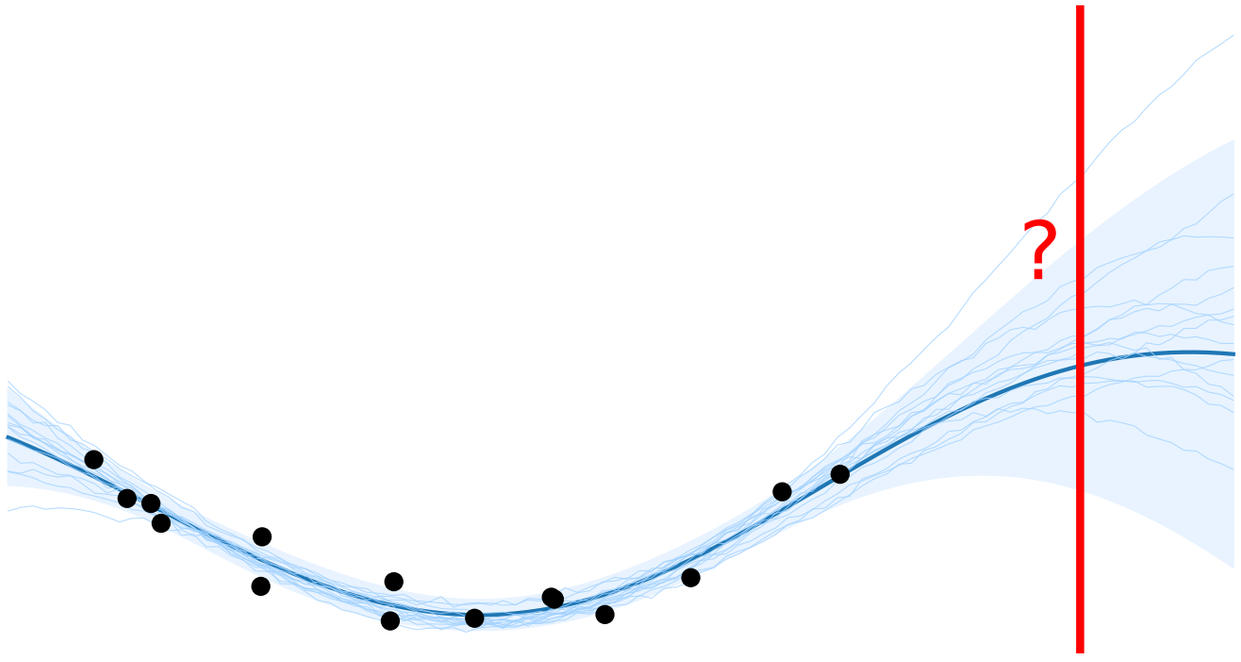
Probabilistic ML - Part II

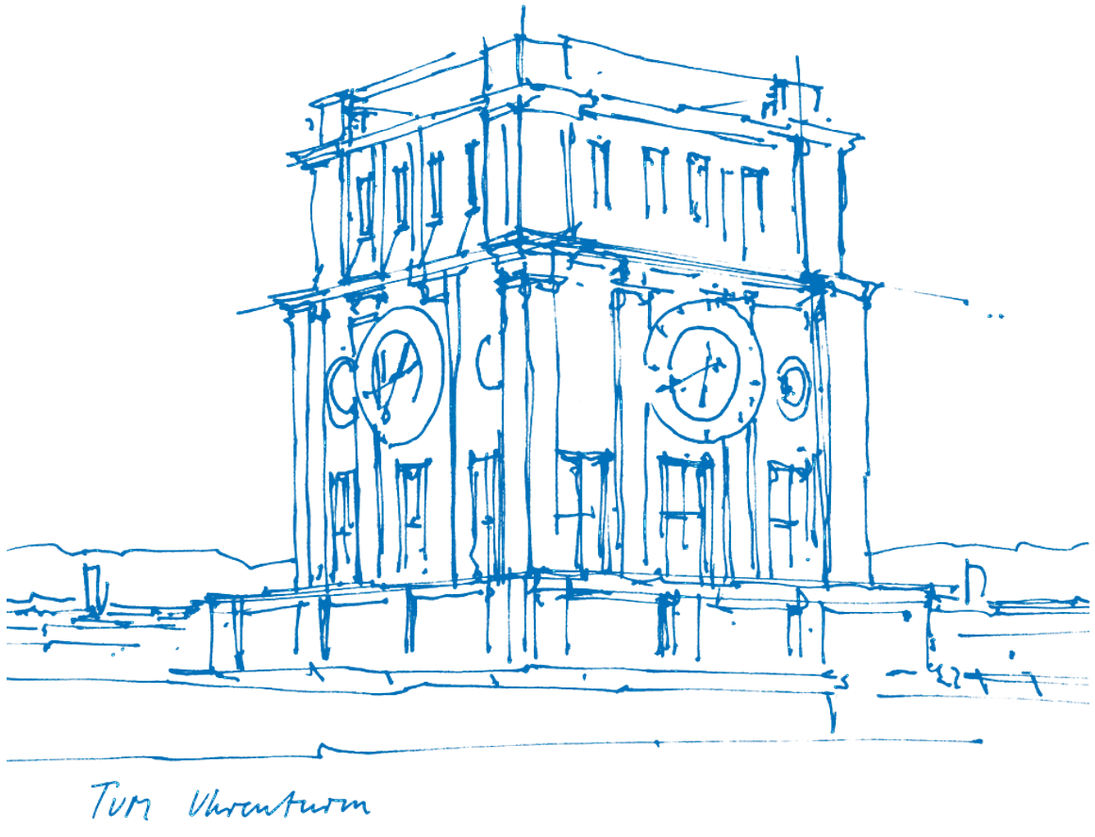


Ai in Medicine Lecture Series

Chair for Ai in Medicine
Faculty of Informatics & Medicine
Technical University Munich

