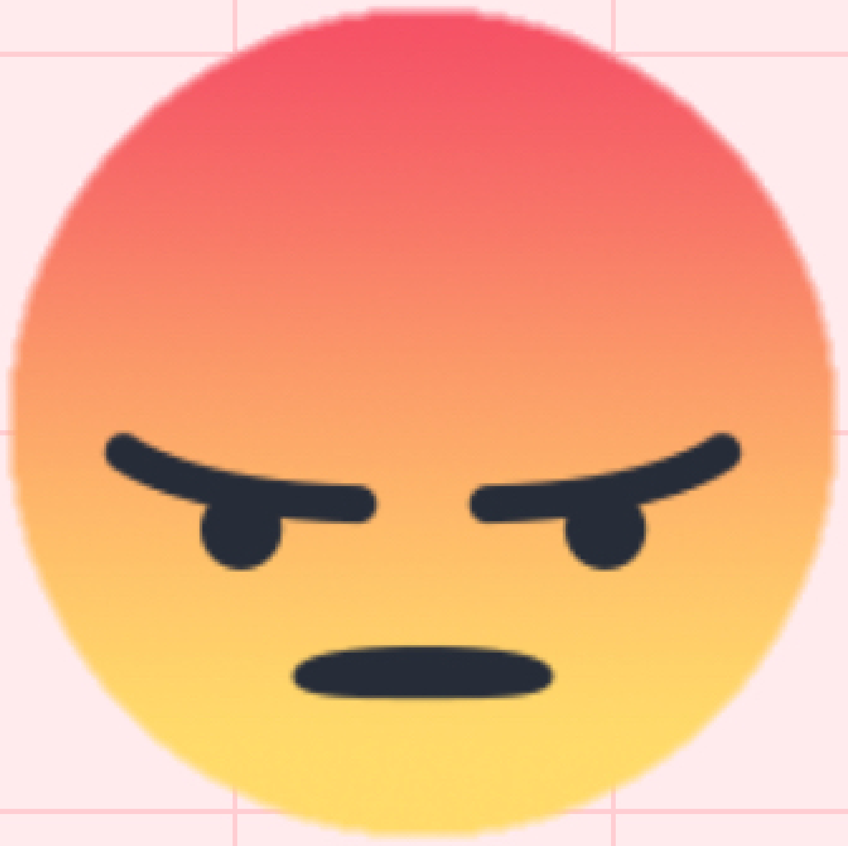
Estimation in Unnormalised Models

Z()

Maximum Likelihood

- write down likelihood as function of θ
- (try to) maximise this function
- easy? (conceptually, in practice)
- hard? (non-standard models)
- variants (penalised, Bayes, ...)

$\hat{\theta}_{MLE}$

Cramér-Rao, Fisher Efficiency

- consider asymptotic (co)variance of estimator
- among unbiased estimators, can't beat MLE
- same for biased (more conditions)
- open and shut case? (asymptotically)

Quick proof of Cramér-Rao

$$
\begin{aligned}
a &= \nabla_\theta \langle a, \theta \rangle \\
&= \nabla_\theta \int P(x|\theta) \langle a, T(x) \rangle \, dx \\
&= \int P(x|\theta) \langle a, T(x) \rangle \nabla_\theta \log P(x|\theta) \, dx \\
&= \left(\int P(x|\theta) \nabla_\theta \log P(x|\theta) T(x)^T dx \right) a
\end{aligned}
$$

$$\Rightarrow I = \mathrm{Cov}(\nabla_\theta \log P(x|\theta), T(x)) \Rightarrow \text{apply CS}$$

Challenges of MLE

- tractability (convexity, computability, ...)
- robustness / sensitivity (misspecification)
- identifiability (parametrisation, symmetries)

When is MLE hard / weird?

- Gamma, Beta (too 🦥 to code up Γ, ψ?)
- Mixture of Gaussians (non-identifiable, non-convex)
- Latent Variable Models (likelihood = integral)
- Non-Regular Models (uniform, constrained)

Alternatives to MLE

- method of moments (GMM, GEE, QL, ...)
- one-step estimator ($\sqrt{n}$ + Newton)
- robustified, Z-/M-estimation
- (model-specific, more fancy stuff, ...)

