comes down to conditioning on the data and computing the posterior $p(z|x)$. In complex Bayesian models, however, computing the posterior often requires approximate inference (because the posterior is intractable).

As mentioned in Section 2.4, Markov Chain Monte Carlo (MCMC) has been the primary paradigm for approximate inference over the past many decades. The basic procedure is straight forward: First, construct a Markov Chain on $\mathbf{z}$ whos stationary distribution $\pi = p(z|x)$, that is, the posterior distribution of interest. Then we sample the chain (for example, with a Gibbs sampler) for a long enough time to get independent samples from the stationary distribution (i.e., after it *mixes*). Finally, we approximate the posterior with an empirical estimate constructed from the collected samples[6].

[5] We know this by the Strong Law of Large Numbers, see Section 8.1.

[6] In particular, the intractable integral for the evidence is estimated using Monte Carlo Integration

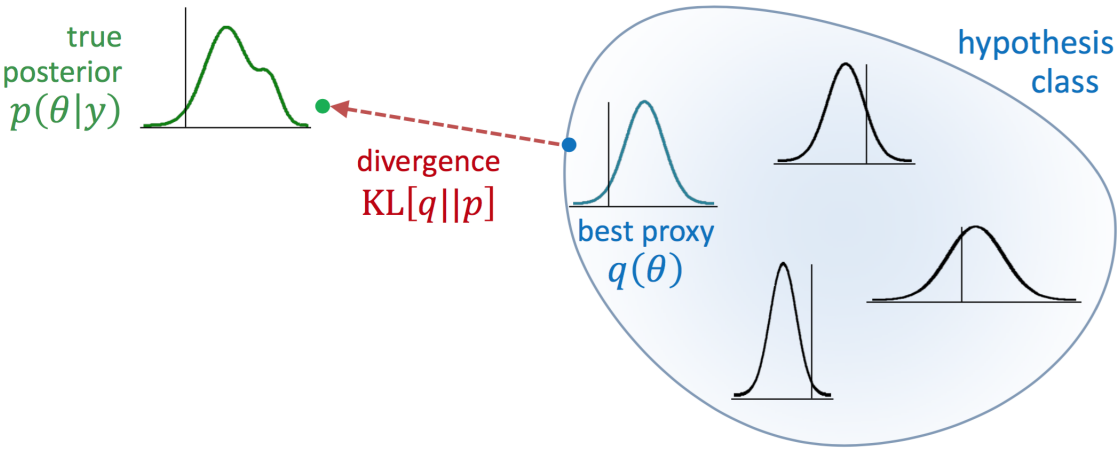

Figure 1: Variational Inference

Variational inference is an alternative approach to approximate inference. Rather than using sampling (e.g., MCMC), the main idea behind variational inference is to use optimization. In particular, first we posit a *family* of approximate distributions of the latent variables $\mathscr{L}$. Then we try (using optimization) to find a member of that family that minimizes the Kullback-Leibler (KL) divergence to extract the posterior (see Figure 1):

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z})\in\mathscr{L}}{\operatorname{argmax}} \, \mathrm{D}_{\mathrm{KL}}[q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})] \tag{37}$$

Finally, we approximate the posterior with $q^*(\mathbf{z})$. The key thing to notice here is that variational techniques turn the inference problem into an optimization problem, using the family $\mathscr{L}$ to manage the complexity of the optimization. One of the key ideas behind variational inference is to choose $\mathscr{L}$ to be flexible enough to be close to $p(\mathbf{z}|\mathbf{x})$ while being simple enough for efficient optimization.

5 Variational Bayes

Recall that Bayesian inference formalizes model inversion, the process of passing from a prior to a posterior in light of data (Figure 2, courtesy Kay Brodersen). However, in practice evaluating the posterior can be difficult because in many cases we cannot easily evaluate the model evidence $p(y)$, including when

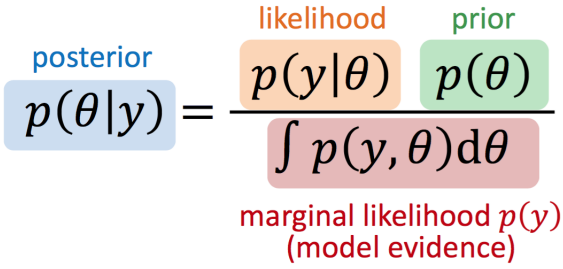

Figure 2: Bayesian Model Inversion

- $p(y)$ is intractable
- analytical solutions are not available
- numerical integration is too expensive

