

TUTORIAL 1 - Scalable Bayesian Inference - P. Dunsdon (Duke)

- Interpretable uncertainty quantification / characterizable (UQ)
→ bayesian stats \Rightarrow one prob. model perspective \rightarrow (butlers 2nd want)
- Var. inf. \Rightarrow crude way of eliciting UQ \rightarrow very constrained posterior estimate
(\hookrightarrow vs. robustness of too flexible model)
(\hookrightarrow less robust inferences as sample size n grows \rightarrow (anything significant))

Notation: $\pi_n(\theta | y^{(n)}) = \frac{x(\theta) L(y^{(n)} | \theta)}{\int x(\theta) L(y^{(n)} | \theta) d\theta}$ $\xrightarrow{L(y^{(n)})} \text{ML/Evidence}$ \rightarrow characterizes uncertainty in θ or $f(\theta)$
 \xrightarrow{b} e.g. UNI (interpretability)

- Problem: approx. of intractable high-dim. integral
→ 'doubly' intractable: also prior (LT not given)
→ 'intercepty' / non-conjugate models
 \hookrightarrow use tractable class of distrs. to approx.:

① Bayesian CLT (van Trees): $\pi_n(\theta | y^{(n)}) \approx N(\mu_n, \Sigma_n)$ as $n \rightarrow \infty$
(large sample Gaussian approx.)

- ass. smooth & differentiable likelihood
- true θ_0 in interior of param space

② Laplace approx.: Good job at 1st + 2nd want approx.

• Alternative: Define approx. class $g(\theta)$

- g charact. of exp. family \Rightarrow minimize divergence from (KL, α , bregman)

→ VI办法: handle away intractable term \Rightarrow UQ!
 \hookrightarrow don't know how accurate

\hookrightarrow Fix-ups \Rightarrow Girolami, Fassoulas, Jordan (2015)

- MCMC methods \rightarrow sequential algo \rightarrow obtain correlated draws from posterior
 \rightarrow samples can be used in very many ways! \Rightarrow flexible inference
 ↳ summaries of posterior of any fct. $f(\theta)$
 - \rightarrow constructs MC with stationary dists. $\pi_n(\theta | y^{(n)}) \Rightarrow$ particle level
 - \rightarrow MHT algo : ① sample proposal : $\theta^* = g(\theta^{(t-1)}) \Rightarrow$ MC
 - ② accept : $\theta^{(t)} = \theta^*$ w.p. $\min \left\{ 1, \frac{\pi(\theta^*) L(y^{(n)}, \theta^*)}{\pi(\theta^{(t-1)}) L(y^{(n)}, \theta^{(t-1)})} \frac{g(\theta^{(t-1)})}{g(\theta^*)} \right\}$
 - \rightarrow how to choose g ?! \Rightarrow Gibbs : draw subchains from θ from cond.
 \Rightarrow Metropolis-Hastings : Tunable variance
 \Rightarrow Metropolis-Hastings : Exploit gradient info

COST PER ITERATION

①. Bottlenecks

- Scalability

COST PER ITERATION

- cost of sampling + evaluate acceptance prob. (\rightarrow eval. Ltt)
 - similar to most optimization procedures

CONVERGENCE PROBLEMS ②. Methods

- CONVERGENCE

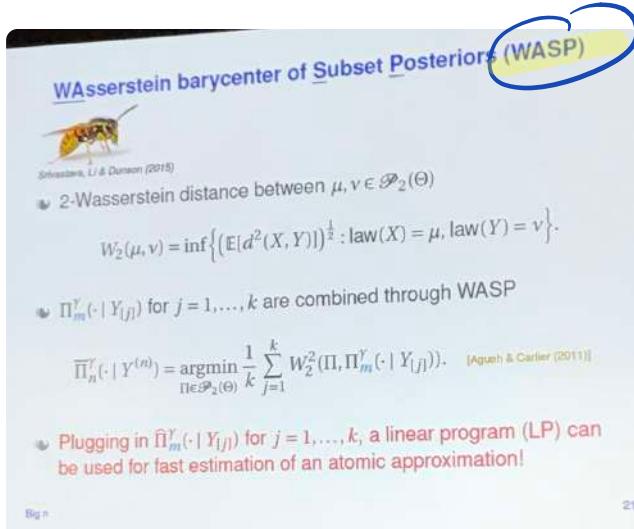
 - No dupl. samples \Rightarrow auto-correlated / slow mixing
 - NLE/rate should not explode with supply edge

