Bayesian Machine Learning

1 Introduction

Let $D = (x_1, y_1, \ldots, x_n, y_n)$ be a dataset. It is seen as a random vector whose subvectors (x_i, y_i) are **independent and identically distributed** according to an unknown joint distribution $p(x, y)$. We can write $p(x, y) = p(y|x)p(x)$.
Given a random variable x^* sampled from $p(x)$ and independent from all (x_i, y_i), our goal is to compute the **conditional distribution** $p(y|x^*, D)$.
We have a **family of models** $p(y|x, w)$ parameterized by w. The parameter w is seen as a random variable. The conditional distribution is obtained by marginalizing the parameter w:

$$p(y|x^*, D) = \int p(y, w|x^*, D)dw.$$

We can expand the integrand:

$$p(y|x^*, D) = \int p(y|x^*, w, D)p(w|x^*, D)dw.$$

Now, we have to make two assumptions:

- The parameter w is independent with x^*. In fact, we will assume that w and x_i are independent for all i. The collection of the data is independent from the training.
- The chain $D \to w \to y|x^*$ is Markov. Intuitively, after training, the parameter w contains all the information about the dataset D that are necessary to infer $p(y|x^*)$.

Thus, we obtain **Bayesian Model Averaging** (BMA):

$$p(y|x^*, D) = \int p(y|x^*, w)p(w|D)dw.$$

The term $p(y|x^*, w)$ corresponds to a model from our family of model. The prediction of this model is weighted by $p(w|D)$. We compute $p(w|D)$ with **Bayesian inference**:

$$p(w|D) = \frac{p(D|w)p(w)}{\int p(D|w')p(w')dw'}.$$

The various terms are:

- The **prior** $p(w)$. It is chosen by us. We will discuss how to choose the prior.
- The **likelihood**:
$$p(D|w) = \prod_{i=1}^{n} p(y_i|x_i, w)p(x_i|w) = \prod_{i=1}^{n} p(y_i|x_i, w)p(x_i).$$
Here, we used the assumption that x_i and w are independent. Thus, $p(x_i|w) = p(x_i)$ does not depend on w. Maximizing $\prod_{i=1}^{n} p(y_i|x_i, w)$ is equivalent to maximizing $p(D|w)$.
The negative log-likelihood $-\sum_{i=1}^{n} \log p(y_i|x_i, w)$ is often used as the loss function in non-Bayesian machine learning (along with regularization terms).
- The **evidence** (or **model evidence**, or **marginal likelihood**) $p(D) = \int p(D|w')p(w')dw'$. It is considered to be intractable except for very simple models. Its value is used for *model selection*: better models should yield higher evidence.

2 Variational Inference

There are two methods to approximate $p(w|D)$ when the evidence $p(D)$ is intractable:

- Markov Chain Monte Carlo (MCMC).
- Variational Inference (VI).

MCMC is more precise (and has asymptotic guarantees) but is more computationally heavy. In the rest of this document, we will focus on Variational Inference.

2.1 Evidence Lower Bound

We consider a **variational family** $q(w|\xi)$ parameterized by ξ. Variational Inference consists in using optimization to find a parameter ξ^* such that $q(w|\xi^*)$ is "close" to $p(w|D)$. Then, we will use $q(w|\xi^*)$ instead of $p(w|D)$ to compute $p(y|x^*, D)$. The "closeness" is measured with the KL divergence:

$$\begin{aligned}\mathrm{KL}\left[q(w|\xi)||p(w|D)\right] &= \mathbb{E}_{w|\xi}\left[\log \frac{q(w|\xi)}{p(w|D)}\right], \\ &= \mathbb{E}_{w|\xi}\left[\log q(w|\xi)\right] - \mathbb{E}_{w|\xi}\left[\log p(w|D)\right], \\ &= H\left(q(w|\xi), p(w|D)\right) - H\left(q(w|\xi)\right),\end{aligned}$$

where $H(p)$ is the entropy, and $H(p, q)$ is the cross-entropy.