Fortunately, there are two complementary approaches to approximate inference which are outlined in Figure 3. As we saw above, Monte Carlo techniques (stochastic approximate inference) provide a numerical approximation to the exact posterior using a set of samples, while Variational Bayesian techniques (Structural approximate inference) provide a locally-optimal, exact analytical solution to an approximation of the posterior.[7]

5.1 The Laplace Approximation

The Laplace approximation provides a way of approximating a density whose normalization constant is either intractable or difficult to evaluate. The idea behind the Laplace Approximation is simple and intuitive: approximate the posterior by fitting a Gaussian to the empirical *mode*. In particular, the set up for the Laplace Approximation is

$$p(z) = \frac{1}{Z} \times f(z) \tag{38}$$

where $\frac{1}{Z}$ is the (likely intractable) unknown normalization constant and $f(z)$ is the main part of the density (typically easy to evaluate). The key observation here is that this is exactly the situation we face when we want to evaluate the posterior distribution while doing Bayesian inference. That is,

[7] Or to quote the famous statistician John Tukey, "An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."

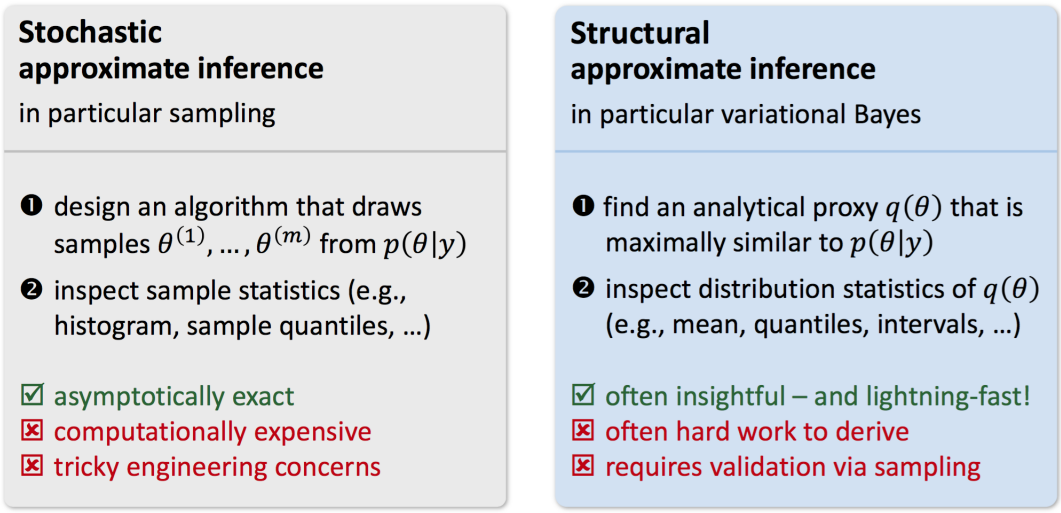

Figure 3: Approaches to Approximate Inference

$$p(\theta|y) = \frac{1}{p(y)} \times p(y, \theta) \tag{39}$$

where $\frac{1}{p(y)}$ is the (unknown) model evidence and $p(y, \theta)$ is the joint density (also easy to evaluate).

One approach to approximating the posterior is with a Taylor series. Note that any function $f(x)$ can be approximated by a series:

$$\begin{aligned} f(x) \approx\ & f(x^*) \\ & + f'(x^*)(x - x^*) \\ & + \frac{1}{2!} f''(x^*)(x - x^*)^2 \\ & + \frac{1}{3!} f'''(x^*)(x - x^*)^3 \\ & + \ldots \end{aligned}$$

To derive the Laplace approximation, we begin by expressing the log-joint density $\mathcal{L}(\theta) \equiv \ln p(y, \theta)$ in terms of a second order Taylor approximation around the mode θ^*. Again

trading inference for optimization (as is the case with variational techniques), we optimize $\mathcal{L}(\theta)$ by setting $\mathcal{L}'(\theta^*) = 0$, which yields

$$\begin{aligned}
\mathcal{L}(\theta) &\approx \mathcal{L}(\theta^*) \\
&\quad + \mathcal{L}'(\theta^*)(\theta - \theta^*) \\
&\quad + \frac{1}{2}\mathcal{L}''(\theta^*)(\theta - \theta^*)^2 \\
&= \mathcal{L}(\theta^*) + \frac{1}{2}\mathcal{L}''(\theta^*)(\theta - \theta^*)^2 \qquad\qquad \# \, \mathcal{L}'(\theta^*) = 0
\end{aligned}$$

