Bayesian deep learning

MASH & IASD

Julyan ARBEL

Feb 5, 2021

Bayesian NN under a Laplace approximation

Chap 5.7 of Bishop

Univariate regression $t \in \mathbb{R}$

Gaussian $p(t \mid x)$ to be Gaussian

the mean $=$ $y(\underbrace{x}_{\text{input}}, \underbrace{w}_{\text{weight}})$ output of NN

variance β^{-1} β precision

Data $\mathcal{D} = \{(x_n, t_n), n = 1 \dots N\}$

$$\begin{cases} \text{Model} & p(t \mid x) = \mathcal{N}(t \mid y(x, w), \beta^{-1}) \\ \text{Prior} & p(w \mid \alpha) = \mathcal{N}(w \mid 0, \alpha^{-1} I) \end{cases}$$

Posterior $p(w \mid \mathcal{D}, \alpha, \beta) \propto \prod_{n=1}^{N} \mathcal{N}(t_n \mid \underline{y(x_n, w)}, \beta^{-1})\, p(w \mid \alpha)$

w_{MAP} : numerically find local optimum.

$$\log p(w \mid \mathcal{D}, \alpha, \beta) = \underbrace{-\frac{\alpha}{2} w^T w}_{\text{penalty to the MLE.}} - \frac{\beta}{2} \underbrace{\sum_{n=1}^{N} (y(x_n, w) - t_n)^2}_{} + c^t$$

Laplace approx. of the posterior :

$$q(w \mid \mathcal{D}, \alpha, \beta) = \mathcal{N}(w \mid w_{MAP}, A^{-1})$$

$$A = -\nabla\nabla \log p = \alpha I + \beta H \quad , \text{ with } H = \text{Hess}(SSE)$$

approx. predictive

$$p(t \mid x, \mathcal{D}) = \int_w \underbrace{p(t \mid x, w)}_{\mathcal{N}(t \mid y(x,w), \beta^{-1})} q(w \mid \mathcal{D}, \alpha, \beta) \, dw \qquad (*)$$

Taylor approximation for NN :

$$y(x,w) \approx y(x,w_{MAP}) + g^T(w - w_{MAP})$$

$$g = \nabla_w \, y(x,w) \big|_{w=w_{MAP}}$$

$$p(t \mid x, w, \beta) \approx \mathcal{N}\left(t \mid y(x,w_{MAP}) + g^T(w - w_{MAP}), \beta^{-1}\right)$$

plug this in (*) :

Plus use

$$\begin{cases} p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1}) \\ p(y \mid x) = \mathcal{N}(y \mid Ax+b, L^{-1}) \end{cases}$$

$$\Rightarrow p(y) = \mathcal{N}(y \mid A\mu + b, \underbrace{L^{-1} + A\Lambda^{-1}A^T})$$

$$p(t \mid x, \mathcal{D}, \alpha, \beta) \approx \mathcal{N}\left(t \mid y(x, w_{MAP}), \sigma^2(x)\right)$$

$$\sigma^2(x) = \underbrace{\beta^{-1}}_{\text{aleatoric}} + \underbrace{g^T A^{-1} g}_{x\text{-dependent : epistemic.}}$$

α, β hyperparameters

Marginal likelihood $p(\mathcal{D} \mid \alpha, \beta)$, aka evidence

$\rightsquigarrow$ w is integrated out.

$$(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha, \beta)}{\operatorname{argmax}}\ p(\mathcal{D} \mid \alpha, \beta).$$

$$p(\mathcal{D} \mid \alpha, \beta) = \int p(\mathcal{D} \mid w, \beta)\, p(w \mid \alpha)\, dw$$