Since $p(w|D)$ is unknown, we cannot compute this KL divergence. Thus, we will compute another quantity whose minimum is achieved at the same ξ^* than the minimum of the KL divergence. Using the law of contional probability:

$$\begin{aligned}\mathrm{KL}\left[q(w|\xi)||p(w|D)\right] &= \mathbb{E}_{w|\xi}\left[\log q(w|\xi)\right] - \mathbb{E}_{w|\xi}\left[\log p(D, w)\right] + \mathbb{E}_{w|\xi}\left[\log p(D)\right], \\ &= \mathbb{E}_{w|\xi}\left[\log q(w|\xi)\right] - \mathbb{E}_{w|\xi}\left[\log p(D, w)\right] + \log p(D).\end{aligned}$$

Here, we used the fact that $\log p(D)$ is a constant wrt to $w|\xi$.
Define the **Evidence Lower Bound** or **negative variational free energy**:

$$\begin{aligned}\mathrm{ELBO}(\xi) &= \mathbb{E}_{w|\xi}\left[\log p(D, w)\right] - \mathbb{E}_{w|\xi}\left[\log q(w|\xi)\right], \\ &= -U\left(q(w|\xi), p(D, w)\right) + H\left(q(w|\xi)\right),\end{aligned}$$

where $U(q(w|\xi), p(D, w)) = -\mathbb{E}_{w|\xi}\left[\log p(D, w)\right]$ is the energy. We obtain a similar relation than for the free energy A in thermodynamics: $A = U - TS$, where U is the energy, S the entropy, and T the temperature (a constant).
The ELBO can also be expressed as:

$$\begin{aligned}\mathrm{ELBO}(\xi) &= \mathbb{E}_{w|\xi}\left[\log p(D|w) + \log p(w) - \log q(w|\xi)\right], \\ &= \mathbb{E}_{w|\xi}\left[\log p(D|w)\right] - \mathrm{KL}\left[q(w|\xi)||p(w)\right],\end{aligned}$$

which corresponds to the expected log-likelihood plus a regularization term.
We have:

$$\mathrm{KL}\left[q(w|\xi)||p(w|D)\right] = \log p(D) - \mathrm{ELBO}(\xi).$$

Minimizing the KL divergence wrt ξ is equivalent to maximizing the ELBO. Using the positivity of KL divergence (i.e., Gibbs' inequality), we have $\log p(D) \geq \mathrm{ELBO}(\xi)$, hence the name. Equality is achieved iff $q(w|\xi) = p(w|D)$.

2.2 Mean-field Approximation

The mean-field approximation consists in partitioning the latent variables $w = (w_1, \ldots, w_h)$, and **assuming** that each factor w_j is conditionally independent with the other factors w_i when conditioning on the data D:

$$p(w_j|D, w_i) = p(w_j|D).$$

Thus:

$$\begin{aligned} p(w|D) &= p(w_1|w_2,\dots,w_h,D)p(w_2|w_3,\dots,w_h,D)\dots p(w_h|D), \\ &= \prod_{j=1}^{h} p(w_j|D). \end{aligned}$$

Hence, we define the **mean-field variational family** as:

$$q(w|\xi) = \prod_{j=1}^{h} q_j(w_j|\xi_j).$$

The general algorithm to maximize the ELBO using mean-field variational family is **coordinate ascent variational inference**. It is an "EM-like" algorithm[1], that is able to solve a system of equations with circular dependence using an iterative procedure. We start with some initial estimates of ξ, and update each factor ξ_j in turns using the estimate of the other factors, denoted ξ_{-j}. The ELBO must be computed after every h steps to check the convergence. This algorithm is guaranteed to converge to a local maximum of the ELBO.

Algorithm 1 Coordinate ascent variational inference

Input: the joint distribution $p(D, w)$, a dataset D.
Output: variational parameters $q(w|\xi)$.
  Initialize: variational factors $q_j(w_j|\xi_j)$.
  while ELBO has not converged **do**
    for $j \in \{1,\dots,h\}$ **do**
      $q_j(w_j|\xi_j) \propto \exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right)$ ▷ Coordinate update
    end for
    Compute $\text{ELBO}(\xi) = \mathbb{E}_{w|\xi}\left[\log p(D,w)\right] - \mathbb{E}_{w|\xi}\left[\log q(w|\xi)\right]$.
  end while
  return $q(w|\xi)$

Here is a proof for the coordinate update. Assume that ξ_{-j} is constant, and write the ELBO as a function of ξ_j:

