

# 1 Preliminary

## 1.1 Assumptions on Posterior

Let  $f(x; \theta)$  a function of  $x$  with parameter  $\theta$ . Let  $y = f(x; \theta)$  an observable, thus the observed value obeys a Gaussian distribution. Thus, for a list of observations  $D := \{(x_i, y_i, \sigma_i) : i = 1, \dots, n\}$  ( $\sigma_i$  is the observational error of  $y_i$ ), we can construct a (logarithmic) likelihood, as

$$\begin{aligned}\ln p(D|\theta) &= \ln \left( \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left\{ -\frac{1}{2} \left( \frac{y_i - f(x_i; \theta)}{\sigma_i} \right)^2 \right\} \right) \\ &= \sum_{i=1}^n \left\{ -\frac{1}{2} \ln(2\pi\sigma_i^2) - \frac{1}{2} \left( \frac{y_i - f(x_i; \theta)}{\sigma_i} \right)^2 \right\}.\end{aligned}$$

If in addition assume a Gaussian prior, for some hyper-parameter  $\sigma$ ,

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{\theta^2}{2\sigma^2} \right),$$

then we have posterior  $p(\theta|D)$

$$\begin{aligned}\ln p(\theta|D) &= -\frac{1}{2} \left\{ \sum_{i=1}^n \left( \frac{y_i - f(x_i; \theta)}{\sigma_i} \right)^2 + \left( \frac{\theta}{\sigma} \right)^2 \right\} \\ &\quad - \frac{1}{2} \left\{ \sum_{i=1}^n \ln(2\pi\sigma_i^2) + \ln(2\pi\sigma^2) \right\},\end{aligned}$$

where the second line is  $\theta$ -independent.

## 1.2 Bayesian Inference

Sample  $m$  samples from  $p(\theta|D)$ ,  $\{\theta_{(s)} : s = 1, \dots, m\}$ . Thus, the Bayesian inference gives prediction from  $x$  to  $y$  as

$$\begin{aligned}\hat{y} &= \mathbb{E}_{\theta \sim p(\theta|D)}[f(x; \theta)] \\ &\approx \left( \frac{1}{m} \sum_{s=1}^m \right) f(x; \theta_{(s)}).\end{aligned}$$

# 2 Neural Network for Posterior

## 2.1 Model

Suppose we have a model,  $f(x, \theta)$ , where  $x$  is the input and  $\theta$  is the set of parameters of this model. Let  $D$  denotes an arbitrarily given dataset, i.e.  $D = \{(x_i, y_i) : i = 1, 2, \dots\}$  wherein for  $\forall i$   $x_i$  is the input and  $y_i$  the target (observed). With some assumption of the dataset, e.g. independency and Gaussianity, we can gain a likelihood  $L(\theta; D) := p(D|\theta)$ . Suppose we have some prior on  $\theta$ ,  $p(\theta)$ , we gain the unnormalized posterior  $L(D, \theta) p(\theta)$ . With  $D$  arbitrarily given, this unnormalized posterior is a function of  $\theta$ , denoted by  $p(\theta; D)$ .

We we are going to do is fit this  $p_D(\theta)$  by ANN for any given  $D$ . To do so, we have to assume that  $\text{supp}\{p(\theta; D)\} = \mathbb{R}^m$  for some  $m \in \mathbb{N}^+$  (i.e. has no compact support) but decrease exponentially fast as  $\|\theta\| \rightarrow +\infty$ . With this assumption, we can use Gaussian function as the activation of the ANN. We propose the fitting function

$$q(\theta; a, b, w) = \sum_i a_i^2 \left\{ \prod_j N(\theta_j, w_{ji}, b_{ji}) \right\},$$

where  $a_i \in \mathbb{R}$  for  $\forall i$  (always  $a_i^2 \geq 0$ ) and

$$N(x, w, b) := \sqrt{\frac{w^2}{2\pi}} \exp\left(-\frac{1}{2}(wx + b)^2\right)$$

While fitting,  $q(\theta; a, b, w)$  has no need of normalization, since  $p(\theta; D)$  is unnormalized.

