2024 Summer Seminar · Final Topic

An Introduction to Bayesian Neural Networks

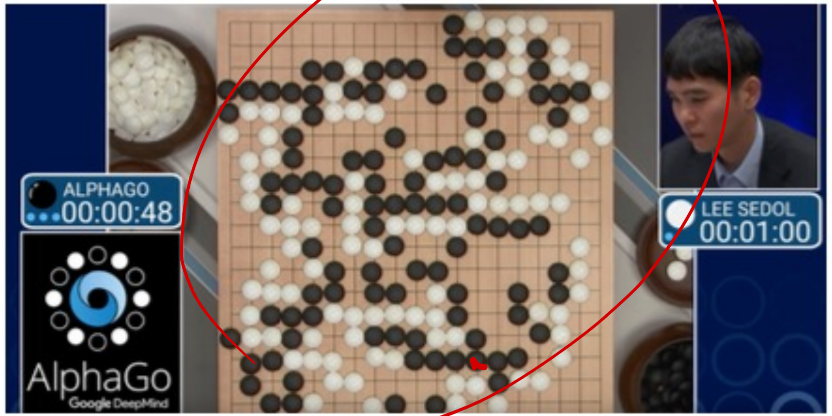
— Wang Ma

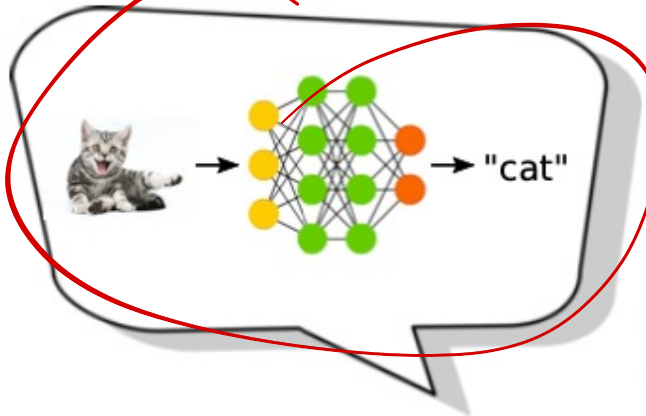
Slides reference: Prof. Yingzhen Li (ProbAI 2022)

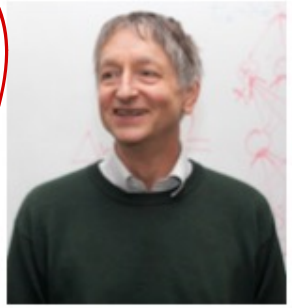
An Introduction to Bayesian Neural Networks

Yingzhen Li

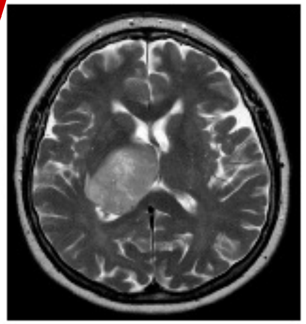
yingzhen.li@imperial.ac.uk







Deep Learning



模型

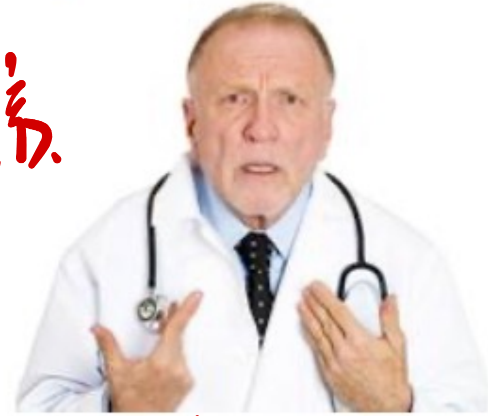
output 多可信

brain tumor 自信

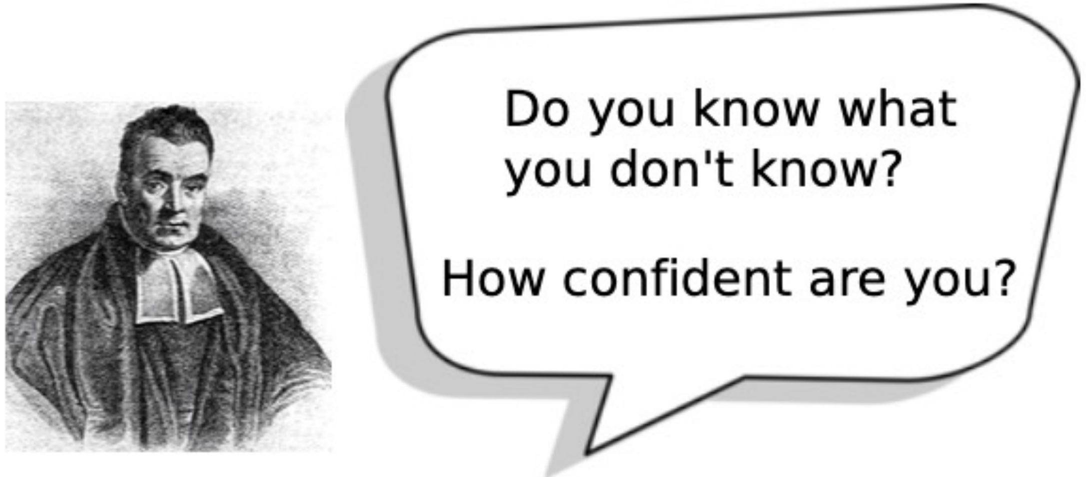
type "A"

are you sure? why?

softmax ← 过度自信



1/0 - 0.



Bayesian Inference

$\pi(\theta) = p(\theta|data)$

θ: para , p(θ) 先验.
p(data|θ) MLE θ* → L
p(θ|data) 后验.
认知 (有了 data)

$$P(\theta|\,data) = \frac{P(\theta)P(data\,|\,\theta)}{P(data)}$$

- $P(\theta)$: prior distribution
- $P(data\,|\,\theta)$: likelihood of θ given $data$ 参数对于数据的支撑.
- $P(\theta|\,data)$: posterior distribution of θ given $data$
- $P(data)$: marginal likelihood/model evidence

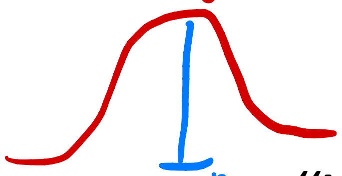
$$P(data) = \int P(\theta)P(data\,|\,\theta)$$

Bayesian Inference

Bayesian prediction distribution

- The central equation for Bayesian inference:

$p(\theta|D) = (0, 0)$



θ^*

MAP

$$\int F(\theta)p(\theta|D)d\theta = E_{p(\theta|D)}[F(\theta)]$$

"What is the prediction distribution of the test output given a test input?"

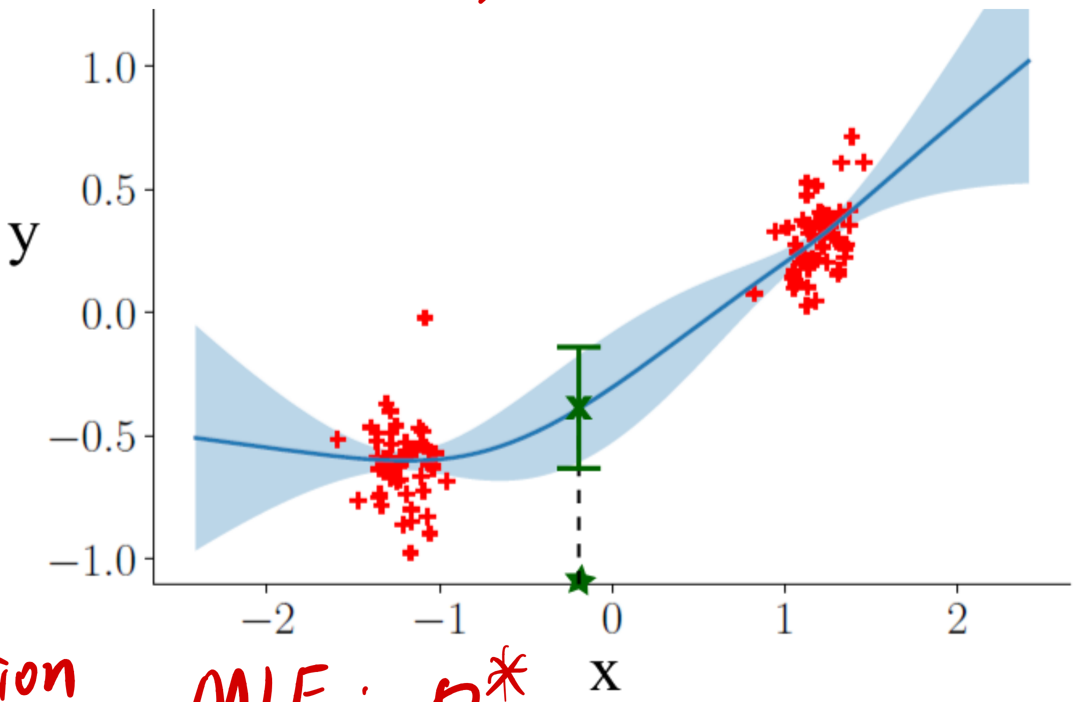
$F(\theta) = p(y|x, \theta),$

D = observed datapoints

$\theta = \theta^*$

$p(y^* | x^*, \theta)$

prediction



MLE: θ^*

Bayesian Neural Network (BNN) 101

分类

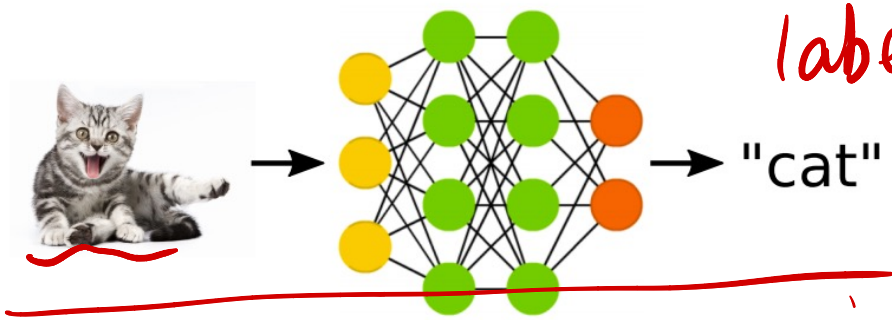
Classifying different types of animals:

- x: input image; y: output label
- Build a neural network with parameters θ:

$$p(y|x,\theta) = softmax(f_\theta(x))$$

f_θ: 参数

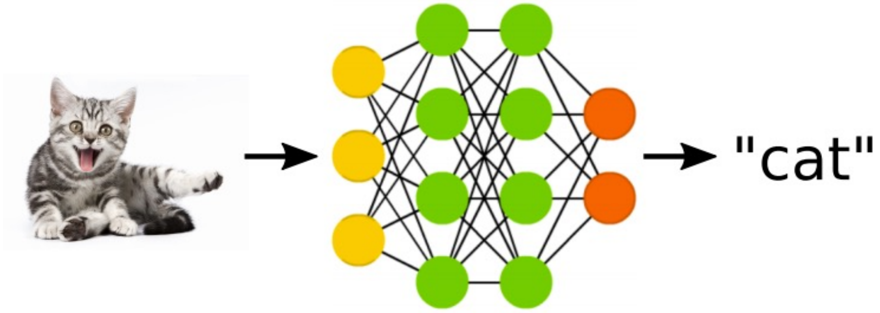
label



Bayesian Neural Network (BNN) 101

Classifying different types of animals:
- x: input image; y: output label
- Build a neural network with parameters θ:

$$p(y|x,\theta) = softmax(f_\theta(x))$$



A typical neural network (with non-linearity $g(\cdot)$):

$$f_\theta(x) = W^L g\big(W^{L-1}\, g\big(\dots g(W^1\, x + b^1)\big) + b^{L-1}\big) + b^L,$$

$$h^l = g(W^l\, h^{l-1} + b^l),\, h^1 = g(W^1\, x + b^1).$$

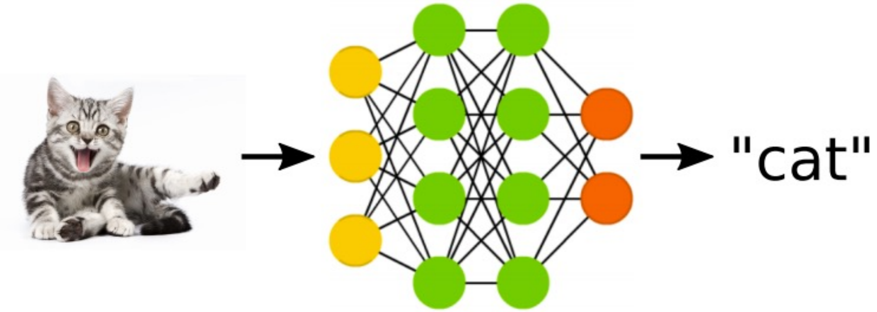
Neural network parameters: $\theta = \{W^l, b^l\}_{l=1}^L$

Bayesian Neural Network (BNN) 101

Classifying different types of animals:

- x: input image; y: output label
- Build a neural network with parameters θ:

$$p(y|x,\theta) = softmax(f_\theta(x))$$



Typical deep learning solution:

- Optimize θ to obtain a point estimates (MLE):

likelihood.

$$\theta^* = argmax \ \log p(D \mid \theta) \ ,$$

$$\log p(D \mid \theta) = \sum_{n=1}^{N} \log p(y_n \mid x_n, \theta) \ , D = \{(x_n, y_n)\}_{n=1}^{N}$$

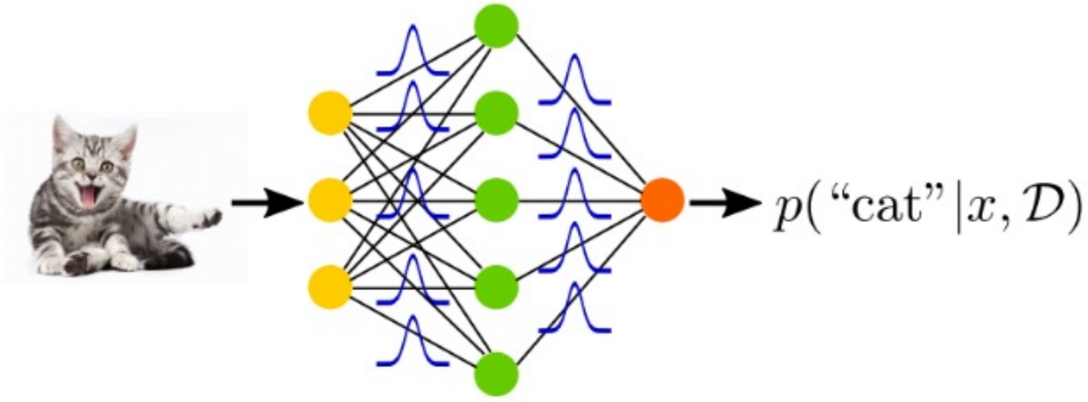
- Prediction: using $p(y^* \mid x^*, \theta^*)$

