Machine Learning and Pattern Recognition - Week 10 - Variational Inference

Antonio León Villares

November 2022

Contents

Based on the online notes here.

1 Recap: Bayesian Logistic Regression

We defined a Bayesian version for ***logistic regression****, with* ***prior*** $P(\underline{w})$ *and* ***likelihood*** $P(\mathcal{D} \mid \underline{w}) = \sigma(\underline{w}^T \underline{x})$*:*

$$P(\underline{w} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \underline{w})P(\underline{w})}{P(\mathcal{D})}$$

The ***posterior*** *is used, for instance, to make* ***predictions*** *using the* ***predictive posterior****:*

$$P(y = 1 \mid \underline{x}, \mathcal{D}) = \int \sigma(\underline{w}^T \underline{x})P(\underline{w} \mid \mathcal{D})d\underline{w}$$

However, sampling from the ***posterior*** *is very hard: we need to be able to compute the* ***marginal likelihood*** $P(\mathcal{D})$*.*

Unless we are working with nicely parametrised distributions (i.e Gaussian), computing $P(\mathcal{D})$ *and sampling from the posterior is extremely difficult. One way of* ***approximating*** *the posterior is to use a* ***Laplace approximation****:*

$$P(\underline{w} \mid \mathcal{D}) \approx \mathcal{N}(\underline{w}; \underline{w}^*, H^{-1})$$

where:

- $\underline{w}^*$ *is the* ***MAP approximation*** *of* $\underline{w}$ *for* $-\log P(\underline{w} \mid \mathcal{D})$
- H *is the* ***Hessian*** *of* $E(\underline{w}) = -\log P(\underline{w} \mid \mathcal{D})$

Using the ***Laplace approximation****, we could also approximate the* ***marginal likelihood****:*

$$P(\mathcal{D}) \approx P(\underline{w}^*, \mathcal{D})|2\pi H^{-1}|^{1/2}$$

which in turn allowed us to ***approximate*** *the* ***predictive posterior****:*

$$P(y = 1 \mid \underline{x}, \mathcal{D}) \approx \int \sigma(a)\mathcal{N}(a; \underline{w}^{*T}\underline{x}, \underline{x}^T H^{-1}\underline{x}da$$

where $a = \underline{w}^T \underline{x}$*.*

2 Bayesian Logistic Regression via Importance Sampling

2.1 Importance Sampling

- **What is importance sampling?**
  - allows us to **sample** from an **arbitrary** distribution when computing the **predictive posterior**
  - if $q(\underline{w})$ is some tractable distribution, then:

$$
\begin{aligned}
P(y = 1 \mid \underline{x}, \mathcal{D}) &= \int \sigma(\underline{w}^T \underline{x}) P(\underline{w} \mid \mathcal{D}) d\underline{w} \\
&= \int \sigma(\underline{w}^T \underline{x}) P(\underline{w} \mid \mathcal{D}) \frac{q(\underline{w})}{q(\underline{w})} d\underline{w} \\
&= \int q(\underline{w}) \frac{\sigma(\underline{w}^T \underline{x}) P(\underline{w} \mid \mathcal{D})}{q(\underline{w})} d\underline{w} \\
&= \mathbb{E}_{\underline{w} \sim q(\underline{w})} \left[\frac{\sigma(\underline{w}^T \underline{x}) P(\underline{w} \mid \mathcal{D})}{q(\underline{w})} \right] \\
&\approx \frac{1}{S} \sum_{s=1}^{S} \sigma(\underline{w}^{(s)T} \underline{x}) \frac{P(\underline{w}^{(s)} \mid \mathcal{D})}{q(\underline{w}^{(s)})}, \qquad \underline{w}^{(s)} \sim q(\underline{w})
\end{aligned}
$$

- **What is the importance weight in importance sampling?**
  - the **importance weight** is the quotient:

$$
r^{(s)} = \frac{P(\underline{w}^{(s)} \mid \mathcal{D})}{q(\underline{w})}
$$

  - **upweights** those parameters which are more likely under the **posterior** $P(\underline{w} \mid \mathcal{D})$ than the **sampling distribution** $q(\underline{w})$

- **Why can't we sample from the posterior even with importance sampling?**
  - to use **importance sampling**, even if we don't **sample** from the **posterior**, we still have to **evaluate** it
  - however, this involves computing $P(\mathcal{D})$
  - we can **approximate** this by using **importance sampling** again:

$$
\begin{aligned}
P(\mathcal{D}) &= \int P(\mathcal{D} \mid \underline{w}) P(\underline{w}) d\underline{w} \\
&= \int P(\mathcal{D} \mid \underline{w}) P(\underline{w}) \frac{q(\underline{w})}{q(\underline{w})} d\underline{w} \\
&= \mathbb{E}_{\underline{w} \sim q(\underline{w})} \left[\frac{P(\mathcal{D} \mid \underline{w}) P(\underline{w})}{q(\underline{w})} \right] \\
&\approx \frac{1}{S} \sum_{s=1}^{S} \frac{P(\mathcal{D} \mid \underline{w}^{(s)}) P(\underline{w}^{(s)})}{q(\underline{w}^{(s)})}
\end{aligned}
$$

