Probabilistic models for big data
Fall 2014: Introductory lecture

Antti Honkela and Arto Klami

3 September 2014

Probabilistic models for big data

Course code:	58314301
Credit points:	3
Teacher:	Dr Antti Honkela and Dr Arto Klami
Contact email:	`first.last@cs.helsinki.fi`
Office hours:	Please make an appointment by email
Prerequisites:	Probabilistic Models, Scientific writing
Moodle:	"58314301 Seminar in Probabilistic Models for Big Data, autumn 2014" Registration code: stochastic

Schedule

- Today: Introduction, papers
- Next week: Choosing the topic, instructions for presentations etc
- Article summaries:
  1. October 19th: Submission deadline
  2. November 5th: Review deadline
  3. November 25th: Final version
- Presentations between early November and Mid December (6-7 sessions)

Data analysis with probabilistic models

The modeling task:

- Construct a (often general-purpose) probabilistic model by specifying a set of probability distributions
- Observe data
- Fit the model to the data, learning the probability distribution of the model parameters given the data
- Make predictions (e.g. class labels of future observations) by averaging over that distribution

Big data

- Big data: Any data collection that is large enough to be difficult to process with "traditional" technique
- A lot of hype in the business world, but the practical challenges are real
- Computational neuroscience: Typical session with fMRI procudes around 3 gigabytes of data (activities of $\sim$1M brain voxels every few seconds)
- Computational biology: A typical high-throughput sequencing run yields 30M-100M sequencing reads of $\sim$100 nt, some GBs per sample
- Big business players (Amazon, FB etc) have hundreds of millions of customers requiring real-time predictions (e.g. recommender engines, search)

Big data at our department (examples)

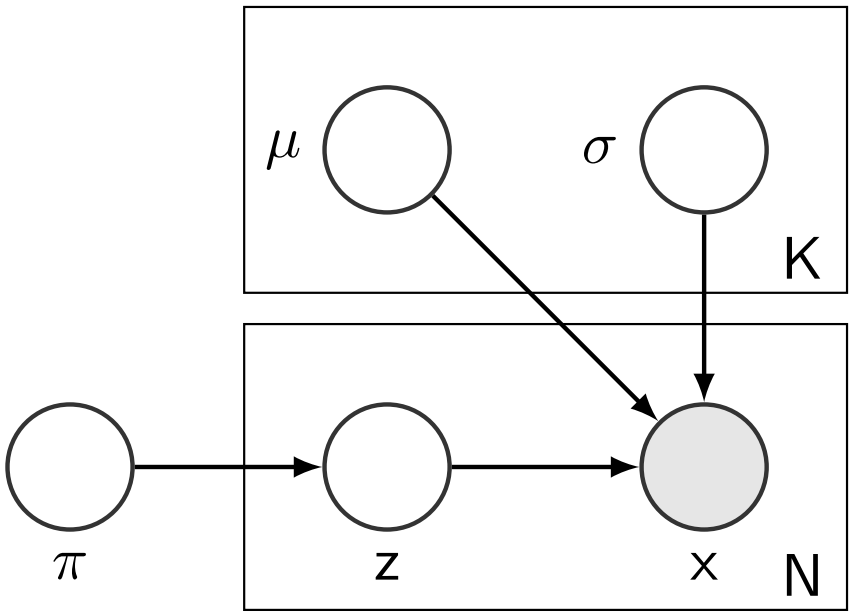
- Helsinki Privacy Experiment (50 terabytes of video and audio of home surveillance)
- 200M text documents covering editorial and social media of the past year, 60M scientific articles covering the history of human scientific progress
- Several genomics data sets with 100s high-throughput sequencing samples, multiple terabytes each set

Probabilistic models for big data

- Probabilistic models often considered to be computationally heavy; classical papers on MCMC often have only a few parameters and the sampling chains are long
- **This course:** What can probabilistic modeling offer for big data applications?
  - Scaling up variational inference
  - More efficient samplers, distributed sampling
  - Implementation issues not covered: GPU computing or other forms of scaling up the computational resources, distributed computing frameworks such as Hadoop, Spark, ...