- big N solubles

- Big N solutes

 - ① Emperically ferred CEP MCMC: run in parallel on small subsets
 - ② Approximate MCMC: approx. exp. to update transition levels
→ trade-off: accuracy vs. computational budget
 - ③ C-Bayes: condition on observed data being for small n_{trials}
 - ④ Hybrid MCMC:

- ① EP-MCMC: combine many subset posteriors
 → diversity reduces noise \Rightarrow stochastic approx. \rightarrow as in SGD
 $\rightarrow \prod_m (\gamma^{(j)}) \Rightarrow \gamma$ weights the respective subset poster.
 → Min Wasserstein = solution to discrete optimal transport problem
 ↳ linear program solves



- \rightarrow Posterior Bilevel Estimation (PBE)
 ↳ 1-d
 ↳ WTB: quantiles = simple averages of quantiles of sub.p.
 ↳ run MCMC for subsets and average quantiles! (for any pct. of interest)
 ↳ big bag of little bootstraps

- Robustness: If model is not exact \Rightarrow model complexity will explode / grow larger than optimal! \rightarrow model misspecification!
 - \rightarrow Coarsened poster.: idealized vs. observed data (corrupted)
- $\pi(\theta | X_{\text{obs}} = x_{\text{obs}})$ vs. $\pi(\theta | d(\hat{p}_{x_{\text{obs}}}, \hat{p}_{x_{\text{obs}}}) < R)$
- ↳ $R \sim \exp(\alpha)$ \Rightarrow poor posterior approx.!
 if d is closer to be relative entropy
- \rightarrow can show robustness to small perturbations to null hypothesis
- \rightarrow MCMC for additive model versions! // also robust to local misspecification

- Big P solutions \Rightarrow Variable selection for $p \gg n$

- ① Independent Screening \Rightarrow select when minimize negative p-value & control for FDR \Rightarrow Benjamini-Hochberg
- ② Fused Lasso Screening \Rightarrow introduce small bias to reduce vertex sparsity

↳ horseshoe, generalized double porosity, Darcy's law: flow envelope and O/D envelope obstacles.

$$p_j \text{ iid } N(0, \sigma_j^2), \gamma_j = f, \delta_j^2$$

local scale global scale

→ problem: Need aggressive shrinkage to combat high - order, but most of water is not sparse!

- Modular Bayes → fresh data and build model for each part!
 - don't have to model everything jointly!
 - separate modeling \oplus combination
 - ↳ Shared Kernel Mixtures Model → Syntex Ma - Rau - Rieffel Method
- approach to terrain network modelling → crosscutting technology!
 - ↳ hydrogeographic connectivity on fractal level!
- Väistö → form of 'Bayesian assistant'
 - 'these question can't be answered' \Rightarrow But there can be ...

TUTORIAL 2 - Unsupervised Deep Learning - A. Graves (DeepMind), M. Ranzato (FB)

- Active vs. passive // Teacher vs. no teacher → not necessarily better but different!

Types of Learning			
	With Teacher	Without Teacher	
Active	Reinforcement Learning / Active Learning	Intrinsic Motivation / Exploration	g
Passive	Supervised Learning	Unsupervised Learning	g

- No teacher: rewards / targets hard to obtain / define
 - ↳ How to determine reward structure
 - ↳ Unsupervised → 'learn' learning
- ⇒ rapid generalization to new levels!
- ↳ transfer learning → retrain existing structure
- ⇒ Problem: Not just or enough generalization
- ⇒ Hypothesis: Not enough structure / info in inputs
Unsupervised → Supervised → RL
~~amount of info~~

→ LEARN SKILLS AND NOT TASKS!

- Supervised learning outputs info in the targets not in the inputs!
- Overfitting not problem w.r.t. flexibility in terms of bits vs. # weights / pars
- Density modelling ⇒ ML on data instead of targets → learn from DBP!
$$D = \sum x_j, L(D) = \sum_{x \in D} -\log p(x)$$
- Problem: Want to learn every pixel → focus on what we believe to be useful!

