Bayesian deep learning

MASH & IASD

Julyan ARBEL

Feb 5, 2021.

Bayesian neural network, predictive distribution:

$$p(t|x,\mathcal{D}) = \int p(t|x,w)\,p(w|\mathcal{D})\,dw$$

Approximations to the posterior $p(w|\mathcal{D})$:

- Laplace (yesterday)
- Variational inference
- Monte Carlo dropout

VI (Blundell, 2015)

parameters are weights $w = (w_1, \dots, w_{\overline{W}})$

Approximate family : Gaussian : $N(\mu, \sigma^2)$

w_i approximated by $\theta_i = (\mu_i, \sigma_i^2)$, $i = 1, \dots, \overline{W}$

$$q_\theta(w) = \prod_{i=1}^{\overline{W}} N(w_i \mid \mu_i, \sigma_i^2)$$

$$Q = \{ q_\theta(w), \ \mu_i \in \mathbb{R}, \ \sigma_i > 0 \}$$

$$KL\left(q_\theta(w) \,\|\, p(w \mid X, Y) \right) \qquad X, Y \begin{cases} X = (X_i) \\ Y = (Y_i) \end{cases} \quad i = 1 \dots N$$

$$= \int q_\theta(w) \log \left(\frac{q_\theta(w)}{p(w \mid X, Y)} \right) dw$$

$$= \int q_\theta(w) \log \left(\frac{q_\theta(w)\, p(Y \mid X)}{p(w)\, p(Y \mid w, X)} \right) dw$$

$$= - \int q_\theta(w) \log \underbrace{p(Y|w,X)}_{\text{likelih.}} + \underbrace{KL\left(q_\theta(w) \| p(w)\right)}_{\text{KL for prior}} + \log \underbrace{p(Y|X)}_{\text{evidence}}$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{\text{expected log-like.}}$$

$$= - \underbrace{\mathcal{L}_{VI}(\theta)}_{} + \underbrace{\log p(Y|X)}_{c^t}$$

Since $KL \geq 0$, $\mathcal{L}_{VI}(\theta) \leq \log p(Y|X)$

so $\mathcal{L}_{VI}$ is called evidence lower bound ELBO

Minimize VI $\iff$ Maximize ELBO

ELBO : $\mathcal{L}_{VI}(\theta) = \mathbb{E}_{q_\theta(w)}\left[\log p(Y|X,w)\right] - \underbrace{KL\left(q_\theta(w) \| p(w)\right)}_{\text{closed-form}}$

$$\mathcal{L}_{VI}(\theta) = \sum_{i=1}^{N} \mathbb{E}_{q_\theta(w)}\left[\log p(Y_i|w,X_i)\right] - KL$$

approximate by sampling : $\hat{w} \sim q_\theta(\cdot)$

In order to estimate the gradient of $\mathcal{L}_{VI}(\theta)$, use the reparametrization trick.

idea : $w = g(\theta, \varepsilon)$, $\begin{cases} g \text{ is deterministic} \\ \varepsilon \perp\!\!\!\perp \theta \end{cases}$

$w_j \sim \mathcal{N}(w_j \mid \mu_j, \sigma_j^2) = q_{\theta_j}(w)$

$w_j = g(\theta_j, \varepsilon_j) = \mu_j + \sigma_j \varepsilon_j$, with $\varepsilon_j \sim \mathcal{N}(0, 1)$.

$$\text{ELBO}(\theta) = \mathcal{L}_{VI}(\theta) \approx \underbrace{\sum_{i=1}^{N} \log p(y_i \mid y(x_i, g(\theta, \hat{\varepsilon}_i)))}_{\text{MC approx of } \mathbb{E}} - KL$$

Algorithm : Stochastic VI

- Given X, Y data, η: learning rate, initialise θ
- **Repeat**: Sample $\begin{cases} \varepsilon_j \sim p(\varepsilon), & j \in S \\ S \text{ subset from } \{1, \dots, N\} \text{ of size } M \end{cases}$