Unnormalised Models

- "energy-based", "doubly-intractable"

$$P(x \mid \theta) = \frac{f(x,\theta)}{Z(\theta)}$$

😇 😈

- usually dim x » 1 (otherwise, could integrate)
- from DAGs to Factor Graphs, causes to interactions

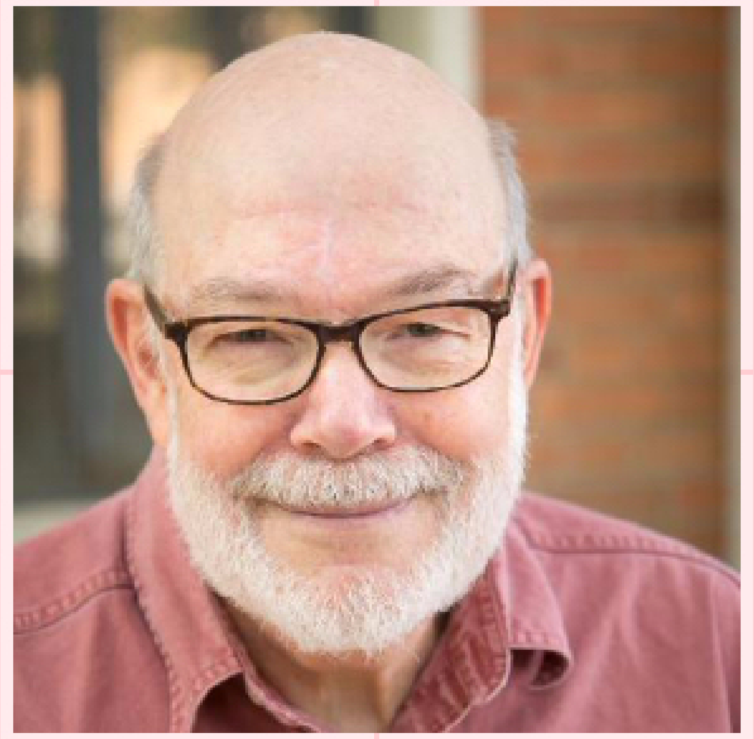
Unnormalised Models

- Ising, (Deep, Restricted) Boltzmann Machine
- (Gaussian, Hidden, Sequential) Markov Random Field
- Text Models, Image Models (Field of Experts)
- ERGMs, Stochastic Block Model, Random Networks
- (Kernel) Exponential Family

Can it get worse?

- latent variables as well! (HMRF, RBM)
- conditionally-unnormalised as well! (LATKES)
- high-dimensional! ($d \asymp n$)
- nonparametric! ($\log f$ = NN, Kernel, ...)
- but, already pretty hard ...

MLE Objective, Algorithms



- (MC)MC-MLE / Importance Sampling

$$\log Z(\theta) \approx \log\left(\frac{1}{M}\sum_{j=1}^{M}\frac{f(y_j,\theta)}{q(y_j)}\right)$$

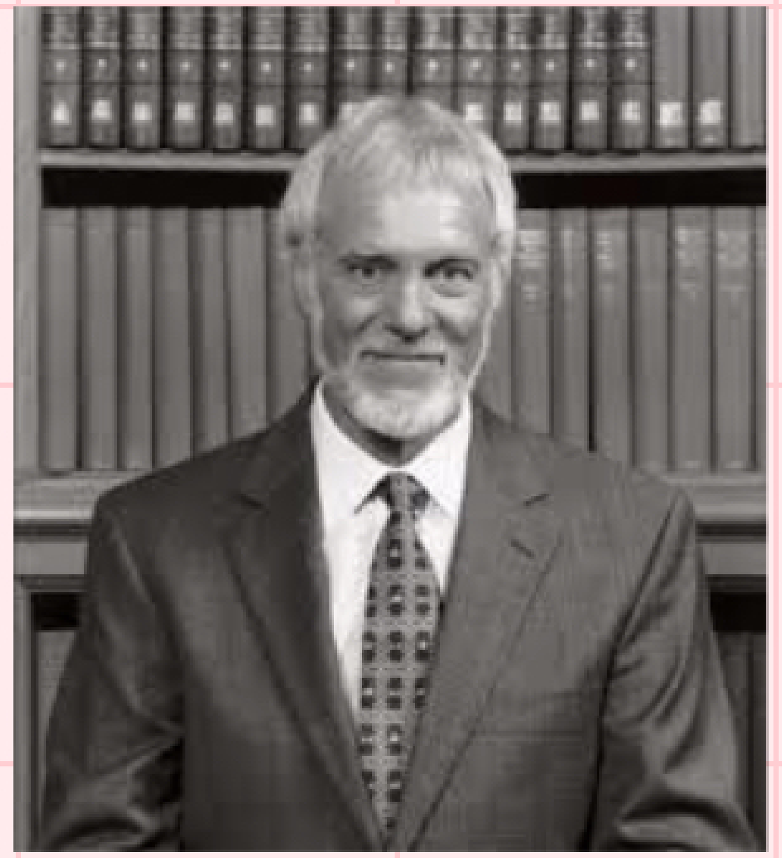
- Stochastic Approximation, Contrastive Divergence

$$\nabla_\theta \log Z(\theta) = \mathbb{E}_\theta\left[\nabla_\theta \log f(x,\theta)\right]$$

