1 Preliminary

1.1 Assumptions on Posterior

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

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

If in addition assume a Gaussian prior, for some hyper-parameter σ,

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

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

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

where the second line is θ-independent.

1.2 Bayesian Inference

Sample m samples from $p(\theta|D)$, $\{\theta_{(s)}: s=1, ..., m\}$. 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 The Model

Suppose we have a model, $f(x,\theta)$, where x is the input and θ the set of parameters of this model. Let D denotes an arbitrarily given dataset, i.e. $D=\{(x_i, y_i): i=1,2,...\}$ 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 θ, $p(\theta)$, we gain the unormalized posterior $L(D,\theta)\,p(\theta)$. With D arbitrarily given, this unormalized posterior is a function of θ, denoted by $p(\theta;D)$.

We we are going to do is fit this $p(\theta;D)$ by ANN for any given D. To do so, we have to assume that $\mathrm{supp}\{p(\theta;D)\}=\mathbb{R}^d$ for some $d\in\mathbb{N}^+$ (i.e. has no compact support) but decrease exponentially fast as $\|\theta\|\to+\infty$. With this assumption, $\ln p(\theta;D)$ is well-defined. For ANN, we propose using Gaussian function as the activation-function. Thus, we have the fitting function

$$q(\theta;a,\mu,\zeta) = \sum_{i=1}^{N_c} w_i(a)\left\{\prod_{j=1}^{d} \Phi(\theta_j - \mu_{ij}, \sigma(\zeta_{ij}))\right\},$$

where

$$\begin{aligned} w_i(a) &= \frac{\exp(a_i)}{\sum_{j=1}^N \exp(a_j)} = \text{softmax}(i; a); \\ \sigma(\zeta_{ij}) &= \ln(1 + \exp(\zeta_{ij})), \end{aligned}$$

and $a_i, \mu_{ij}, \zeta_{ij} \in \mathbb{R}$ for $\forall i, \forall j$ and

$$\Phi(x - \mu, \sigma) := \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

being the Gaussian PDF. The introduction of ζ is for numerical consideration, see below.

2.1.1 Numerical Consideration

If, in q, we regard w, μ, and σ as independent variables, then the only singularity appears at $\sigma = 0$. Indeed, σ appears in Φ (as well as the derivatives of Φ) as denominator only, while others as numerators. However, once doing numerical iterations with a finite step-length of σ, the probability of reaching or even crossing 0 point cannot be surely absent. This is how we may encounter this singularity in practice.

Introducing the ζ is our trick of avoiding this singularity. Precisely, using a singular map that pushes the singularity to infinity solves the singularity. In this case, using softplus(.) that pushes $\sigma = 0$ to $\zeta \to -\infty$, so that, with finite steps of iteration, singularity (at $-\infty$) cannot be reached.

This trick (i.e. pushing a singularity to infinity) is the same as in avoiding the horizon-singularity of Schwarzschild solution of black hole.

2.2 Interpretation

2.2.1 As a Mixture Distribution

$q(\theta;\, a,\, \mu,\, \zeta)$ has a probablitic interpretation. $\prod_{j=1}^d \Phi(\theta_j - \mu_{ij},\, \sigma(\zeta_{ij}))$ corresponds to multi-dimensional Gaussian distribution (denote $\mathcal{N}$), with all dimensions independent with each other. The $\{w_i(a)\}$ is a categorical distribution, randomly choosing the Gaussian distributions. Thus $q(\theta; a, \mu, \zeta)$ is a composition: categorical $\to$ Gaussian. This is the *mixture distribution*.

2.2.2 As a Generalization

This model can also be intrepreted as a direct generalization of mean-field variational inference. Indeed, let $N_c = 1$, this model reduces to mean-field variational inference. Remark that mean-field variational inference is a mature algorithm and has been sucessfully established on many practical applications.

2.3 Cost-Function

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

where $\{\theta_{(s)}: s = 1, ..., n\}$ is sampled from $q(\theta; a, \mu, \zeta)$ as a distribution. Since there's no compact support for both $p(\theta; D)$ and $q(\theta; a, \mu, \zeta)$, ELBO is well-defined, as the cost-function (or say loss-function, performance, etc) of the fitting.