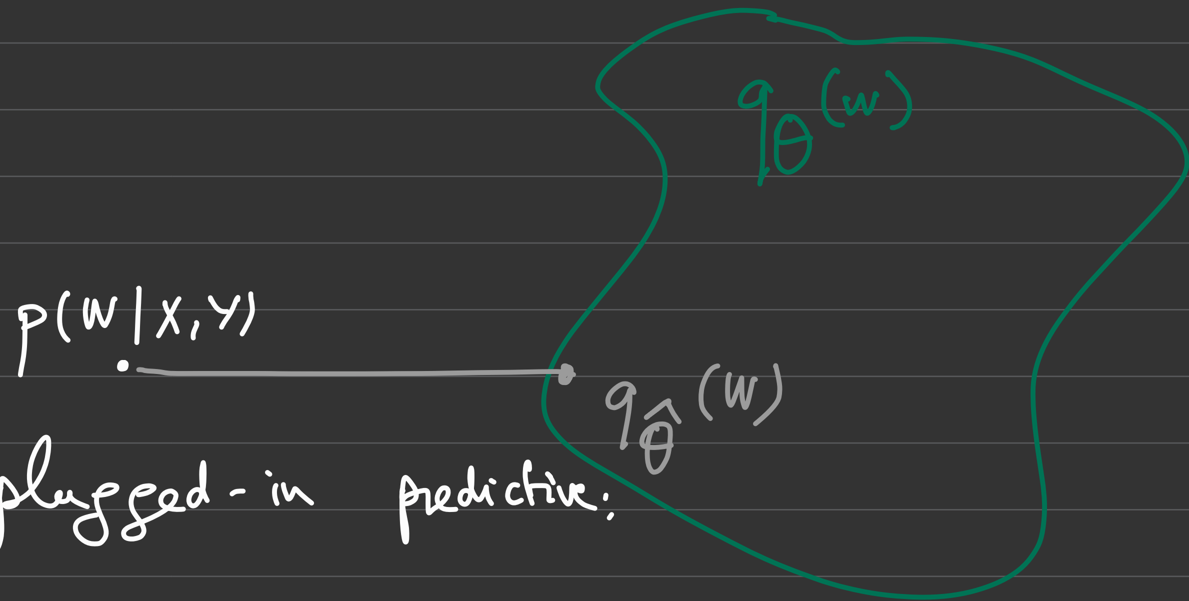
  Stochastic derivative estimator wrt θ:

  $$\widehat{\Delta\theta} = -\frac{N}{M} \sum_{i \in S} \underbrace{\frac{\partial}{\partial\theta} \log p(Y_i \mid y(g(\theta, \hat{\varepsilon}_i), X_i))}_{\text{avail. by reparam. trick.}} + \underbrace{\frac{\partial KL}{\partial\theta}}_{\text{closed-form.}}$$

  $\theta \leftarrow \theta + \eta \widehat{\Delta\theta}$.

  Until (some) convergence.

$\mathcal{L}_{VI}(\theta)$



$q_{\hat{\theta}}(w)$ can be plugged-in predictive:

$$p(t \mid x, \mathcal{D}) \approx \int \underbrace{p(t \mid x, w)}_{\substack{\text{can be made Gaussian} \\ \text{by some Taylor expansion} \\ \text{as yesterday}}} \; \underbrace{q_{\hat{\theta}}(w)}_{\text{Gaussian approx by VI}} \, dw$$

Monte Carlo dropout (Gal, Ghahramani, ICML, 2016)

based on: Dropout (Hinton, 2012)

- training a NN with dropout
  $\Longleftrightarrow$ training a BNN with variational posterior $q_\theta(w)$
- MC dropout:
  sampling several passes of NN with dropout
  $\Longleftrightarrow$ MC approx. inference with $q_\theta(w)$.