$$\begin{aligned} \text{ELBO}(\xi_j) &= \mathbb{E}_{w|\xi}\left[\log p(D,w)\right] - \mathbb{E}_{w|\xi}\left[\log q(w|\xi)\right], \\ &= \mathbb{E}_{w_j|\xi_j}\left[\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right] - \sum_{i=1}^{k}\mathbb{E}_{w|\xi}\left[\log q_i(w_i|\xi_i)\right], \\ &= \mathbb{E}_{w_j|\xi_j}\left[\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right] - \sum_{i=1}^{k}\mathbb{E}_{w_i|\xi_i}\left[\log q_i(w_i|\xi_i)\right], \\ &= \mathbb{E}_{w_j|\xi_j}\left[\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right] - \mathbb{E}_{w_j|\xi_j}\left[\log q_j(w_j|\xi_j)\right] + C, \\ &= \mathbb{E}_{w_j|\xi_j}\left[\log \frac{\exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right)}{Z(\xi_{-j},D)}\right] + \log Z(\xi_{-j},D) - \mathbb{E}_{w_j|\xi_j}\left[\log q_j(w_j|\xi_j)\right] + C, \\ &= -\text{KL}\left[q_j(w_j|\xi_j)\,\middle\|\,\frac{\exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right)}{Z(\xi_{-j},D)}\right] + C'. \end{aligned}$$

where the constants C and C' do not depend on ξ_j, and:

$$Z(\xi_{-j},D) = \int \exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right) dw_j.$$

[1]However, there is no distinction between expectation and maximization steps, because there is no distinction between latent variables and parameters. In some sense, every step is an expectation step.

We went from line 2 to line 3 by remarking that the multiple integrals become single integrals, since each term of the sum depends only on one variational factor. Since the KL divergence is non-negative, the ELBO is maximized when the KL divergence is zero, i.e., when:

$$q_j(w_j|\xi_j) = \frac{\exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right)}{Z(\xi_{-j}, D)}.$$

2.3 Conditionally Conjugate Models

In order to apply the coordinate ascent variational inference, we must be able to derive an analytical formula for $\exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right)$. This is the hard part of the variational inference method. Sometimes, it is impossible. However, if the complete conditionals are in an **exponential family**, then it is always possible (but still hard).

Assume that the **complete conditionals** are in an exponential family:

$$p(w_j|w_{-j}, D) = \exp\left(\langle\theta_j(w_{-j}, D), w_j\rangle - \psi(\theta_j(w_{-j}, D))\right).$$

The coordinate update becomes[2]:

$$\begin{aligned} q_j(w_j|\xi_j) &= \exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\langle\theta_j(w_{-j}, D), w_j\rangle - \psi(\theta_j(w_{-j}, D))\right]\right), \\ &\propto \exp\left(\left\langle\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\theta_j(w_{-j}, D)\right], w_j\right\rangle\right). \end{aligned}$$

We state this result in the following proposition:

Proposition 2.1. *For all $j \in \{1, \ldots, h\}$, assume that the complete conditional $p(w_j|w_{-j}, D)$ belongs to an exponential family with natural parameter $\theta_j(w_{,j}, D)$.*
Then, the optimal variational factor $q_j(w_j|\xi_j)$ computed with the coordinate update belongs to the same exponential family. Its natural parameter is $\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\theta_j(w_{-j}, D)\right]$, i.e., the expectation of the parameter of the complete conditional.

The **conditionally conjugate models** (CCM) are special cases of exponential family models. Here, we have two types of latent variables:

- *Global latent variables* β that govern all the data.
- *Local latent variables* z_i that only governs the i-th data point.

The joint density of CCM is:

$$p(\beta, z, D) = p(\beta)\prod_{i=1}^{n} p(z_i, D_i|\beta).$$

We see that, given β and D_i, this model assumes that z_i is conditionally independent of all z_j and D_j for $j \neq i$. In order to be an exponential family model, we must choose $p(\beta)$ and $p(z_i, D_i|\beta)$ such that each complete conditional is in the exponential family. To do so, we **assume** that:

$$p(z_i, D_i|\beta) = \exp\left(\langle\beta, T(z_i, D_i)\rangle - \psi(\beta)\right),$$

for some sufficient statistic T. Now, we can choose the prior $p(\beta)$ to be the corresponding **conjugate prior**:

$$p(\beta) = \exp\left(\langle\alpha_1, \beta\rangle - \langle\alpha_2, \psi(\beta)\rangle - \chi(\alpha_1, \alpha_2)\right).$$

[2]Note that we can replace $\exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right)$ by $\exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(w_j|w_{,j}, D)\right]\right)$ because $\exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(D,w)\right]\right) \propto \exp\left(\mathbb{E}_{w_{-j}|\xi_{-j}}\left[\log p(w_j|w_{,j}, D)\right]\right)$, where the proportional constant does not depend on w_j. This is because $p(D,w) = p(w_j|w_{,j}, D)p(w_{-j}, D)$.