 1. curse of dimensionality
 2. Not all bits created equal ⇒ low-level details (pixel artifacts)
 3. How to use learned structure ⇒ representation learning

- Generative Model ⇒ Visual inspection allows us to see what model has learned! no imagine / simulate
- Intrinsic Motivation ⇒ incentivize curiosity
- driven by compression progress - Schmidhuber (2008)
 - ↳ Incentivizes agents to make their actions act like compressed data!
- Empowered Agents ⇒ each. M / behavior actions act compressed!
↳ more control of dynamics!

① Autoregressive Models

- split high-dim data into sequence of small pieces \Rightarrow predict each piece from those before \Rightarrow the cond
 \hookrightarrow conditioning on network state \rightarrow LSTM / GRU

$$\Rightarrow D = \{x_t\}, p(x) = \prod_{t=1}^T p(x_t | x_{\leq t}), \mathcal{L}(D) = \sum_{x \in D} \sum_{t=1}^T -\log p(x_t | x_{\leq t})$$

- + - simple to define \Rightarrow pick ordering, easy to sample \Rightarrow sequential feeding of network output to next time step \rightarrow (forward)
- - very expensive \Rightarrow training via parallelisation, but necessarily slow because of sequential generate!
- - Order dependence! // Teacher forcing \rightarrow only leaves one-step forecasts
 \hookrightarrow no layer dependence generation?!
- - WaveNets, PixelRNN \Rightarrow language model for images \rightarrow softmax output for each channel
 \hookrightarrow local vs ~~global~~ structure
 \hookrightarrow can be induced over label conditioning!

② Representation Learning

- Internal language to describe data! \Rightarrow language can emerge from tasks \Rightarrow Wittgenstein
- E.g. C. Olah Distill visualizer \Rightarrow visual vocabulary
- Unsupervised L. should more general than task constrained
- Latent variables flat ('described') \rightarrow re-use to plan, reason, generalize
- Autoencoder \Rightarrow Variational version \rightarrow inferable cost of reconstruction!
- Latent variable models \rightarrow latent variables structure from high-dim
- \hookrightarrow Eval representability with dann gans
- Contrastive Predictive Coding \rightarrow van den Oord et al
 \hookrightarrow voice contrastive estimation \rightarrow mutual info (2018)

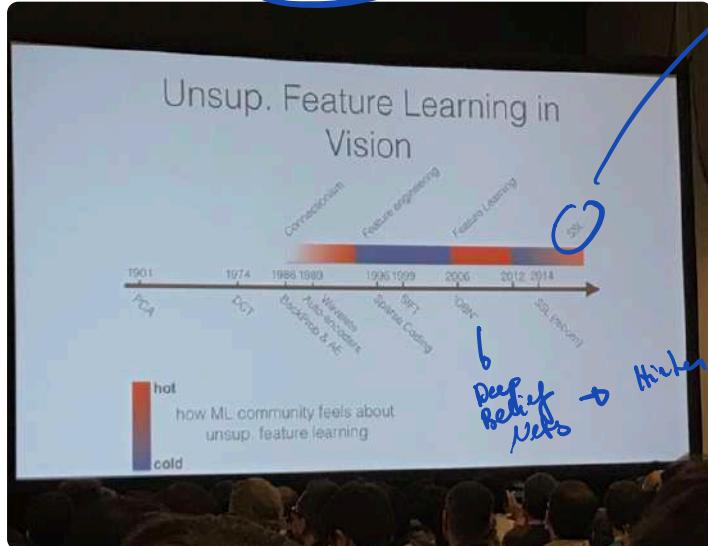
③ Unsupervised RL

- auxiliary task \Rightarrow mark. reward + uses unsupervised task!