Bayesian Neural Network (BNN) 101

Classifying different types of animals:

- x: input image; y: output label
- Build a neural network with parameters θ:

$$p(y|x,\theta) = softmax(f_\theta(x))$$



Bayesian solution:

- Put a prior $p(\theta)$ on network parameters θ, e.g. Gaussian prior

$$p(\theta) = N(\theta; 0, \sigma^2 I)$$

- Compute the posterior distribution $p(\theta \mid D)$:

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta) / D$$

P(D) evidence (input, lable)

- Bayesian predictive inference:

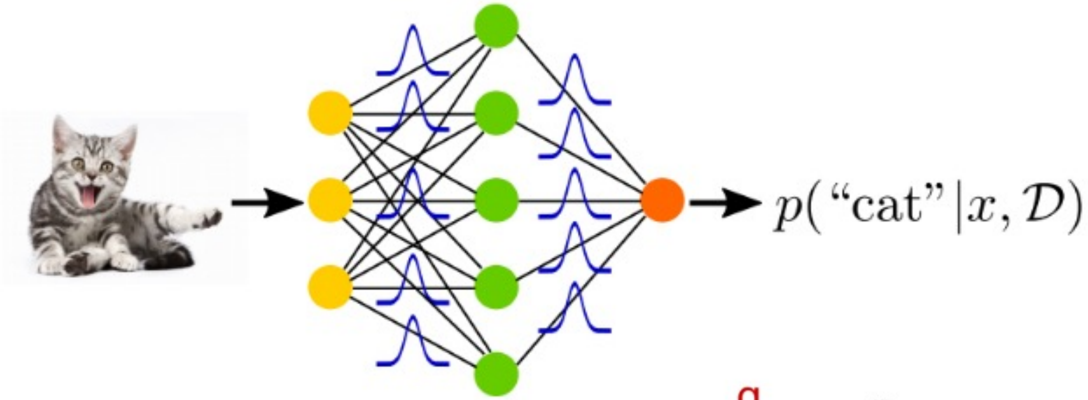
$$p(y^* \mid x^*, D) = E_{p(\theta \mid D)}[p(y^* \mid x^*, \theta)]$$

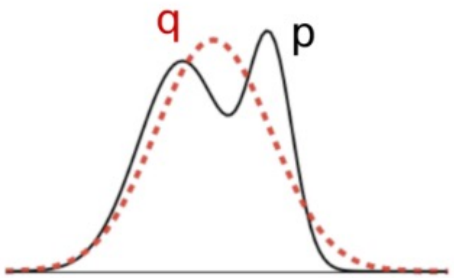
Bayesian Neural Network (BNN) 101

Classifying different types of animals:

- x: input image; y: output label
- Build a neural network with parameters θ:

$$p(y|x,\theta) = softmax(f_\theta(x))$$





Approximate (Bayesian) inference solution:

- Exact posterior intractable, use approximate posterior:

$$q(\theta) \approx p(\theta \mid D)$$

- Approximate Bayesian predictive inference:

$$p(y^* \mid x^*, D) \approx E_{q(\theta)}[p(y^* \mid x^*, \theta)]$$

$q(\theta)$

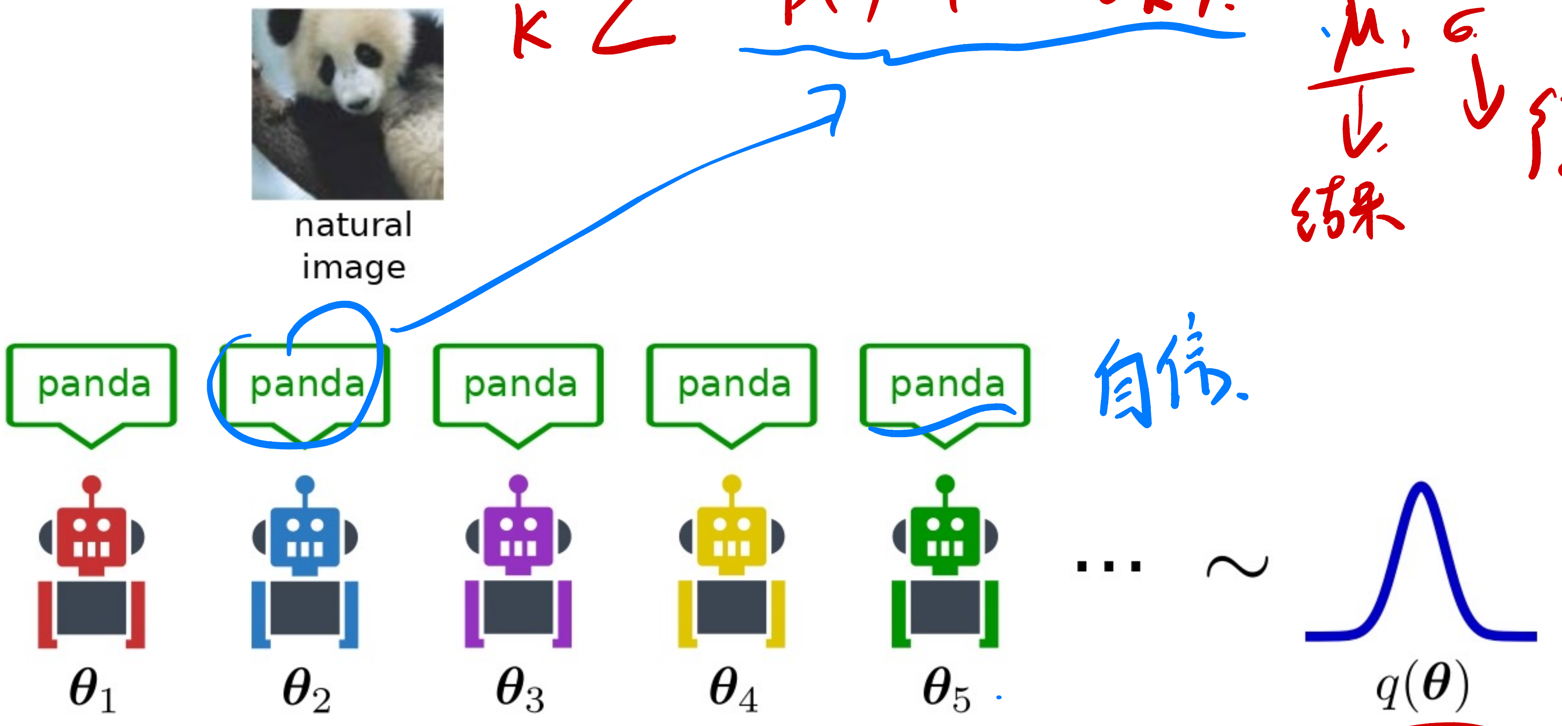
- Monte Carlo approximation:

$$p(y^* \mid x^*, D) \approx \frac{1}{K} \sum_{k=1}^{K} p(y^* \mid x^*, \theta_k), \quad \theta_k \sim q(\theta)$$

Bayesian Neural Network (BNN) 101

$$\frac{1}{K}\sum p(y^* \mid x^*, \theta_K).$$

μ, σ



$\sigma^2 \to 0$

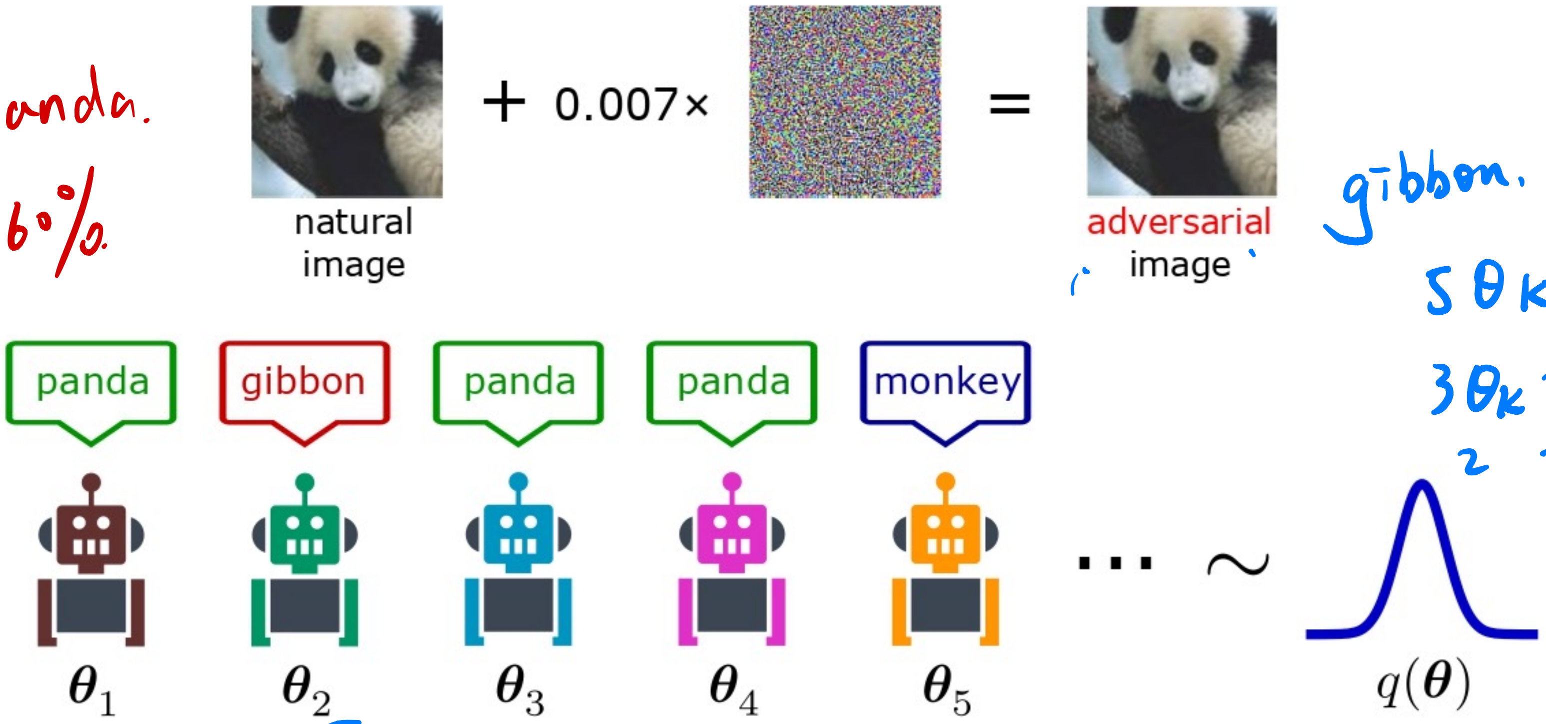
Prediction on in-distribution data:
ensemble over networks, using weights sampled from $q(\theta)$

Bayesian Neural Network (BNN) 101

Output → (μ_{panda} , $\sigma^2 I$).

output : panda.

置信性 : 60%



gibbon.

5 θ_k

3 θ_k → Panda

2 → x.

Prediction on OOD/noisy/adversarial data:

Disagreement (i.e. uncertainty) exists over networks sampled from $q(\theta)$

传统 DL: ① panda - 90% ② gibbon - 90%.

$\theta_1 \{w, b\}$ $\theta_2 \{w, b\}$ $q(\theta)$

$\theta \to \{w, b\} \leftarrow \theta$

Bayesian Neural Network (BNN) 101

$p(y^* | x^*, \theta = \theta_1)$ =

$\theta \sim q(\theta)$

$\theta \sim$

θ_1 θ_2 θ_3

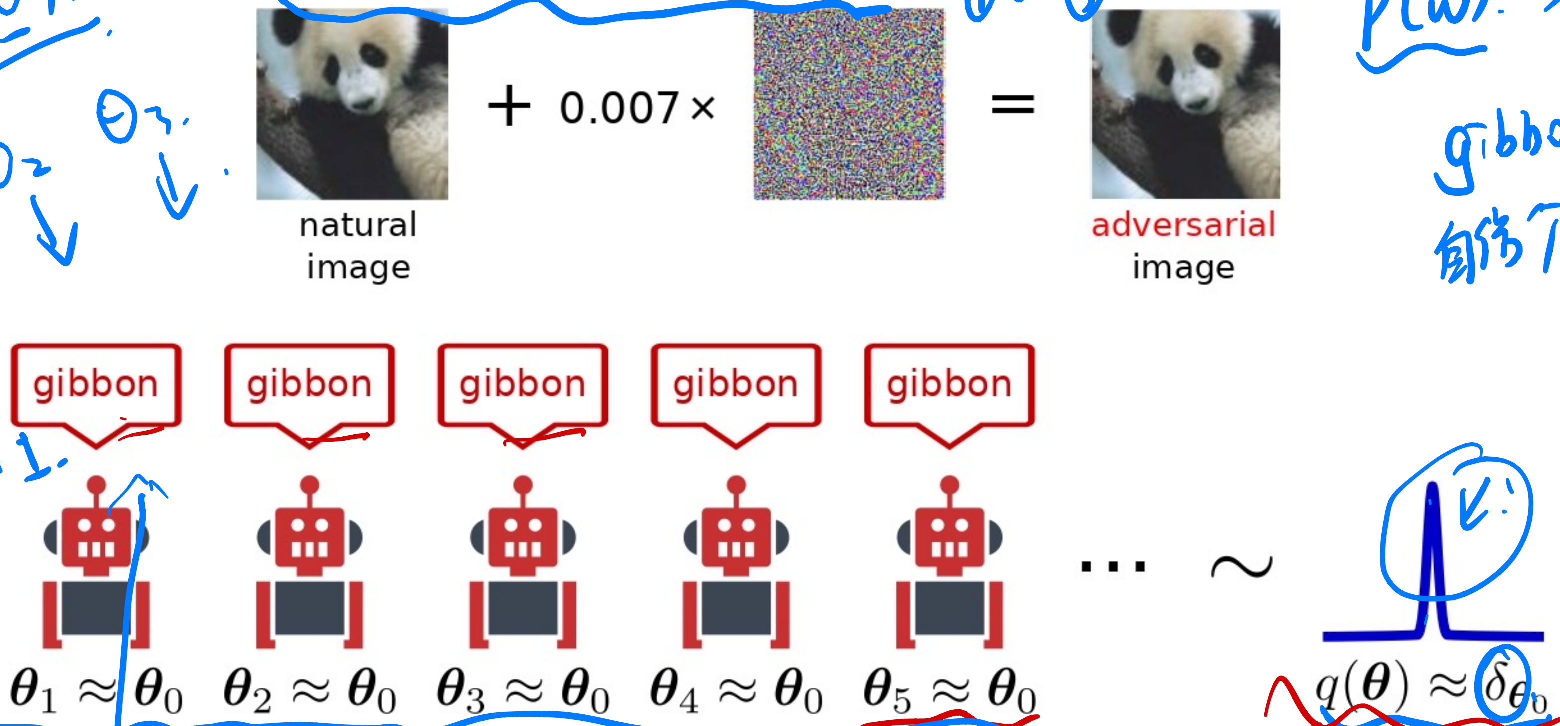
$p(w) \to p(w | \text{data})$

$q(w)$

gibbon, 倒了.

w

inference,



output 1.