$q(\theta; a, b, w)$  has probabilistic illustration.  $N(x, w, b)$  is realized as a one-dimensional Gaussian distribution (denote  $\mathcal{N}$ ). Indeed,  $N(x, w, b) = \mathcal{N}(x - \mu, \sigma)$  if  $\mu = b/w$  and  $\sigma = 1/|w|$ . Thus  $\prod_j N(\theta_j, w_{ij}, b_{ij})$  is a multi-dimensional Gaussian distribution, with all dimensions independent. The  $\{a_i^2\}$  is an empirical distribution, randomly choosing the Gaussian distributions. Thus  $q(\theta)$  is a composition: Empirical  $\rightarrow$  Gaussian. This is the *mixture distribution*.

Since there's no compact support, for both  $p(\theta; D)$  and  $q(\theta; a, b, w)$ , KL-divergence (equivalently, ELBO) can be safely employed as the cost-function of the fitting.

## 2.2 Numerical Consideration

For numerical consideration, instead of fitting  $p(\theta; D)$  by  $q(\theta; a, b, w)$ , we fit  $\ln p(\theta; D)$  by  $\ln q(\theta; a, b, w)$ . To compute  $\ln q(\theta; a, b, w)$ , we have to employ some approximation method. Let

$$\begin{aligned} \beta_i &:= \ln \left( a_i^2 \left\{ \prod_j N(\theta_j, w_{ji}, b_{ji}) \right\} \right) \\ &= \ln a_i^2 + \sum_j \left\{ -\frac{1}{2}(\theta_j w_{ji} + b_{ji})^2 + \frac{1}{2} \ln \left( \frac{w_{ji}^2}{2\pi} \right) \right\}, \end{aligned}$$

thus  $\ln q = \ln (\sum_i \exp(\beta_i))$ . We first compute all the  $\beta_i$  and pick the maximum  $\beta_{\max}$ . Then pick out other  $\beta_i$  for which  $\exp(\beta_i)$  has the same order as  $\exp(\beta_{\max})$ , collected as set  $M$  (excluding  $\beta_{\max}$ ). We have

$$\begin{aligned} \ln q &= \ln \left( \exp(\beta_{\max}) + \sum_{i \neq \max} \exp(\beta_i) \right) \\ &= \ln \left( 1 + \sum_{i \neq \max} \exp(\beta_i - \beta_{\max}) \right) + \beta_{\max}. \end{aligned}$$

Since  $\exp(\beta_i) \sim \exp(\beta_{\max})$  iff  $i \in M_{\beta}$  and  $\exp(\beta_i) \ll \exp(\beta_{\max})$  iff  $i \notin M_{\beta}$ ,  $\sum_{i \in M_{\beta}} \exp(\beta_i - \beta_{\max}) \sim 1$ , and others are negligible (comparing to 1). Thus,

$$\ln q \approx \ln \left( 1 + \sum_{i \in M_{\beta}} \exp(\beta_i - \beta_{\max}) \right) + \beta_{\max}.$$

This makes  $\ln q$  numerically computable. (And if  $\sum_{i \in M_{\beta}} \exp(\beta_i - \beta_{\max}) < 0.1$ , we can further approximate  $\ln q \approx \beta_{\max} + \sum_{i \in M_{\beta}} \exp(\beta_i - \beta_{\max})$ , thus no logarithm is to be computed.)

## 2.3 Cost-Function (Performance)

$$\begin{aligned} \text{ELBO}(a, b, w) &:= \mathbb{E}_{\theta \sim q(\theta; w, b)} [\ln p(\theta; D) - \ln q(\theta; a, b, w)] \\ &\approx \left( \frac{1}{n} \sum_{\theta^{(s)}} \right) \{ \ln p(\theta^{(s)}; D) - \ln q(\theta^{(s)}; a, b, w) \}, \end{aligned}$$

where  $\{\theta_{(s)}: s = 1, \dots, n\}$  is sampled from  $q(\theta; a, b, w)$  as a distribution.