Practical Service

- Visual representations

history

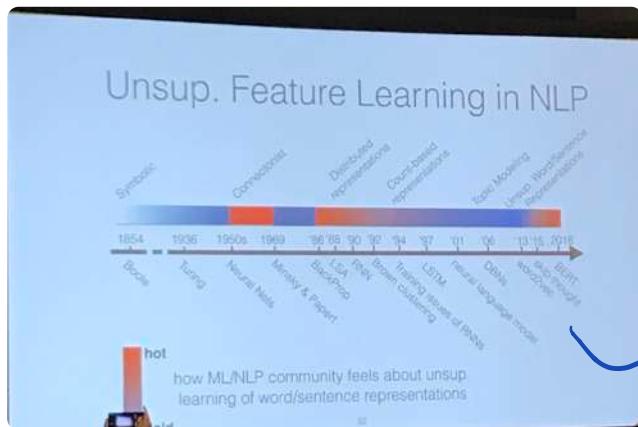


- Text representations

→ word carries a lot more info than pixel

↳ discrete signal → word search ↔ every word is modellable

→ word rec: Meaning of word is determined by its context



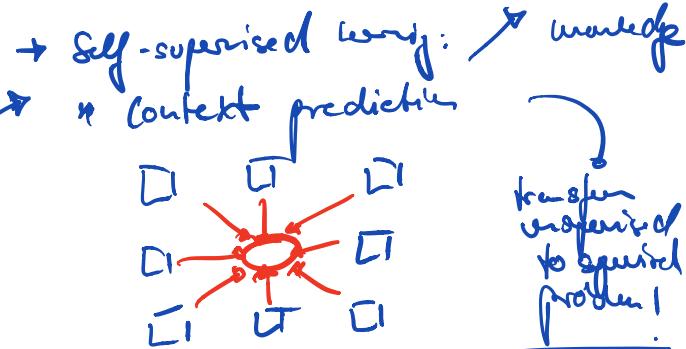
word-embeddings

two things are similar if they have similar affect
↳ discrete to continuous conversion!

→ BERT: sentences consist of very corp. blocks → transfer learning by hiding blocks and predicting = de-noising
↳ GLUE benchmark (for texts)

- Vision Sampling

→ GANs, autoregressive, OLO ⇒ crucial: architecture!



→ learning by clustering:

Random filters have selective property

↳ init random CNN

- ↳ ① Extract features \Rightarrow non-linear assignments
② Train network on linear assignments

BERT

This accomplished by using **attention**
(each block is a Transformer)

- For each layer and for each block in a layer do (simplified version)
- 1) let each current block representation at this layer be: h_j
 - 2) compute dot products: $h_i \cdot h_j$
 - 3) normalize scores: $\alpha_i = \frac{\exp(h_i \cdot h_j)}{\sum_k \exp(h_k \cdot h_j)}$
 - 4) compute new block representation as in: $h_j \leftarrow \sum_i \alpha_i h_i$



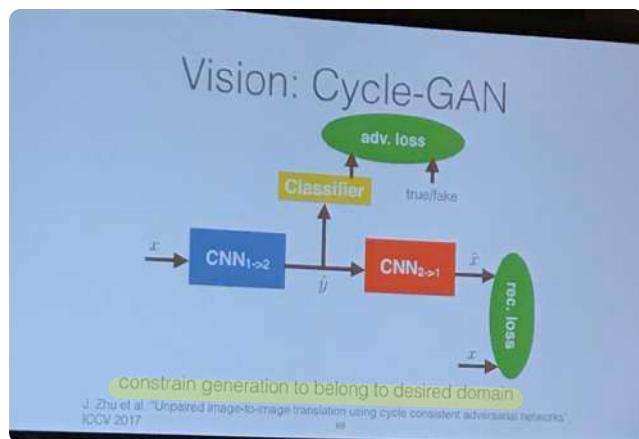
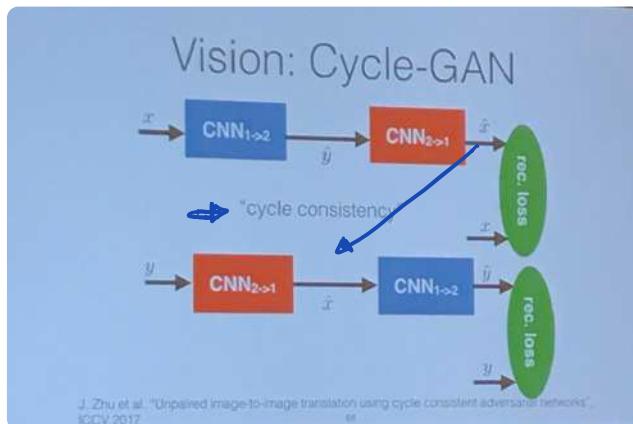
<sep> The cat sat on the mat <sep> It fell asleep soon after

A. Vaswani et al. "Attention is all you need", NIPS 2017

Text Simplify

→ Integrative, Retrieved-based, order of both

- Learning to map between domains
- Analogy
- Zhu et al (2017) \Rightarrow Cycle-GAN :



- Unsupervised Machine Translation : MUSE \rightarrow few words
- ① learn separate word embeddings per language
- ② learn joint space word embeddings jointly + refinement
- ↳ Does not work for sentences \Rightarrow seq2seq as in cycle-GAN w. constraint

Open Research

- ① Definition of metrics & domain-driven tools & task-specific metrics
- ② General principles \Rightarrow Energy-based models \rightarrow energy: contrastive feat., lower value denotes less high density \Rightarrow hard to tell if to count as positive/neg. samples?!