$\theta_0 \to +\infty$

$\pm\theta_0 \to 0$

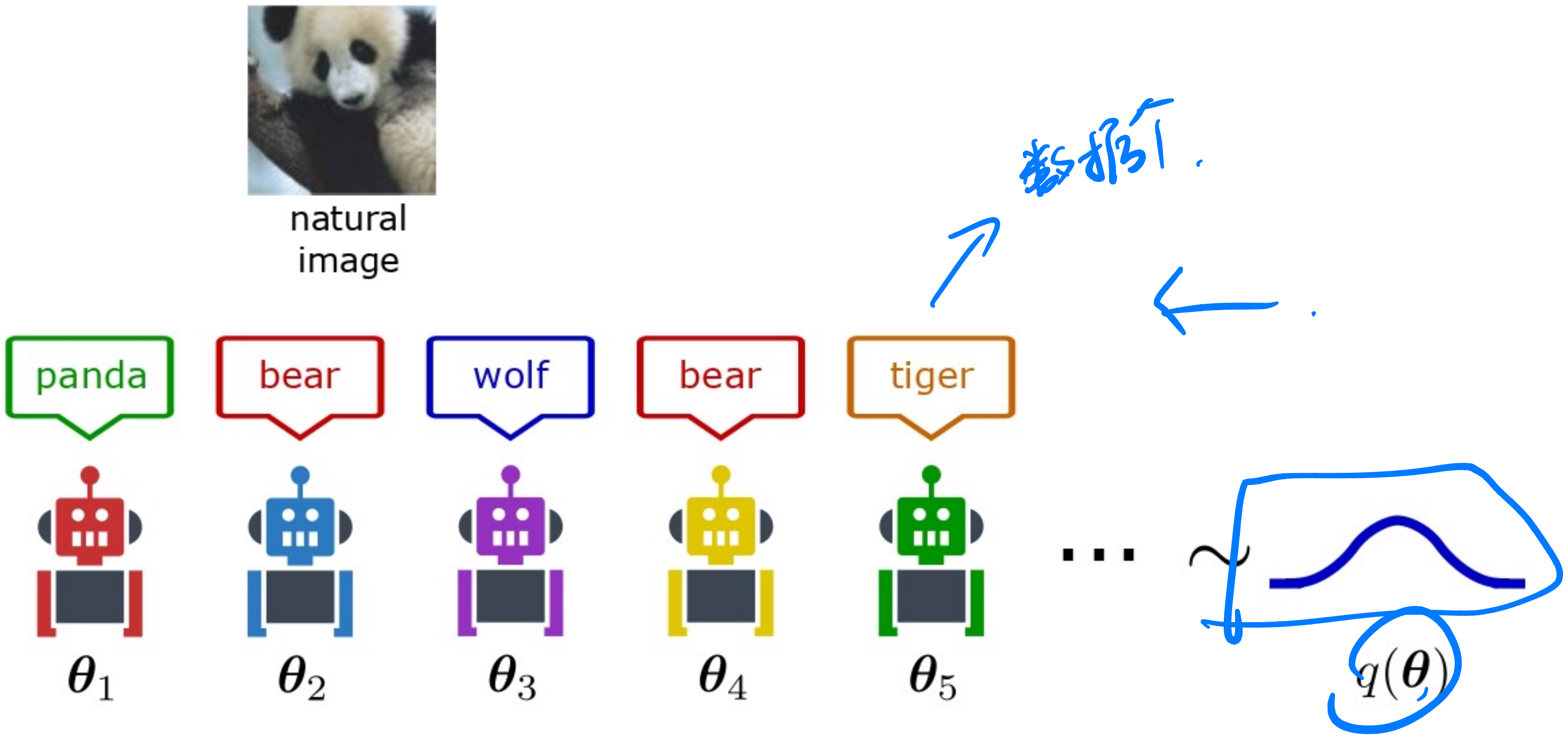
DL.

Prediction on OOD/noisy/adversarial data when $q(\theta)$ is over-confident: $N(\theta_0, \sigma^2 I)$

Return confidently wrong answers (close to point estimate)

θ,

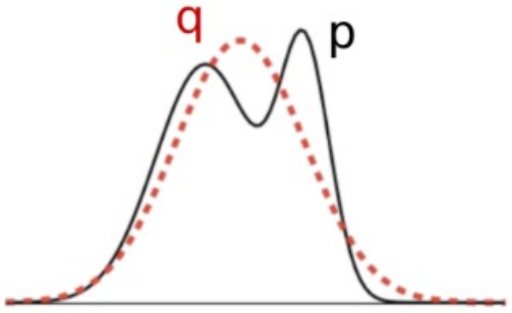
Bayesian Neural Network (BNN) 101



Prediction on in-distribution data when $q(\theta)$ is under-confident:
Low accuracy in prediction tasks (less desirable)

Approximate Inference in BNNs

- Key steps of approximate inference in BNNs
  1. Construct the $q(\theta) \approx p(\theta \mid D)$ distribution
     - Simple distributions: e.g. Mean-field Gaussian
     - Structured approximations, e.g. low-rank Gaussians
     - Others (non-Gaussian)



$\theta = \{W_l, b_l\}_{l=1}^{L}$ ← 超大的联合分布

μ: 50个 μ

σ: 2500/2

$q(\theta) = \prod q(W_l)\, q(b_l)$

$q(W_l) = q(w_1) \cdot q(w_2) \cdots$

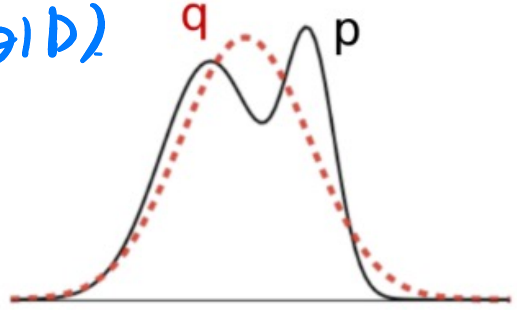
$w, b \Rightarrow$ 1个 μ, 1个 σ

See my NeurIPS 2020 tutorial on approximate inference

Approximate Inference in BNNs

- Key steps of approximate inference in BNNs
  1. Construct the $q(\theta) \approx p(\theta \mid D)$ distribution
     - Simple distributions: e.g. Mean-field Gaussian
     - Structured approximations, e.g. low-rank Gaussians
     - Others (non-Gaussian)
  2. Fit the $q(\theta)$ distribution
     - E.g. with variational inference

$q(\theta) \Rightarrow p(\theta|D)$



$q(\theta)$

prediction:

$$p(y^* | x^*, D) = \int p(y^* | x^*, \theta)\, p(\theta|D)\, d\theta$$

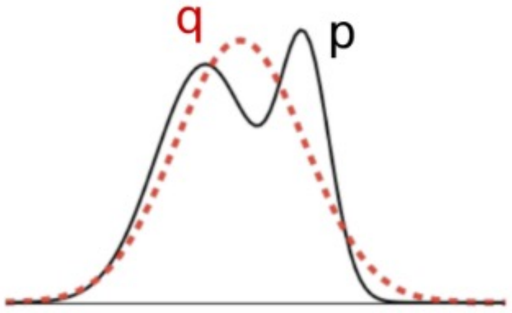
$$= \underline{E_{\underline{p(\theta|D)}} [\, p(y^* | x^*, \theta)\,]}$$

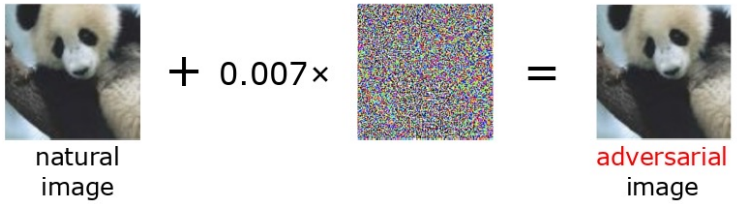
$$= \frac{1}{K} \sum_k p(y^* | x^*, \theta_k), \quad \theta_k \sim p(\theta|D)$$

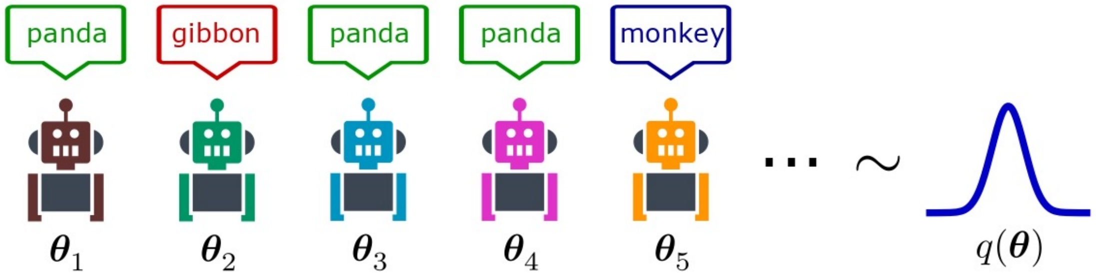
See my NeurIPS 2020 tutorial on approximate inference

Approximate Inference in BNNs

- Key steps of approximate inference in BNNs
  1. Construct the $q(\theta) \approx p(\theta \mid D)$ distribution
     - Simple distributions: e.g. Mean-field Gaussian
     - Structured approximations, e.g. low-rank Gaussians
     - Others (non-Gaussian)
  2. Fit the $q(\theta)$ distribution
     - E.g. with variational inference
  3. Compute prediction with Monte Carlo approximations







See my NeurIPS 2020 tutorial on approximate inference

Today's agenda

- Lecture on Basics: MFVI for BNNs
- Hands-on tutorial on BNNs
  - i.e., programming exercises
  - Also some case studies

Part I: Basics

- **Variational inference** ← $q(\theta) \to p(\theta|D)$
- **Bayes-by-backprop**

Bayesian Inference

更新的认知　已有的认知

$$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{P(D)}$$

- $P(\theta)$: prior
- $P(D \mid \theta)$: likelihood
- $P(\theta \mid D)$: posterior
- $P(D)$: marginal

Image courtesy of Sebastian Nowozin
Re-use of the image for any other purpose is not allowed

Variational Inference (VI)

The posterior

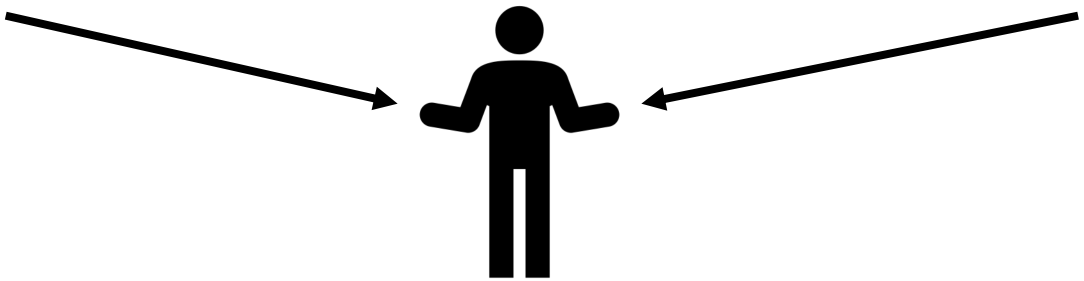
$$p(\theta|D) = p(D\,|\theta)p(\theta)/p(D)$$

The variational distribution

$$\boxed{q_\phi(\theta)}$$

$q_\phi(\theta)$

Inference as Optimization

$p(\theta|D)$  $q_\phi(\theta)$

Kullback-Leibler (KL) divergence

$KL[p(\theta|D) \,||\, q_\phi(\theta)] \downarrow$

Kullback-Leibler Divergence

$$KL[q(\theta)||p(\theta)] = \int q(\theta)\log\frac{q(\theta)}{p(\theta)}d\theta = E_{q(\theta)}[\log\frac{q(\theta)}{p(\theta)}]$$

Jenssen - shannon.

- When $p = q$, KL is 0
- Otherwise, KL > 0
- It measures how similar are these two distributions

Let’s Derive the Objective of VI

- Minimize $KL[q(\theta)||p(\theta|D)]$

— log-likeli

$$KL[q(\theta)||p(\theta|D)] = -\mathrm{E}_{q(\theta)}\left[\log\frac{p(\theta|D)}{q(\theta)}\right]$$

Let's Derive the Objective of VI

- Minimize $KL[q(\theta)||p(\theta|D)]$

$p(\theta,D)/p(D)$

$$KL[q(\theta)||p(\theta|D)] = -E_{q(\theta)}\left[\log\frac{p(\theta|D)}{q(\theta)}\right]$$

$$= -E_{q(\theta)}\left[\log\frac{\mathrm{p}(\theta,D)}{p(D)q(\theta)}\right] = -E_{q(\theta)}\left[\log\frac{\mathrm{p}(\theta,D)}{q(\theta)} - \log p(D)\right]$$

Let’s Derive the Objective of VI

- Minimize $KL[q(\theta)||p(\theta|D)]$

$$KL[q(\theta)||p(\theta|D)] = -E_{q(\theta)}\left[\log \frac{p(\theta|D)}{q(\theta)}\right]$$

$$= -E_{q(\theta)}\left[\log \frac{p(\theta,D)}{p(D)q(\theta)}\right] = -E_{q(\theta)}\left[\log \frac{p(\theta,D)}{q(\theta)} - \log p(D)\right]$$

$$= \boxed{\log p(D)} - E_{q(\theta)}\left[\log \frac{p(\theta,D)}{q(\theta)}\right]$$

Model Evidence

数据.