Moritz Knolle

15.12.2022

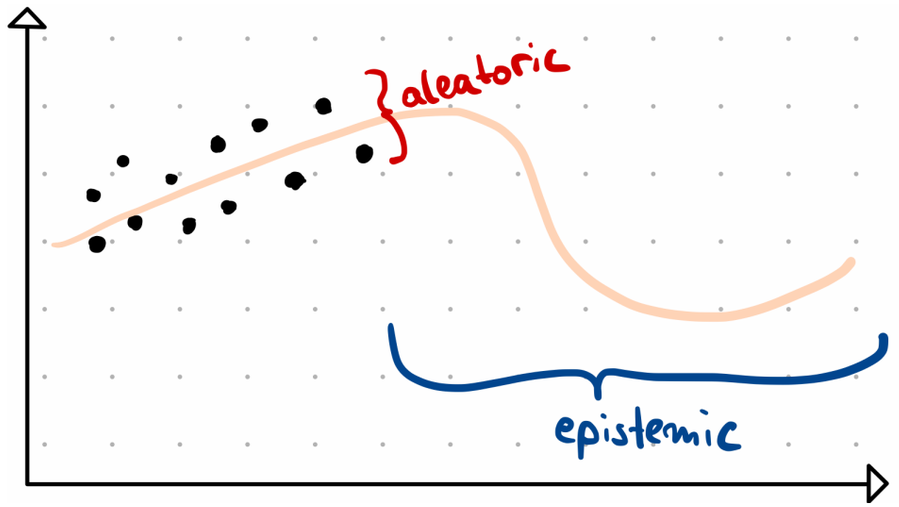


Last time:

- *Pdf*, *cdf*, random variables
- Joint & conditional probability
- Probabilistic process > Bayes Theorem

$$p(H \mid D) = \frac{p(D \mid H)\, p(H)}{p(D)}$$

- Epistemic vs. Aleatoric uncertainty
- Sampling-based inference



Outline

I. Overview approximate inference

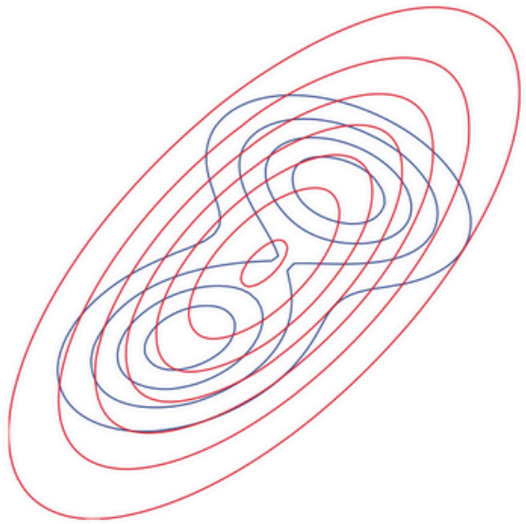
II. From rejection sampling to Markov Chain Monte Carlo

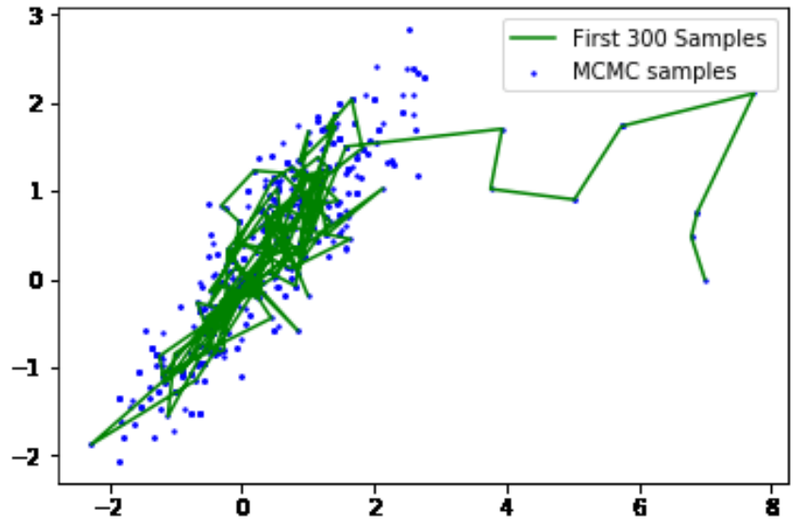
III. Variational methods

IV. Measuring the quality of probabilistic predictions

V. Practical methods for uncertainty quantification in deep learning

Today: Approximate Inference





Variational Methods

Approximation as Optimisation

Monte Carlo Methods

Sampling based

Recap: Sampling-based approximate inference

cheap to evaluate

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}$$

expensive to evaluate

$$\int p(D \mid \theta)\, p(\theta) d\theta$$

Monte Carlo sampling:

$$\int p(D \mid \theta) p(\theta) d\theta = \frac{1}{N} \sum_{i}^{N} p(D \mid \theta_i)$$

$$\theta_i \sim p(\theta)$$

Rejection sampling in disguise

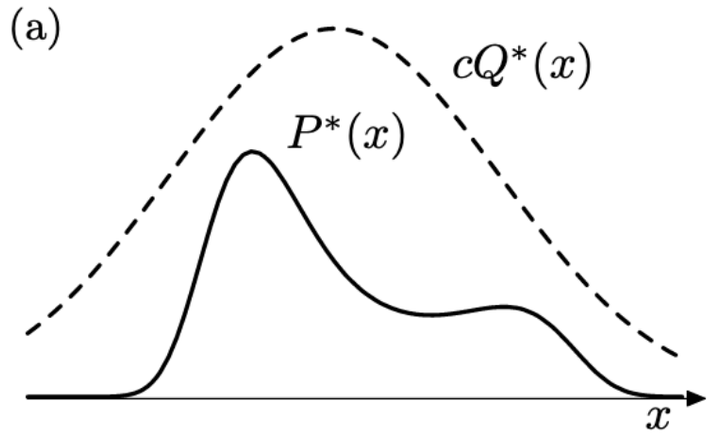
- want to evaluate $P(x)$
- **BUT** can only evaluate an unnormalised version $P^*(x)$

