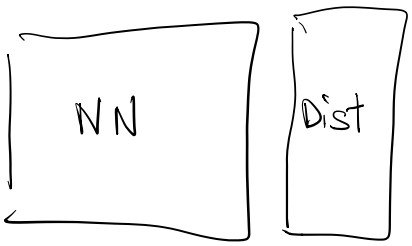
Maximum likelihood



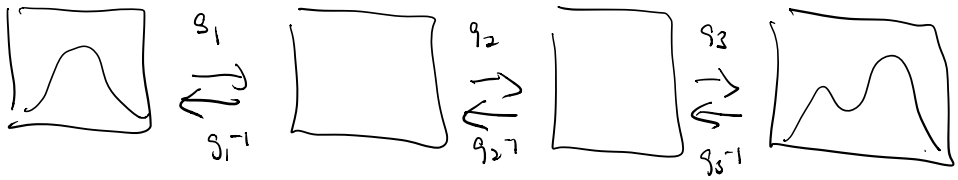
NN outputs parameters of chosen distribution

Loss function is negative log likelihood of distribution

E.g. for a classification model,

the distribution would be a multinomial.

Normalizing flows



↓

bijections → i.e. reversible functions

↓

these functions have parameters θ

↓

θ estimated using maximum likelihood

$g = g_1 \circ g_2 \circ g_3$

is also a bijecti

Math Review

Transformations of RV

$$x = g(y)$$

$$f_y(y) = f_x(g(y))\, g'(y)$$

Replace $|\cdot|$ Jacobian determinant in $\mathbb{R}^n$

Variational Bayes

$$D_{KL}(P\|Q) = \sum P(x) \log \frac{P(x)}{Q(x)} \quad \text{or} \quad \int P(x) \log \frac{P(x)}{Q(x)} dx$$

In VB, we have a family of distributions parameterized by λ, $Q_\lambda(\theta)$ that we want to use to approximate $P(\theta|D)$

$$D_{KL}\left[Q_\lambda(\theta) \| P(\theta|D)\right]$$

note flip!

$$= \int Q_\lambda(\theta) \log \frac{Q_\lambda(\theta)}{P(\theta|D)} d\theta$$

$$= \int Q_\lambda(\theta) \log \frac{Q_\lambda(\theta) P(D)}{P(D|\theta) P(\theta)} d\theta$$

$$= \int Q_\lambda(\theta) \log P(D) d\theta - \int Q_\lambda(\theta) \log \frac{P(D|\theta) P(\theta)}{Q_\lambda(\theta)} d\theta$$

$$= \log P(D) \int Q_\lambda(\theta) d\theta + \int Q_\lambda(\theta) \log \frac{Q_\lambda(\theta)}{P(\theta)} d\theta - \int Q_\lambda(\theta) \log P(D|\theta) d\theta$$

$$= \log P(D) + D_{KL}\left[Q_\lambda(\theta) \| P(\theta)\right] - E_{\theta \sim Q_\lambda}\left[\log P(D|\theta)\right]$$

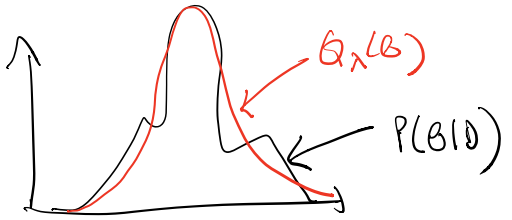
Custom loss fn!