Let's Derive the Objective of VI

Minimize $KL[q(\theta)||p(\theta|D)]$

$$KL[q(\theta)||p(\theta|D)] = \log p(D) - E_{q(\theta)}\left[\log \frac{p(\theta, D)}{q(\theta)}\right]$$

Maximize $E_{q(\theta)}\left[\log \frac{p(\theta,D)}{q(\theta)}\right]$

Let's Derive the Objective of VI

Minimize $KL[q(\theta)||p(\theta|D)]$

$$KL[q(\theta)||p(\theta|D)] = \boxed{\log p(D)} - E_{q(\theta)}\left[\log \frac{p(\theta, D)}{q(\theta)}\right]$$

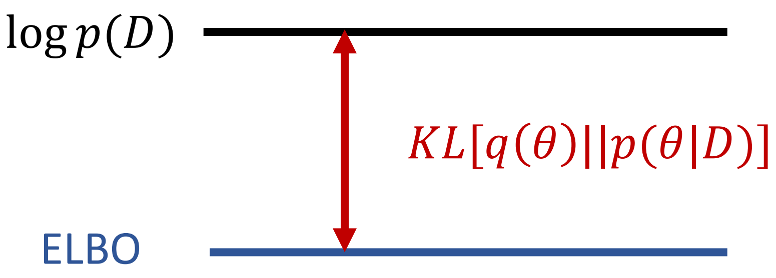
Model Evidence

≥ 0

$E + \underbrace{KL}_{} = \log p(D)$ $\Rightarrow$ $\log p(D) \geq E$.

Maximize $L = E_{q(\theta)}\left[\log \frac{p(\theta,D)}{q(\theta)}\right]$

Evidence Lower Bound (ELBO)



"Model Evidence = ELBO + KL"

Variational Inference (VI)

KL : $q(\theta) \,||\, p(\theta|D)$

The posterior

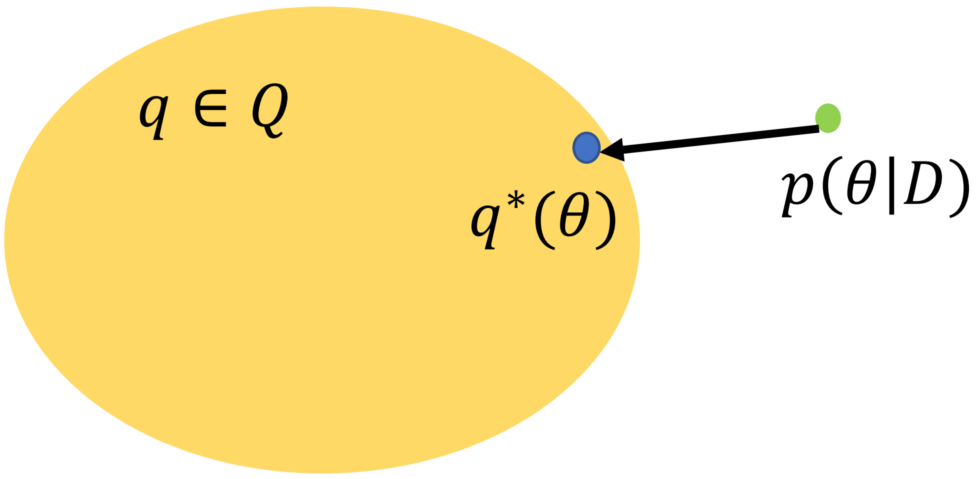
$$p(\theta|D) = p(D|\theta)p(\theta)/p(D)$$

The variational distribution

$$q_\phi(\theta)$$

$p(D,\theta) = p(D|\theta) \cdot \boxed{p(\theta)}$

$$L = E_{q_\phi(\theta)}\left[\log \frac{\mathrm{p}(\mathrm{D}, \theta)}{q_\phi(\theta)}\right] = \log p(D) - KL[\, q_\phi(\theta)||p(\theta)]$$



Variational Inference (VI)

KL ⇓

p. q. 近似.

N.θ. SGD.

q(θ)∼p(θ|D)

q_ϕ → p(θ)

不复杂

不overfit...

- Rewriting the ELBO: max

evidence.

$$\log p(D) \geq L = \underbrace{E_{q_\phi(\theta)}[\log p(D|\theta)]}_{\text{Data fitting term}} - \beta \underbrace{KL[q_\phi(\theta) \| p(\theta)]}_{\text{KL regulariser}}$$

↑大 ↓

p(θ|D) = L × prior.

MLE.

(Negative) Data fitting term:

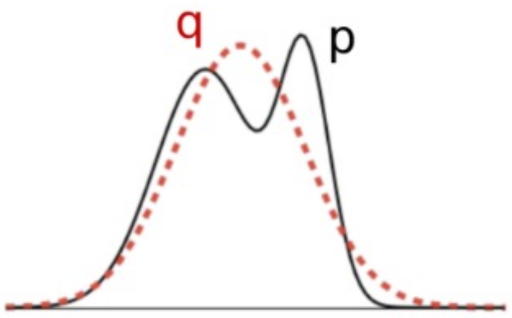
- Like the usual DL loss you'll use for training neural networks
- ...except that now the network's weights are sampled from q

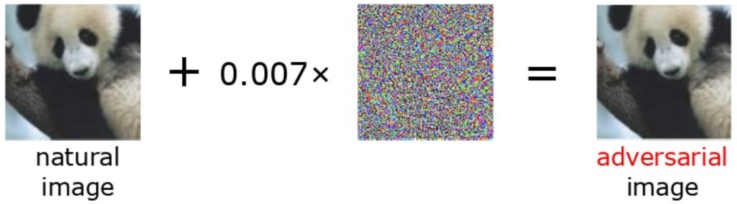
KL regulariser:

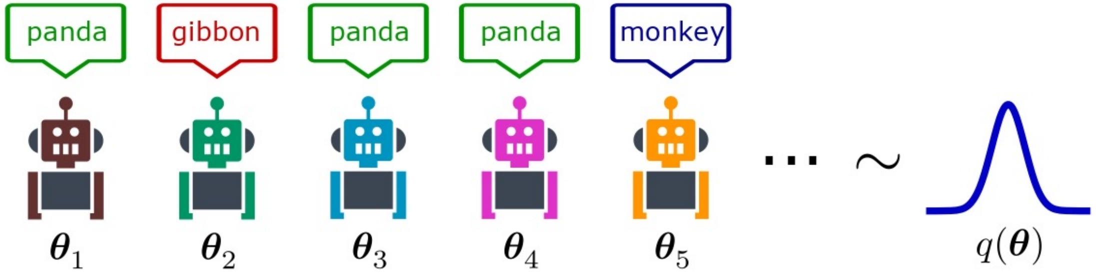
- Make the q distribution closer to the prior
- Regularises the approximate posterior, especially when using e.g., Gaussian prior

Approximate Inference in BNNs

- Key steps of approximate inference in BNNs
  1. Construct the $q(\theta) \approx p(\theta \mid D)$ distribution
     - Simple distributions: e.g. Mean-field Gaussian
     - Structured approximations, e.g. low-rank Gaussians
     - Others (non-Gaussian)
  2. Fit the $q(\theta)$ distribution
     - E.g. with variational inference
  3. Compute prediction with Monte Carlo approximations







See my NeurIPS 2020 tutorial on approximate inference

Approximate Inference in BNNs

- Step 1: construct the $q(\theta) \approx p(\theta \mid D)$ distribution
  - Example: Mean-field Gaussian distribution:

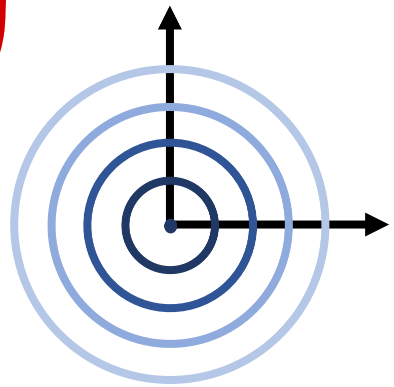
SGD.

$$q(\theta) = \prod_{l=1}^{L} q(W^l)\, q(b^l)$$

$$q(W_l) = \prod_{ij} q(W_{ij}^l), \qquad q(W_{ij}^l) = N(W_{ij}^l; M_{ij}^l, V_{ij}^l)$$

b.

$$q(b^l) = \prod_i q(b_i^l), \quad q(b_i^l) = N(b_i^l; m_i^l, v_i^l)$$



- Variational parameters: $\phi = \{M_{ij}^l, \log V_{ij}^l, m_i^l, \log v_i^l\}_{l=1}^{L}$

R. SGD

Blundell et al. Weight Uncertainty in Neural Networks. ICML 2015

Approximate Inference in BNNs

- Step 2: fit the $q(\theta)$ distribution:
  - Variational inference: $\phi^* = argmax\ L(\phi)$

$$L(\phi) = E_{q_\phi(\theta)}[\log p(D \mid \theta)] - KL\left[q_\phi(\theta) \parallel p(\theta)\right]$$

Blundell et al. Weight Uncertainty in Neural Networks. ICML 2015

Approximate Inference in BNNs

- Step 2: fit the $q(\theta)$ distribution:
  - Variational inference: $\phi^* = argmax\, L(\phi)$ Likeli
  $$L(\phi) = \boxed{E_{q_\phi(\theta)}[\log p(D \mid \theta)]} - KL[q_\phi(\theta) \parallel p(\theta)]$$
  - First scalable technique: Stochastic optimization
    - i.i.d. assumption of data: $\log p(D \mid \theta) = \sum_{n=1}^{N} \log p(y_n \mid x_n, \theta)$
    - Enable mini-batch training with $\{(x_m, y_m)\} \sim D^M$:

$$L(\phi) \approx \frac{N}{M} \sum_{m=1}^{M} E_{q(\theta)}[\log p(y_m \mid x_m, \theta)] - KL[q(\theta) \,||\, p(\theta)]$$

Blundell et al. Weight Uncertainty in Neural Networks. ICML 2015

Approximate Inference in BNNs

- Step 2: fit the $q(\theta)$ distribution:
  - Variational inference: $\phi^* = argmax\ L(\phi)$

$$L(\phi) = E_{q_\phi(\theta)}[\log p(D \mid \theta)] - KL\big[q_\phi(\theta) \,\|\, p(\theta)\big]$$

  - First scalable technique: Stochastic optimization
    - i.i.d. assumption of data: $\log p(D \mid \theta) = \sum_{n=1}^{N} \log p(y_n \mid x_n, \theta)$
    - Enable mini-batch training with $\{(x_m, y_m)\} \sim D^M$:

$$L(\phi) \approx \boxed{\frac{N}{M}} \sum_{m=1}^{M} E_{q(\theta)}[\log p(y_m \mid x_m, \theta)] - KL[q(\theta) \,||\, p(\theta)]$$

reweighting to ensure calibrated posterior concentration

$w, b.$ $w \begin{cases} M_w \\ V_w. \end{cases}$ $b \sim \begin{cases} M_b \\ V_b. \end{cases}$

$\frac{\partial D}{\partial w}$, $\frac{\partial D}{\partial b}$

$\frac{\partial D}{\partial w} \cdot \frac{\partial w}{\partial M_w}$,

$\frac{\partial D}{\partial w} \cdot \frac{\partial w}{\partial V_w}$.

Approximate Inference in BNNs

- Step 2: fit the $q(\theta)$ distribution:
  - 2nd scalable technique: Monte Carlo sampling
    - $E_{q(\theta)}[\log p(y \mid x, \theta)]$ intractable even with Gaussian $q(\theta)$
    - Solution: Monte Carlo estimate:

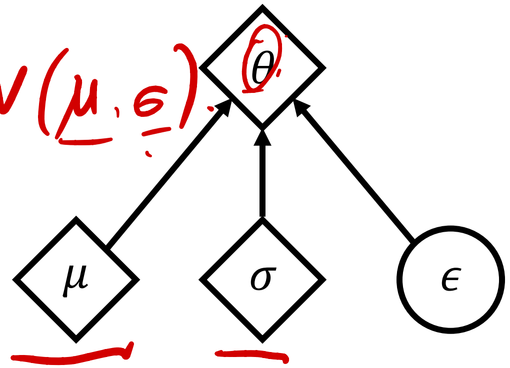
$$E_{q(\theta)}[\log p(y \mid x, \theta)] \approx \frac{1}{K} \sum_{k}^{K} \log p(y \mid x, \theta_k), \quad \theta_k \sim q(\theta)$$

Approximate Inference in BNNs

- Step 2: fit the $q(\theta)$ distribution:
  - 2nd scalable technique: Monte Carlo sampling
    - $E_{q(\theta)}[\log p(y \mid x, \ \theta)]$ intractable even with Gaussian $q(\theta)$
    - Solution: Monte Carlo estimate:

$$E_{q(\theta)}[\log p(y \mid x, \theta)] \approx \frac{1}{K} \sum_{k}^{K} \log p(y \mid x, \theta_k), \qquad \theta_k \sim q(\theta)$$

$\sim N(\mu, \sigma)$



    - Reparameterization trick to sample mean-field Gaussians:

$$\theta_k \sim q(\theta) \Leftrightarrow \theta_k = m_\theta + \sigma_\theta \odot \epsilon_k, \ \epsilon_k \sim N(0, I)$$

θ_k.

$\partial \theta_k$. $\frac{\partial \theta_k}{\partial m_\theta}$ $\frac{\partial \theta_k}{\partial \sigma_\theta}$

Approximate Inference in BNNs

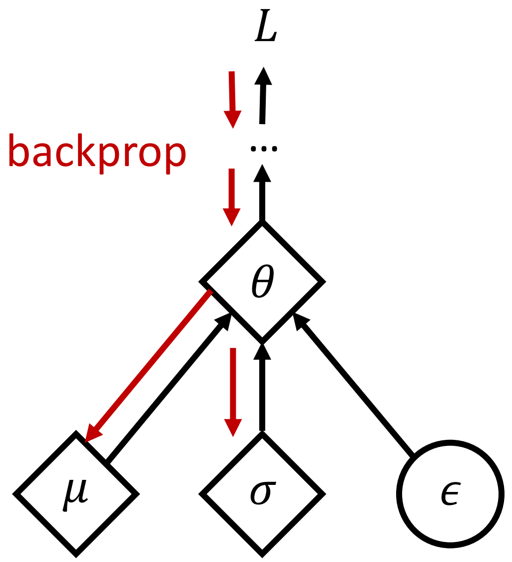
- Step 2: fit the $q(\theta)$ distribution:
  - 2nd scalable technique: Monte Carlo sampling
    - $E_{q(\theta)}[\log p(y \mid x,\ \theta)]$ intractable even with Gaussian $q(\theta)$
    - Solution: Monte Carlo estimate:

$$E_{q(\theta)}[\log p(y\,|x,\theta)] \approx \frac{1}{K}\sum_{k}^{K}\log p(y\,|x,\theta_k)\,, \qquad \theta_k \sim q(\theta)$$

    - Reparameterization trick to sample mean-field Gaussians:

$$\theta_k \sim q(\theta) \Leftrightarrow \theta_k = m_\theta + \sigma_\theta \odot \epsilon_k,\ \ \epsilon_k \sim N(0, I)$$

$$\Rightarrow\ E_{q(\theta)}\,[\log p(y\,|x,\theta)] \approx \frac{1}{K}\sum_{k}^{K}\log p(y\,|x,\theta_k = m_\theta + \sigma_\theta\epsilon_k)\,, \epsilon_k \sim N(0, I)$$



Blundell et al. Weight Uncertainty in Neural Networks. ICML 2015

Approximate Inference in BNNs

- Combining both steps:

$$L(\phi) \approx \frac{N}{M} \sum_{m=1}^{M} \frac{1}{K} \sum_{k=1}^{K} \log p(y_m \mid x_m, \theta_k) - \underline{KL[\overset{\mathcal{G}}{q(\theta)} \,||\, \overset{?}{p(\theta)}]}, \theta_k \sim q(\theta)$$

analytic between two Gaussians
(if not, can also be estimated with Monte Carlo)

In regression:
$p(y \mid x, \theta) = N(f_\theta(x), \sigma^2)$

In classification:
$p(y \mid x, \theta) = Categorical(logit = f_\theta(x))$

Blundell et al. Weight Uncertainty in Neural Networks. ICML 2015

Approximate Inference in BNNs

MC.