  - if we define the **unnormalised importance weights** as:

$$
\tilde{r}^{(s)} = \frac{P(\mathcal{D} \mid \underline{w}^{(s)}) P(\underline{w}^{(s)})}{q(\underline{w}^{(s)})}
$$

we can **approximate** the **predictive posterior**:

$$\begin{aligned}
P(y=1 \mid \underline{x}, \mathcal{D}) &\approx \frac{1}{S}\sum_{s=1}^{S} \sigma(\underline{w}^{(s)T}\underline{x}) \frac{P(\underline{w}^{(s)} \mid \mathcal{D})}{q(\underline{w}^{(s)})}, \qquad \underline{w}^{(s)} \sim q(\underline{w}) \\
&= \frac{1}{S}\sum_{s=1}^{S} \sigma(\underline{w}^{(s)T}\underline{x}) \frac{P(\mathcal{D} \mid \underline{w}^{(s)})P(\underline{w})}{P(\mathcal{D})q(\underline{w}^{(s)})} \\
&= \frac{1}{S}\sum_{s=1}^{S} \sigma(\underline{w}^{(s)T}\underline{x}) \frac{\tilde{r}^{(s)}}{P(\mathcal{D})} \\
&= \frac{1}{S}\sum_{s=1}^{S} \sigma(\underline{w}^{(s)T}\underline{x}) \frac{\tilde{r}^{(s)}}{\frac{1}{S}\sum_{s'=1}^{S}\tilde{r}^{(s')}} \\
&= \sum_{s=1}^{S} \sigma(\underline{w}^{(s)T}\underline{x})\rho^{(s)}
\end{aligned}$$

where in the last step we have defined the **normalised importance weights**

$$\rho^{(s)} = \frac{\tilde{r}^{(s)}}{\sum_{s'=1}^{s}\tilde{r}^{(s')}}$$

- **What are the unnormalised importance weights if we use the prior as our sampling distribution?**
  - we have that the **unnormalised importance weights** are:

    $$\tilde{r}^{(s)} = \frac{P(\mathcal{D} \mid \underline{w}^{(s)})P(\underline{w}^{(s)})}{q(\underline{w}^{(s)})}$$

  - if we use $q(\underline{w}) = P(\underline{w})$, then:

    $$\tilde{r}^{(s)} = P(\mathcal{D} \mid \underline{w}^{(s)})$$

    so the **unnormalised importance weight** will be the **likelihood**

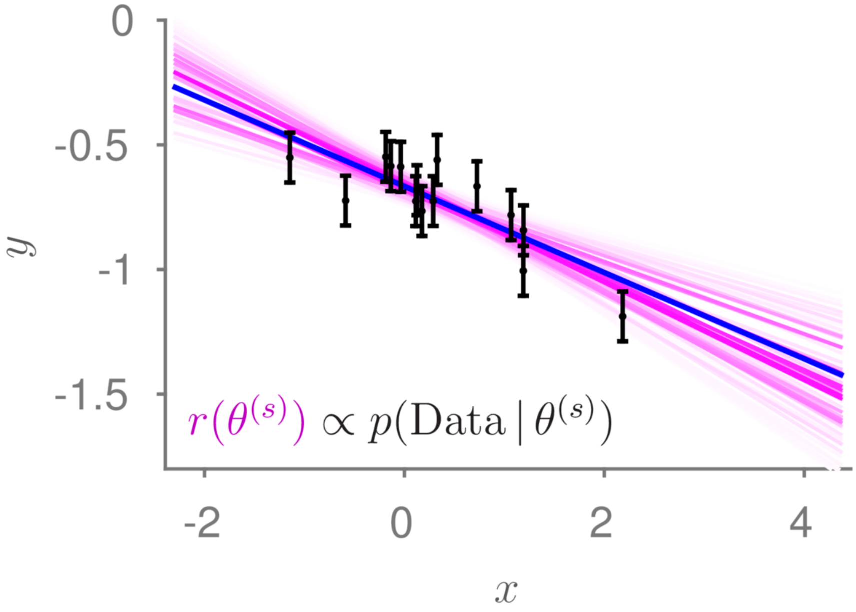


Figure 1: Importance sampling applied to linear regression. Data (black) is generated by the dark, blue line. We apply linear regression, and draw 10,000 samples from the **prior** (purple). The **intensity** of the purple corresponds to the **importance weight** (which is proportional to likelihood). Some models are so unlikely that they appear close to white.

2.2 Choosing the Sampling Distribution

- **How should the sampling distribution $q(\underline{w})$ be restricted to make sure importance sampling is reasonable?**

  1. $q(\underline{w}) > 0$: since we divide by $q(\underline{w})$
  2. We can't have $q(\underline{w}) << P(\underline{w} \mid \mathcal{D})$: if we did, the importance weight $r^{(s)}$ would be large for many different weight settings, so the estimator will have high variance

- **When won't importance sampling work well?**

  - in principle, **importance sampling** works well, so long as we can **sample** from $q(\underline{w})$ and can **evalute** the **likelihood** (as is the case for **logistic regression**)
  - importance sampling might not work well if there are a lot of parameters
  - for example, if we use $q(\underline{w}) = P(\underline{w})$, when we draw the S samples it is unlikely that any of the weights match the data well (since the weights are based on our prior knowledge), so we will have poor estimates
  - we could try using $q(\underline{w})$ to approximate the posterior, but with a lot of parameters this becomes very difficult (at least if we want an approximation which is sufficiently useful fro **importance sampling)**
  - in such cases, **Markov Chain Monte Carlo** can b used (this has been used successfully for neural networks)