Interestingly, this has exactly the same form as a Gaussian density, namely

$$\begin{aligned}
\ln \mathcal{N}(\theta|\mu, \eta^{-1}) &= \frac{1}{2}\ln \eta - \frac{1}{2}\ln 2\pi - \frac{\eta}{2}(\theta - \mu)^2 \\
&= \frac{1}{2}\ln \frac{\eta}{2\pi} + \frac{1}{2}(-\eta)(\theta - \mu)^2
\end{aligned}$$

and so we have an approximate posterior:

$$q(\theta) = \mathcal{N}(\theta|\mu, \eta^{-1})$$

where $\mu = \theta^*$ is the mode of the log posterior and and $\eta = \mathcal{L}''(\theta^*)$ is negative curvature at the mode.

Now, given a model with parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$, the Laplace Approximation reduces to the simple three step procedure shown in Figure 4.

Unfortunately, the Laplace approximation is often to much of a simplification, since it

- ignores the the global properties of the posterior
- is only directly applicable to real-valued parameters
- becomes brittle when the true posterior is multi-modal

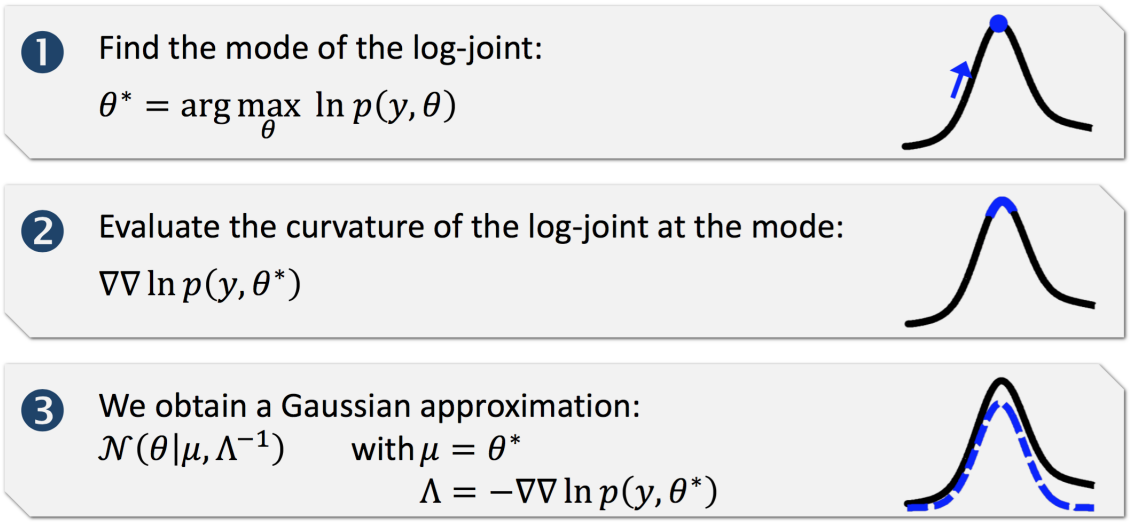

Figure 4: Laplace Approximation Procedure

6 The Problem of Approximate Inference, Redux

Let $\mathbf{x} = x_{1:N}$ be a set of observed variables and let $\mathbf{z} = z_{1:M}$ be a set of latent variables with joint distribution $p(\mathbf{z}, \mathbf{x})$. Then the inference problem is to compute the conditional distribution of latent variables given the observations, that is, $p(\mathbf{z}|\mathbf{x})$.

We can write the conditional distribution as

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})} \tag{40}$$

The denominator of 40 is the marginal distribution of the observiations (also called the *evidence*), and is calculated by *marginalizing* out the latent variables from the joint distribution, i.e.,

$$p(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{z}, \mathbf{x}) \tag{41}$$

In many cases of interest this integral is not available in closed form or is intractable (requires exponential time to compute). However, the model *evidence* is just the quantity

we need to compute the conditional ($p(\mathbf{z}|\mathbf{x})$) from the joint ($p(\mathbf{z}, \mathbf{x})$). This is why inference in these cases can be hard.