- Step 3: compute prediction with Monte Carlo approximations:

$$p(y^* \mid x^*, D) \approx \frac{1}{K} \sum_{k=1}^{K} p(y^* \mid x^*, \theta_k), \quad \theta_k \sim q(\theta)$$

$p(\theta|D)$.

Mean-field Gaussian case:
$\theta_k = m_\theta + \sigma_\theta \odot \epsilon_k, \ \epsilon_k \sim N(0, I)$

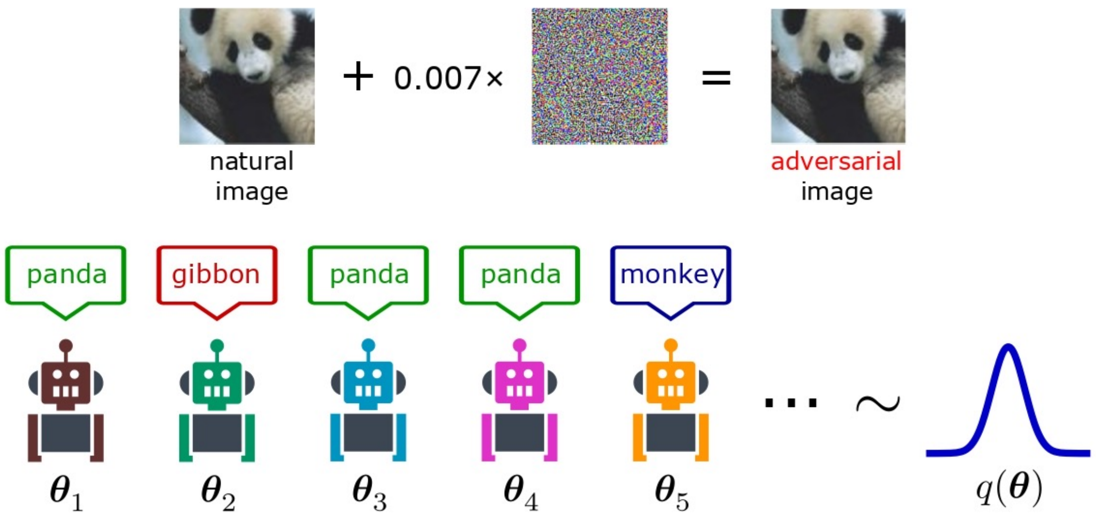
$m_\theta \ \ \sigma_\theta$.

$q(\theta)$

$q(w) \sim N(M, V)$

$w = M + V \cdot \varepsilon$

$\varepsilon \sim N(0,1)$



Part II: Bayesian MLPs

- **Implement various BNN methods for MLP architectures**
- **Regression example test**
- **Case study 1: Bayesian Optimisation with UCB**

https://bit.ly/3zF1zvA

Instructions for using this Google Colab notebook:

- Make sure you have signed in with your Google account;
- Click "File > Save a copy in Drive" to create your own copy;
- Let's play around with the demo using your own copy!

Findings with MFVI for Bayesian MLPs

Variation Infer-

mean field. $\Rightarrow$ $q(\theta) = \underbrace{q(W)}_{\prod q(w_{ij}^{l})} \cdot q(b).$

Findings with MFVI for Bayesian MLPs

Mixture Gaussian.

$W, b \leftarrow m_b, v_b$. prior.

- MFVI tends to underfit
  - Initialisation matters ←
  - Tuning the beta parameter also helps

$$\text{ELBO} = \text{data fitting} - \beta\, KL[q_\phi(\theta) \| p(\theta)]$$

No free lunch theorem.

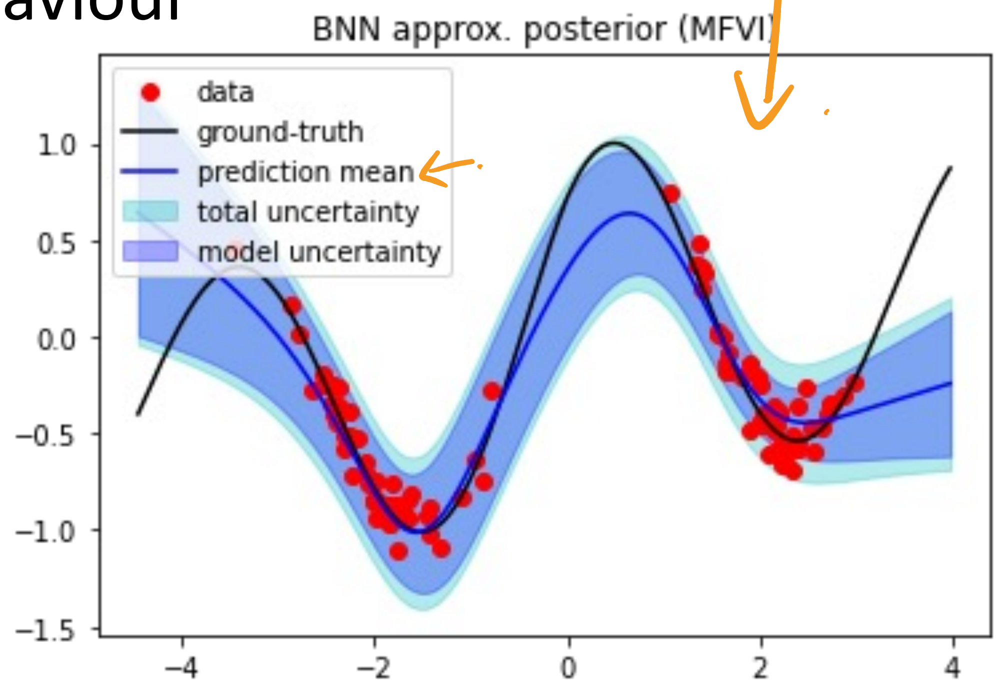
没有一个模型可以在所有任务上都表现得最好.

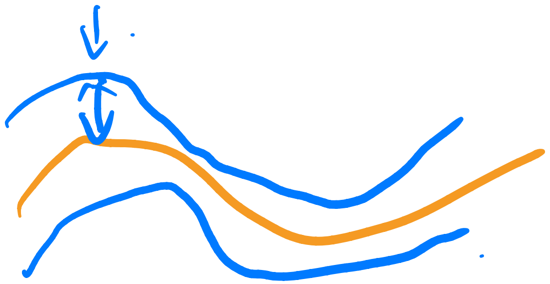
模型 ← 内含先验.

bayes: 显式的先验 $p(\theta)$. 没有一个通用的最好的prior.

Findings with MFVI for Bayesian MLPs

- MFVI tends to underfit
  - Initialisation matters
  - Tuning the beta parameter also helps
- Uncertainty behaviour





uncertainty

Using other q distributions?

- Using more complicated q distributions? q : mean field.
  - Pros: more flexible approximations ⇒ better posterior approximations (?)
  - Cons: higher time & space complexities

$q(\underline{W}) = \prod q(w_{ij}^{l})$.

full

$q\begin{pmatrix} W \\ B \end{pmatrix} = \begin{pmatrix} \quad \end{pmatrix}$.

Using other q distributions?

- Using more complicated q distributions?
  - Pros: more flexible approximations ⇒ better posterior approximations (?)
  - Cons: higher time & space complexities

deterministic | θ → random variable.

- We will look at 2 alternatives:
  - "Last-layer BNN": Full covariance Gaussian approximations for the last layer
  - MC-Dropout: adding dropout layers and run them in both train & test time

layer.

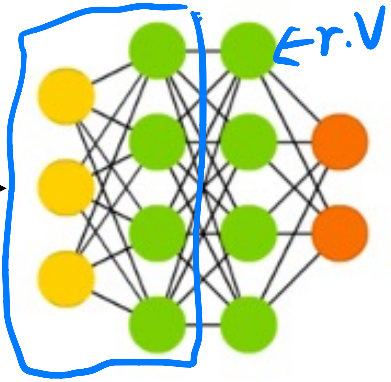
“Last-layer BNN”

- Use deterministic layers for all but the last layer
- For the last layer: Use Full-covariance Gaussian approximate posterior:

$$
\begin{cases}
q(\theta^l) = \delta(W^l = M^l, b^l = m^l), l = 1, \dots, L-1, \\
q(\theta^L) = N(vec(\theta^L); vec(\mu^L), \Sigma), \theta^L = \{W^L, b^L\}
\end{cases}
$$

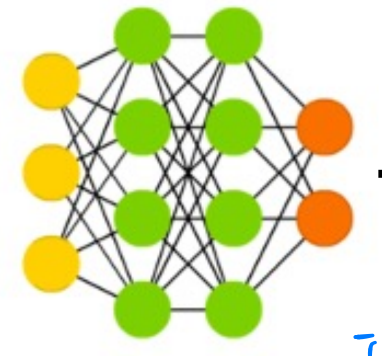
- For regression this is equivalent to Bayesian linear regression (BLR) with NN-based non-linear features

r.v.

$x \longrightarrow$  $\longrightarrow f_{\theta^{1:L-1}}(x) \longrightarrow$ BLR $\longrightarrow p(y \mid x, \theta)$

"Last-layer BNN"

- Use deterministic layers for all but the last layer
- For the last layer: Use Full-covariance Gaussian approximate posterior
- For regression this is equivalent to Bayesian linear regression (BLR) with NN-based non-linear features

$x \longrightarrow$  $\longrightarrow f_{\theta^{1:L-1}}(x) \longrightarrow$ BLR $\longrightarrow p(y \mid x, \theta)$

前: $L-1$

$q(\theta) = 10$

$p(\theta) = 8$ $+\infty$

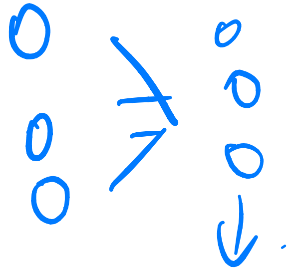
Inference

$$L = E_q[\log p(D \mid \theta)] - KL[q(\theta^L) \parallel p(\theta^L)]$$

Use deterministic weights for $l = 1, \ldots, L-1$
Sample $W^L \sim q(W^L)$

KL regulariser for the last layer only

MC-Dropout



train → ou
test

$KL[q(\theta) \| p(\theta)]$

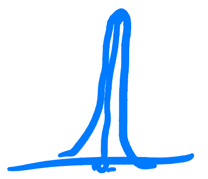
- Add dropout layers to the network
- Perform dropout during training

L2 regulariser on the variational parameters

KL.

$W_0 \quad z \sim Bin$

$W_0 \sim N(0, \sigma)$



$$L = E_q[\log p(D|\theta)] - (1-\pi)\ell_2(\phi)$$

The MC sampling procedure is implicitly defined

Dropout rate

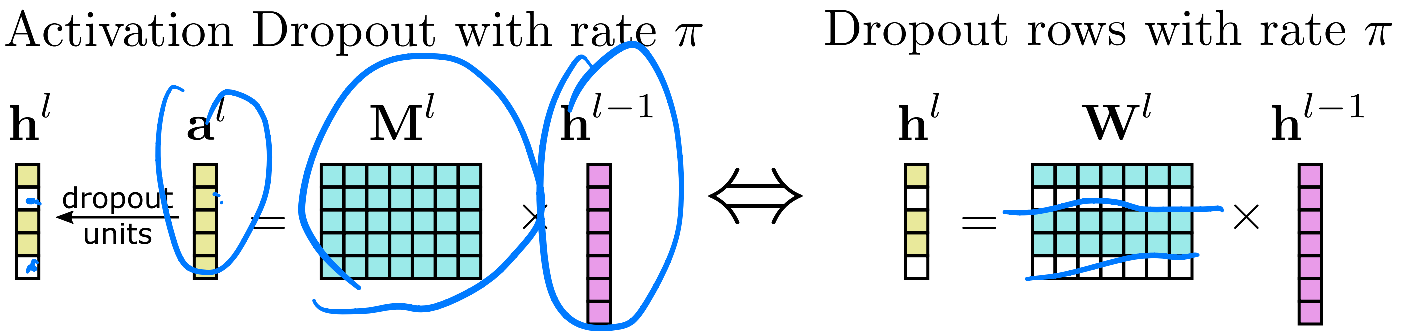
- In test time, run multiple forward passes with dropout

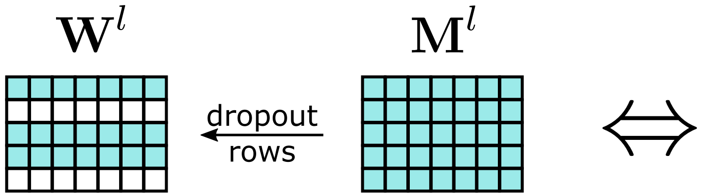
$$p(y^* \mid x^*, D) \approx \frac{1}{K}\sum_{k=1}^{K} p(y^* \mid x^*, \theta_k), \quad \theta_k \sim q(\theta)$$

The MC sampling procedure is implicitly defined

MC-Dropout

- Two equivalent ways to implement MC-Dropout:





Sample rows $\mathbf{W}_i^l$ from

$$q(\mathbf{W}_i^l) = (1-\pi)\mathcal{N}(\mathbf{M}_i^l, \eta \mathbf{I}) + \pi \mathcal{N}(\mathbf{0}, \eta \mathbf{I})$$

$$\eta \to 0$$

(Similar logic applies when including the bias terms, see lecture notes.)
(Notice that pytorch's nn.Linear layer uses formats like xW^T instead of Wx.)