The complete conditional of β is the posterior distribution:

$$\begin{aligned}
p(\beta|z, D) &\propto p(z, D|\beta)p(\beta), \\
&\propto p(\beta)\prod_{i=1}^{n} p(z_i, D_i|\beta), \\
&\propto \exp\left(\langle\alpha_1, \beta\rangle - \langle\alpha_2, \psi(\beta)\rangle\right)\exp\left(\sum_{i=1}^{n}\langle\beta, T(z_i, D_i)\rangle - n\psi(\beta)\right), \\
&\propto \exp\left(\left\langle\alpha_1 + \sum_{i=1}^{n} T(z_i, D_i), \beta\right\rangle - \langle\alpha_2 + n, \psi(\beta)\rangle\right).
\end{aligned}$$

We see that the complete conditional of β is in the same exponential family as the prior $p(\beta)$. This is why $p(\beta)$ is called the conjugate prior of $p(z_i, D_i|\beta)$. The natural parameters of $p(\beta|z, D)$ are $(\alpha_1 + \sum_{i=1}^{n} T(z_i, D_i), \alpha_2 + n)$.

Now, we consider the complete conditional of z_i. By the conditional independence, we have:

$$p(z_i|\beta, z_{-i}, D) = p(z_i|D_i, \beta).$$

We **assume** that this is again an exponential family:

$$p(z_i|D_i, \beta) = \exp\left(\langle\theta(\beta, D_i), z_i\rangle - \Xi(\theta(\beta, D_i))\right).$$

In order to apply the coordinate ascent variational inference, we choose a mean-field variational family:

$$q(\beta, z|\xi) = q(\beta|\xi_0)\prod_{i=1}^{n} q(z_i|\xi_i).$$

According to Proposition 2.1, the optimal variational factors belongs to the same exponential families as the complete conditionals, and their natural parameters are the expectation of the parameters of the complete conditionals. Thus:

$$q(\beta|\xi_0) \propto \exp\left(\left\langle\alpha_1 + \sum_{i=1}^{n}\mathbb{E}_{z_i|\xi_i}\left[T(z_i, D_i)\right], \beta\right\rangle - \langle\alpha_2 + n, \psi(\beta)\rangle\right),$$

and:

$$q(z_i|\xi_i) \propto \exp\left(\left\langle\mathbb{E}_{\beta|\xi_0}\left[\theta(\beta, D_i)\right], z_i\right\rangle\right).$$

We can also compute the ELBO:

$$\begin{aligned}
\text{ELBO}(\xi) &= \mathbb{E}_{\beta,z|\xi}\left[\log p(D, \beta, z)\right] - \mathbb{E}_{\beta,z|\xi}\left[\log q(\beta, z|\xi)\right], \\
&= \mathbb{E}_{\beta,z|\xi}\left[\log p(\beta) + \sum_{i=1}^{n}\log p(z_i, D_i|\beta)\right] - \mathbb{E}_{\beta,z|\xi}\left[\log q(\beta|\xi_0) + \sum_{i=1}^{n}\log q(z_i|\xi_i)\right], \\
&= \left\langle\alpha_1 + \sum_{i=1}^{n}\mathbb{E}_{z_i|\xi_i}\left[T(z_i, D_i)\right], \mathbb{E}_{\beta|\xi_0}\left[\beta\right]\right\rangle - \left\langle\alpha_2 + n, \mathbb{E}_{\beta|\xi_0}\left[\psi(\beta)\right]\right\rangle + C + ...
\end{aligned}$$

I am extremely confused. Blei et al. do not include the normalizing terms of the left term, but do include the normalizing terms of the right term. Moreover, the terms from $p(\beta|z, D)$ seems to cancel with the terms from $q(\beta|\xi_0)$ leaving only the normalizing terms and the terms from $q(z_i|\xi_i)$. Blei et al. never wrote the complete expression, such that this cancellation never appears. This is dubious. Using the abuse of notation $\psi(p(\cdot))$ for the log-normalization terms, it seems to me that the final expression is:

$$\text{ELBO}(\xi) = -\mathbb{E}_{\beta,z|\xi}\left[\psi(p(D, \beta, z))\right] + \psi(q(\beta|\xi_0) + \sum_{i=1}^{n}\psi(q(z_i|\xi_i)) - \sum_{i=1}^{n}\left\langle\mathbb{E}_{\beta|\xi_0}\left[\theta(\beta, D_i)\right], \mathbb{E}_{z_i|\xi_i}\left[z_i\right]\right\rangle.$$

2.4 Stochastic variational inference

In order to address large dataset, we would like to avoid summing over all data "$(\sum_{i=1}^n)$" in the expressions introduced in the previous subsection. Thus, we must find an alternative to coordinate ascent variational inference. One alternative is **gradient-based optimization**, where we use the gradient of the ELBO to do gradient ascent. Instead of using the entire dataset to compute the gradient, we can sample a subset of the data points. The gradient ascent is used only for the global parameters β. The local parameters z_i are optimized using the current estimate of β and the subsampled data points.