Recall that in variational inference we specify a family $\mathscr{L}$ of distributions over the latent variables. Each $q(\mathbf{z}) \in \mathscr{L}$ is a candidate approximation to the exact posterior. The goal is to find the best approximation, e.g., the one that satisfies the following optimization problem:

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z})\in\mathscr{L}}{\operatorname{argmax}} \, \mathrm{D}_{\mathrm{KL}}[q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})] \tag{42}$$

Unfortunately, this objective is still not computable because it requires computing the evidence $\log p(x)$ in Equation 41 (that the evidence is hard to compute is why we appeal to approximate inference in the first place). You can see why pretty easily. The general form of the KL divergence is

$$D_{\mathrm{KL}}\big[P||Q\big] = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{43}$$

Now, let $X_i = \log \frac{P(i)}{Q(i)}$. Then

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\big[P||Q\big] &= \sum_i P(i) X_i \\
&= \mathbb{E}_p[X] && \text{\# expectation taken with respect to } P \\
&= \mathbb{E}_p\Big[\log \frac{P(i)}{Q(i)}\Big] && \text{\# } X_i = \log \frac{P(i)}{Q(i)} \\
&= \mathbb{E}_p\big[\log P(i)\big] - \mathbb{E}\big[\log Q(i)\big]
\end{aligned}$$

Going back to our case, we have

$$\begin{aligned}
D_{\mathrm{KL}}[q(z)||p(z|x)] &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z|x)] \\
&= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z, x)] + \log p(x)
\end{aligned}$$

noting that $\mathbb{E}[X + c] = \mathbb{E}[X] + c$, where $X = \log p(z, x)$ and $c = \log p(x)$ ($\log p(x)$ is a constant with respect to $q(z)$). So we see the dependence on the difficult if not impossible to calculate evidence $p(x)$. Because we cannot compute the KL, we optimize an alternative that is equivalent to the KL up to an added constant:

$$\text{ELBO}(q) = \mathbb{E}[\log p(z, x)] - \mathbb{E}[\log q(z)] \tag{44}$$

This function is called the evidence lower bound (ELBO). The ELBO is the negative KL-divergence minus the log of the evidence, $\log p(x)$ (which is constant with respect to $q(z)$). Importantly, note that maximizing the ELBO is equivalent to minimizing the KL-divergence.

The ELBO also gives some information about the optimal variational distribution. In particular, we can rewrite the elBO as follows

$$\begin{aligned}\text{ELBO} &= \mathbb{E}[\log p(z)] + \mathbb{E}[\log p(x|z)] - \mathbb{E}[\log q(z)] \\ &= \mathbb{E}[\log p(x|z)] - \mathrm{D}_{\mathrm{KL}}[q(z)||p(z)]\end{aligned}$$

An interesting and important observation here is that the ELBO is a lower bound on the log evidence. That is, $\log p(x) \geq \text{ELBO}(q), \forall q(z)$. You can see this by noticing that $\log p(x) = \mathrm{D}_{\mathrm{KL}}[q(z)||p(z|x)] + \text{ELBO}(q)$; applying *Jensen's Inequality* (Figure 5) gives $\mathrm{D}_{\mathrm{KL}}(\cdot) \geq 0$. The (lower) bound follows directly from this fact. More below...

6.1 Varational Bayes

Variational Bayesian (VB) inference generalizes the idea behind the Laplace approximation. In VB, we wish to find an approximate density that is maximally similar to the true posterior. This is shown in Figure 1.

Note that Variational Bayesian inference is based on Variational Calculus (hence the name). A high-level comparision of standard and variational calculus is shown in Figure 6. Notice that Variational calculus lends itself naturally to Bayesian inference:

$$\begin{aligned}\ln p(y) &\equiv \ln \frac{p(y,\theta)}{p(\theta|y)} \\ &= \int q(\theta) \ln \frac{p(y,\theta)}{p(\theta|y)} d\theta \\ &= \int q(\theta) \ln \frac{p(y,\theta)}{p(\theta|y)} \frac{q(\theta)}{q(\theta)} d\theta \\ &= \int q(\theta) \Big(\ln \frac{q(\theta)}{p(\theta|y)} - \ln \frac{p(y,\theta)}{q(\theta)} \Big) d\theta \\ &= \int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} d\theta + \int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} d\theta \\ &= \mathrm{D}_{\mathrm{KL}}[q||p] + F(q,y)\end{aligned}$$