3 Bayesian Logistic Regression via Variational Inference

3.1 Variational Inference: KL-Divergence

3.1.1 The KL-Divergence

- **What is a major pitfall of the Laplace approximation?**
  - the **Laplace Approximation** finds a **Gaussian** which best approximates a distribution at its **mode**
  - however, the behaviour at the **mode** of a distribution might be **misleading** about the rest of the distribution (i.e multimodal distributions)
- **What are variational methods for inference?**
  - fit a **target distribution**, by using **optimisation**
  - given a **family** of distributions $\{q(\underline{w}; \alpha)\}$ (parametrised by a **variational parameter** α), seeks to minimise a **variational cost function** (which measures the discrepancy between the **target** and $q(\underline{w}; \alpha)$), by varying α
  - for example, if we have a family of **Gaussians**, α would encompass the **mean** and **covariance** of the distributions:
  $$q(\underline{w}; \alpha = \{\underline{m}, mV\}) = \mathcal{N}(\underline{w}; \underline{m}, V)$$
  - for this course, we only consider **Gaussian** families, but naturally this method applies to other families aswell
- **Why is the Kullback-Leibler Divergence a good variational cost function?**
  - the **KL-Divergence** for 2 distributions $P(\underline{z}), Q(\underline{z})$ gives a measure of how distinct the distribution are:
  $$D_{KL}(P||Q) = \int P(\underline{z}) \log \frac{P(\underline{z})}{Q(\underline{z})} d\underline{z}$$
  - being a **divergence**, **KL-Divergence** satisfies:
    1. $$D_{KL}(P||Q) \geq 0$$
    2. $$D_{KL}(P||Q) \geq 0 \iff P(\underline{z}) = Q(\underline{z})$$
    3. it isn't **symmetric**:
    $$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$
  - these make the **KL-Divergence** appropriate as a **variational cost function**: minimising **KL-Divergence** implies moving our target family Q closer to P

> *The **KL-Divergence** appears in the context of **information theory**: it gives the **average** storage wasted by using 2 different compression systems.*

---

*At this point, there are 2 ways of minimising the **KL-Divergence**: we can either try to match our family to the target **directly** (i.e minimise $D_{KL}(P||Q)$), or **indirectly** (i.e minimise $D_{KL}(Q||P)$).*

3.1.2 Minimising KL-Divergence: Directly Matching the Posterior

- **How is KL-Divergence minimised, if we directly match the family to the target?**
  - to **minimise**:
  $$D_{KL}(P||Q)$$
  we set $\underline{m}, V$ in our **Gaussian** approximation to the **mean** and **covariance** of the **target**

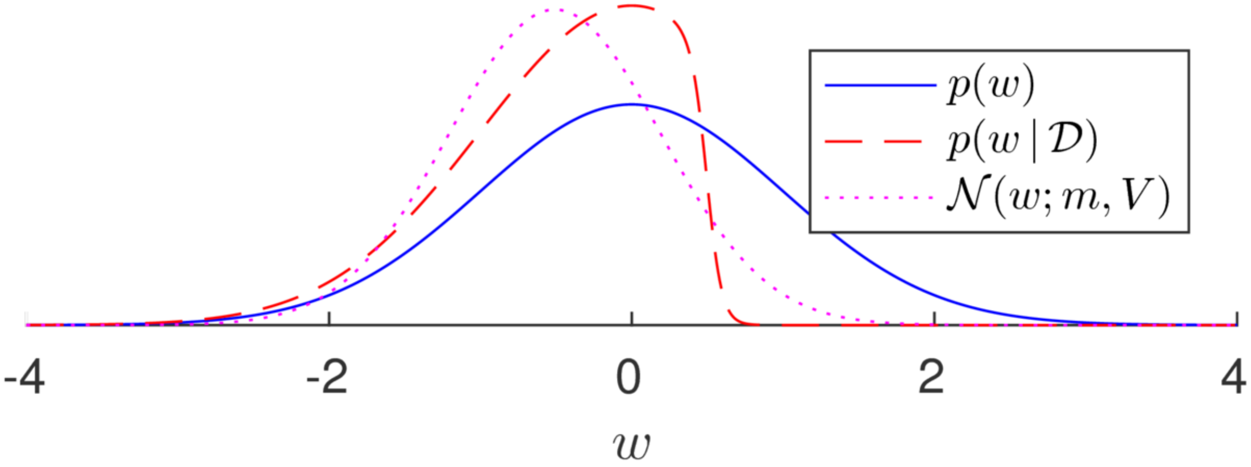
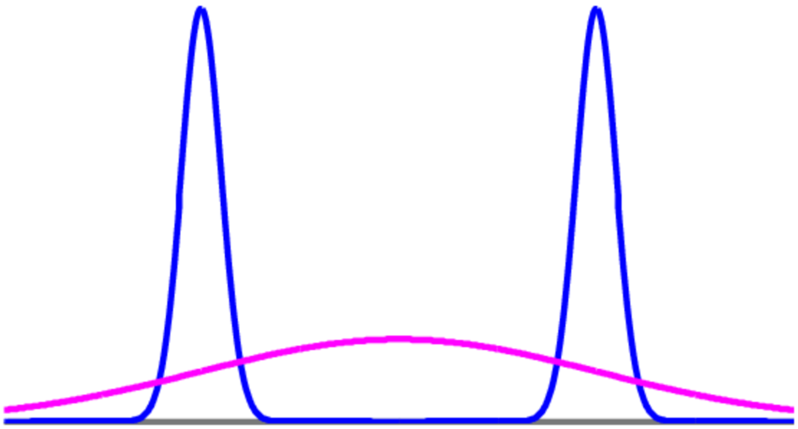