Gal and Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. ICML 2016.

Using other q distributions?

- What you'll do for the next part of the tutorial:
  - Implement MC-Dropout in 2 ways
  - Run the regression sample with the 2 approximation methods discussed
  - Compare with MFVI

Case study 1: Bayesian Optimisation

- Imagine you'd like to solve the following task:

$$x^* = argmax_x f_0(x)$$ 黑盒

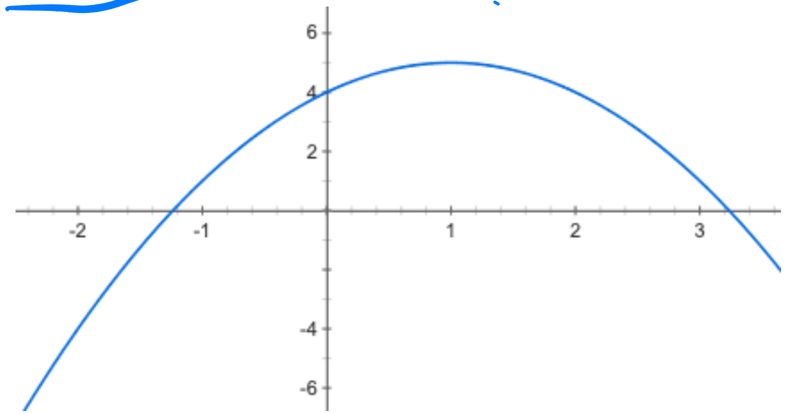
Case study 1: Bayesian Optimisation

- Imagine you'd like to solve the following task:

$$x^* = argmax_x \, f_0(x)$$

gradient

Known functional form of f_0:



Gradient descent, Newton's method,

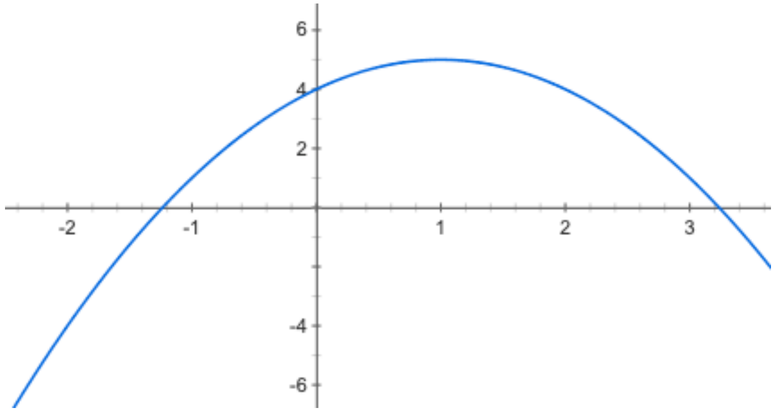
…

Case study 1: Bayesian Optimisation

- Imagine you'd like to solve the following task:

$$x^* = argmax_x \, f_0(x)$$

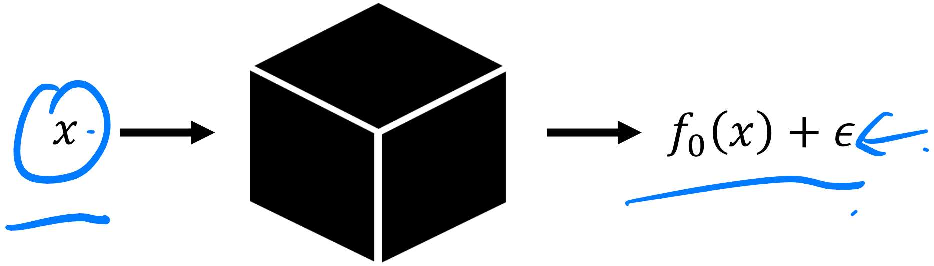
Known functional form of f_0:



Gradient descent, Newton's method,

…

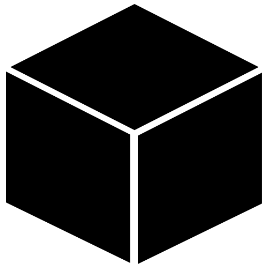
Unknown functional form of f_0:



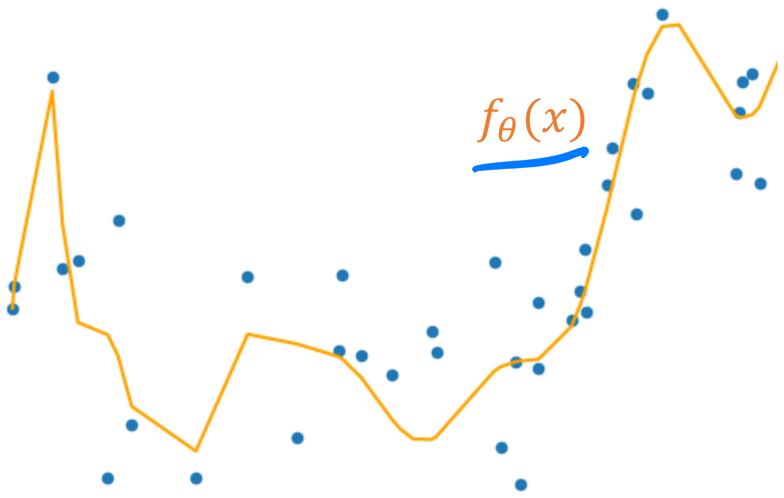
(can only query (noisy) function values)

Case study 1: Bayesian Optimisation

- Idea 1: fit a surrogate function $f_\theta \approx f_0$

$x_i \longrightarrow$  $\longrightarrow y_i = f_0(x_i) + \epsilon_i$

Fit f_θ using D $\longrightarrow$



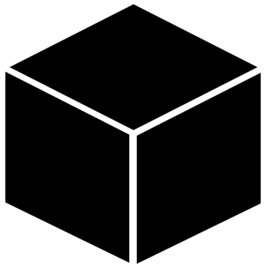
Collect a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$

$f_\theta(x)$ has a known (parametric) form

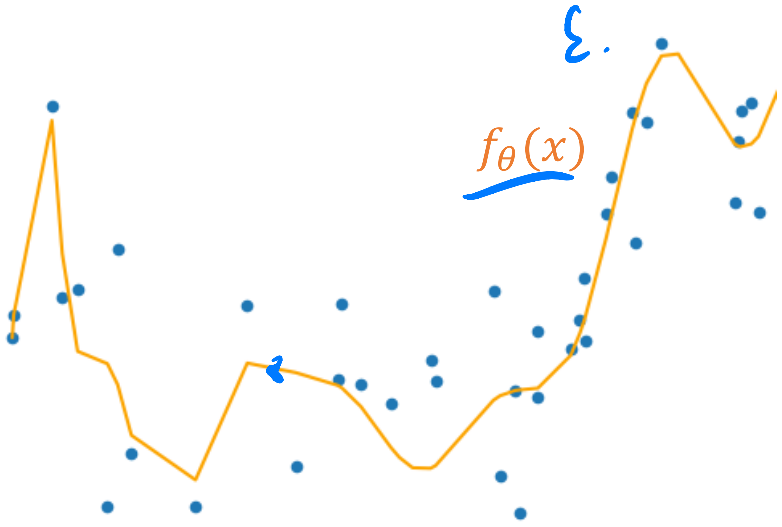
$\Rightarrow$ find maximum using e.g., Newton's method

Case study 1: Bayesian Optimisation

- Idea 1: fit a surrogate function $f_\theta \approx f_0$

$x_i \rightarrow$  $\rightarrow y_i = f_0(x_i) + \epsilon_i$

Fit f_θ using D $\longrightarrow$



Collect a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$

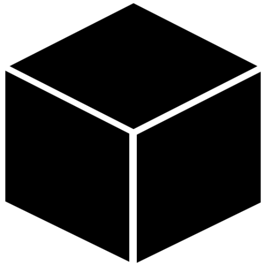
- Issues of this approach:
  - Need to collect a lot of datapoints for accurate fitting of f_θ
  - Do not consider uncertainty at unseen locations

Case study 1: Bayesian Optimisation

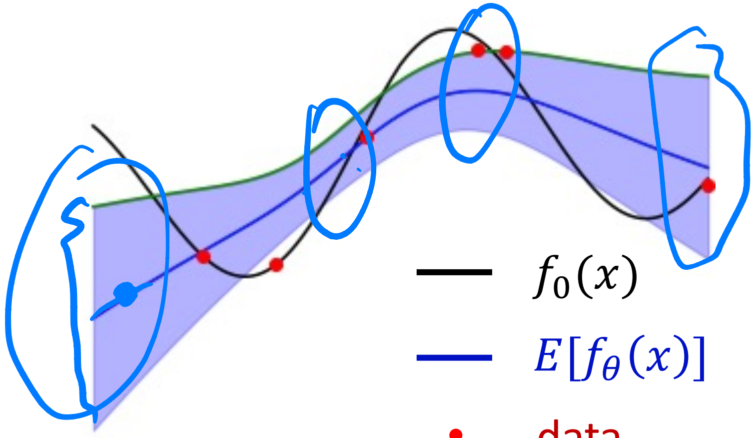
① 均
② 不确定/方差

- Idea of BO: iterate the following steps
  - fit a surrogate function f_θ with uncertainty estimates
  - Use the surrogate function to guide the dataset collection process

Update the dataset $D = D \cup \{(x_*, y_*)\}$

x_* →  → $y_* = f_0(x_*) + \epsilon_*$

Fit f_θ using D
with uncertainty estimates



Propose next point to query: $x_* = argmax_x\, a(x)$ ← acquisition function

Srinivas et al. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. ICML 2010.
Snoek et al. Practical Bayesian Optimization of Machine Learning Algorithms. NeurIPS 2012.

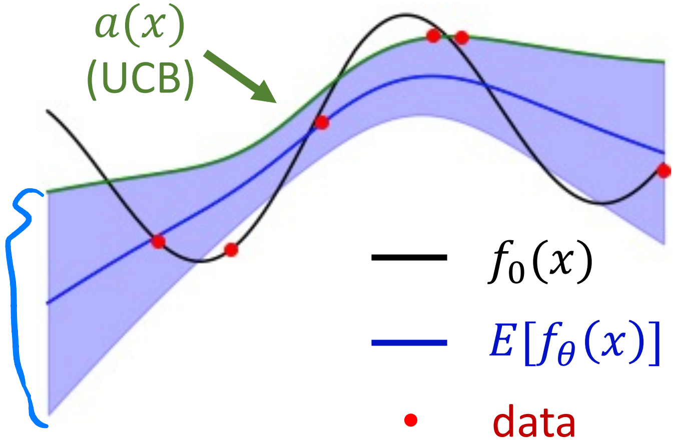
Case study 1: Bayesian Optimisation

- Upper confidence bound (UCB): a widely used acquisition function

$$a(x) = m(x) + \beta\sigma(x)$$

Mean of $f_\theta(x)$ over $\theta \sim q(\theta)$

Std of $f_\theta(x)$ over $\theta \sim q(\theta)$



BNN - infinite wide → GP.

Srinivas et al. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. ICML 2010.
Snoek et al. Practical Bayesian Optimization of Machine Learning Algorithms. NeurIPS 2012.

Case study 1: Bayesian Optimisation

- What you'll do for the case study part of the tutorial:
  - Implement UCB acquisition function
  - Run the BO example
  - Play around with hyper-parameters and other settings

Example answers of the tutorial demos:
Regression: https://bit.ly/39eZHit

Part III: Bayesian ConvNets

- **Classification example test**
- **Case study 2: Detecting adversarial examples**

https://bit.ly/3Hd1Ass

Instructions for using this Google Colab notebook:

- Make sure you have signed in with your Google account;
- Click “File > Save a copy in Drive” to create your own copy;
- Use GPU: in “Runtime > Change runtime type”, choose “GPU” for “hardware accelerator”
- Let’s play around with the demo using your own copy!

Case study 2: Detecting adversarial examples

□ → BNN → (lable, 可信度).

- Hypothesis:
  - Adversarial examples are regarded as OOD data 非 OOD, noisy, 非对抗
  - BNNs become uncertain about their prediction on OOD data
  - ⇒ uncertainty measures can be used for detecting adversarial examples

Uncertainty measures

$y = y^* + \varepsilon$

$x = x^* + \varepsilon$

(y, x)

固有

Total uncertainty = epistemic uncertainty + aleatoric uncertainty

数据

Due to lack of "knowledge"
(reducible when having more data)

Due to inherent stochasticity in data
(non-reducible)

Imagine flipping a coin:

- Epistemic uncertainty: "How much do I believe the coin is fair?"
  - Model's belief after seeing the population
  - Reduces when having more data
- Aleatoric uncertainty: "What's the next coin flip outcome?"
  - Individual experiment outcome
  - Non-reducible

Uncertainty measures

Total uncertainty = epistemic uncertainty + aleatoric uncertainty

Due to lack of "knowledge"
(reducible when having more data)

Due to inherent stochasticity in data
(non-reducible)

Computing uncertainty in classification models:

C 类别

$$H[p] = -\sum_{c=1}^{C} p_c \log p_c, \qquad p = (p_1, \dots, p_C), \sum_{c=1}^{C} p_c = 1$$

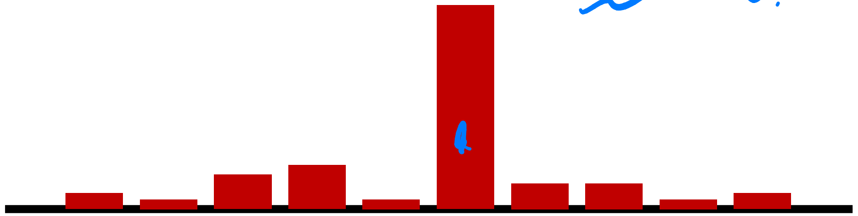
High entropy: $H[p] \to \log C$

$p_c = \frac{1}{C}$,

Low entropy: $H[p] \to 0$

$p_c = 1$

$\sim = 0$.



Uncertainty measures

Total uncertainty = epistemic uncertainty + aleatoric uncertainty

Due to lack of “knowledge”
(reducible when having more data)

$\boxed{y^*}$, $\boxed{\theta}$.

Due to inherent stochasticity in data
(non-reducible)

Computing uncertainty in classification models:

Recall for Bayesian predictive distribution:

(y^*, x^*).

$$p(y^* \mid x^*, D) = \int p(y^*|x^*, \theta) p(\theta|D) d\theta$$

$$H[y^* \mid x^*, D] = I[y^*; \theta \mid x^*, D] + E_{p(\theta \mid D)}[H[y^* \mid x^*, \theta]]$$

Total entropy of the predictive distribution

Mutual information between y^* and θ

Conditional entropy under posterior

数据 y^*, θ 相关性度.