Probabilistic graphical models

- Graph illustrates independencies, data generation described with a set of probability distributions
- Notation: $\mathcal{D} = \{\mathbf{x}\}$ is data, θ are the parameters



$$\pi \sim \text{Dirichlet}(\alpha)$$
$$z_n \sim \text{Categorical}(\pi)$$
$$\mu_k \sim \text{N}(\mu_0, \sigma_0)$$
$$\sigma_k^{-2} \sim \text{Gamma}(\alpha_0, \beta_0)$$
$$x_n \sim \text{N}(\mu_k, \sigma_k)$$

Bayesian inference

- Model: $p(\mathcal{D}, \theta) = p(\mathcal{D}|\theta)p(\theta) = p(\theta)\prod_n p(\mathbf{x}_n|\theta)$
- Given a model, we are typically interested in predictions $p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta$
- The posterior distribution $p(\theta|\mathcal{D})$ hence summarizes the model
- Bayes' rule $p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$
- In principle easy, in practice the quantity $p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)\mathrm{d}\theta$ necessitates approximative inference
- On this course: (Mostly) Markov chain Monte Carlo and variational inference

Variational inference basics

- Idea: approximate the posterior distribution $p(\theta|\mathcal{D})$ with another distribution $q(\theta)$ that is analytically tractable
- Learn the approximation by minimizing the distance between $q(\theta)$ and $p(\theta|\mathcal{D})$
- The distance is measured by the Kullback-Leibler divergence $D(q||p) = \int q(\theta) \log \frac{q(\theta)}{p(\theta|\mathcal{D})} \mathrm{d}\theta$
- ...and the minimization is often converted into maximizing a lower bound on the marginal likelihood (ELBO): $L = \int q(\theta) \log \frac{p(\mathcal{D},\theta)}{q(\theta)} \mathrm{d}\theta = p(\mathcal{D}) - D(q||p)$
- Predictions then made by replacing $\int p(\mathbf{x}|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta$ by $\int p(\mathbf{x}|\theta)q(\theta)\mathrm{d}\theta$

Mean-field variational inference

- Often $q(\theta)$ is factorized as $\prod_i q(\theta_i)$, so that we can optimize one factor at a time
- Differentiating wrt to $q(\theta_i)$ and setting the derivative to zero provides a closed-form update
  $\log q(\theta_i) = \int q(\theta_{-i}) \log p(\mathcal{D}, \theta) \mathrm{d}\theta_{-i} + C$
- The expectation over all other factors is typically easy to compute for exponential family distributions with conjugate priors (and much harder for everything else)
- Leads to an algorithm closely resembling expectation maximization

Towards more scalable variational inference

- Given a parametric form $q(\theta|\phi)$ VB is an optimization problem: $L(\phi) = \int q(\theta|\phi) \log \frac{p(\mathcal{D},\theta)}{q(\theta|\phi)} \mathrm{d}\theta$
- Gradient-based optimization generally applicable: $\phi \leftarrow \phi + \delta \nabla L(\phi)$
- Natural gradient speeds up convergence: Replace $\nabla L(\phi)$ with $F^{-1}(\phi)\nabla L(\phi)$, where $F(\phi)$ is the Fisher information matrix consisting of expectations of second derivatives of $\log q(\theta|\phi)$ (Honkela et al., JMLR 2010)
- Stochastic gradients applicable (Hoffman et al., JMLR 2013); more about this during the seminar