3 Example: Probit Regression

4 Uncertainty Propagation

Now, we consider the **logistic regression**. In this case, the joint distribution is not an exponential family (why?). Thus, we cannot apply coordinate ascent variational inference because we cannot compute the coordinate update, as well as the first term of the ELBO: $\mathbb{E}_{w|\xi}\left[\log p(D, w)\right]$.

The model is:

$$p(y = c|x, w) = \sigma^c(wx),$$

where $x \in \mathbb{R}^d$, $w \in \mathbb{R}^{m\times d}$, and $\sigma^c : \mathbb{R}^m \to \mathbb{R}$ is the c-th component of the softmax function.
The log-likelihood of the entire dataset is:

$$\begin{aligned}
\log p(D|w) &= \sum_{i=1}^n \log p(y_i, x_i|w), \\
&= \sum_{i=1}^n \log p(y_i|x_i, w) + \log p(x_i), && (x_i \text{ and } w \text{ are independent})
\end{aligned}$$

We can remove the constant term $\sum_{i=1}^n \log p(x_i)$ in the ELBO and consider:

$$\sum_{i=1}^n \log p(y_i|x_i, w) = \sum_{i=1}^n \delta(y_i)^T \log \sigma(wx_i),$$

where $\delta(y_i) \in \mathbb{R}^m$ is the one-hot encoding of y_i. We obtain:

$$\begin{aligned}
\text{ELBO}(\xi) &= \mathbb{E}_{w|\xi}\left[\sum_{i=1}^n \delta(y_i)^T \log \sigma(wx_i)\right] - \text{KL}\left[q(w|\xi)||p(w)\right], \\
&= \sum_{i=1}^n \delta(y_i)^T \mathbb{E}_{w|\xi}\left[\log \sigma(wx_i)\right] - \text{KL}\left[q(w|\xi)||p(w)\right].
\end{aligned}$$

5 Bayesian priors

The choice of a prior in Bayesian inference depends on the end goal of the researcher. Here is a typology of different types of priors:

- **Informative priors** contain knowledge. This knowledge can be subjective, it can be expert knowledge. Or it can come from previous data, or other sources of information. *Prior elicitation* aims at transforming this knowledge into priors. Current data can also be used to learn the prior. In this case, we deviate from the Bayesian paradigm, since the prior should not depend on the data. Thus, there is a risk of overfitting. *Model selection* aims at using data to choose the best prior. It relies on tools such that Bayesian information criterion (BIC) or marginal likelihood (i.e., Bayesian evidence $p(z|M) = \int p(z|w)p(w|M)dw$). Remark than model selection is an "hypothesis testing problem" as opposed to an "estimation problem".

- **Regularization priors** enforce some pre-determined properties, such as smoothness or sparsity.
- **Conjugate priors** reduce the computational complexity of the posterior analysis.
- **Uninformative priors** assume as little as possible about the data in order to minimize its effect on the posterior.
  Jeffreys prior is the most famous uninformative prior. It is *invariant under reparametrization of the parameter space* w. However, it has several drawbacks. It is hard to compute, it may be improper (i.e., normalization is infinite), it is not defined for singular models, and it may not be as uninformative as it claims to be. Jeffreys prior is uninformative only in the limit of infinite data. This is weird because with infinite data, the prior has no impact on the posterior (assuming that the prior is supported on the entire parameter space).
  People tend to get rid of the prior choice problem by choosing Gaussian priors. Gaussian priors are often said to be "uninformative", which may be false (why?).

Our goal is to focus on uninformative priors. Here are the motivations:

- Avoid overfitting and overconfidence.
- Favor "simple" models, in order to avoid model unstability (i.e., adversarial attacks).
- Learn from limited data (few-shot learning, semi-supervised learning, transfer learning).