Uncertainty measures

Total uncertainty = epistemic uncertainty + aleatoric uncertainty

Due to lack of "knowledge"
(reducible when having more data)

Due to inherent stochasticity in data
(non-reducible)

Computing uncertainty in classification models:

Recall for Bayesian predictive distribution with approximation:

$$p(y^* \mid x^*, D) \approx \frac{1}{K} \sum_{k=1}^{K} p(y^* \mid x^*, \theta_k), \quad \theta_k \sim q(\theta)$$

Total entropy (for total uncertainty):

$$H[y^* \mid x^*, D] \approx H[\frac{1}{K} \sum_{k=1}^{K} p(y^* \mid x^*, \theta_k)]$$

Uncertainty measures

Total uncertainty = epistemic uncertainty + aleatoric uncertainty

Due to lack of “knowledge”
(reducible when having more data)

Due to inherent stochasticity in data
(non-reducible)

Computing uncertainty in classification models:

Recall for Bayesian predictive distribution with approximation:

$$p(y^* \mid x^*, D) \approx \frac{1}{K} \sum_{k=1}^{K} p(y^* \mid x^*, \theta_k), \quad \theta_k \sim q(\theta)$$

Conditional entropy (for aleatoric uncertainty): 固有 uncertainty.

$$E_{p(\theta \mid D)}[H[y^* \mid x^*, \theta]] \approx \frac{1}{K} \sum_{k=1}^{K} H[p(y^* \mid x^*, \theta_k)] \quad \theta_k \sim p(\theta|D)$$

Uncertainty measures

Total uncertainty = epistemic uncertainty + aleatoric uncertainty

Due to lack of “knowledge”
(reducible when having more data)

Due to inherent stochasticity in data
(non-reducible)

Computing uncertainty in classification models:
Recall for Bayesian predictive distribution with approximation:

$$p(y^* \mid x^*, D) \approx \frac{1}{K} \sum_{k=1}^{K} p(y^* \mid x^*, \theta_k), \quad \theta_k \sim q(\theta)$$

Mutual information (for epistemic uncertainty):

$$I[y^*; \theta \mid x^*, D] \approx \underbrace{H[\frac{1}{K} \sum_{k=1}^{K} p(y^* \mid x^*, \theta_k)]}_{\text{Total}} - \underbrace{\frac{1}{K} \sum_{k=1}^{K} H[p(y^* \mid x^*, \theta_k)]}_{\text{固有}}$$

Uncertainty measures

Total uncertainty = epistemic uncertainty + aleatoric uncertainty

Due to lack of “knowledge”
(reducible when having more data)

Due to inherent stochasticity in data
(non-reducible)

Computing uncertainty in classification models:
Recall for Bayesian predictive distribution with approximation:

$$p(y^* \mid x^*, D) \approx \frac{1}{K} \sum_{k=1}^{K} p(y^* \mid x^*, \theta_k), \quad \theta_k \sim q(\theta)$$

多观测了1次
x^*, y^*

$p(\theta)$ 认知改变.

Mutual information (for epistemic uncertainty) if you can do exact inference:

$$I[y^*; \theta \mid x^*, D] = E_{p(y^* \mid x^*, D)}[KL[p(\theta \mid D, x^*, y^*) \,\|\, p(\theta \mid D)]]$$

“What the model thinks the posterior is going to change if we add new observation at location x^*”

$\boxed{}$ $\Rightarrow y^*$, uncertainty. OOD/noisy

input

Case study 2: Detecting adversarial examples

- What you'll do for the case study part of the tutorial:
  - Implement the uncertainty measures
    - Total entropy, conditional entropy, and mutual info
  - Run adversarial attacks on various trained networks
  - See how diversity helps in detecting adversarial examples
    - Detection by thresholding the uncertainty measures
    - We consider best TPR with FPR $\leq$ 5%

Ensemble BNNs

$0 \quad 0. \quad p(y^* | x^*, \theta^I)$.

- Define q distribution as mixture of mean-field Gaussian:

$$q(\theta) = \frac{1}{S}\sum_{s=1}^{S} q(\theta|s), \quad q(\theta|s) = N(\theta; \mu_s, diag(\sigma_s^2))$$

- Objective is still a valid lower-bound to $\log p(D)$:

$$L = \frac{1}{S}\sum_{s=1}^{S} ELBO[q(\theta|s)]\,, ELBO[q(\theta|s)] = E_{q(\theta|s)}[\log p(D \mid \theta)] - KL[q(\theta|s) \,||\, p(\theta)]$$

- The parameters of $q(\theta|s)$ for different s are independent
  $\Rightarrow$ train S number of MFVI-BNNs independently

Ensemble BNNs are winning solutions of NeurIPS 2021 Approximate Inference in BDL Competition

VAE

bayes w(μ,σ).

latent variable, bayes.

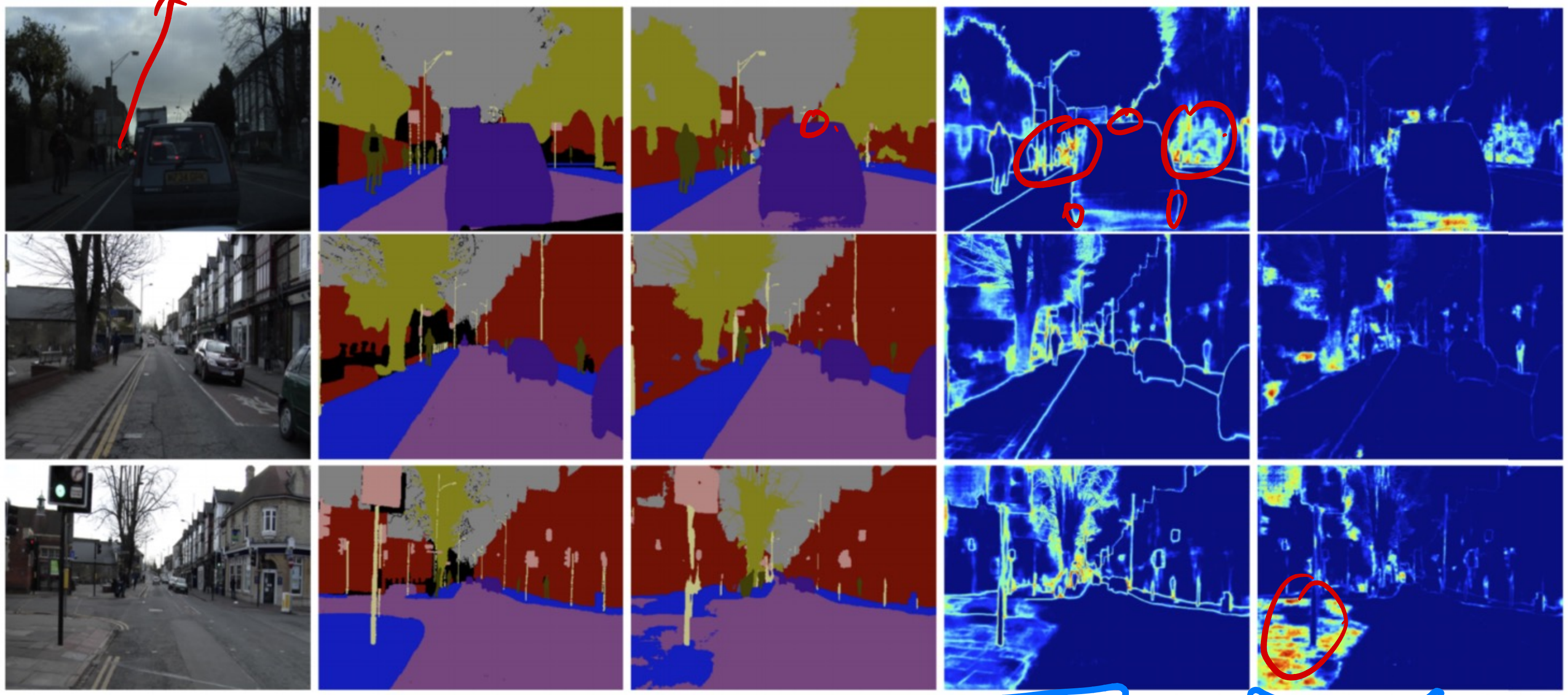
prior. → VI

Part IV: Advances & Future Works

- **Various applications**
- **Overview of recent progresses**
- **Future directions**

Applications of BNNs: Image Segmentation

模糊的/noisy 高级.

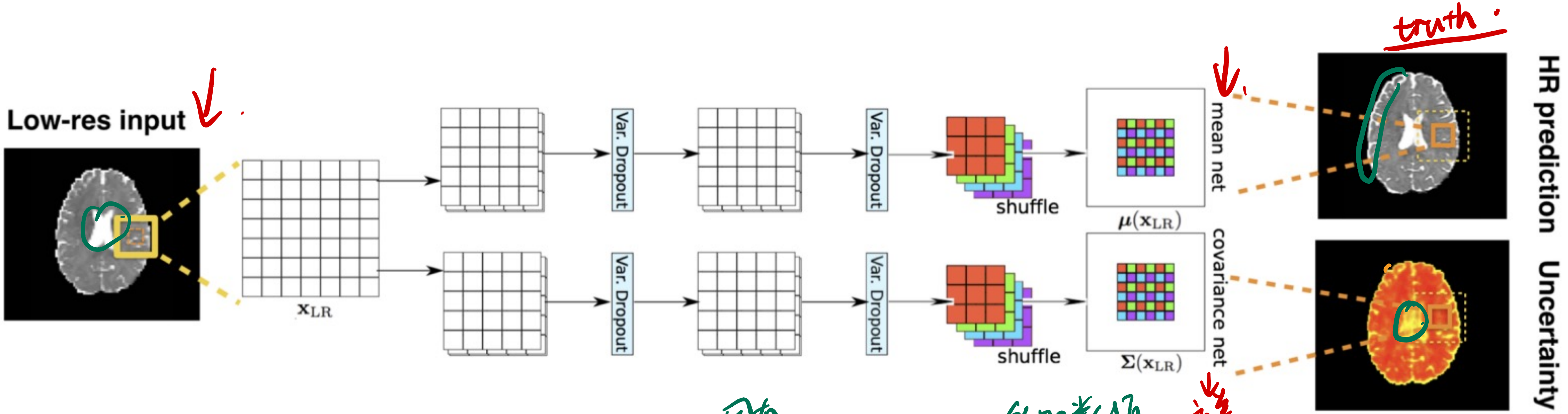


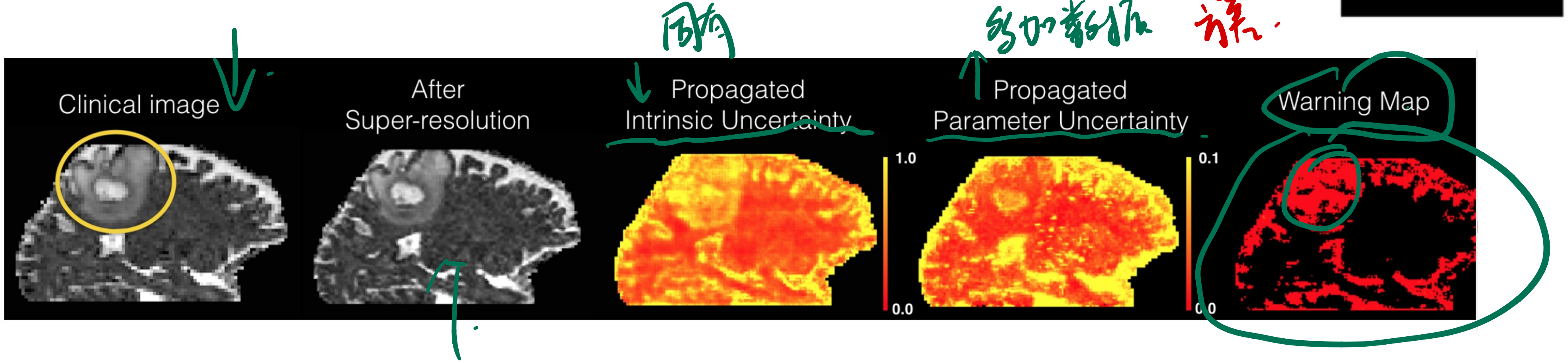
(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty 数据

因

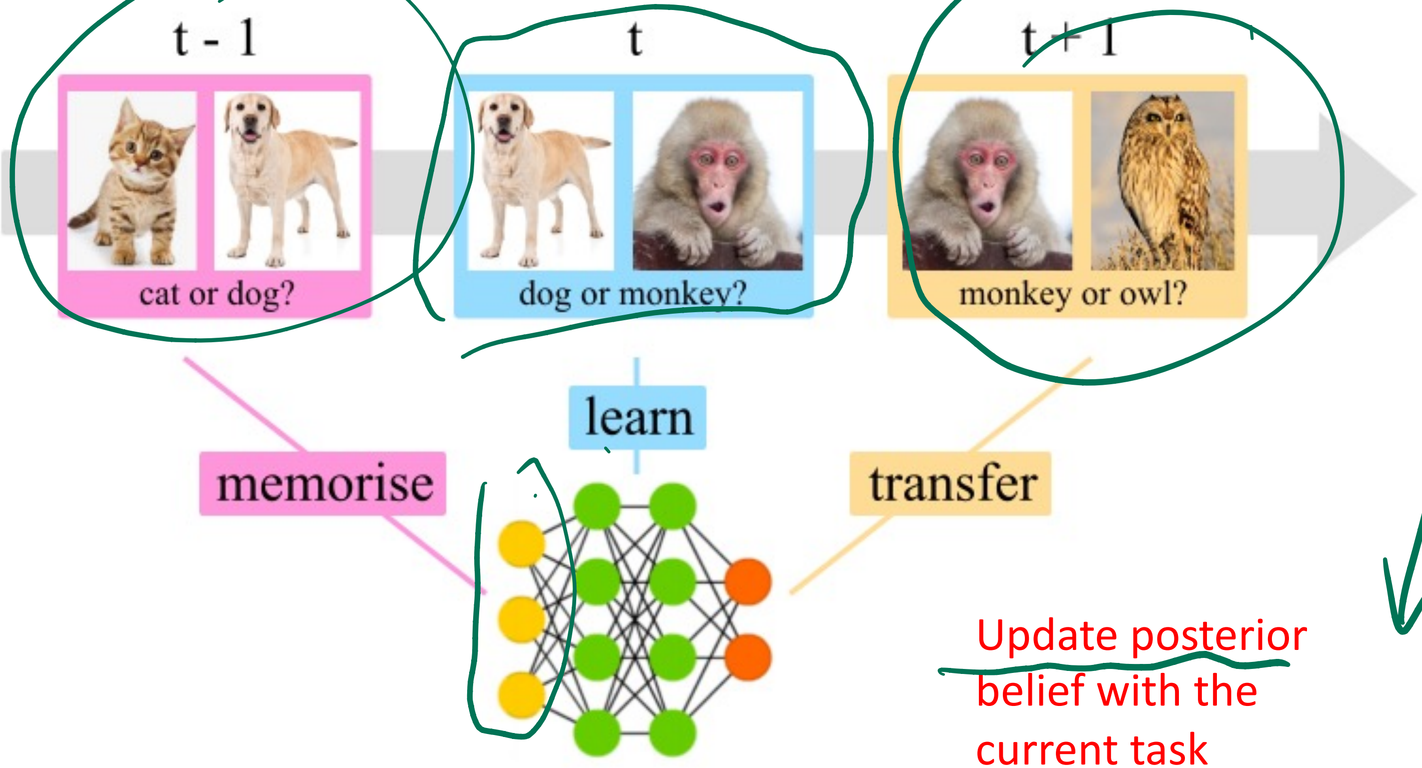
Kendall and Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NeurIPS 2017