Laplace approximation $\quad z_0$ argmax, $A = -\nabla\nabla \log f(z)\big|_{z=z_0}$

$$\int f(z)\,dz = f(z_0) \int \exp\left(-\frac{1}{2}(z-z_0)^T A (z-z_0)\right) dz$$

$$= f(z_0) \frac{(2\pi)^{M/2}}{|A|^{1/2}} \qquad M = \#(z)$$

$$\Rightarrow \log p(\mathcal{D} \mid \alpha, \beta) = -\underbrace{E(w_{MAP})} - \frac{1}{2}\log|A| + \frac{W}{2}\log\alpha + \frac{N}{2}\log\beta + ct$$

regularized SSE : $\underbrace{\frac{\alpha}{2} w_{MAP}^T w_{MAP}}_{\text{pen.}} + \frac{\beta}{2}\sum_{n=1}^{N}\left(y(x_n, w_{MAP}) - t_n\right)^2$

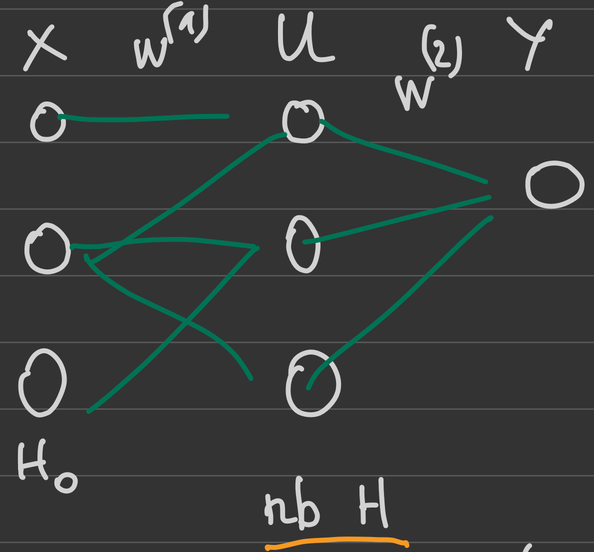
optimization in (α, β) is done by analogy with linear regression. $\rightarrow$ Bishop 5.7.

Other techniques for practical BNN:

variational inference: with various assumptions on approx family.

Distributional properties of BNN

Neal's result : 1-hidden layer NN, fully-connected



$W^{(1)}_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

$W^{(2)}_{ij} \sim N(0, \sigma_H^2)$

ϕ

$\|X\|_2^2 = \sum X_h^2$

$m = 1 \dots H$

$g_m(x)$, pre non linearity $= \sum_{k=1}^{H_0} W_{mk} \cdot X_k \overset{iid}{\sim} N(0, \|X\|_2^2 \sigma^2)$

post non linearity : $h_m(X) = \phi(g_m(x))$

$$y = \sum_{m=1}^{H} \underbrace{W^{(2)}_m \, h_m(x)}_{Y_m}$$

Y_m are iid

$\forall m : \mathbb{E}[Y_m] = \mathbb{E}[W^{(2)}_m] \, \mathbb{E}[h_m(X)] = 0 \times \mathbb{E}[h_m(X)] = 0$

$$\mathbb{V}[Y_m] = \underbrace{\mathbb{E}\left[W_m^{(2)\,2} \right]}_{\sigma_H^2} \underbrace{\mathbb{E}\left[h_m^2(x) \right]}_{c^F}$$

