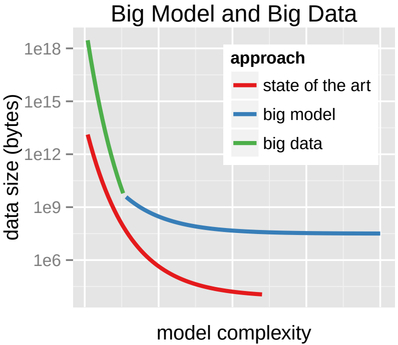
Section 5.
Stan for “Big Data”

Bob Carpenter

Columbia University

Part I
Overview

Scaling and Evaluation



- Types of Scaling: data, parameters, **models**

Riemannian Manifold HMC

- Best mixing MCMC method (fixed # of continuous params)
- Moves on Riemannian manifold rather than Euclidean
  - adapts to position-dependent curvature
- **geoNUTS** generalizes NUTS to RHMC (Betancourt *arXiv*)
- **SoftAbs** metric (Betancourt *arXiv*)
  - eigendecompose Hessian and condition
  - computationally feasible alternative to original Fisher info metric of Girolami and Calderhead (*JRSS, Series B*)
  - requires third-order derivatives and implicit integrator
- Code complete; awaiting higher-order auto-diff

Adiabatic Sampling

- Physically motivated alternative to “simulated” **annealing and tempering** (not really simulated!)
- Supplies external **heat bath**
- Operates through **contact manifold**
- System relaxes more naturally between energy levels
- Betancourt paper on *arXiv*

- Prototype complete

“Black Box” Variational Inference

- **Black box** so can fit any Stan model
- Multivariate **normal approx to unconstrained** posterior
  - covariance: diagonal mean-field or full rank
  - not Laplace approx — around posterior mean, not mode
  - transformed back to constrained space (built-in Jacobians)
- Stochastic **gradient-descent** optimization
  - ELBO gradient estimated via Monte Carlo + autdiff
- Returns **approximate posterior** mean / covariance
- Returns **sample** transformed to constrained space

“Black Box” EP

- Fast, approximate inference (like VB)
  - VB and EP minimize divergence in opposite directions
  - especially useful for Gaussian processes
- Asynchronous, data-parallel **expectation propagation** (EP)
- Cavity distributions control subsample variance

- Prototypte stage
- collaborating with Seth Flaxman, Aki Vehtari, Pasi Jylänki, John Cunningham, Nicholas Chopin, Christian Robert

Maximum Marginal Likelihood

- Fast, approximate inference for hierarchical models
- Marginalize out lower-level parameters
- Optimize higher-level parameters and fix
- Optimize lower-level parameters given higher-level
- Errors estimated as in MLE
- aka “empirical Bayes”
    - but not fully Bayesian
    - and no more empirical than full Bayes
- Design complete; awaiting parameter tagging

Part II

Posterior Modes & Laplace Approximation

Laplace Approximation

- Maximum (penalized) likelihood as approximate Bayes
- Laplace approximation to posterior
- Compute posterior mode via optimization

$$\theta^* = \operatorname{arg\,max}_\theta \; p(\theta|y)$$

- Estimate posterior as

$$p(\theta|y) \approx \mathsf{MultiNormal}(\theta^* | -H^{-1})$$

- H is the Hessian of the log posterior

$$H_{i,j} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\theta|y)$$

Stan's Laplace Approximation

- L-BFGS to compute posterior mode $\theta*$
- Automatic differentiation to compute H
  - current R: finite differences of gradients
  - soon: second-order automatic differentiation

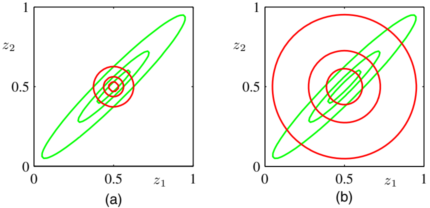
Part III
Variational Bayes

VB in a Nutshell

- y is observed data, θ parameters
- Goal is to approximate posterior $p(\theta|y)$
- with a convenient approximating density $g(\theta|\phi)$
  - ϕ is a vector of parameters of approximating density
- Given data y, VB computes ϕ^* minimizing KL-divergence
  - from approximation $g(\theta \mid \phi)$ to posterior $p(\theta \mid y)$

$$
\begin{aligned}
\phi^* &= \arg\min_\phi \mathrm{KL}[g(\theta|\phi) \;||\; p(\theta|y)] \\
&= \arg\max_\phi -\int_\Theta \log\left(\frac{p(\theta \mid y)}{g(\theta \mid \phi)}\right) g(\theta|\phi)\, \mathrm{d}\theta
\end{aligned}
$$

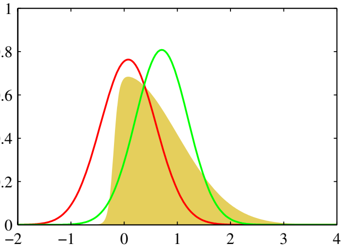
KL-Divergence Example



- Green: true distribution p; Red: approx. distribution g
  (a) VB-like: $\text{KL}[g \,||\, p]$; (b) EP-like: $\text{KL}[p \,||\, g]$
- VB systematically **understimates posterior variance**

— Bishop (2006) *Pattern Recognition and Machine Learning*, fig. 10.2

VB vs. Laplace



- solid yellow: target;   red: Laplace;   green: VB
- VB approximates posterior mean; Laplace posterior mode

— Bishop (2006) *Pattern Recognition and Machine Learning*, fig. 10.2

Stan's "Black-Box" VB

- Typically custom $g()$ per model
  - based on conjugacy and analytic updates
- Stan uses "black-box VB" with multivariate Gaussian g

$$g(\theta|\phi) = \text{MultiNormal}(\theta \mid \mu, \Sigma)$$

  for the **unconstrained posterior**
  - e.g., scales σ log-transformed with Jacobian
- Stan provides two versions
  - Mean field: Σ diagonal
  - General: Σ dense

Stan's VB: Computation

- Use L-BFGS optimization to optimize θ^*
- Requires differentiable $\mathrm{KL}[g(\theta|\phi) \mid\mid p(\theta|y)]$
  - only up to constant (i.e., use evidence lower bound (ELBO))
- Approximate KL-divergence and gradient via Monte Carlo
  - KL divergence is an expectation w.r.t. approximation $g(\theta|\phi)$
  - Monte Carlo draws i.i.d. from approximating multi-normal
  - only need approximate gradient calculation for soundness
  - so only a few Monte Carlo iterations are enough

Stan's VB: Computation (cont.)

- To support compatible plug-in inference
  - draw Monte Carlo sample $\theta^{(1)}, \ldots, \theta^{(M)}$ with

$$\theta^{(m)} \sim \mathsf{MultiNormal}(\theta \mid \mu^*, \Sigma^*)$$

  - inverse transfrom from unconstrained to constrained scale
  - report to user in same way as MCMC draws

- Future: reweight $\theta^{(m)}$ via importance sampling
  - with respect to true posterior
  - to improve expectation calculations

Near Future: Stochastic VB

- Data-streaming form of VB
  - Scales to billions of observations
  - Hoffman et al. (2013) Stochastic variational inference. *JMLR* 14.
- Mashup of stochastic gradient (Robbins and Monro 1951) and VB
  - subsample data (e.g., stream in minibatches)
  - upweight each minibatch to full data set size
  - use to make unbiased estimate of true gradient
  - take gradient step to minimimize KL-divergence
- Prototype code complete

The End (Section 5)