# $\Delta$-STN on BNN
# Research notes

Yuchen Wang

June 17, 2020

## Contents

# 1 Bayesian Linear Regression

**Motivation** Avoid the over-fitting problem of maximum likelihood (equivalent to MSE) and will also lead to automatic methods of determining model complexity using the training data alone.

**Notation 1.1.** Inputs: $X = \{x_1, \ldots, x_N\}$.
Targets: $t_1, \ldots, t_N$, which we group into a column vector $\mathbf{t} \in \mathbb{R}^N$.
Basis functions: $\phi_j(\mathbf{x})$, groups into a column vector $\phi \in \mathbb{R}^M$.

Design matrix: $\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$

**Assumptions**:

1. The target variable $t$ is given by a deterministic function $y(\mathbf{x}, \mathbf{w})$ with additive Gaussian noise so that
$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon = \mathbf{w}^T \phi(\mathbf{x}) \tag{1.1}$$
2. The inputs are drawn independently from the distribution
$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}) \tag{1.2}$$
where $\beta$ is the **precision** (inverse variance).

The likelihood function (a function of the adjustable parameters $\mathbf{w}$ and $\beta$):

$$p(\mathbf{t}|X, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}) \tag{1.3}$$

## 1.1 Parameter Distribution

**Introduction** Introduce a prior probability distribution over the model parameters $\mathbf{w}$.
For the moment, treat $\beta$ as a known constant.
The corresponding conjugate prior of the likelihood function $p(\mathbf{t}|\mathbf{w})$ defined by (1.2) is of the form

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \tag{1.4}$$

having mean $\mathbf{m}_0$ and covariance $\mathbf{S}_0$.
The posterior is in the form

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S_N}) \tag{1.5}$$

where

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t}) \tag{1.6}$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi} \tag{1.7}$$

**Remark 1.1.** Note that the mode of a Gaussian distribution coincides with its mean. Thus

$$\mathbf{w}_{MAP} = \mathbf{m}_N$$

**Remark 1.2.** If we consider an infinitely broad prior $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$ with $\alpha \to 0$, the mean $\mathbf{m}_N$ of the posterior distribution reduces the maximum likelihood value $\mathbf{w}_{ML}$ given by

$$\mathbf{w}_{ML} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t} \tag{1.8}$$

**Remark 1.3.** Similarly, if $N = 0$, then the posterior distribution reverts to the prior.

**Remark 1.4.** Furthermore, if data points arrive sequentially, then the posterior distribution at any stage acts as the prior distribution for the subsequent data point, such that the new posterior distribution is again given by (1.5).

**Simplification** Consider a zero-mean isotropic Gaussian governed by a single precision parameter $\alpha$ so that

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{1.9}$$

and the corresponding posterior distribution is then (1.5) with

$$\mathbf{m}_N = \beta S_N \mathbf{\Phi}^T\mathbf{t} \tag{1.10}$$

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^T\mathbf{\Phi} \tag{1.11}$$

The log of the posterior distribution is of the form

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + const \tag{1.12}$$

Maximization of this function is equivalent to the minimization of MSE plus L2 regularization with $\lambda = \alpha/\beta$.

## 1.2 Predictive distribution

In practice, we are interested in making predictions of $y$ for new values of $\mathbf{x}$ rather than the value of $\mathbf{w}$. This requires that we evaluate the **predictive distribution** define by

$$p(y|\mathbf{t}, \alpha, \beta) = \int p(y, \mathbf{w}|\mathbf{t}, \alpha, \beta)\, d\mathbf{w} \tag{1.13}$$

$$= \int p(y|\mathbf{w}, \mathbf{t}, \alpha, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta)\, d\mathbf{w} \qquad \text{(by Chain Rule)} \tag{1.14}$$

$$= \int p(y|\mathbf{w}, \alpha, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta)\, d\mathbf{w} \qquad (y \perp \mathbf{t}) \tag{1.15}$$

where $\mathbf{t}$ is the vector of target values from the training set. We see that the above equation involves the convolution of two Gaussian distributions, and so making use of the result in (Figure 1) we can see that the predictive distribution takes the form

$$p(y|\mathbf{t}, \mathbf{x}, \alpha, \beta) = \mathcal{N}(y|\mathbf{m}_N^T\phi(\mathbf{x}), \underbrace{\beta^{-1} + \phi(\mathbf{x})^T\mathbf{S}_N\phi(\mathbf{x})}_{\sigma_N^2(\mathbf{x})}) \tag{1.16}$$

**Marginal and Conditional Gaussians**

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (2.113)$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \quad (2.114)$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \quad (2.115)$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \quad (2.116)$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \quad (2.117)$$

Figure 1:

**Remark 1.5.** Note that, as additional data points are observed, the posterior distribution becomes narrower. It can be shown that $\sigma^2_{N+1}(\mathbf{x}) \leq \sigma^2_N(\mathbf{x})$. In the limit $N \to \infty$, the second term of $\sigma^2_N(\mathbf{x})$ goes to zero, and the variance of the predictive distribution arises solely from the additive noise governed by the parameter $\beta$.

## 1.3 Equivalent kernel

The posterior mean solution for the linear basis function model has an interesting interpretation that will set the stage for kernel methods, including Gaussian process.