Figure 2: Recall this example from last week, where we considered the posterior for a Bayesian logistic regression model.
If we use a **Laplace approximation**, the approximate posterior would be practically identical to the prior (blue) (the posterior [red] has a nearly identical mode to the prior, since the logistic likelihood is nearly 1 in that region; because of this, the curvature at the mode will also be nearly identical; since the Lapalce approximation assumes a Gaussian distribution, it'll produce a posterior which looks nearly identical to the prior).
On the other hand, if we minimise $D_{KL}(P||Q)$ by picking $\underline{m}, V$ to match the true posterior, we get a more reasonable approximation (purple).

- **Why don't we typically minimise KL-Divergence by matching the family to the target directly?**
  1. **Evaluating Posterior**: we might not have a **parametric** posterior, so it is hard to compute the **KL-Divergence**; even if it were parametric, the **integral** might be very hard to compute
  2. **Posterior Parameters**: we'd need to have access to the **mean** and **covariance** of $P(\underline{w} \mid \mathcal{D})$, which we don't always have
  3. **Not Always Sensible**: even if we could evaluate the divergence, and had the posterior parameters, the result won't always be **sensible** (i.e a bimodal distribution will have its mean at a trough of the distribution, so we wouldn't match any of the modes, and we'd have a high variance distribution)



*We can justify why minimising $D_{KL}(P||Q)$ with a **Gaussian** Q is equivalent to matching the mean and covariance of P.*
*Whilst this can be justified using calculus, multivariate gaussians can be messy, so we choose a more general approach. In particular, let Q be some **exponential family** parametrised by θ:*

$$Q(\underline{w}) = \frac{1}{Z(\theta)} \exp(\theta^T \phi(\underline{w}))$$

where:

$$Z(\theta) = \int \exp(\theta^T \phi(\underline{w})) d\underline{w}$$

*and $\phi(\underline{w})$ is a **vector** containing certain **statistics** (for example, ones defining a Gaussian).*
*If we plug in the **approximating family** into the **KL-Divergence**, then:*

$$\begin{aligned} D_{KL}(P||Q) &= \int P(\underline{w}) \log \frac{P(\underline{w})}{Q(\underline{w})} d\underline{w} \\ &= \int P(\underline{w}) \log \frac{P(\underline{w})}{\frac{1}{Z(\theta)} \exp(\theta^T \phi(\underline{w}))} d\underline{w} \\ &= \int P(\underline{w}) \left[\log P(\underline{w}) - \theta^T \phi(\underline{w}) + \log Z(\theta) \right] d\underline{w} \end{aligned}$$

If we then differentiate this with respect to θ and set to 0:

$$\begin{aligned} 0 &= -\int P(\underline{w}) \phi(\underline{w}) d\underline{w} + \frac{1}{Z(\theta)} \int \phi(\underline{w}) \exp(\theta^T \phi(\underline{w})) d\underline{w} \\ \implies 0 &= -\int P(\underline{w}) \phi(\underline{w}) d\underline{w} + \int \phi(\underline{w}) Q(\underline{w}) d\underline{w} \end{aligned}$$

which implies that:

$$\mathbb{E}_{\underline{w} \sim P(\underline{w})}[\phi(\underline{w})] = \mathbb{E}_{\underline{w} \sim Q(\underline{w})}[\phi(\underline{w})]$$

In other words, the statistics $\phi(\underline{w})$ defining the distribution must match.
*If Q is **Gaussian**, this implies that the **mean** and **covariance** must be equal to those of P.*

3.1.3 Minimising KL-Divergence: Indirectly Matching the Posterior

- **Why is often chosen to minimise KL-Divergence by matching our target indirectly?**

1. **Optimisation**: it is easier to optimise $D_{KL}(Q||P)$
2. **Parameter Choice**: optimising $D_{KL}(Q||P)$ encourages that the chosen α is more plausible (i.e instead of fitting to the trough, we'll fit to one of the modes)

- **How is KL-Divergence minimised when matching the posterior indirectly?**
  - if we tried to naively minimise $K_{DL}(Q||P)$, we'd still have the problem of being able to evaluate the **posterior**
  - instead, we can try to "break down" the terms in the divergence:

$$\begin{aligned} D_{KL}(Q||P) &= \int Q(\underline{w};\alpha)\log\frac{Q(\underline{w};\alpha)}{P(\underline{w}\mid\mathcal{D})}d\underline{w} \\ &= \underbrace{\int Q(\underline{w};\alpha)\log Q(\underline{w};\alpha)d\underline{w}}_{\text{negative entropy: } -H(Q)} \underbrace{- \int Q(\underline{w};\alpha)\log P(\underline{w}\mid\mathcal{D})d\underline{w}}_{\text{cross-entropy}} \end{aligned}$$