Applications of BNNs: Super Resolution





Applications of BNNs: Continual Learning



θ

$q(\theta)$

$p(\theta \mid D)$

Update posterior belief with the current task

Posterior from the previous tasks as prior for the current task

$$L_{VCL}^{t}(q_t(\theta)) = E_{q_{t(\theta)}}[\log p(D_t \mid \theta)] - KL[\boxed{q_t(\theta)} \,||\, \boxed{q_{t-1}(\theta)}]$$

$q(\theta)$

There are more! (See Miguel's lecture)

Nguyen et al. Variational Continual Learning. ICLR 2018
Pan et al. Continual Deep Learning by Functional Regularisation of Memorable Past. NeurIPS 2020

Recent Progress in BNNs: Inference

SGD: $\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \widetilde{U}(\theta_t)$

SGLD: $\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \widetilde{U}(\theta_t) + \sqrt{2\eta}\epsilon, \qquad \epsilon \sim N(0, I)$

Stochastic gradient MCMC

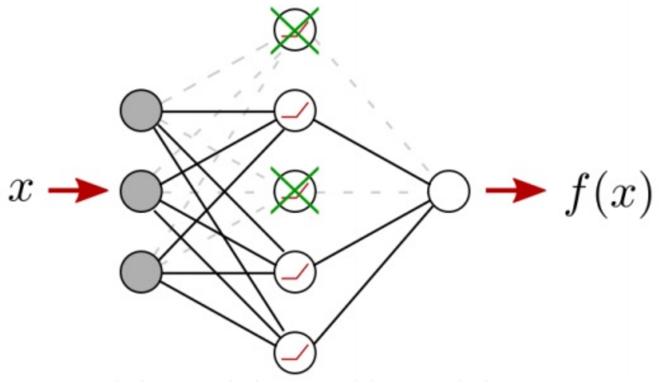
Li et al. Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. AAAI 2016
Zhang et al. Cyclical Stochastic Gradient MCMC for Bayesian Deep Learning. ICLR 2020

Recent Progress in BNNs: Inference

SGD: $\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \widetilde{U}(\theta_t)$

SGLD: $\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \widetilde{U}(\theta_t) + \sqrt{2\eta}\epsilon, \qquad \epsilon \sim N(0, I)$

Stochastic gradient MCMC



Monte Carlo dropout

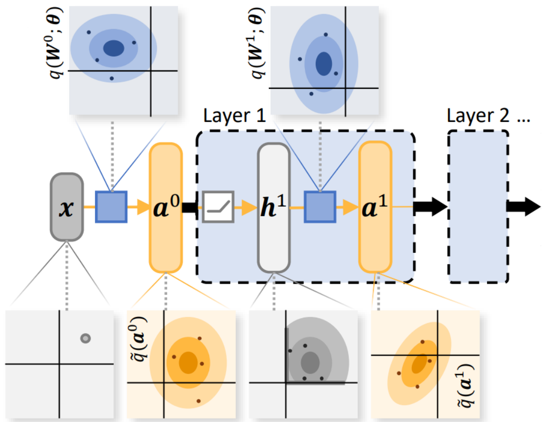
Gal and Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. ICML 2016

Recent Progress in BNNs: Inference

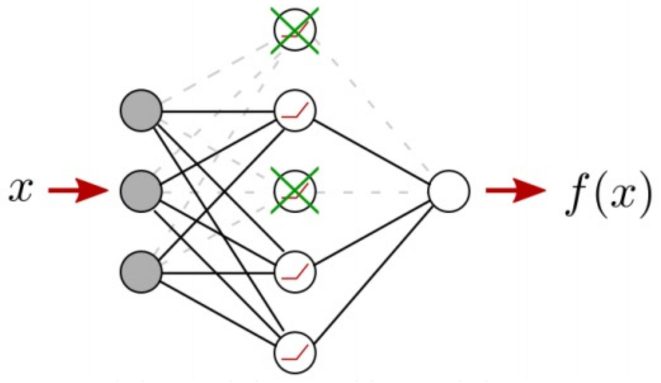
SGD: $\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \widetilde{U}(\theta_t)$

SGLD: $\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \widetilde{U}(\theta_t) + \sqrt{2\eta}\epsilon, \quad \epsilon \sim N(0, I)$

Stochastic gradient MCMC



Deterministic approximations



Monte Carlo dropout

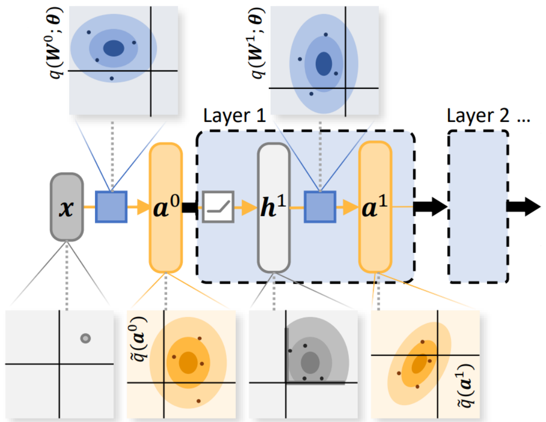
Hernandez-Lobato and Adams. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. ICML 2015
Wu et al. Deterministic Variational Inference for Robust Bayesian Neural Networks. ICLR 2019

Recent Progress in BNNs: Inference

SGD: $\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \tilde{U}(\theta_t)$ ← ① bottk ② stable

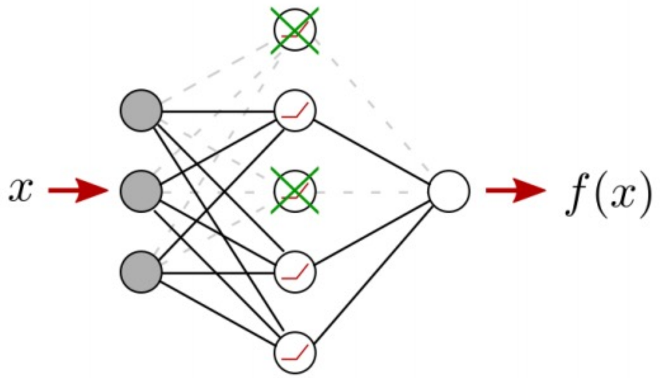
SGLD: $\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \tilde{U}(\theta_t) + \sqrt{2\eta}\epsilon,\quad \epsilon \sim N(0, I)$

Stochastic gradient MCMC



Deterministic approximations

VI



Monte Carlo dropout

$f_l \sim$ prior.

$\boxed{f_l}\ \sigma(W_l x + b_l)$

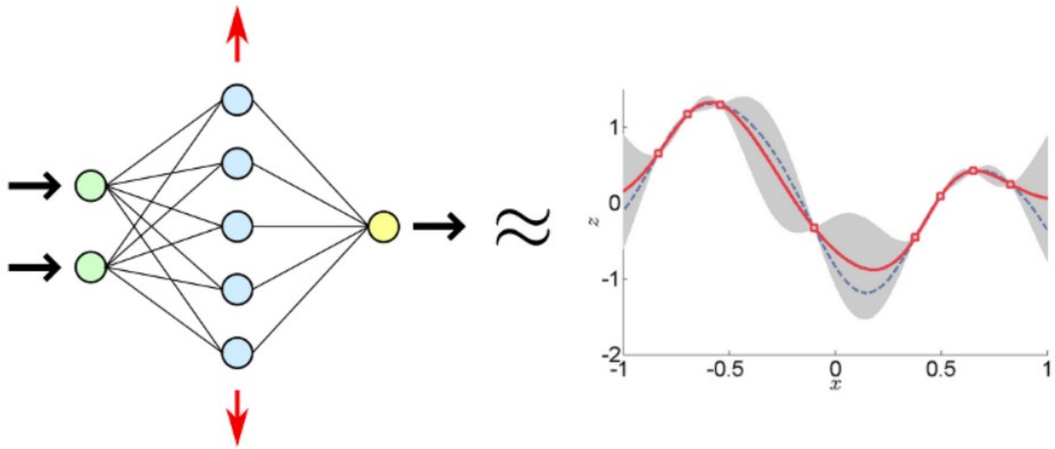
Function space approximate inference

There are more! (See Miguel's lecture)

Ma et al. Variational Implicit Processes. ICML 2019

Sun et al. Functional Variational Bayesian Neural Networks. ICLR 2019

Recent Progress in BNNs: Theory

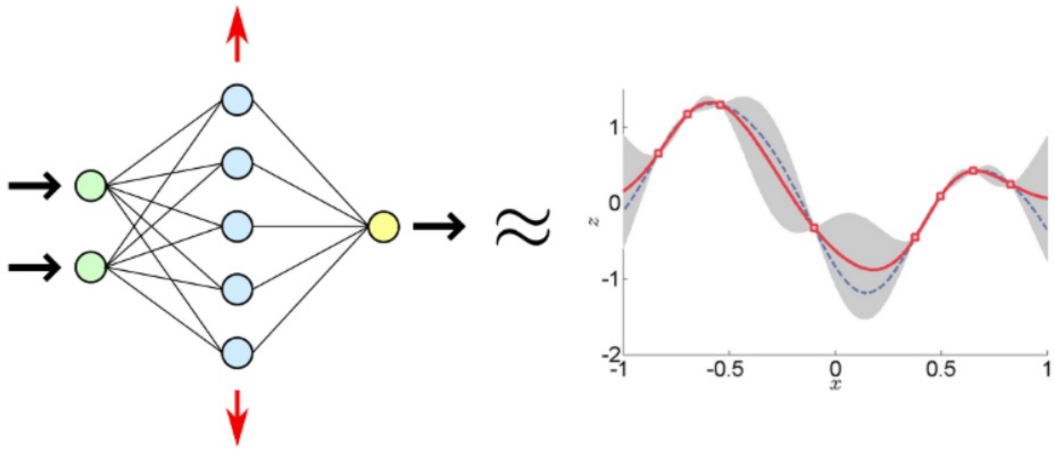


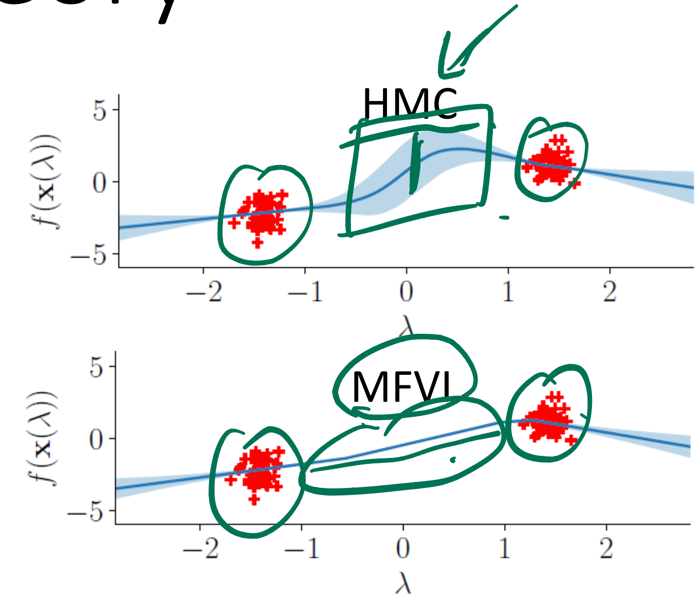
Connections to GPs:

- BNN with very wide hidden layers ≈ Gaussian process
- Width limit convergence: in both prior (Neal's result) and posterior

Neal. Bayesian Learning for Neural Networks. PhD Thesis, 1996
Matthews et al. Gaussian Process Behaviour in Wide Deep Neural Networks. ICLR 2018
Lee et al. Deep Neural Networks as Gaussian Processes. ICLR 2018
Hron et al. Exact posterior distributions of wide Bayesian neural networks. 2020

Recent Progress in BNNs: Theory





Connections to GPs:
- BNN with very wide hidden layers $\approx$ Gaussian process
- Width limit convergence: in both prior (Neal's result) and posterior

Approx. vs exact inference: underfit.
- Theoretical limitation of MFVI in shallow BNNs with ReLU activations
- Empirically deep BNNs with MVFI still fails in certain cases

Foong et al. On the Expressiveness of Approximate Inference in Bayesian Neural Networks. NeurIPS 2020
Farquhar et al. Liberty or Depth: Deep Bayesian Neural Nets Do Not Need Complex Weight Posterior Approximations. NeurIPS 2020
Coker et al. Wide Mean-Field Bayesian Neural Networks Ignore the Data. AISTATS 2022

Future directions

- Understanding BNN behaviour: $\leftrightarrow$ function space
  - How would $q(f)$ behave given a particular form of $q(W)$?
  - Is weight-space objective appropriate for MFVI? $q(w)$. MFVI
  - We don't understand very well the optimisation properties of VI-BNN VI
- Computational complexity overhead: worth it?
  - How can we make the approximate posterior more efficient in both time and space complexities? 节约.
- Priors for BNNs
  - "Default" Gaussian prior $N(W; 0, \sigma^2 I)$: the right prior? f, w.
  - How to think about priors in function space?
- Applications
  - Improve for applications that require good uncertainty estimates

Ritter et al. Sparse Uncertainty Representation in Deep Learning with Inducing Weights. NeurIPS 2021
Fortuin. Priors in Bayesian deep learning: A review. International Statistical Review, 2022.

Thank You!

Questions? Ask NOW or email:
yingzhen.li@imperial.ac.uk

Example answers of the tutorial demos:
Regression: https://bit.ly/39eZHit
Classification: https://bit.ly/3QikcLO