**Definition 1.1.** equivalent kernel We see that the predictive mean can be written in the form

$$y(\mathbf{x}, \mathbf{m}_N) = \mathbf{m}_N^T \phi(\mathbf{x}) = \beta\phi(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\Phi}^T \mathbf{t} = \sum_{n=1}^{N} \beta\phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}_n) t_n \quad (1.17)$$

Thus the mean of the predictive distribution at a point $\mathbf{x}$ is given by a linear combination of the training set target variables $t_n$, so that we can write

$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n \quad (1.18)$$

where

$$k(\mathbf{x}, \mathbf{x}') = \beta\phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}') \quad (1.19)$$

is known is the **equivalent kernel (smoother matrix)**.

**Remark 1.6.** The mean is obtained by forming a weighted combination of the target values in which data points close to $x$ are given higher weight than point further removed from $x$.

**Property 1.1.**

$$\begin{aligned}\text{Cov}\left[y(\mathbf{x}), y(\mathbf{x}')\right] &= \text{Cov}\left[\phi(\mathbf{x}^T\mathbf{w}, \mathbf{w}^T\phi(\mathbf{x}')\right] && (1.20)\\ &= \phi(\mathbf{x})^T\mathbf{S}_N\phi(\mathbf{x}') && (1.21)\\ &= \beta^{-1}k(\mathbf{x}, \mathbf{x}') && (1.22)\end{aligned}$$

**Remark 1.7.** The predictive mean at nearby points will be highly correlated, whereas for more distant pairs of points the correlation will be smaller.

**Remark 1.8.** Instead of introducing a set of basis functions which implicitly determines an equivalent kernel, we can instead define a localized kernel directly and use this to make predictions for new input vectors $\mathbf{x}$, given the observed training $\mathbf{x}$. This leads to **Gaussian process**.

**Property 1.2.** For all values of $\mathbf{x}$,

$$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = 1 \tag{1.23}$$

**Property 1.3.** The equivalent kernel can be expressed in the form of an inner product with respect to a vector $\psi(\mathbf{x})$ of nonlinear functions, so that

$$k(\mathbf{x}, \mathbf{z}) = \psi(\mathbf{x})^T\psi(\mathbf{z}) \tag{1.24}$$

where $\psi(\mathbf{x}) = \beta^{1/2}\mathbf{S}_N^{1/2}\phi(\mathbf{x})$.

# 2 Bayesian Model Comparison

There are two types of models

1. The distribution is defined over the set of target values $\mathbf{t}$, while the set of input values $X$ is assumed to be known.
2. Define a joint distribution over $X$ and $\mathbf{t}$.

Suppose the data is generated from one of the above but we are uncertain about which one. The uncertainty is expressed through a prior probability distribution $p(\mathcal{M}_i)$. Given a training set $\mathcal{D}$, we wish to evaluate the posterior distribution

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i) \tag{2.1}$$

**Definition 2.1** (mixture distribution)**.** Once we know the posterior distribution over models, the predictive distribution is given, from the sum and product rules, by

$$p(t|\mathbf{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{D})p(\mathcal{M}_1|D) \tag{2.2}$$

This is an example of a **mixture distribution** in which the overall predictive distribution is obtained by averaging the predictive distributions $p(t|x, \mathcal{M}_i, D)$ of individual models, weighted by the posterior probabilities $p(\mathcal{M}_i|D)$ of those models.

**Definition 2.2** (model selection)**.** A simple approximation to model averaging is to use the single most probable model alone to make predictions. This is known as **model selection**.

**Definition 2.3** (model evidence)**.** For a model governed by a set of parameters $\mathbf{w}$, the **model evidence** is given, from the sum and product rules of probability, by

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i)\, d\mathbf{w} \tag{2.3}$$

We can obtain some insight into the model evidence by making a simple approximation to the integral over parameters. Consider first the case of a model having a single parameter $w$. The posterior distribution over parameters is proportional to $p(\mathcal{D}|w)p(w)$, where we omit the dependence on the model $\mathcal{M}_i$ to keep the notation uncluttered. If we assume that the posterior distribution is sharply peaked around the most probable value $w_{\mathrm{MAP}}$, with width $\Delta w_{\text{posterior}}$ , then we can approximate the integral by the value of the integrand at its maximum times the width of the peak. If we further assume that the prior is flat with width $\Delta w_{\text{prior}}$ so that $p(w) = 1/\Delta w_{\text{prior}}$ then we have

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\mathrm{d}w \simeq p\left(\mathcal{D}|w_{\mathrm{MAP}}\right) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \tag{2.4}$$

where we omit the $\mathcal{M}_i$ to keep the notation uncluttered.

**Definition 2.4** (evaluating the posterior distribution)**.**

$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)} \tag{2.5}$$

**Definition 2.5** (Bayes factor)**.** The ratio of model evidences

$$K = \frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)} \tag{2.6}$$

for two models is known as a **Bayes factor**.
A value of $K > 1$ means that $M_1$ is more strongly supported by the data under consideration than $M_2$.

**Remark 2.1.** The use of Bayes factors is a Bayesian alternative to classical hypothesis testing.