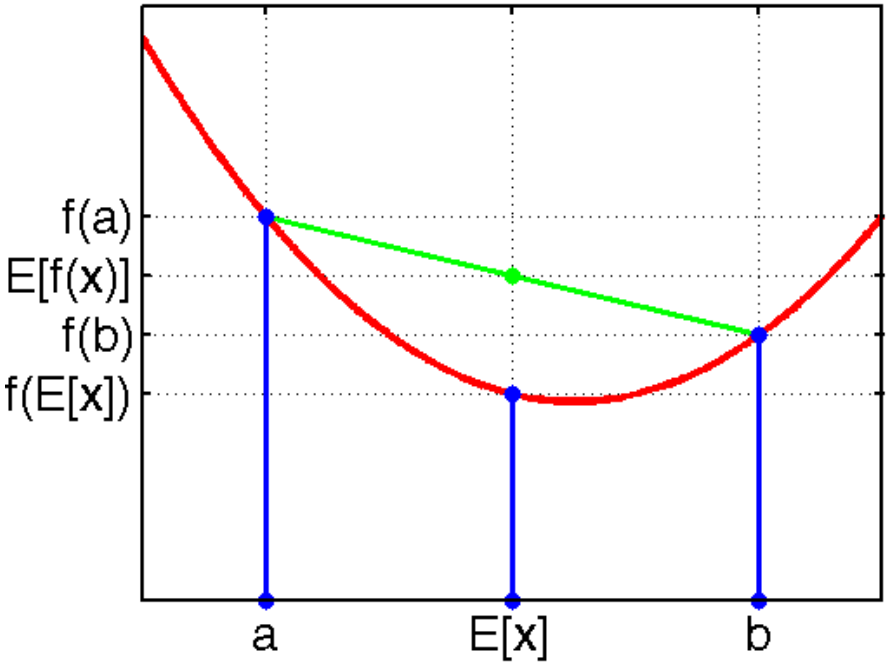

Figure 5: Jensen's Inequality: $\mathrm{f}(\mathbb{E}[\mathrm{X}]) \leq \mathbb{E}[\mathrm{f}(\mathrm{X})]$

Standard calculus
Newton, Leibniz, and others

- functions $f: x \mapsto f(x)$
- derivatives $\frac{\mathrm{d}f}{\mathrm{d}x}$

Example: maximize the likelihood expression $p(y|\theta)$ w.r.t. θ

Variational calculus
Euler, Lagrange, and others

- functionals $F: f \mapsto F(f)$
- derivatives $\frac{\mathrm{d}F}{\mathrm{d}f}$

Example: maximize the entropy $H[p]$ w.r.t. a probability distribution $p(x)$

Leonhard Euler
(1707 – 1783)
Swiss mathematician, 'Elementa Calculi Variationum'

Figure 6: Comparison of Standard and Variational Calculus

where $\mathrm{D}_{\mathrm{KL}}[q||p] = \int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} d\theta$, the Kullback-Leibler divergence between $q(\theta)$ and $p(\theta|y)$ (i.e., the cost of estimating $p(\theta|y)$ with $q(\theta)$), and $F(q, y) = \int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} d\theta$, the Free Energy.

So we see that the log model evidence $\ln p(y)$ can be expressed as

$$\ln p(y) = \mathrm{D}_{\mathrm{KL}}[q||p] + F(q, y) \tag{45}$$

and that maximizing $F(q, y)$ is equivalent to

- minimizing $\mathrm{D}_{\mathrm{KL}}\big[q||p\big]$
- tightening $F(q, y)$ as a lower bound to the log model evidence

See Figure 7 for a depiction of the relationship between $\ln p(y)$, $\mathrm{D}_{\mathrm{KL}}[q||p]$ and $F(q, y)$.

We can further decompose the free energy $F(q, y)$ as follows:

$$\begin{aligned} F(q, y) &= \int q(\theta) \ln \frac{p(y, \theta)}{q(\theta)} d\theta && (46) \\ &= \int q(\theta) \ln p(y, \theta) d\theta - \int q(\theta) \ln q(\theta) d\theta && (47) \\ &= \big\langle \ln p(y, \theta) \big\rangle_q + \mathrm{H}[q] && (48) \end{aligned}$$

where $\big\langle \ln p(y, \theta) \big\rangle_q$ is the expected log-joint density and $\mathrm{H}[q]$ is the Shannon Entropy.