cheap to evaluate

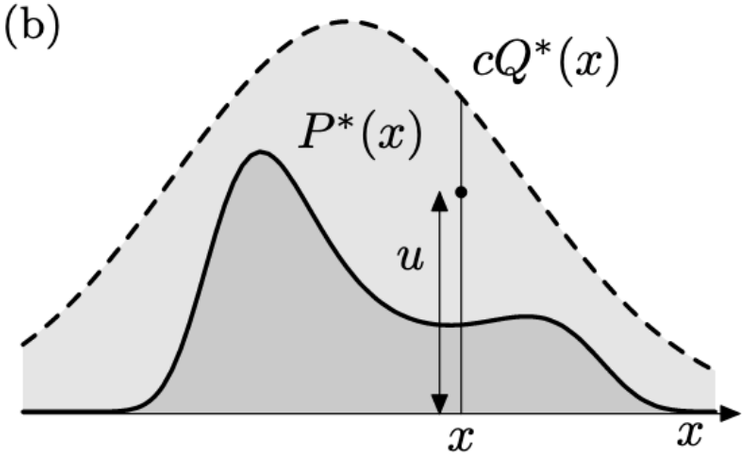
$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)} \quad \Longleftrightarrow \quad P(x) = \frac{P^*(x)}{Z}$$

expensive to evaluate

$$\int p(D \mid \theta)\, p(\theta) d\theta$$



(b)



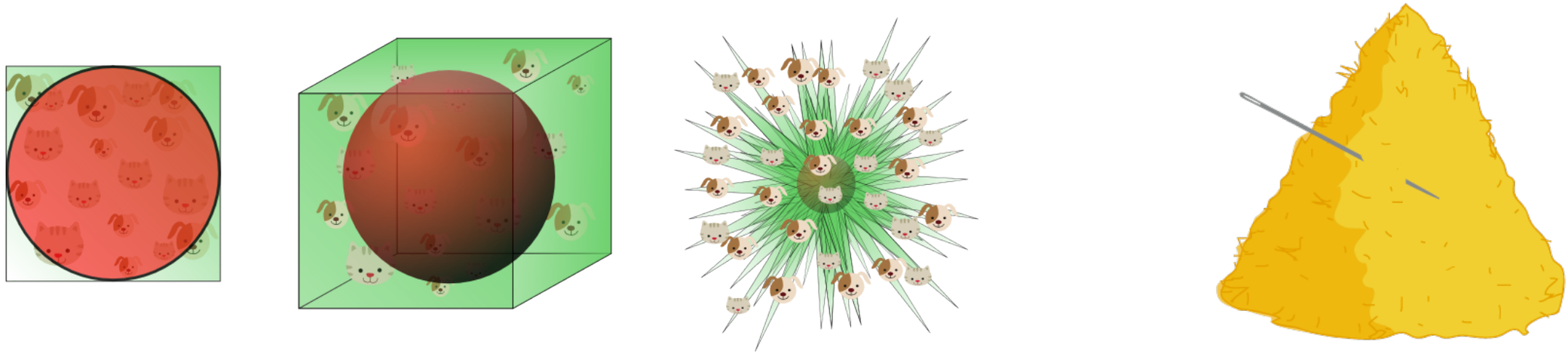
Rejection Sampling:

1. Construct proposal density $Q(x)$ (un-normalised version $Q^*(x)$)
2. Generate a sample x from $Q(x)$
3. Sample uniformly from
$$u \sim [0, cQ^*(x)]$$
4. If $u > P^*(x)$ discard x otherwise add x to list of samples $[\ldots, x^{(r)}]$

Problems: Rejection sampling

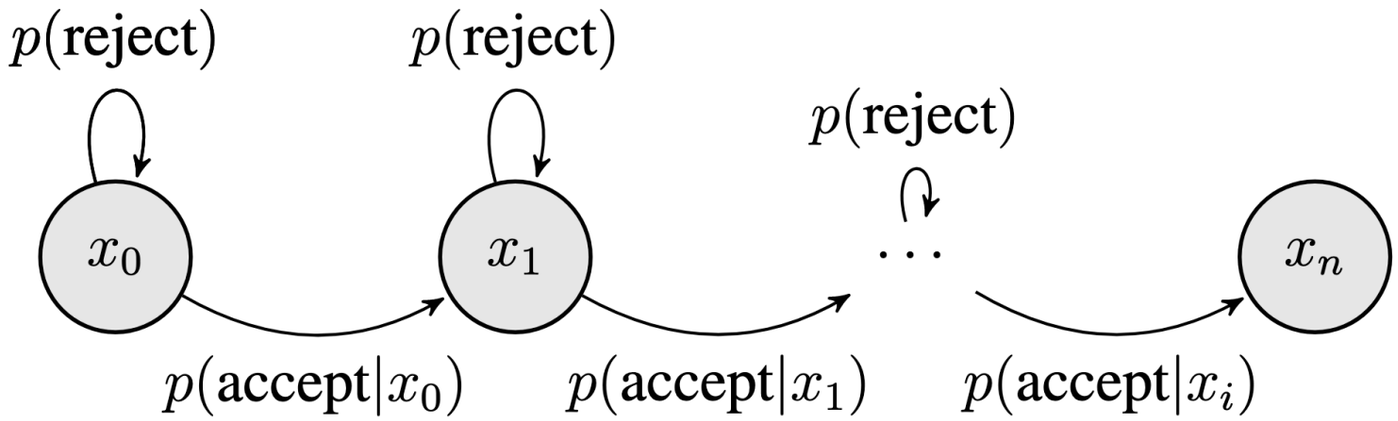
- $Q(x)$ needs to be similar to $P(x)$
- Rejection rate r is a function of dimensionality d:
  - As $d \to \infty$; $r \to 0$

- Samples are uncorrelated
  - repeatedly finding needle in haystack



Can we improve rejection sampling?

- What if we could create samples that are correlated?