  - hence, to **minimise** the **KL-Divergence**:
    1. we need to **minimise** $-H(Q)$, or equivalently **maximise** the **entropy** $H(Q)$: in other words, we want a distribution Q which is close to **uniform** - spread out, without any significant peaks
    2. we need to **maximise** the **cross-entropy**: to do this, we need to select α, such that Q and P match well (as an extreme example, if we have some implausible weight $\underline{w}$ with $P(\underline{w}\mid\mathcal{D})\to 0$, unless α makes it so that $Q(\underline{w};\alpha)\to 0$, the cross-entropy term will diverge to infinity)
  - intuitively, the above tells us that by minimising $D_{KL}(Q||P)$, we will pick Q which fits well to one of the modes of P, and then spreads out as much as possible to encompass as much of the probability mass of the mode as possible

- **What is the evidence lower bound?**
  - we can further break down $D_{KL}(Q||P)$ by using the definition of the posterior:

$$\begin{aligned} D_{KL}(Q||P) &= \int Q(\underline{w};\alpha)\log\frac{Q(\underline{w};\alpha)}{P(\underline{w}\mid\mathcal{D})}d\underline{w} \\ &= \underbrace{\int Q(\underline{w};\alpha)\log Q(\underline{w};\alpha)d\underline{w}}_{\text{negative entropy: } -H(Q)} \underbrace{- \int Q(\underline{w};\alpha)\log P(\underline{w}\mid\mathcal{D})d\underline{w}}_{\text{cross-entropy}} \\ &= \mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log Q(\underline{w})] - \int Q(\underline{w};\alpha)\log\frac{P(\mathcal{D}\mid\underline{w})P(\underline{w})}{P(\mathcal{D})}d\underline{w} \\ &= \underbrace{\mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log Q(\underline{w})] - \mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log P(\mathcal{D}\mid\underline{w})] - \mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log P(\underline{w})]}_{J(Q)} + \log P(\mathcal{D}) \end{aligned}$$

  - the term $\log P(\mathcal{D})$ is:
    * the **log-marginal likelihood**
    * the **model evidence**
  - the term $-J(Q)$ is known as the **ELBO** (**Evidence Lower Bound**), since:

$$D_{KL}(Q||P)\geq 0 \implies \log P(\mathcal{D})\geq -J(Q)$$

  - notice, **minimising** the divergence is equivalent to maximising the **ELBO** $-J(Q)$, which is a term dependent on distributions which we know:

∗ the **variational distribution** $Q(\underline{w}; \alpha)$
∗ the **likelihood** $P(\mathcal{D} \mid \underline{w})$
∗ the **prior** $P(\underline{w})$

We compare our 2 methods for Gaussian approximation.

Laplace Approximation:

- ***straightforward:*** *compute MAP weight & Hessian*
- *the* ***Hessian*** *gives certainty of parameters*
- *incrementally improves MAP estimate*
- *gives an approximation for the marginal $P(\mathcal{D})$*

Variational Methods:

- *optimise using* ***distribution family***
- *typically optimise indirectly (i.e using $D_{KL}(Q||P)$, not $D_{KL}(P||Q)$)*
- *gives bound for the marginal $P(\mathcal{D}$*
- *hard to optimise (see next section)*

3.2 Stochastic Variational Inference to Minimise KL-Divergence

3.2.1 The Optimisation Problem

We want to **minimise** the **KL-Divergence**:

$$D_{KL}(Q||P) = \mathbb{E}_{\underline{w} \sim Q(\underline{w};\alpha)}[\log Q(\underline{w})] - \mathbb{E}_{\underline{w} \sim Q(\underline{w};\alpha)}[\log P(\mathcal{D} \mid \underline{w})] - \mathbb{E}_{\underline{w} \sim Q(\underline{w};\alpha)}[\log P(\underline{w})] + \log P(\mathcal{D})$$

with respect to $\alpha = \{\underline{m}, V\}$.

Notice, the **model evidence** doesn't depend on α, so we seek to **minimise**:

$$J(\underline{m}, V) = \mathbb{E}_{\underline{w} \sim Q(\underline{w};\alpha)}[\log Q(\underline{w})] - \mathbb{E}_{\underline{w} \sim Q(\underline{w};\alpha)}[\log P(\mathcal{D} \mid \underline{w})] - \mathbb{E}_{\underline{w} \sim Q(\underline{w};\alpha)}[\log P(\underline{w})]$$

with respect to $\underline{m}, V$. Notice, since we have:

$$\log P(\mathcal{D}) \geq -J(\underline{m}, V)$$

minimising J will increase the model likelihood $P(\mathcal{D})$ (ideally we'd like to maximise this as well, but doing so exactly is hard).

The **entropy** terms (entropy + cross-entropy):

$$\mathbb{E}_{\underline{w} \sim Q(\underline{w};\alpha)}[\log Q(\underline{w})] \qquad \mathbb{E}_{\underline{w} \sim Q(\underline{w};\alpha)}[\log P(\underline{w})]$$

are easy to compute in closed form. The **log-likelihood**:

$$\mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log P(\mathcal{D} \mid \underline{w})] = \sum_{n=1}^{N} \mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log P(y^{(n)} \mid \underline{x}^{(n)}, \underline{w})]$$

on the other hand is a sum of integrals, which is hard to optimise in some closed-form way.