Variational inference example

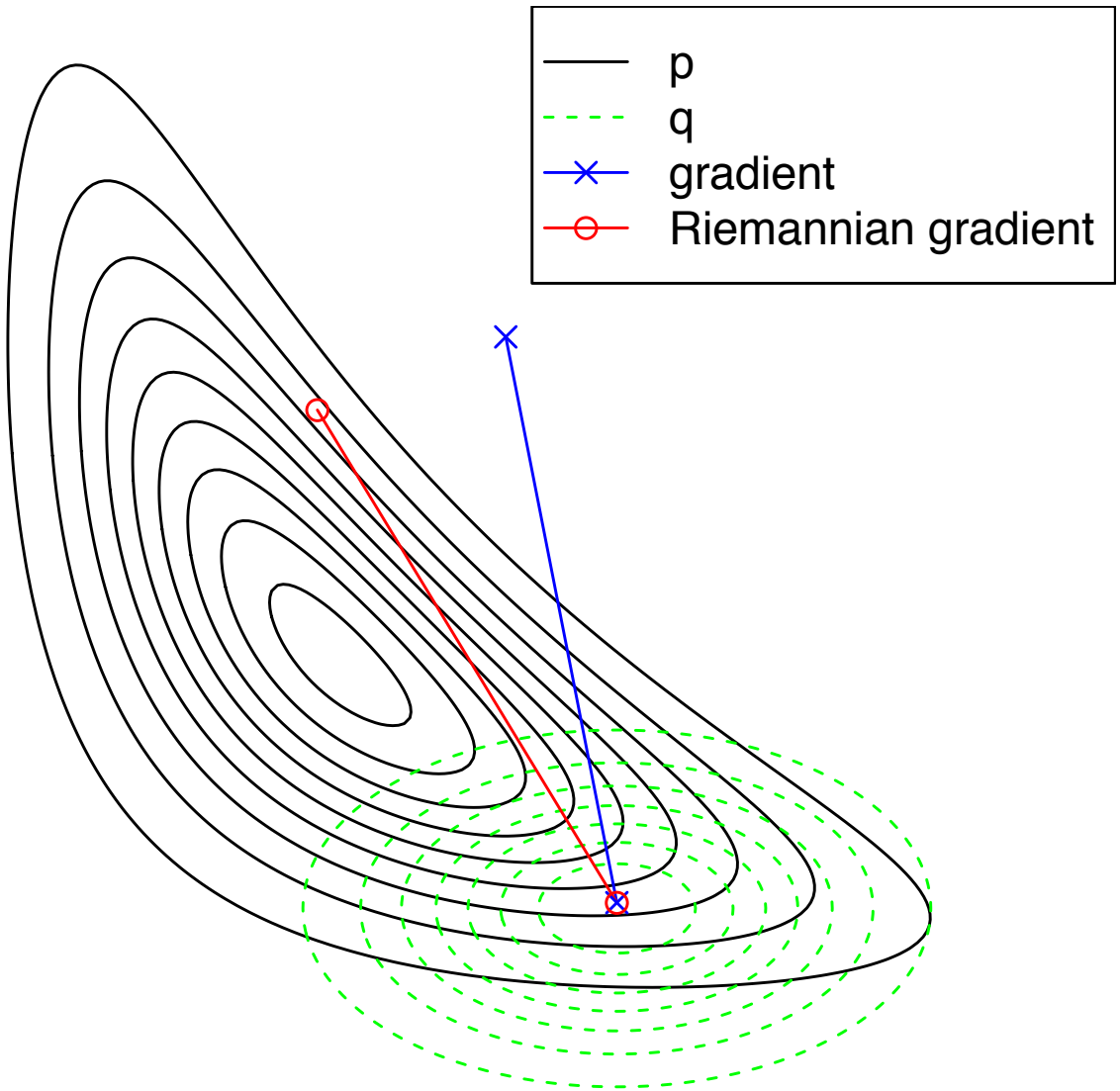


Figure from Honkela et al. (JMLR 2010)

Approximate inference by sampling

- A lot of Bayesian inference boils down to computing integrals

$$E[f(\theta)] = \int_{\theta} f(\theta) p(\theta|\mathcal{D}) \mathrm{d}\theta$$

  - Model predictions, posterior statistics of parameters, ...
- θ is often high-dimensional which makes these very difficult
- Stochastic approximation:

$$E[f(\theta)] \approx \frac{1}{N} \sum_{i=1}^{N} f(\theta_i),$$

  when $\theta_i \sim p(\theta|\mathcal{D})$
- How to simulate samples following a given distribution?

MCMC basics (Metropolis et al., 1953; Hastings, 1970)

- Idea: construct a Markov chain, whose stationary distribution is the distribution of interest $p(\theta|\mathcal{D})$
- Requires an *unnormalised* $p^*(\theta|\mathcal{D}) \propto p(\theta|\mathcal{D})$
- In the Bayesian setting typically

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

  which easily yields the unnormalised density

$$p^*(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)$$

- To define the Markov chain, we need to specify a transition distribution $q(\theta'|\theta)$
- The Markov chain is guaranteed to converge if it satisfies sufficient regularity conditions and the *detailed balance* condition

$$q(\theta'|\theta)p(\theta|\mathcal{D}) = q(\theta|\theta')p(\theta'|\mathcal{D})$$

Metropolis–Hastings algorithm

- The most widely used MCMC algorithm is the Metropolis–Hastings algorithm
- Accept–reject mechanism, proposals are accepted with probability

$$f(\theta'|\theta) = \min\left(1, \frac{q(\theta|\theta')p(\theta'|\mathcal{D})}{q(\theta'|\theta)p(\theta|\mathcal{D})}\right)$$