5.1 Reference priors

One promising direction is the *reference prior*. Consider a machine learning architecture (e.g., a neural network architecture) characterized by its likelihood $p(z|w)$. For the moment, we ignore the distinction between x and y and denote the data as z. We are looking for the prior $\pi(w)$ that is as far as possible from the posterior $p(w|z)$. We use the KL divergence:

$$\mathrm{KL}[p(w|z)||\pi(w)] = \int p(w|z) \log \frac{p(w|z)}{\pi(w)} dw.$$

Intuitively, we want the posterior to be dominated by the data z and not by the prior $\pi(w)$. Since, we haven't observed the data, we maximize the average of the KL divergence:

$$\begin{aligned}
\mathbb{E}_z \left[\mathrm{KL}[p(w|z)||\pi(w)]\right] &= \int \mathrm{KL}\left[p(w|z)||\pi(w)\right] p(z) dz, \\
&= \int p(w|z)p(z) \log \frac{p(w|z)}{\pi(w)} dw dz, \\
&= \int p(w,z) \log \frac{p(w,z)}{\pi(w)p(z)} dw dz, \\
&= \mathrm{KL}\left[p(w,z)||\pi(w)p(z)\right], \\
&= I_\pi(w,z),
\end{aligned}$$

where $I_\pi(w,z)$ is the mutual information between $\pi(w)$ and $p(z)$.
Remark. The mutual information measures the amount of information that $\pi(w)$ contains about $p(z)$ (and vice versa since it is symmetric). It is the reduction of uncertainty of $p(z)$ due to the knowledge of $\pi(w)$. It measures the inefficiency of assuming that $\pi(w)$ and $p(z)$ are independent. Imagine that we want to construct a code for messages sampled from (w,z). We assume that w and z are independent and we use the optimal code of the distribution $\pi(w)p(z)$. However, the true distribution is $p(w,z)$. If we had used the true distribution $p(w,z)$, we would have needed $I_\pi(w,z)$ less bits on average to code a sample from (w,z).

Moreover, we have:

$$\begin{aligned}
I_\pi(w,z) &= \int p(w,z)\log\frac{p(w|z)}{\pi(w)}dwdz,\\
&= \int p(w,z)\log p(w|z)dwdz - \int p(w,z)\log\pi(w)dwdz,\\
&= -\int\left(\int p(w,z)dz\right)\log\pi(w)dw + \int p(z')\left(\int p(w|z=z')\log p(w|z=z')dw\right)dz',\\
&= -\int\pi(w)\log\pi(w) + \int p(z')H(w|z=z')dz',\\
&= H(w) - H(w|z),
\end{aligned}$$

where $H(w)$ is the entropy of $\pi(w)$ and $H(w|z)$ is the conditional entropy[3] of w given z. It seems natural for an *uninformative* prior to maximize the mutual information, i.e., maximizing the information brought by the data about the parameters.

Since the mutual information is symmetric, we can write:

$$\begin{aligned}
I_\pi(w,z) &= H(z) - H(z|w),\\
&= \int p(z|w)\pi(w)\log(p(z|w))dwdz - \int p(z|w)\pi(w)\log\left(\int p(z|w)\pi(w)dw\right)dwdz,\\
&= \int p(z|w)\pi(w)\log\frac{p(z|w)}{p(z)}dwdz,\\
&= \int p(z|w)\pi(w)\log\frac{p(z|w)}{\int p(z|w)\pi(w)dw}dwdz.
\end{aligned}$$

Here is another interpretation of reference prior. In order to maximize $I_\pi(w,z)$, we have to maximize $H(z)$ and minimize $H(z|w)$. Maximizing $H(z)$ means that the various models $p(z|w)$ should be as diverse as possible, i.e., we are looking at the average predictive distribution over z by averaging over all models w, and we want this distribution to have the highest entropy possible. Minimizing $H(z|w)$ means that each model w should be as confident as possible in its prediction. Thus, the reference prior is an ensemble of confident models that are spread as much as possible across the parameter space.

5.2 Practical implementation

It has been shown that if z is a fixed and finite iid sample, then the solution of $\arg\max_\pi I_\pi(w,z)$ is a discrete distribution with a finite number of atoms. Hence, in practice, we will choose a finite number K of atoms w_k and maximize $I_\pi(w,z)$ with gradient ascent.