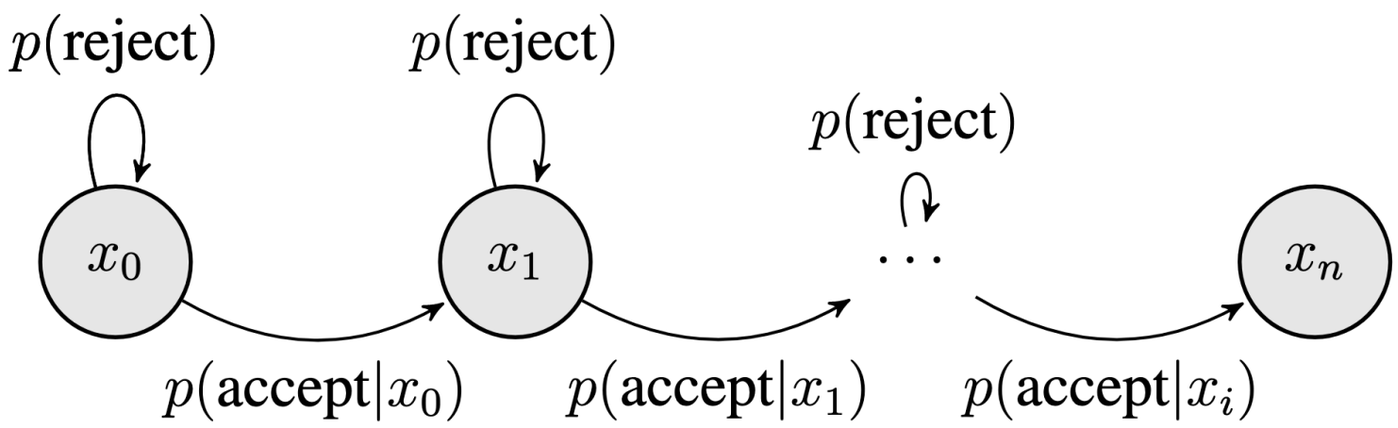


Can we improve rejection sampling?

- For every x_i, we need to model the transition probability $p(x_{i+1} \mid x_i)$
- next state only depends on current state!

Transition operator:

- $a = \frac{P^*(x')/Z}{P^*(x^{(t)})/Z} = \frac{P^*(x')}{P^*(x^{(t)})}$;
- If $a > 1$ likely to accept x', otherwise likely to reject x'



Markov Chain (Monte Carlo)

Metropolis-Hastings Algorithm

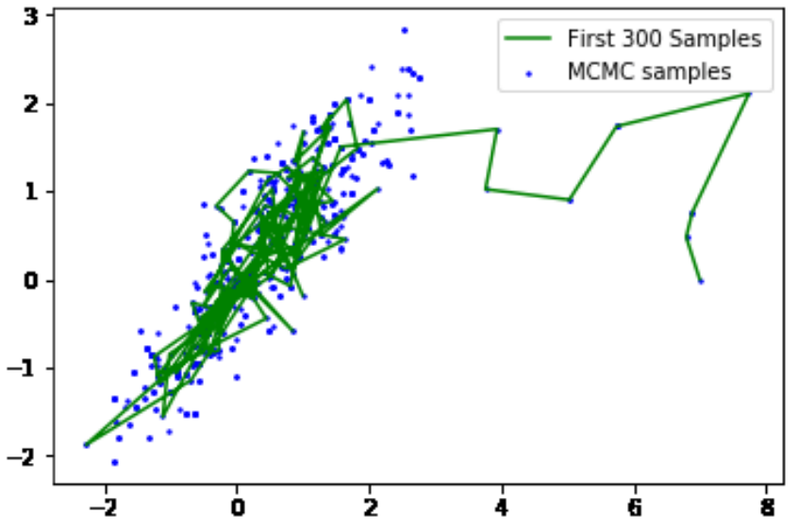
- A Markov Chain Monte Carlo (MCMC) method

Metropolis Hastings:

1. Sample $x' \sim Q(x'|x^{(t)})$
2. $a = \dfrac{P^*(x')}{P^*(x^{(t)})}, u \sim U(0,1)$
3. If $u \leq a$: **accept**

   then $x_{t+1} = x'$
4. Else: **reject**

   $x_{t+1} = x_t$



Cool Demo

Can we come up with a similarity measure to see how close we are to the exact solution?

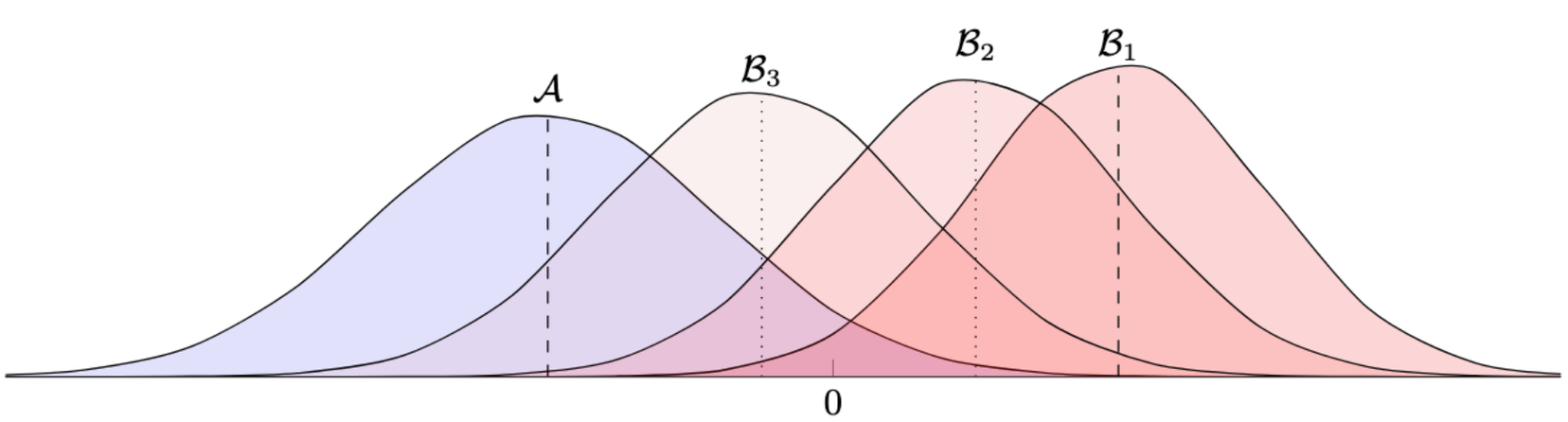
(To build an optimisation based inference approach)