6.2 Variational Methods for Maximum Likelihood Learning

Variational methods have also been used to approximate maximum likelihood learning for probabilistic graphical models with hidden variables. A great example of this is the derivation of the Expectation Maximization (EM) algorithm for MLE.

Consider the following setup: we have a graphical model with hidden variables $\mathbf{x}$, observable variables $\mathbf{y}$, and parameters $\boldsymbol{\theta}$. Maximum Likelihood (ML) learning then seeks to maximize the likelihood (or equivalently the log likelihood) of a data set $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ as a function of $\boldsymbol{\theta}$:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \ln P(Y|\boldsymbol{\theta}) && (49) \\ &= \sum_{i=1}^{n} \ln P(y_i|\boldsymbol{\theta}) && (50) \\ &= \sum_{i=1}^{n} \ln \int dx \; P(\mathbf{y}_i, \mathbf{x}|\boldsymbol{\theta}) && (51) \end{aligned}$$

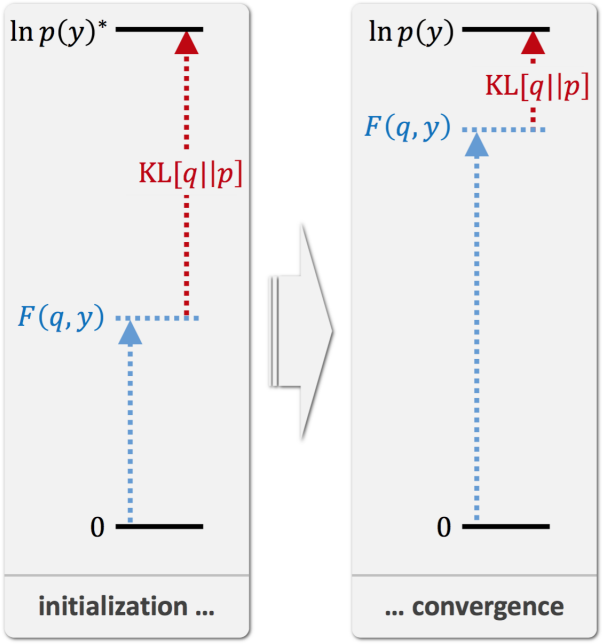

Figure 7: Relationship between log model evidence, KL divergence, and Free Energy

Note that we assume the data Y is independent and identically distributed (iid). The integral (or sum) over $\mathbf{x}$ is needed to get the marginal probability of the data (Y). Maximizing Equation 49 directly can be hard because the log of the integral can potentially couple all of the parameters of the model. In addition, for models with many hidden variables the integral (sum) over $\mathbf{x}$ can be intractable. However, one can simplify the problem of maximizing $\mathcal{L}$ with respect to $\boldsymbol{\theta}$ by making use of the following observation: Any distribution $Q_x(x)$ over the hidden variables defines a lower bound on $\mathcal{L}$. More precisely, for each data point $\mathbf{y}_i$ we use a distinct distribution $Q_{x_i}(\mathbf{x}_i)$ over the hidden variables to get the lower bound:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_i \ln \int d\mathbf{x}_i \; P(\mathbf{y}_i, \mathbf{x}|\boldsymbol{\theta}) \tag{52}$$

$$= \sum_i \ln \int d\mathbf{x}_i \; Q_{\mathbf{x}_i}(\mathbf{x}_i) \frac{P(\mathbf{y}_i, \mathbf{x}_i|\boldsymbol{\theta})}{Q_{x_i}(\mathbf{x}_i)} \tag{53}$$

$$\geq \sum_i \int d\mathbf{x} \; Q_{\mathbf{x}_i}(\mathbf{x}_i) \; \ln \frac{P(\mathbf{y}_i, \mathbf{x}_i|\boldsymbol{\theta})}{Q_{x_i}(\mathbf{x}_i)} \tag{54}$$

$$= \mathcal{F}(Q_{\mathbf{x}_1}(x_1), \ldots, Q_{x_n}(\mathbf{x}_n), \boldsymbol{\theta}) \tag{55}$$

where the inequality is known as Jensen's inequality and follows from the fact that the function is concave.