TUTORIAL 3 - Counterfactual Inference - S. Athey (Stanford)

Artificial Intelligence/Machine Learning Desired Properties for Applications	
DESIRED PROPERTIES	CAUSAL INFERENCE FRAMEWORK
Interpretability	Goal: learn model of how the world works <ul style="list-style-type: none">Impact of interventions can be context-specific
Stability/Robustness	<ul style="list-style-type: none">Model maps contexts and interventions to outcomes
Transferability	<ul style="list-style-type: none">Formal language to separate out correlates and causes
Ideal causal model is by definition stable, interpretable	
Fairness/Non-discrimination	Transferability: straightforward for new context dist'n
"Human-like" AI	Fairness: Many aspects of discrimination relate to correlation v. causation <ul style="list-style-type: none">Performance may depend on physical and mental ability, psychological factors (e.g. risk taking)Gender and race may be correlated with factors that shift these distributions, relatively limited direct causal effects
Reasonable decisions in never-experienced situations	

- Example: Contextual Doubts
→ context influences what test arm is!
↳ fundamentally counterfactual problem!
- Program Eval., Treatment effect estimation
- Randomized experiments / A/B Test.

①

Counterfactual Inference Approaches

"PROGRAM EVALUATION", "TREATMENT EFFECT ESTIMATION"

Goal: estimate the impact of interventions or treatment assignment policies

- Low dimensional intervention

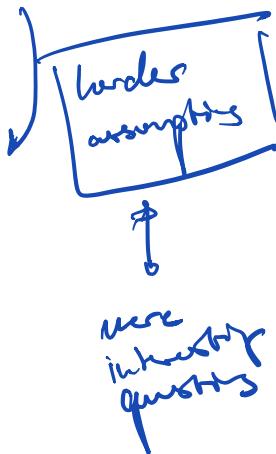
Estimands

- Average effect
- Heterogeneous effects
- Optimal policy

Confidence intervals

Designs that enable identification and estimation of these effects

- (Alternative treatments observed in historical data in relevant contexts)
- Randomized experiments
- "Natural" experiments (Unconf., IV)
- Regression discontinuity
- Difference-in-difference
- Longitudinal data



②

Counterfactual Inference Approaches

"STRUCTURAL ESTIMATION", "GENERATIVE MODELS" & COUNTERFACTUALS

Goal: estimate impact on welfare/profits of participants in alternative counterfactual regimes

- Counterfactual regimes may not have ever been observed in relevant contexts
- Need behavioral model of participants

Still need designs that enable identification and estimation, now of preference parameters

- E.g. need to see changes in prices to understand price sensitivity

Use "revealed preference" to uncover preference parameters

Rely on behavioral model to estimate behavior in different circumstances

Also may need to specify equilibrium selection

Dynamic structural models

Learn about value function from agent choices in different states

III

Counterfactual Inference Approaches

"CAUSAL DISCOVERY", "LEARNING THE CAUSAL GRAPH"

Goal: uncover the causal structure of a system
• Many observed variables
• Analyst believes that there is an underlying structure where some variables are causes of others, e.g. a physical stimulus leads to biological responses

Applications
◦ Understanding software systems
◦ Biological systems

Focus on ways to test for causal relationships

Take note
↳ a lot more complex discovery for what is often possible in practice

- Identification \Rightarrow Can Q be answered with infinite data? \rightarrow Estimation with finite data
 - Regularization \Rightarrow Induces omitted variable bias
 - Semi-Parametric efficiency theory \Rightarrow Cross-fitting / out-of-bag estimation of missed penalties
- \Rightarrow orthogonal events / double robustness

Estimating Average Treatment Effects under Confounders

$$Y_i(1), Y_i(0) \perp W_i | X_i \quad \bullet \text{ Unconfounders}$$

- Double Robust: $\begin{array}{c} \text{Fitting by propensity scores} \\ \oplus \end{array}$ Outcome modelling
- flexible model with regularization w.r.t. small variables which together moderate full effect

Left & right counterfactual