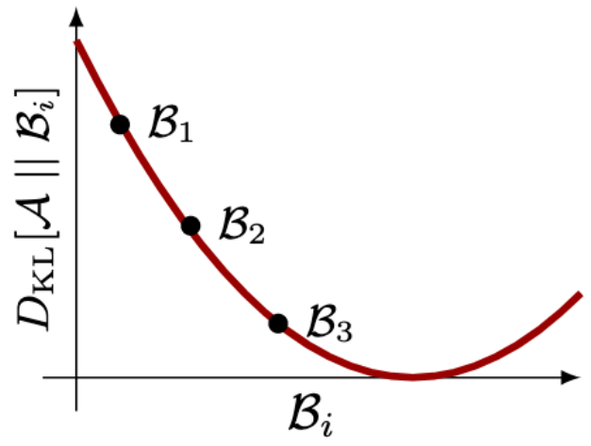
The KL divergence - a measure of distributional similarity

- similarity between two probability distributions A and B

$$D_{\mathrm{KL}}[A\,||\,B] = \mathbb{E}_{x\sim A}[\log \frac{A(x)}{B(x)}]$$

- $D_{\mathrm{KL}}[A\,||\,B] = 0$ only if A and B are identical (under x)
- Nice error-coding interpretation





Variational Inference - Approximation through **Optimisation**:

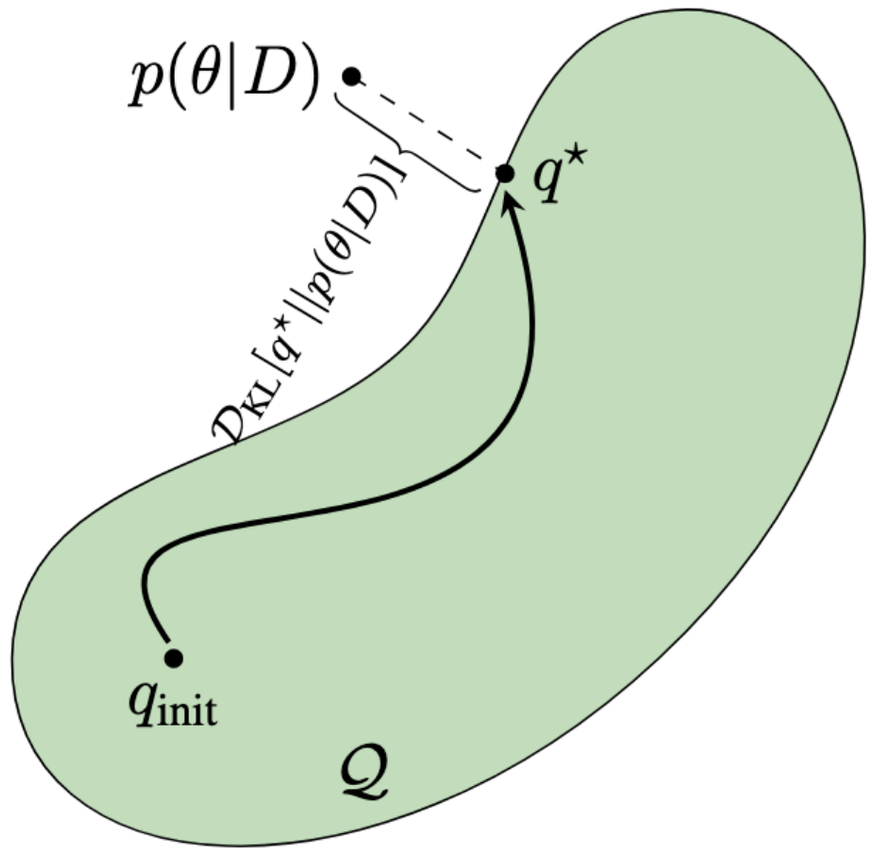
1. Define variational distribution family $\mathcal{Q}$, parameterised by v
2. Measure quality of approximation through:

$$M(q_v^{(i)}) = D_{\mathrm{KL}}[q_v^{(i)} \,||\, p(\theta|D)]$$

3. Update v

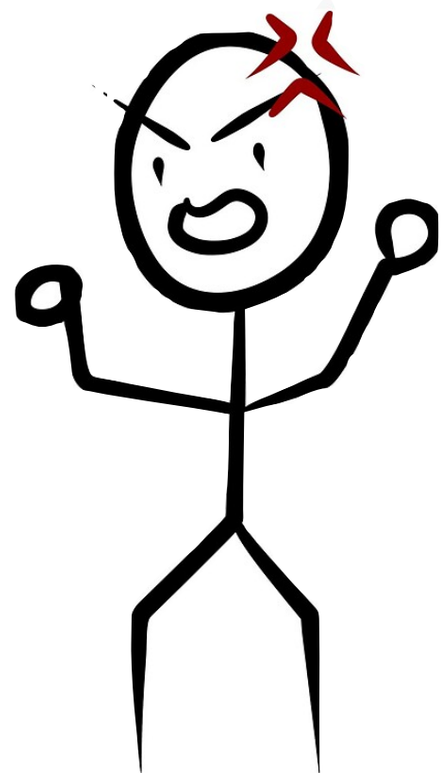
$$v^{(i+1)} = v^{(i)} - \nabla_v M(q_v^{(i)})$$

4. Repeat 2. & 3. until convergence



Can you spot the problem with this approach?

$D_{\mathrm{KL}}[q_v^{(i)} \,||\, p(\theta|D)]$ involves the expensive posterior $p(\theta|D)$ again



The ELBO Trick

- Maximise lower bound on the objective instead (ELBO, $\mathscr{L}$)

$$\mathscr{L}(q_v) = \mathbb{E}_q[\log p(D \mid \theta)] - \underbrace{D_{KL}[q(\theta) \mid\mid p(\theta)]}$$

KL-divergence between q and prior on θ

$$\arg\max_v \mathscr{L}(q_v)) = \arg\min_v D_{KL}[q_v \mid\mid p(\theta \mid D)]$$

- Maximising $\mathscr{L}$, monotonically decreases our objective!