$$y \longrightarrow \mathcal{N}(0,\ H\,\sigma_H^2) \quad \text{by CLT}$$

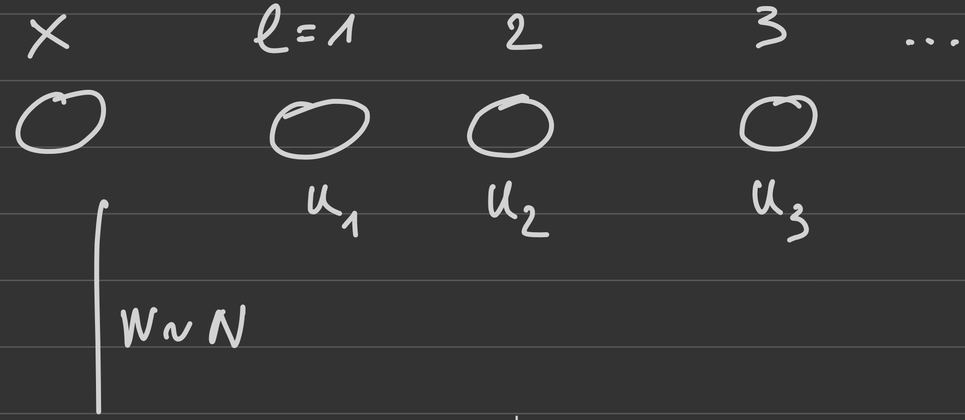
$$\sigma_H^2 = \frac{1}{H}$$

Gaussian limit of output when width $\longrightarrow \infty$.

Extended to any deeper NN in 2018.

Another distributional property of BNN

what is the distrib. of units under Gaussian prior on weights ?



Assumption on prior :

indep priors, centered

eg : $W_{ij}^{\ell} \overset{iid}{\sim} N(0, \sigma^2)$

Assumption on nonlinearity ϕ

enveloppe condition for ϕ

Theorem Conditional on x, the prior for unit at layer ℓ is Sub Weibull $(\ell/2)$.

More on Sub-Weibull : generalizes sub-Gaussian & sub Exp properties

Let X be r.v., let $\|X\|_k = \left(\mathbb{E}[X^k]\right)^{1/k}$

X is sub-Gaussian if $\exists C > 0$ s.t.

$$\forall k, \quad \|X\|_k \leq C\sqrt{k} = C k^{1/2}$$

X is sub-Exponential if $\exists C > 0$ s.t.

$$\forall k, \quad \|X\|_k \leq C k$$

X is sub-Weibull(θ) $\forall k$, $\boxed{\|X\|_k \leq C k^{\theta}}$

ℓ subW$\left(\frac{\ell}{2}\right)$.

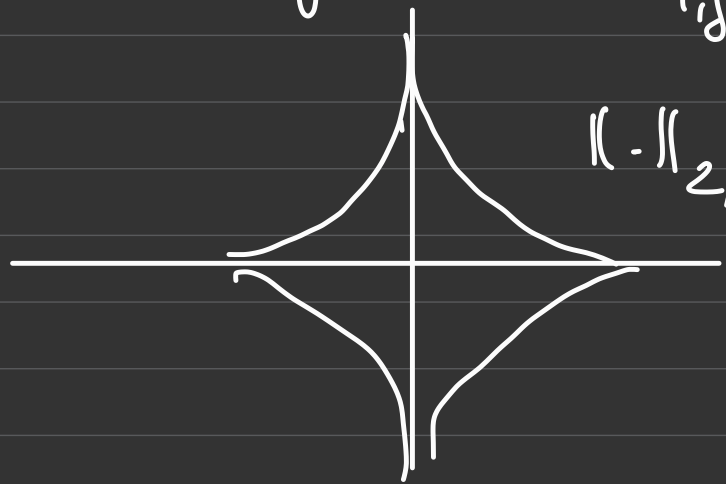
subW$\left(\frac{2}{2}\right)$ = subW(1) = subE

N = subG = subW$\left(\frac{1}{2}\right)$

density for Weibull(θ) $g(w) \propto e^{-|w|^{1/\theta}}$

induces a penalty $\longrightarrow$ $|w|^{1/\theta}$ $\qquad \theta = \frac{\ell}{2}$

Layer ℓ $\qquad \sum_{i,j} |w_{ij}^{\ell}|^{2/\ell} = \| W^{\ell} \|_{2/\ell}$



$\|\cdot\|_{2/\ell}$ unit ball in $\mathbb{R}^2$.

Vladimirova, 2018

PyTorch, pyro.