- This satisfies the detailed balance because

$$\begin{aligned} f(\theta'|\theta)q(\theta'|\theta)p(\theta|\mathcal{D}) &= \min(q(\theta'|\theta)p(\theta|\mathcal{D}), q(\theta|\theta')p(\theta'|\mathcal{D})) \\ &= \min(q(\theta|\theta')p(\theta'|\mathcal{D}), q(\theta'|\theta)p(\theta|\mathcal{D})) \\ &= f(\theta|\theta')q(\theta|\theta')p(\theta'|\mathcal{D}) \end{aligned}$$

Gradients in MCMC

- Standard MCMC is based on proposal distributions whose shape is essentially *independent of the target*
  - E.g. fixed multivariate Gaussian proposals
- Target distribution gradients would allow utilising local shape
- Common algorithms:
  - Langevin dynamics MCMC
  - Hamiltonian Monte Carlo (a.k.a. hybrid Monte Carlo)
- Both based on constructing a suitable dynamical system and simulating it

Demo time

http://nbviewer.ipython.org/630ec3bc0d4bbaa94d03

Stochastic gradients

Papers: MCMC I

M1. S. Ahn, A. Korattikara, M. Welling, **Bayesian posterior sampling via stochastic gradient Fisher scoring**, ICML 2012 and M. Welling, Y.W.Teh, **Bayesian Learning via Stochastic Gradient Langevin Dynamics**, ICML 2011.

M2. S.Patterson, Y. W. Teh, **Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex**, NIPS 2013.

M3. S. Ahn, B. Shahbaba, M. Welling, **Distributed stochastic gradient MCMC**, ICML 2014.

M4. T. Chen, E. Fox, C. Guestrin, **Stochastic Gradient Hamiltonian Monte Carlo**, ICML 2014.

Papers: MCMC II

M5. A. Korattikara, Y. Chen, M. Welling, **Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget**, ICML 2014.

M6. D. Maclaurin, R.P. Adams, **Firefly Monte Carlo: Exact MCMC with Subsets of Data**, UAI 2014.

M7. W. Neiswanger, E. Xing, C. Wang, **Asymptotically Exact, Embarrassingly Parallel MCMC**, UAI 2014; S.L. Scott, A.W. Blocker, F.V. Bonassi, H.A. Chipman, E.I. George, R.E. McCulloch, **Bayes and Big Data: The Consensus Monte Carlo Algorithm**, Bayes 250, 2013; and T. Campbell, J. How, **Approximate Decentralized Bayesian Inference**, UAI 2014.

Papers: Variational

V1. M.D. Hoffman, D.M. Blei, C. Wang, J. Paisley. **Stochastic Variational Inference**, JMLR 2013 (Sections 1-2 and 5)

V2. ... (Sections 3-4) and R. Ranganath, C. Wang, D.M. Blei, E. Xing, **An Adaptive Learning Rate for Stochastic Variational Inference**, ICML 2013.

V3. M. Titsias, M. Lazaro-Gredilla, **Doubly Stochastic Variational Bayes for non-Conjugate Inference**, ICML 2014.

V4. D.J. Rezende, S. Mohamed, D. Wierstra, **Stochastic Backpropagation and Approximate Inference in Deep Generative Models**, ICML 2014.

V5. J. Hensman, N. Fusi, N.D. Lawrence, **Gaussian Processes for Big Data**, UAI 2013.

V6. J. M. Hernandez-Lobato, N. Houlsby, Z. Ghahramani, **Stochastic Inference for Scalable Probabilistic Modeling of Binary Matrices**, ICML 2014.

Papers: Other

O1. H. Rue, S. Martino, **Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations**, JRSS:B, 2009.

O2. M. Schmidt, N. Le Roux, F. Bach, **Minimizing Finite Sums with the Stochastic Average Gradient**, arXiv 2013.

- Own suggestions?

Next steps

Moodle: "58314301 Seminar in Probabilistic Models for Big Data, autumn 2014"
Registration code: stochastic

For next week:

1. Check the paper list (Moodle or course web page) and have a look at the papers
2. Mark all papers you are interested in on Moodle by Tuesday 9 September
   - Expressed preferences may be used to pre-allocate papers
3. Come to the next session on 10 September with a list of preferred papers and open mind for non-favourites too!