One remaining question is how to deal with the input x in a predictive setting $p(y|x,w)$. In [1], the authors use unlabeled inputs. Here, we see the input x as a "nuisance" parameter for the predictive likelihood $p(y|x,w)$. We assume that x and w are independent such that the prior can be decomposed as $\pi(x,w) = \pi(w)p(x)$. We also define N atoms for the inputs x_i (the number N is an hyperparameter called the "order" of the reference prior[4]). For simplicity, we assume that the N atoms are equiprobable (i.e. $p(x_i) = 1/N$ for all i). The functional to be maximized is:

$$I_\pi((w,x),y) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left\{\mathbb{E}_{y|x_i,w_k}\left[\log p(y|x_i,w_k)\right] - \mathbb{E}_{y|x_i,w_k}\left[\log\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\pi(w_k)p(y|x_i,w_k)\right]\right\}\pi(w_k),$$

If the number of classes is small, the expectations can be computed exactly. Otherwise, Monte Carlo approximation can be used. Let m be the number of classes. First, we compute the M-dimensional vector

[3] Be careful not to confuse the entropy of the distribution $p(w|z=z')$ where the value of z is fixed at z', and the conditional entropy $H(w|z)$ where we average over z (i.e., z is *not* fixed).

[4] Is it really the order as defined in Bernardo's paper? The order should be the sample size of y (or is it?)

$p(y)$ as:

$$p(y) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} p(y|x_i, w_k)\pi(w_k),$$

where each $p(y|x_i, w_k)$ is the M-dimensional vector of probabilities for each class according to model w_k with input x_i. Then, we compute the mutual information:

$$I_\pi((w, x), y) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{j=1}^{M} p(y_j|x_i, w_k)\pi(w_k) \log p(y_j|x_i, w_k) - \sum_{j=1}^{M} p(y_j) \log p(y_j).$$

The rationales for seeing x as parameters rather than using data are the following:

- The prior should not depend on the data in order to avoid overfitting and overconfidence.
- The prior may be able to "learn" a notion of simplicity for the inputs that will avoid adversarial attacks when the model is tested on unseen inputs.

However, this prior might lead to poor accuracy of the posterior, or be unable to learn with limited data. This will be determined experimentally.

Once the maximum (w, x) is obtained (using backpropagation), no more optimization is required to complete the actual training (with the data). The optimum x is discarded, and we compute the log-likelihoods $\frac{1}{n} \sum_{j=1}^{n} \log p(y_j|x_j, w_k)$ over the entire dataset, for all w_k. Then, we can:

- Select the w_k with the highest log-likelihood and using it as predictive model, i.e., MAP. Or use any other point summary of the posterior, such as the mean of the median. This will reduce the memory complexity and computational complexity during inference.
- Use the entire posterior to do Bayesian model averaging. The probability of some class c is:

  $$p(y = c|x^*, D) = \sum_{k=1}^{K} \lambda_k p(y = c|x^*, w_k),$$

  where:

  $$\lambda_k = \frac{\prod_{i=1}^{n} p(y_i|x_i, w_k)}{\sum_{k'=1}^{K} \prod_{i=1}^{n} p(y_i|x_i, w_{k'})}.$$

The training only requires one "inference" pass through the dataset for each of the K atoms. The computational heavy part is the computation of the prior. However, once the prior is computed, it can be used for any task (i.e., any dataset) that is adapted to the network architecture. For example, we need to compute the prior once, then the model can be quickly trained on MNIST, CIFAR-10, ImageNet, Fashion-MNIST etc. The drawback is that each k in $1, \ldots, K$ is a neural network. The memory complexity may be very high depending on K. This may also slow down the training. For the moment, K is seen as an hyperparameter that can be optimized experimentally. Some theoretical results might help to choose K. In [2], the authors argues that $K \sim n^{4/3}$, which seems enormous even for small datasets with tens of thousands of training examples (I may have confuse something here. The true expression is $K \sim N^{4/3}$, where the order N can be chosen to be much smaller than n). On the other hand, the authors of [1] found in their experiments that $K = 2$ to $K = 16$ was sufficient, because of some "low-dimensional structure" (i.e., the sloppiness hypothesis). Remember that they were using unlabeled data to obtain the prior, so we might need larger K. Moreover, they have not evaluated the robustness of their model against adversarial attacks.

References

[1] Y. Gao, R. Ramesh, and P. Chaudhari, "Deep Reference Priors: What is the best way to pretrain a model?," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[2] H. H. Mattingly, M. K. Transtrum, M. C. Abbott, and B. B. Machta, "Maximizing the information learned from finite data selects a simple model," *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, 2018.

6 Remarks

6.1 Introduction

Remarks from section 1.

1. We said that x_i and w are independent. But D and w cannot be independent, since "training" is precisely the process of computing $p(w|D)$, which must be different from $p(w)$ if we want to learn anything. Since the collection of data is independent from the training, w is also independent with y_i. But w is not independent with D, because w and $y_i|x_i$ are *not* independent. This makes sense since w parameterizes the models $p(y|x, w)$. Somehow, it is possible for w to be independent with both x_i and y_i, but not with $y_i|x_i$. I guess it is possible, but it is not intuitive.