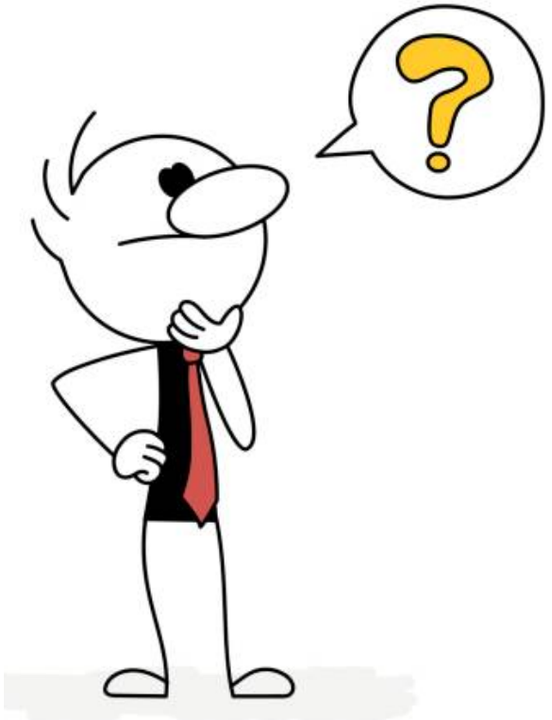
ELBO Derivation: https://chrisorm.github.io/VI-ELBO.html
Blei et al. (2018), Variational Inference: A Review for Statisticians

Quiz: MCMC vs. Variational Inference

	Variational Inference	MCMC Methods
• Approximation Quality Measure	✓	✗
• Recover exact posterior	(✓)	✓
• Fast?	✓	✗

We covered how to approximate posteriors ….

But how to measure the quality of a predictive distribution?



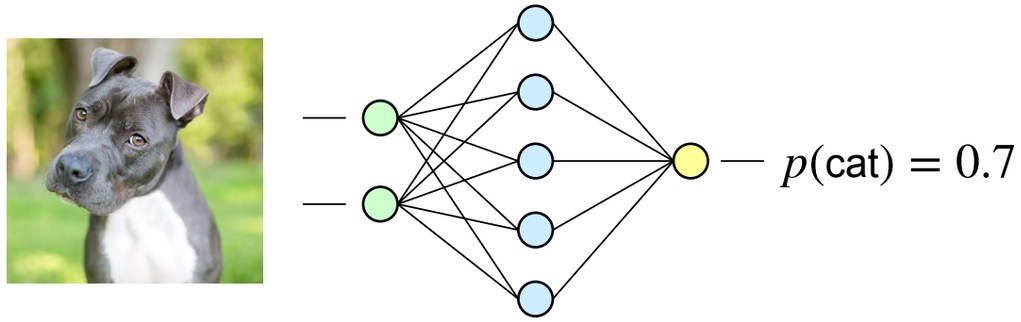
Assessing the quality of a probabilistic prediction?

Calibration

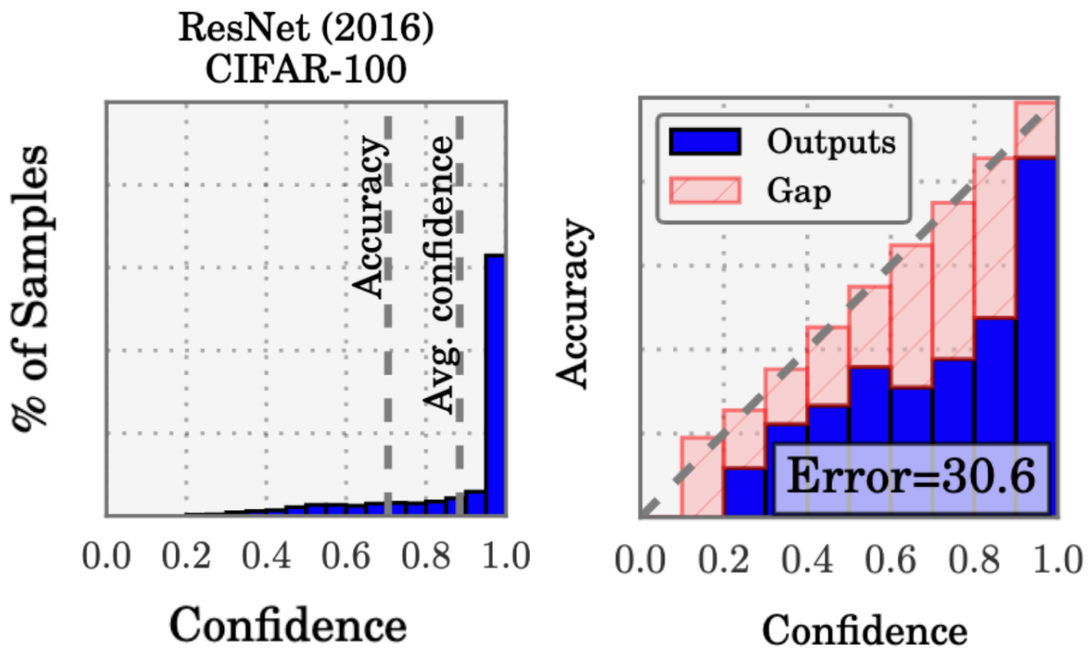
$$P(y \mid \hat{p}) = p \;\forall p \in [0,1]$$

Intuitively (model confidence $\simeq$ probability)

Prediction with confidence 20% should be correct 20% of the time.



- Softmax confidence for modern neural networks tend do be ***mis-calibrated***
- Expected calibration error (ECE)
- ***Calibrated*** $\neq$ ***accurate*** prediction



Guo et al. (2017), On the calibration of modern neural networks

Metrics for probabilistic predictions (**scoring rules**)

- Brier Score (mean squared error)

$$\frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$$

- Log likelihood:

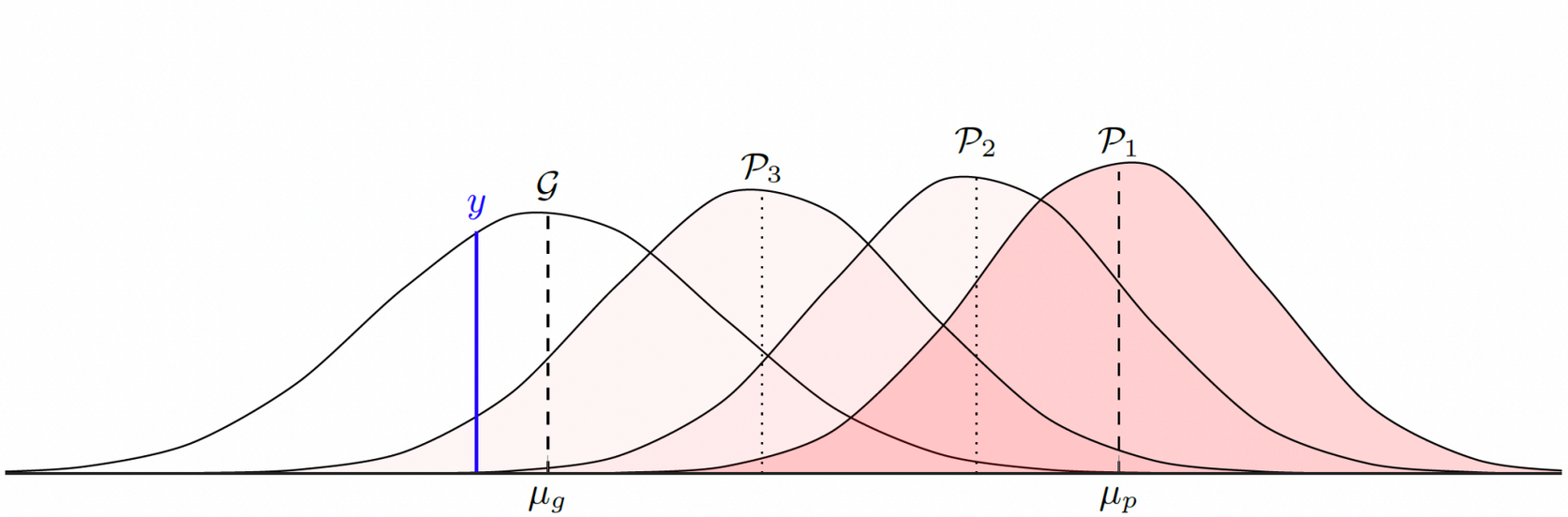
$$\frac{1}{N}\sum_{i=1}^{N}\log p_\theta(y_i)$$

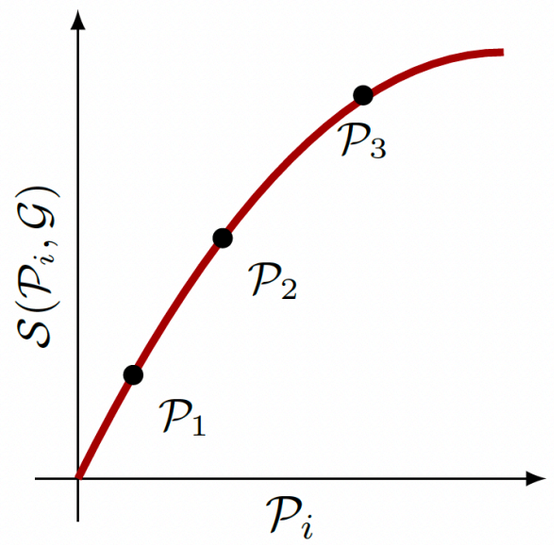
- Also called ***scoring rules***

Proper scoring rule

A **scoring rule** $\mathcal{S}(P, G)$ for a predictive distribution $P = (p_1, \ldots, p_k)$ and underlying ground truth distribution G is called **proper** iff:

$$\mathcal{S}(P, P) \leq \mathcal{S}(P, G) \quad \forall P, G$$

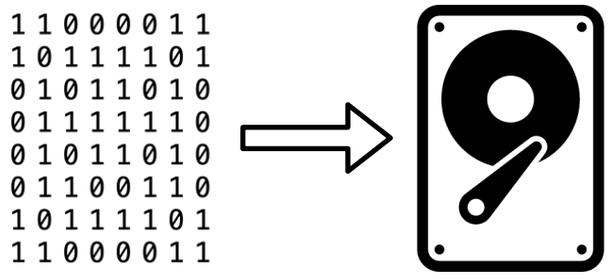




Scoring rules and decision making

- Close connection to *decision* and *information theory*

- Dual form (see Gneiting & Raftery):
  - **Primal**: scoring function maximisation
  - **Dual**: utility maximisation



$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

- **Log likelihood**, direct interpretation as growth rate when betting on mutually exclusive outcomes (Cover & Thomas)

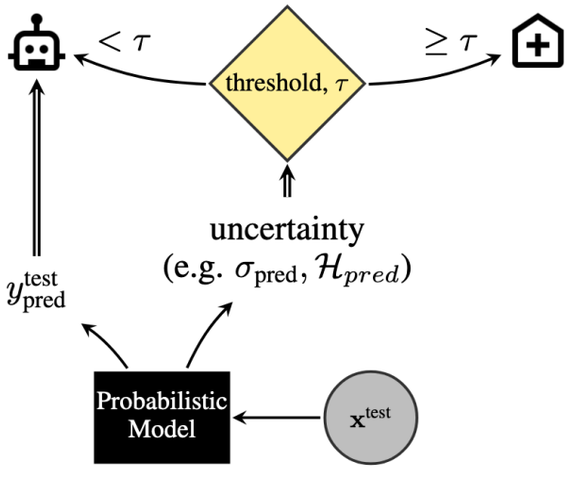
Gneiting & Raftery (2014), Strictly proper scoring rules, Prediction and Estimation

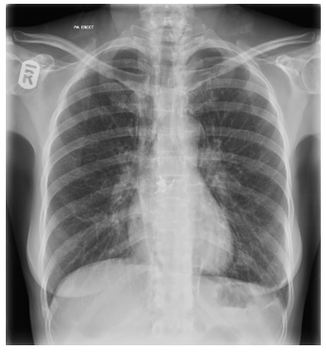
Chapter 6, Cover & Thomas (2012), Elements of Information Theory

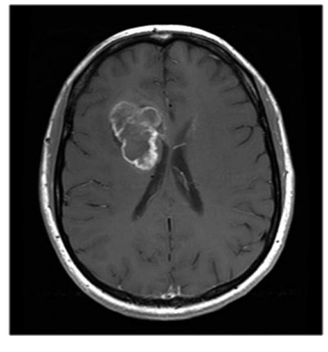
Now onto some practical uncertainty methods for Ai in Medicine