Non-MLE Objectives

- that damn normalising constant!
- can we make it cancel out?
- tricks: differences, ratios, components
- _some_ part of the model is tractable?
- safety check: support of P

Graphical Methods



- $\mathbb{P}(X_A \mid X_B)$ tractable for some A, B
- Pseudo-Likelihood (A = { i }, B = V \ { i })
- Composite Likelihood (arbitrary A, B)
- applicable for MRFs (no sparsity needed!)
- often convex

Pseudo-Likelihood Example

- Consider pairwise Markov Random Field

$$P(X|\theta) = \frac{\exp\left(\sum_{i\sim j} \theta_{ij} x_i x_j\right)}{Z(\theta)}$$

$$\Rightarrow P(x_i | x_{-i}, \theta) = \text{Ber}\left(x_i \mid \sigma((\Theta X)_i)\right)$$

- In practice: constrain / regularise θ
- Remark: Belief Propagation for Sub-Trees

Difference Methods

- $\nabla_x \log P(x \mid \theta) = \nabla_x \log f(x, \theta)$
  - $\rightsquigarrow$ no $Z(\theta)$! (c.f. OLD / MCMC)
- requires smoothness of model
- Score Matching, Stein Discrepancies
- different complexities, both often convex

Score Matching Objective

- Score Matching Objective (Q = Data)

$$\mathbb{E}_Q\left[|\nabla_x \log Q(x) - \nabla_x \log P(x|\theta)|^2\right]$$

$$\rightsquigarrow \mathbb{E}_Q\left[2\Delta_x \log P(x|\theta) + |\nabla_x \log P(x|\theta)|^2\right] + c$$

$$B \quad (\nabla_x \log P \rightarrow \nabla_x \log Q)$$

KSD Objective

- Kernel Stein Discrepancy Objective (Q = Data)

$$\mathbb{D}_P(Q)^2 = \sup_{h \in \mathcal{H}} \mathbb{E}_Q\left[(\mathcal{L}^P h)(x)\right]^2$$

$$= \mathbb{E}_{Q \otimes Q}\left[(\mathcal{L}^P_x \mathcal{L}^P_y K)(x,y)\right]$$

Rational Methods

- $P(y \mid \theta) / P(x \mid \theta) = f(y, \theta) / f(x, \theta)$
- Ratio Matching
- Let $Q(x)$ be known, classify $Q(x)$ vs $P(x \mid \theta)$
  - Optimal: $\sigma(\log Q(x) - \log f(x, \theta) - \log Z(\theta))$
  - Noise-Contrastive Estimation
- Stein Density Ratio Estimation (a bit of both)

Ratio Matching Objective

$$\mathbb{E}_Q\left[\left(\beta\left(\frac{Q(\phi x)}{Q(x)}\right) - \beta\left(\frac{P(\phi x|\theta)}{P(x|\theta)}\right)\right)^2\right] + \text{symmetrise}$$

$$= \mathbb{E}_Q\left[\left(1 - \beta\left(\frac{P(\phi x|\theta)}{P(x|\theta)}\right)\right)^2\right] + \text{const.}$$

Noise-Contrastive Estimation

$x_1, \ldots, x_N \overset{iid}{\sim} P(x|\theta)$ $\quad z = 1$

$y_1, \ldots, y_M \overset{iid}{\sim} Q(y)$ $\quad z = 0$ $\quad \nu = m/n$

$$\mathbb{P}(z = 1 \mid u) = \frac{1}{1 + \nu Q(u) / P(u|\theta)}$$

$$= \sigma\left(\log\{Q(u)/f(u,\theta)\} + c\right)$$

unknown (→ c)

- often convex in (θ, c)

Related / Frontiers

- Denoising Autoencoders, Denoising Score Matching
- Score Estimation (SBGM, DDPM, ...)
- Learned Stein Discrepancies (beyond Kernels)
- Hybrids with other approaches