Defining the *energy* of a global configuration $(\mathbf{x}, \mathbf{y})$ to be $-\ln P(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta})$, the lower bound $\mathcal{F} \leq \mathcal{L}(\boldsymbol{\theta})$ is the negative of the quantity known in statistical physics as *free energy*, that is, the expected energy of Q minus the entropy of Q.

The Expectation Maximum (EM) algorithm alternates between maximizing $\mathcal{F}$ with respect to Q_{x_i} and $\boldsymbol{\theta}$ while holding the other fixed. Starting from some initial parameters $\boldsymbol{\theta}^0$, the EM algorithm can be expressed in terms of the **E step** and the **M step**, as follows:

$$\begin{aligned} \textbf{E step:} \quad & Q_{x_i}^{k+1} \leftarrow \operatorname*{argmax}_{Q_{x_i}} \mathcal{F}(Q, \boldsymbol{\theta}^k), \ \forall i \\ \textbf{M step:} \quad & \boldsymbol{\theta}^{k+1} \leftarrow \operatorname*{argmax}_{\boldsymbol{\theta}} \mathcal{F}(Q^{k+1}, \boldsymbol{\theta}) \end{aligned}$$

It is easy to see that the maximum in the E step is obtained by setting $Q_{x_i}^{k+1} = P(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}^k)$, at which point the lower bound becomes an equality: $\mathcal{F}(Q^{k+1}, \boldsymbol{\theta}^k) = \mathcal{L}(\boldsymbol{\theta}^k)$. The maximum in the M step is obtained by minimizing the expected energy since the entropy of Q does not depend on $\boldsymbol{\theta}$:

$$\textbf{M step: } \boldsymbol{\theta}^{k+1} \leftarrow \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_i \int dx \ P(x|y, \boldsymbol{\theta}^k) \ \ln P(\boldsymbol{x}, \boldsymbol{y}_i, \boldsymbol{\theta})$$

Since $\mathcal{F} = \mathcal{L}$ at the beginning of each M step, and since the E step does not change $\boldsymbol{\theta}$, we are guaranteed not to decrease the likelihood after each combined EM step.

BTW, maximizing $\mathcal{F}$ with respect to Q_{x_i} is equivalent to minimizing

$$\int d\mathbf{x} \, Q_{x_i}(x_i) \ \ln \frac{Q_{x_i}(x_i)}{P(\boldsymbol{x}|\boldsymbol{y}_i, \boldsymbol{\theta})} \tag{56}$$

with is the Kullback-Leibler (KL) divergence measuring the (asymmetric) difference between Q_{x_i} and the turn posterior. Choosing Q_{x_i} to have easily computed moments, and if $\ln P$ is polynomial in $\mathbf{x}$ we can compute the KL-divergence up to a constant and more importantly we can take its derivative to minimize it with respect to the parameters of Q_{x_i}.

6.3 Variational methods for Bayesian learning

Maximum likelihood methods suffer from the problem that they fail to consider model complexity, which is, from an information theoretic point of view, the cost of coding the model parameters. Not penalizing more complex models leads to *overfitting* and the inability to determine the best model size and structure. An example would be if the search is limited to a single parameter that controls the model complexity then for more general search cross-validation becomes computationally infeasible. Bayesian approaches overcome overfitting and learn model structure by treating the parameters $\boldsymbol{\theta}$ as unknown random variables and averaging over the ensemble of models that one would obtain by sampling from $\boldsymbol{\theta}$:

$$P(Y|\mathcal{M}) = \int d\boldsymbol{\theta}\ P(Y|\boldsymbol{\theta}, \mathcal{M})\ P(\boldsymbol{\theta}|\mathcal{M}) \tag{57}$$

$P(Y|\mathcal{M})$ is the *evidence* or *marginal likelihood* for a data set Y assuming model $\mathcal{M}$, and $P(\boldsymbol{\theta}|\mathcal{M})$ is the prior distribution over the parameters. Interestingly, integrating out parameters penalizes models with more degrees of freedom since these models can *a priori* model a large range of data sets. This property of Bayesian inference has been called Ockham's Razor since it favors simpler explanations (models) for the data over complex ones. The overfitting problem is avoided simply because no parameters in the pure Bayesian approach are actually *fit* to the data. Basically, having more parameters provides an advantage in terms of being able to model the data, however, the cost is having to code those parameters under the prior.