## 2.4 Gradient

Let  $z := (a, b, w)$ . Then,

$$\begin{aligned}\frac{\partial \text{ELBO}}{\partial z}(z) &= \frac{\partial}{\partial z} \int d\theta q(\theta; z) \{\ln p(\theta; D) - \ln q(\theta; z)\} \\ &= \int d\theta q(\theta; z) \frac{\partial \ln q}{\partial z}(\theta; z) \{\ln p(\theta; D) - \ln q(\theta; z) - 1\} \\ &\approx \frac{1}{n} \sum_{\theta_{(s)}} \frac{\partial \ln q}{\partial z}(\theta_{(s)}; z) \{\ln p(\theta_{(s)}; D) - \ln q(\theta_{(s)}; z) - 1\}\end{aligned}$$

where  $\{\theta_{(s)}: s = 1, \dots, n\}$  is sampled from  $q(\theta; z)$  as a distribution. Next, since  $\ln q = \ln(\sum_i \exp(\beta_i))$ , we have

$$\begin{aligned}\frac{\partial \ln q}{\partial z}(\theta; z) &= \sum_i \frac{\exp(\beta_i)}{\sum_j \exp(\beta_j)} \frac{\partial \beta_i}{\partial z} \\ &= \sum_i \frac{\exp(\beta_i - \beta_{\max})}{\sum_j \exp(\beta_j - \beta_{\max})} \frac{\partial \beta_i}{\partial z}.\end{aligned}$$

Since  $\partial \beta_i / \partial z$  is polynomial-like, thus

$$\frac{\partial \ln q}{\partial z}(\theta; z) \approx \sum_i \frac{\exp(\beta_i - \beta_{\max})}{\sum_{j \in M_\beta} \exp(\beta_j - \beta_{\max})} \frac{\partial \beta_i}{\partial z} \delta_{i \in M_\beta},$$

where  $M_\beta$  is defined as previous. To calculate  $\partial \beta_i / \partial a_k$ ,  $\partial \beta_i / \partial b_{jk}$  and  $\partial \beta_i / \partial w_{jk}$ , recall

$$\beta_i = \ln a_i^2 + \sum_j \left\{ -\frac{1}{2}(\theta_j w_{ji} + b_{ji})^2 + \frac{1}{2} \ln \left( \frac{w_{ji}^2}{2\pi} \right) \right\},$$

we thus have

$$\begin{aligned}\frac{\partial \beta_i}{\partial a_k} &= \delta_{ik} \frac{2}{a_k}; \\ \frac{\partial \beta_i}{\partial b_{jk}} &= -\delta_{ik} \{\theta_j w_{jk} + b_{jk}\}; \\ \frac{\partial \beta_i}{\partial w_{jk}} &= -\delta_{ik} \left\{ (\theta_j w_{jk} + b_{jk}) \theta_j + \frac{1}{w_{jk}} \right\}.\end{aligned}$$

And recall

$$\frac{\partial \text{ELBO}}{\partial z}(z) \approx \left( \frac{1}{n} \sum_{\theta_{(s)}} \right) \{\ln p(\theta_{(s)}; D) - \ln q(\theta_{(s)}; z) - 1\} \sum_i \frac{\exp(\beta_i - \beta_{\max})}{\sum_{j \in M_\beta} \exp(\beta_j - \beta_{\max})} \frac{\partial \beta_i}{\partial z} \delta_{i \in M_\beta},$$

For stability, using  $\sigma'$  instead of  $\sigma = 1/|w|$  where  $\sigma = \ln(1 + \exp(\sigma'))$ ??? (C.f. [here](#).)