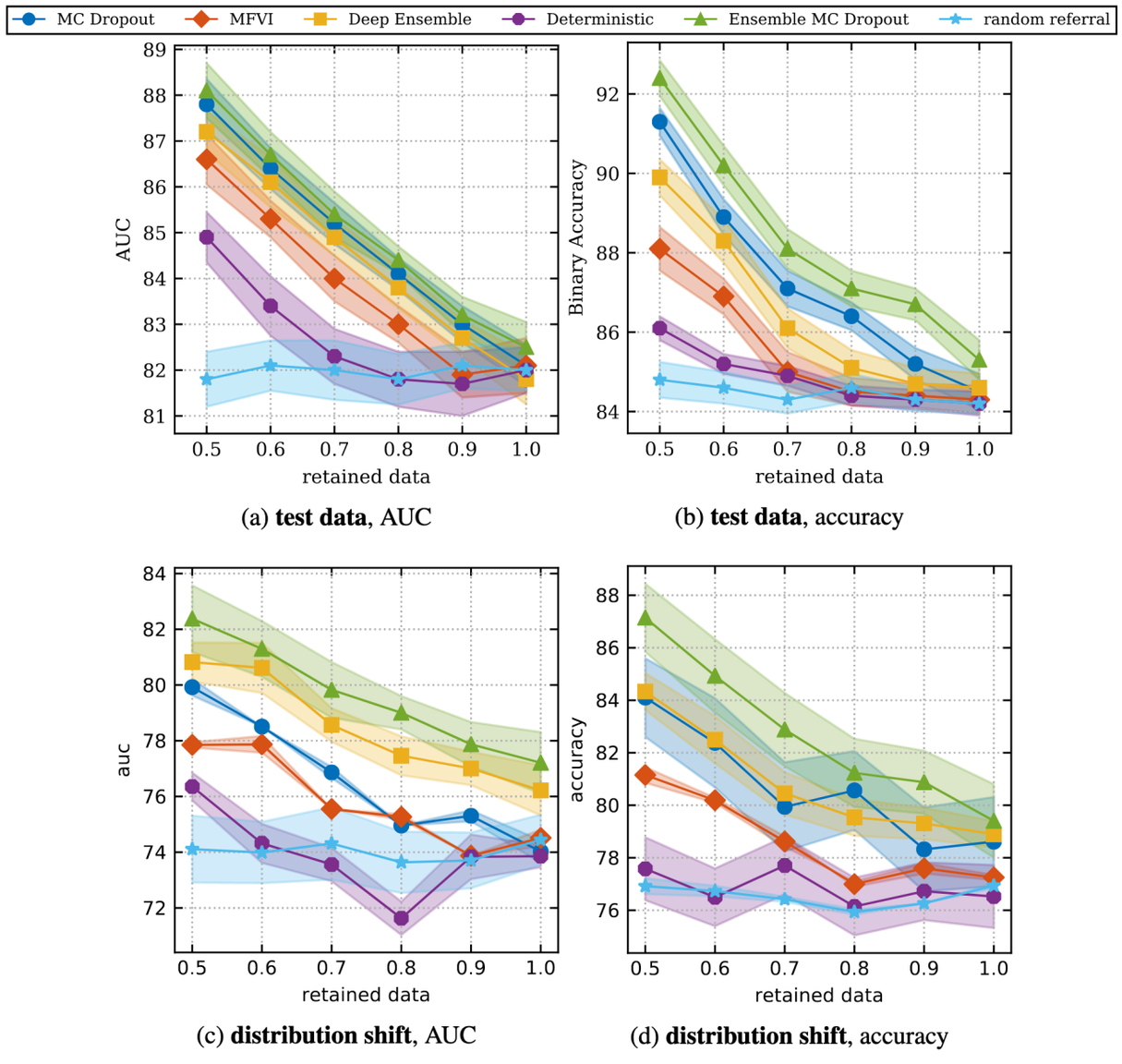
Do VI and MCMC scale to the tasks we care about?

- Sadly not! Computationally prohibitive with large NNs
- Mean field VI gives sub-optimal posterior approximation
- In practice: simply use ensembles of deep networks









(a) **test data**, AUC

(b) **test data**, accuracy

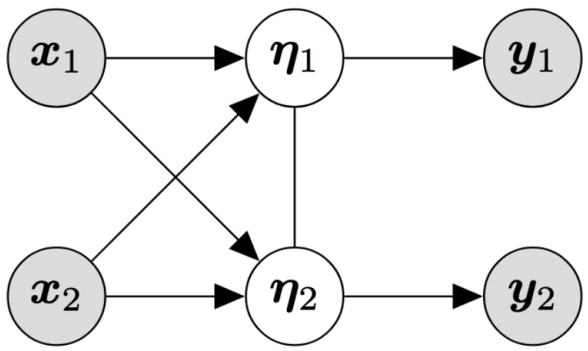
(c) **distribution shift**, AUC

(d) **distribution shift**, accuracy

Filos et al. (2019), A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks

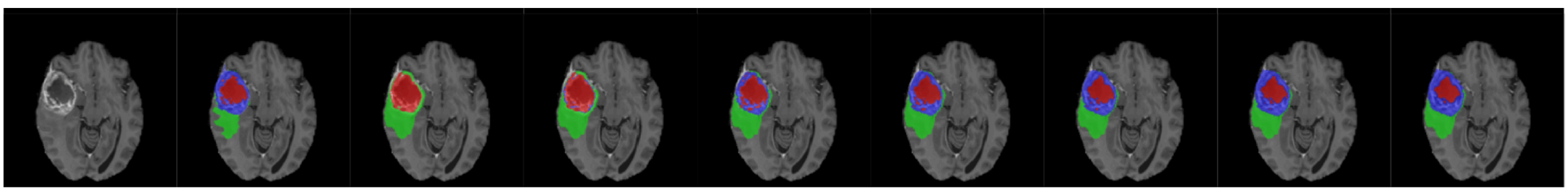
Stochastic Segmentation Networks (SSNs)

- Segmentation targets often ambiguous and can vary between ground truth raters
- Can we model aleatoric inter-rater segmentation uncertainty?
- Network output parameterises a low-rank Gaussian over image space



(c)

$$\boldsymbol{\eta}^{(m)} | \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x})) \,;$$



Monteiro et al. (2020), Stochastic Segmentation Networks: Modelling Spatially Correlated Uncertainty

THE END

Questions?