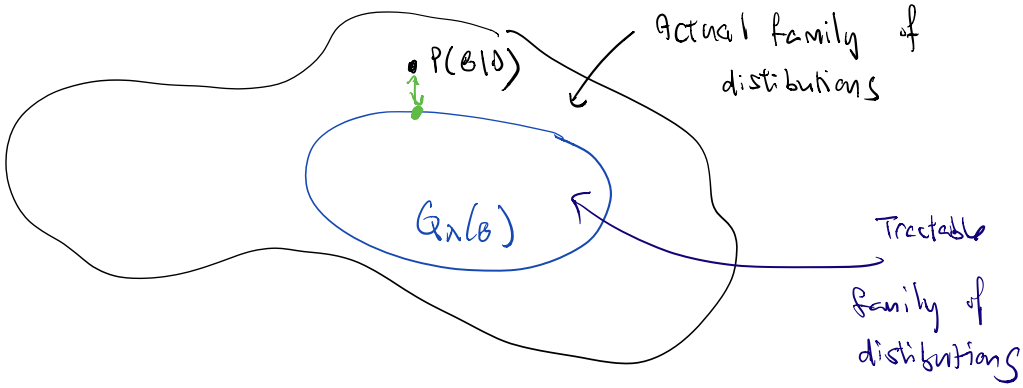
$\log P(D)$: No role in minimization

$D_{KL}\left[Q_\lambda(\theta) \| P(\theta)\right]$: regularizer

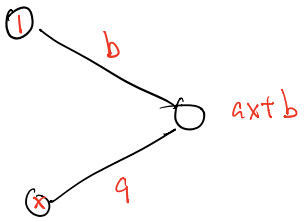
$E_{\theta \sim Q_\lambda}\left[\log P(D|\theta)\right]$: mean NLL when θ is sampled from Q_λ

$P(\theta)$ chosen to be small & centered around 0

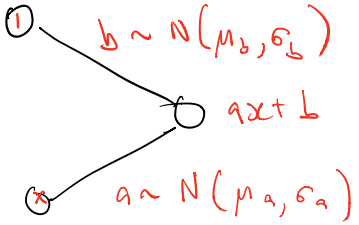




Bayesian NN



Regular NN



Bayesian NN

Variational parameters

$$\lambda = \{\mu_a, \sigma_a, \mu_b, \sigma_b\}$$

Q_λ ~ Product of independent Gaussian distributions

Note: $\sigma_a, \sigma_b > 0$

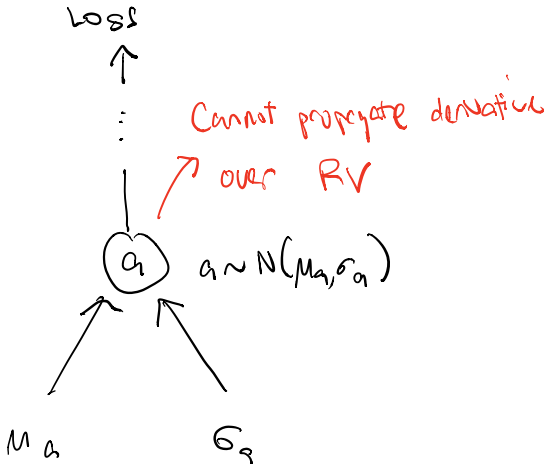
use softplus

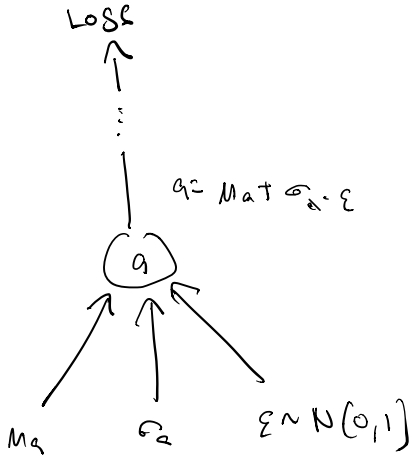
$$f(x) = \log(1 + e^x)$$

$$\text{Loss} = \underbrace{D_{KL}\left[Q_\lambda(\theta) \,\|\, P(\theta) \right]}_{-\frac{1}{2}\left(1 + \log(\sigma^2) - \mu^2 - \sigma^2\right)} - \underbrace{E_{\theta \sim Q_\lambda}\left[\log P(D|\theta) \right]}_{\text{single sample of } \theta \text{ is used in practice!}}$$

$P(\theta) \sim N(0,1)$

$Q_\lambda(\theta) \sim N(\mu, \sigma)$





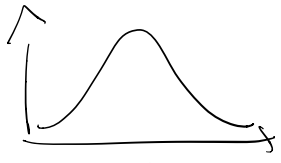
MC dropout

In normal dropout, turn ON during training & OFF during test } prevents overfitting

In MC dropout, turn ON during train AND test

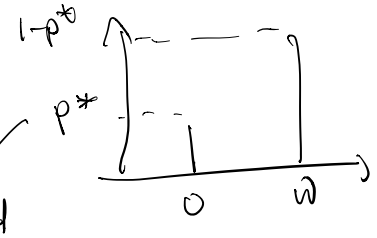
During test → evaluate each sample multiple times

Averaging these outputs gives the Bayesian predictive distribution.



↓

VB uses μ, σ to model distribution of w



Fixed probability of dropout