In practice, for optimising we:

1. for any variance σ_w^2, we optimise the unconstrained quantity $\log \sigma_w$ instead
2. to optimise the covariance matrix V, we apply the Cholesky decomposition:

   $$V = LL^T$$

   (easier to optimise triangular matrices). Since the diagonal elements are positive, we take the log of the diagonal elements, and then optimise the resulting unconstrained matrix

3.2.2 Closed-Form Expression for Entropy and Cross-Entropy

- **What is the expectation of the log of a Gaussian?**
  - consider a general Gaussian $\mathcal{N}(\underline{w}; \underline{\mu}, \Sigma)$, where:

    $$\underline{w} \sim \mathcal{N}(\underline{m}, V)$$

  - then:

    $$\begin{aligned}&\mathbb{E}_{\underline{w}\sim\mathcal{N}(\underline{m},V)}\left[\log \mathcal{N}(\underline{w}; \underline{\mu}, \Sigma)\right]\\ =&\mathbb{E}_{\underline{w}\sim\mathcal{N}(\underline{m},V)}\left[\log\left(\frac{1}{|2\pi\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\underline{w}-\underline{\mu})^T\Sigma^{-1}(\underline{w}-\underline{\mu})\right)\right)\right]\\ =&\mathbb{E}_{\underline{w}\sim\mathcal{N}(\underline{m},V)}\left[-\frac{1}{2}(\underline{w}-\underline{\mu})^T\Sigma^{-1}(\underline{w}-\underline{\mu})\right] - \frac{1}{2}\log|2\pi\Sigma|\end{aligned}$$

*To deal with expectations of quadratic forms, it is useful to use the **trace trick**.*
Since the trace of a scalar is itself, and it is a linear operator:

$$-\frac{1}{2}(\underline{w}-\underline{\mu})^T\Sigma^{-1}(\underline{w}-\underline{\mu}) = -\frac{1}{2}Tr\left((\underline{w}-\underline{\mu})^T\Sigma^{-1}(\underline{w}-\underline{\mu})\right)$$

Moreover, a property of the trace is that:

$$Tr(AB) = Tr(BA)$$

so:

$$-\frac{1}{2}(\underline{w}-\underline{\mu})^T\Sigma^{-1}(\underline{w}-\underline{\mu}) = -\frac{1}{2}Tr\left((\underline{w}-\underline{\mu})(\underline{w}-\underline{\mu})^T\Sigma^{-1}\right)$$

Hence:

$$\begin{aligned}&\mathbb{E}_{\underline{w}\sim\mathcal{N}(\underline{m},V)}\left[-\frac{1}{2}(\underline{w}-\underline{\mu})^T\Sigma^{-1}(\underline{w}-\underline{\mu})\right]\\ =&\mathbb{E}_{\underline{w}\sim\mathcal{N}(\underline{m},V)}\left[-\frac{1}{2}Tr\left((\underline{w}-\underline{\mu})(\underline{w}-\underline{\mu})^T\Sigma^{-1}\right)\right]\\ =&-\frac{1}{2}Tr\left(\mathbb{E}_{\underline{w}\sim\mathcal{N}(\underline{m},V)}\left[(\underline{w}-\underline{\mu})(\underline{w}-\underline{\mu})^T\right]\Sigma^{-1}\right)\end{aligned}$$

- **What is the closed-form expression for the negative entropy?**
  - we have the **negative entropy**:

$$-H(Q) = \mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log Q(\underline{w})]$$

    where:

$$Q(\underline{w};\alpha) = \mathcal{N}(\underline{w};\underline{m},V)$$

  - thus, and using the **trace trick**:

$$\begin{aligned}&\mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log Q(\underline{w})]\\ =&\mathbb{E}_{\underline{w}\sim\mathcal{N}(\underline{m},V)}[\log\mathcal{N}(\underline{w};\underline{m},V)]\\ =&\mathbb{E}_{\underline{w}\sim\mathcal{N}(\underline{m},V)}\left[-\frac{1}{2}(\underline{w}-\underline{m})^TV^{-1}(\underline{w}-\underline{m})\right]-\frac{1}{2}\log|2\pi V|\\ =&-\frac{1}{2}Tr\left(\mathbb{E}_{\underline{w}\sim\mathcal{N}(\underline{m},V)}\left[(\underline{w}-\underline{m})(\underline{w}-\underline{m})^T\right]V^{-1}\right)-\frac{1}{2}\log|2\pi V|\\ =&-\frac{1}{2}Tr\left(VV^{-1}\right)-\frac{1}{2}\log|2\pi V|\\ =&-\frac{D}{2}-\frac{1}{2}\log|2\pi V|\end{aligned}$$

- **What is the closed-form expression for the cross-entropy?**
  - we have the **cross-entropy**:

$$-\mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log P(\underline{w})]$$