2. Not sure what changes in the "fixed design" (i.e., the x_i are chosen deterministically in advance) wrt to the "random design" used here (i.e., the *pairs* (x_i, y_i) are sampled iid).

3. I am not sure how to obtain the maximum likelihood. I guess, in the MLE, the posterior distribution $p(w|D)$ is a dirac mass at w^*, where $w^* = \arg\max p(D|w)$. What priors can lead to this result? Is it the uniform prior, which is an improper prior (total mass is infinite) in general? Or another "uninformative prior"? No, I think I am confusing the *distribution* $p(w|D)$ with the *maximum a posteriori* estimator, which is a real number. If the prior is uniform, then the MAP and the MLE are equal. However, the MAP is only one method to summarize $p(w|D)$, which consists in taking the mode of $p(w|D)$. But, we could also take the expectation, or the median. Or compute the variance (which is what we will do later). In all cases, the complete information is given by $p(w|D)$ and any summarizing methods will loose information.

6.2 Variational Inference

Remarks from section 2.
Question: why not using $\mathrm{KL}\left[p(w|D)||q(w|\xi)\right]$? Is it because we cannot get rid of $p(w|D)$ with this expression? What about the statistical interpretation:

- If we minimize $\mathrm{KL}\left[p(w|D)||q(w|\xi)\right]$, it corresponds to maximum likelihood, where $p(w|D)$ are the "data" (or the "truth"), and $q(w|\xi)$ is the fitting model. This corresponds better to the VI setting, and it would avoid the under-estimation of the variance. Here, $q(w|\xi)$ must cover the entire $p(w|D)$, even if it also cover some areas that are not in $p(w|D)$ ("zero avoidance"). ξ^* is chosen such that samples from $p(w|D)$ are as likely as possible to have been sampled from $q(w|\xi^*)$.

  When using this version, VI corresponds to the **expectation propagation** algorithm.

- When we minimize $\mathrm{KL}\left[q(w|\xi)||p(w|D)\right]$, we may neglect some areas of $p(w|D)$ while strongly fitting other areas of $p(w|D)$ ("zero forcing"). Here, we assume that $q(w|\xi)$ is wrong, and we are looking for a ξ^* such that the samples from $q(w|\xi^*)$ are as likely as possible to have been sampled from $p(w|D)$.

  When using this version, VI corresponds to the **expectation-maximization** algorithm. When using VI in a Bayesian setting, there is no distinction between the expectation step and the maximization step, because Bayesian methods do not distinguish the parameters and the latent variables (i.e., parameters are seen as latent variables). In the frequentist setting, the *maximization step* finds the maximum likelihood point-estimate of the parameters given the current estimate of the latent variables. The *expectation step* finds a new estimate of the latent variables given the current point-estimate of the parameters.

6.3 Uncertainty Propagation

Remarks from section 4.
The purpose of variational inference was to avoid computing:

$$p(D) = \int p(D|w')p(w')dw' = \mathbb{E}_w\left[p(D|w)\right],$$

which is said to be intractable. However, to compute the ELBO, we need to compute:

$$\int (\log p(D|w))\, q(w|\xi) dw = \mathbb{E}_{w|\xi} \left[\log p(D|w)\right].$$

Why should the first expression be intractable but not the second one?

Other confusion. Assume that we want to compute the second expression $\mathbb{E}_w \left[p(D|w)\right]$. What is the method exactly? Are we relying on some kind of Monte Carlo approximation?

$$\mathbb{E}_{w|\xi} \left[\log p(D|w)\right] \approx \frac{1}{k} \sum_{j=1}^{k} \log p(D|w_k),$$

where w_k is sampled from $q(w|\xi)$. I was thinking that the advantage of VI was to avoid Monte Carlo.

The purpose of propagating the first two moments is to approximate $p(D|w)$ as a Gaussian pdf. I should not confuse $\mathbb{E}_{w|\xi} \left[\log p(D|w)\right]$ that requires to know the pdf $p(D|w)$, and $\mathbb{E}_{w|\xi} \left[\log(D|w)\right]$ that doesn't. There are two assumptions for the propagation of the first two moments:

- The variational distribution $q(w|\xi)$ has diagonal covariance matrix, i.e., the components of w are uncorrelated (and even independent if we assume that $q(w|\xi)$ is Gaussian).
- The non-linear functions of the model are linearized with Taylor-series for the propagation of the covariance matrix.