Along with the prior over the parameters, a Bayesian approach to learning starts with some prior knowledge (or assumptions) abou the model structure, essentially the set of edges in the Bayesian network. This knowledge is represented in the form of a prior probability distribution over model structures and is updated using the data to obtain a posterior distribution over models and parameters. In particular, assuming a prior distribution over model structures $P(\mathcal{M})$ and a prior distribution over the parameters for each model structure $P(\boldsymbol{\theta}|\mathcal{M})$, observing the data set Y induces a posterior distribution over the models given by Bayes rule:

$$P(\mathcal{M}|Y) = \frac{P(\mathcal{M})P(Y|\mathcal{M})}{P(Y)} \tag{58}$$

and the most probable mode or model structure is the one that maximizes $P(\mathcal{M}|Y)$.

For a given model structure, we can also compute the posterior distribution over the parameters:

$$P(\boldsymbol{\theta}|Y, \mathcal{M}) = \frac{P(Y|\boldsymbol{\theta}, \mathcal{M})P(\boldsymbol{\theta}|\mathcal{M})}{P(Y|\mathcal{M})} \tag{59}$$

which allows us to quantify our uncertainty about parameter values after observing the data. The density at a new data point $\mathbf{y}$ is obtained by averaging over both the uncertainty in the model structure and the parameters, namely,

$$P(\mathbf{y}|Y) = \int d\boldsymbol{\theta}\; d\mathcal{M}\; P(\mathbf{y}|\boldsymbol{\theta}, \mathcal{M}, Y)P(\boldsymbol{\theta}|\mathcal{M}, Y)P(\mathcal{M}|Y) \tag{60}$$

which is known as the *predictive distribution.*

So the interesting thing is that we can use variational methods to approximate the integrals required for Bayesian learning. Other methods we've seen include *Monte Carlo integration* and *Laplace approximation.* The basic idea behind variational methods is to simultaneously approximate the distribution over both the hidden states and the parameters with a simpler distribution.[8] Like in the case of the EM algorithm, we lower bound the log evidence by applying Jensen's inequality:

$$\ln P(Y|\mathcal{M}) = \int d\boldsymbol{\theta}\; P(P, \boldsymbol{\theta}|\mathcal{M}) \tag{61}$$

$$\geq \int\int d\boldsymbol{\theta}\; dX\; Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})Q_X(X) \ln \frac{P(Y, X, \boldsymbol{\theta}|\mathcal{M})}{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})Q_X(X)} \tag{62}$$

$$= \int d\boldsymbol{\theta}\; Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \left[\int dX\; Q_X(X)\; \ln \frac{P(Y, X|\boldsymbol{\theta}, \mathcal{M})}{Q_X(X)} \;+\; \ln \frac{P(\boldsymbol{\theta}|\mathcal{M})}{Q_{\boldsymbol{\theta})}(\boldsymbol{\theta})} \right] \tag{63}$$

$$= \mathcal{F}(Q_X(X), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) \tag{64}$$

$$= \mathcal{F}(Q_{x_1}(x_1), \ldots, Q_{x_n}(x_n), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) \tag{65}$$

where the last equality follow from the fact that Y is assumed to be iid. The variational Bayesian approach iteratively maximizes $\mathcal{F}$ as a function of the free distributions $Q_X(X)$ and $Q(\boldsymbol{\theta})$. We can also see that this is equivalent to minimizing $\mathrm{D}_{\mathrm{KL}}\big[Q_X(X)||Q(\boldsymbol{\theta})\big]$.

7 Acknowledgements

[8] Usually done by assuming the hidden states and the parameters are independent.

References

8 Appendix

8.1 Strong Law of Large Numbers

Let $X_1, X_2, \ldots, X_M$ be a sequence of **independent** and **identically distributed** random variables, each having a finite mean $\mu_i = E[X_i]$.

Then with probability 1

$$\frac{1}{M}\sum_{i=1}^{M} X_i \to E[X] \tag{66}$$

as $M \to \infty$.

8.2 Ergodic Theorem

Let $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}$ be M samples from a Markov chain that is *aperiodic*, *irreducible*, and *positive recurrent*[9], and $E[g(\theta)] < \infty$.

Then with probability 1

$$\frac{1}{M}\sum_{i=1}^{M} g(\theta_i) \to E[g(\theta)] = \int_{\Theta} g(\theta)\, \pi(\theta)\, d\theta \tag{67}$$

as $M \to \infty$ and where π is the stationary distribution of the Markov chain.

[9] In this case, the chain is said to be *ergodic*.