

# Variational Inference: Introduction

Philip Schulz and Wilker Aziz

[https:  
//github.com/philschulz/VITutorial](https://github.com/philschulz/VITutorial)



# Problems

Supervised problems: “learn a mapping from this to that”

- ▶ e.g. machine translation, syntactic parsing, semantic role labelling, image captioning, ...

Unsupervised problems: “learn a distribution that generates the data with high probability”

- ▶ sentences in natural language, images, videos, ...

# Maximum likelihood estimation

We have data  $x^{(1)}, \dots, x^{(N)}$  e.g.

- ▶ sentences, images, ...

generated by some **unknown** procedure

# Maximum likelihood estimation

We have data  $x^{(1)}, \dots, x^{(N)}$  e.g.

- ▶ sentences, images, ...

generated by some **unknown** procedure  
which we assume can be captured by a probabilistic model

- ▶ with **known** probability (mass/density) function e.g.

$$X \sim \text{Cat}(\pi_1, \dots, \pi_K) \quad \text{or} \quad X \sim \mathcal{N}(\mu, \sigma^2)$$

and proceed to **estimate parameters** that assign maximum likelihood to observations

# Where does deep learning kick in?

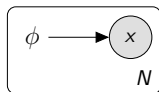
Let  $\phi$  be all side information available  
e.g. deterministic *inputs/features*

Have neural networks predict parameters of our probabilistic model

$$X|\phi \sim \text{Cat}(\pi_w(\phi)) \quad \text{or} \quad X|\phi \sim \mathcal{N}(\mu_w(\phi), \sigma_w(\phi)^2)$$

and proceed to **estimate parameters**  $w$  of the NNs

# Multiple problems, same language



(Conditional) Density estimation

Parsing	$\phi$ a sentence	$x$ its syntactic/semantic parse tree/graph
Translation	a sentence	its translation
Captioning	an image	caption in English
Entailment	a text and hypothesis	entailment relation

# Task-driven feature extraction

Often our side information  $\phi$  is itself some high dimensional data

- ▶  $\phi$  is a sentence and  $x$  a tree
- ▶  $\phi$  is the source sentence and  $x$  is the target
- ▶  $\phi$  is an image and  $x$  is a caption

and part of the job of the NNs that parametrise our models is to also **deterministically** encode that input in a low-dimensional space



# NN as efficient parametrisation

From the statistical point of view NNs do not generate data

- ▶ they parametrise distributions that *by assumption* govern data

Compact and efficient way to map from complex side information to parameter space

Prediction is done by a decision rule outside the statistical model

- ▶ e.g. beam search

# MLE via gradient-based optimisation

The probability of an observation  $X = x$  is given by some **differentiable** probability function

- ▶ the parameters of which are predicted by  $f_w$   
(*also differentiable*)

Example:  $K$  classes

$$\text{Cat}(X = x | f_w(\phi)) = \prod_{i=1}^K f_w(\phi)^{[x=i]}$$

Given a dataset of i.i.d. observations, SGD gives us a local optimum of the log-likelihood

# DL in NLP recipe

Maximum likelihood estimation

- ▶ tells you which **loss** to optimise (i.e. negative log-likelihood)

Automatic differentiation (*backprop*)

- ▶ “give me a tractable forward pass and I will give you **gradients**”

Stochastic optimisation powered by backprop

- ▶ general purpose gradient-based optimisers

# Tractability is central

Likelihood gives you a differentiable objective to optimise for

- ▶ but we need to stick with **tractable** likelihood functions

# When do we have intractable likelihood?

Latent variables: assessing the likelihood requires marginalisation

- ▶ too many forward passes

$$P_X(x|\phi) = \sum_{c=1}^K \text{Cat}(c|\pi_1, \dots, \pi_K) \underbrace{\mathcal{N}(x|\mu_w(c), \sigma_w(c)^2)}_{\text{forward pass}}$$

- ▶ even infinitely many

$$P_X(x|\phi) = \int \mathcal{N}(z|0, I) \underbrace{\text{Cat}(x|\pi_w(z))}_{\text{forward pass}} dz$$

# But I know approximations!

## Beam-search

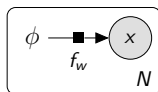
- ▶ stochastic optimisation requires **unbiased gradient estimates**

## Monte Carlo sampling

- ▶ **breaks differentiability** (bye bye backprop)

# What do we do then?

Vast majority of papers published at ACL



- ▶ encode more and more inductive bias through the design of the architecture
  - ▶ and what if we want to learn clusters?
  - ▶ or segmentation?
  - ▶ or sparse embeddings?
  - ▶ or sparse models?
  - ▶ or latent factors?

# Deep Generative Models

Probabilistic models parametrised by neural networks

- ▶ better modelling assumptions  
one of the reasons why there's so much interest
- ▶ but requires efficient inference  
which is the reason why we are here today



# Literature I