- if we assume a **spherical Gaussian Prior**:

$$P(\underline{w}) = \mathcal{N}(\underline{w}; \underline{0}, \sigma_w^2)$$

then the **cross-entropy** becomes (again, using the trace trick, and adapting to higher dimensions the fact that $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$):

$$\begin{aligned}
&- \mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log P(\underline{w})] \\
=& - \mathbb{E}_{\underline{w}\sim \mathcal{N}(\underline{m},V)}[\log \mathcal{N}(\underline{w}; \underline{0}, \sigma_w^2)] \\
=& - \mathbb{E}_{\underline{w}\sim \mathcal{N}(\underline{m},V)}\left[-\frac{1}{2}\underline{w}^T\left(\frac{1}{\sigma_w^2}\mathbb{I}\right)\underline{w}\right] - \frac{1}{2}\log|2\pi\sigma_w^2\mathbb{I}| \\
=& \frac{1}{2\sigma_w^2}\mathbb{E}_{\underline{w}\sim \mathcal{N}(\underline{m},V)}\left[\underline{w}^T\underline{w}\right] + \frac{D}{2}\log(2\pi\sigma_w^2) \\
=& \frac{1}{2\sigma_w^2}\mathbb{E}_{\underline{w}\sim \mathcal{N}(\underline{m},V)}\left[\underline{w}^T\underline{w}\right] + \frac{D}{2}\log(2\pi\sigma_w^2) \\
=& \frac{1}{2\sigma_w^2}Tr(\mathbb{E}_{\underline{w}\sim \mathcal{N}(\underline{m},V)}\left[\underline{ww}^T\right]) + \frac{D}{2}\log(2\pi\sigma_w^2) \\
=& \frac{1}{2\sigma_w^2}(Tr(V) + \underline{m}^T\underline{m}) + \frac{D}{2}\log(2\pi\sigma_w^2)
\end{aligned}$$

- **How can we compute the gradients of the entropy terms?**
  - both the entropy and cross entropy are **differentiable functions** of $\underline{m}$ and V
  - the terms involving V are functions of the Cholesky decomposition:

$$\frac{1}{2}\log|V|)\sum_i \log L_{ii}$$

$$Tr(V) = \sum_{ij} L_{ij}^2$$

  - hence, we can easily compute gradients

3.2.3 Approximating the Log-Likelihood

- **How is the log-likelihood approximated?**
  - the **log-likelihood** we have is a sum of integrals:

$$\mathbb{E}_{\underline{w}\sim Q(\underline{w};\alpha)}[\log P(\mathcal{D} \mid \underline{w})] = \mathbb{E}_{\underline{w}\sim \mathcal{N}(\underline{m},V)}[\log P(\mathcal{D} \mid \underline{w})] = \sum_{n=1}^{N}\mathbb{E}_{\underline{w}\sim \mathcal{N}(\underline{m},V)}[\log P(y^{(n)} \mid \underline{x}^{(n)}, \underline{w})]$$

  - approximating each integral (as a 1D integral) is very expensive
  - instead, we use **Monte Carlo**, to compute an **unbiased estimate** by sampling a random weight $\underline{w}$:

$$\mathbb{E}_{\underline{w}\sim \mathcal{N}(\underline{m},V)}[\log P(\mathcal{D} \mid \underline{w})] \approx \sum_{n=1}^{N}\log P(y^{(n)} \mid \underline{x}^{(n)}, \underline{w}), \qquad \underline{w} \sim \mathcal{N}(\underline{m}, V)$$

  - alternatively, we can randomly select a sample, and scale up its contribution:

$$\mathbb{E}_{\underline{w}\sim \mathcal{N}(\underline{m},V)}[\log P(\mathcal{D} \mid \underline{w})] \approx N\log P(y^{(n)} \mid \underline{x}^{(n)}, \underline{w}), \quad \underline{w} \sim \mathcal{N}(\underline{m}, V), n \sim Uniform[1, N]$$

- finally, we could consider the **average** log-likelihood, for a **random minibatch** of S sample weights, and then scale this up by N:

$$\mathbb{E}_{\underline{w} \sim \mathcal{N}(\underline{m}, V)}[\log P(\mathcal{D} \mid \underline{w})] \approx \frac{N}{S} \sum_{s=1}^{S} \log P(y^{(n)} \mid \underline{x}^{(n)}, \underline{w}^{(s)}), \quad \underline{w}^{(s)} \sim \mathcal{N}(\underline{m}, V), n \sim Uniform[1, N]$$

- **How can we compute the gradients for the log-likelihood?**
  - we begin by using a **reparametrisation**: let $\underline{\nu} \sim \mathcal{N}(\underline{0}, \mathbb{I})$, and define:

    $$\underline{w} = \underline{m} + L\underline{\nu}$$

    (where $V = LL^T$ according to the Cholesky decmposition)
  - then:

    $$\mathbb{E}_{\underline{w} \sim \mathcal{N}(\underline{m}, V)}[\log P(\mathcal{D} \mid \underline{w})] = \mathbb{E}_{\underline{\nu} \sim \mathcal{N}(\underline{0}, \mathbb{I})}[\log P(\mathcal{D} \mid (\underline{m} + L\underline{\nu}))]$$
  - since the **expectation** is now over a constant distribution, we can compute (or approximate) the gradients much more easily. If we write:

    $$f(\underline{w}) = \log P(\mathcal{D} \mid \underline{w})$$

    then:

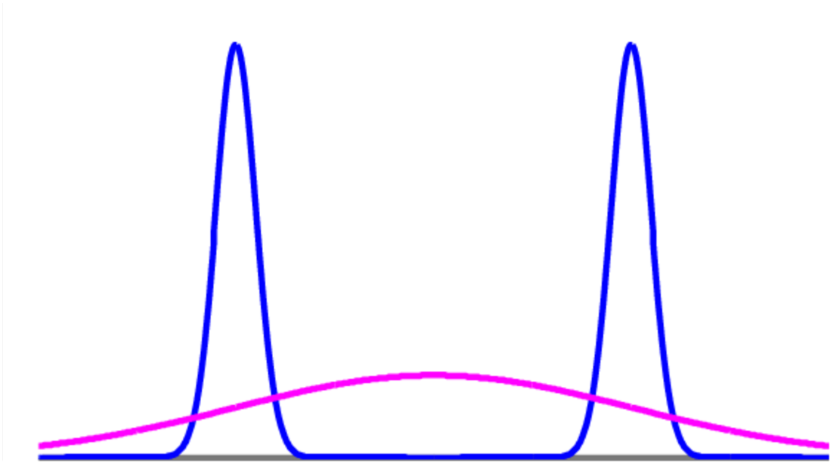
    $$\begin{aligned} \nabla_m \mathbb{E}_{\underline{\nu} \sim \mathcal{N}(\underline{0}, \mathbb{I})}[f(\underline{w})] &= \mathbb{E}_{\underline{\nu} \sim \mathcal{N}(\underline{0}, \mathbb{I})}[\nabla_m f(\underline{w})] \\ &\approx \nabla_m f(\underline{m} + L\underline{\nu}) \end{aligned}$$

    $$\begin{aligned} \nabla_L \mathbb{E}_{\underline{\nu} \sim \mathcal{N}(\underline{0}, \mathbb{I})}[f(\underline{w})] &= \mathbb{E}_{\underline{\nu} \sim \mathcal{N}(\underline{0}, \mathbb{I})}[\nabla_L f(\underline{w})] \\ &\approx \nabla_L f(\underline{m} + L\underline{\nu}) \\ &= \nabla_w f(\underline{w})(\nabla_L \underline{w}) \\ &= \nabla_w f(\underline{w})\underline{\nu}^T \end{aligned}$$
  - hence, to estimate the gradients we just need to be able to find the gradient of **log-likelihood** (which we know from MLE)

4 Question

4.1 Notes Questions

1. **Consider a 1-dimensional bimodal posterior, alongside a Gaussian sampling distribution $q(\underline{w})$ centered at the trough of the posterior:**



Would you expect a large or small importance weight at the mode of $q(\underline{w})$? What about the importance weights at the modes of $P(\underline{w} \mid \mathcal{D})$?

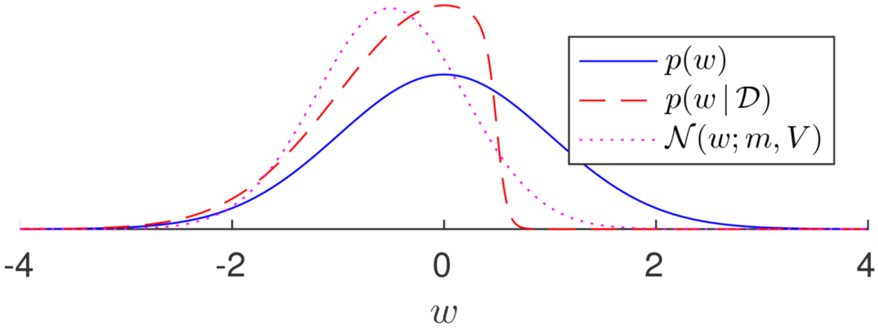
- recall, the importance weight is given by:

$$r^{(s)} = \frac{P(\underline{w}^{(s)} \mid \mathcal{D})}{q(\underline{w}^{(s)})}$$

- at the mode of q we have the the trough of the posterior; here, the posterior is near 0, so the importance weight will be near 0 too
- at the modes, q is small, whilst the posterior is the largest, so we will have the highest importance weight

2. **For logistic regression, if you just wanted to make hard decisions (i.e just report if $P(y = 1 \mid \underline{x}, \mathcal{D}) > 0.5$), would there be any point in doing variational inference rather than just using the Laplace approximation?**

- we saw above that with the Laplace approximation, the approximate posterior will be nearly identical to the prior:



- as such, the MAP weights would be close to 0
- on the other hand, with **variational inference**, and assuming that the observations used to define the model are representative, the approximate posterior better matches the true posterior, so the decision boundary formed by the weights sampled using variational inference could yield better hard decisions

3. **If we have derivatives of a cost with respect to the prior variance σ_w^2, how do we convert them into derivatives with respect to:**

$$s_w = \log \sigma_w$$

   - we have that:

$$e^{s_w} = \sigma_w \implies e^{2s_w} = \sigma_w^2$$

   - thus:

$$\frac{dc}{ds_w} = \frac{dc}{d\sigma_w^2}\frac{d\sigma_w^2}{ds_w} = \frac{dc}{d\sigma_w^2}(2e^{2s_w}) = 2\sigma_w^2\frac{dc}{d\sigma_w^2}$$

4. **When attempting to fit $\underline{m}$ and V by stochastic gradient descent, you might find that the updates are quite noisy, and the variational parameters don't converge. What is a way that you might fix this issue?**

   - reduce the learning rate
   - instead of drawing a single **Gaussian** weight for each step, use a **minibatch**, and compute the **average gradient**; this should average out the noise and give less erratic updates