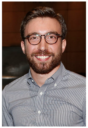
Challenges in Variational Inference: Optimization, Automation, and Accuracy

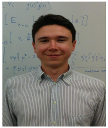
Rajesh Ranganath

December 11, 2015

Collaborators

- Dave Blei
- Alp Kucukelbir
- Stephan Mandt
- James McInerney
- Dustin Tran




Review: Variational Inference

Goal: Fit a distribution to the posterior with optimization

Model:

- Model: $p(x, z)$
- Latent Variables: z
- Data: x

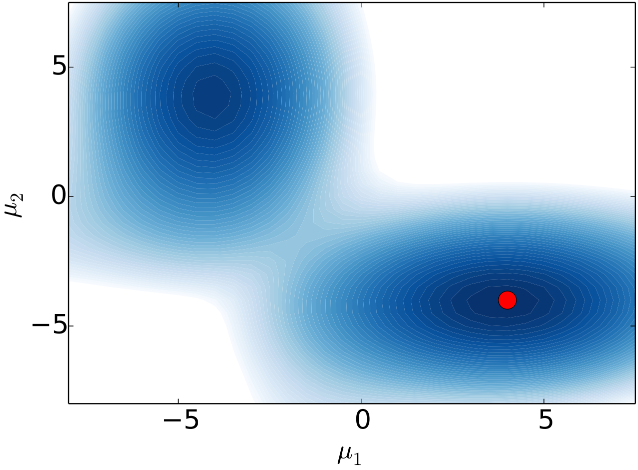
Variational Inference:

- Approximating Family: $q(z; \lambda)$
- Minimize $KL(q||p(z \mid x))$ or maximize ELBO:

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(x, z) - \log q(z; \lambda)]$$

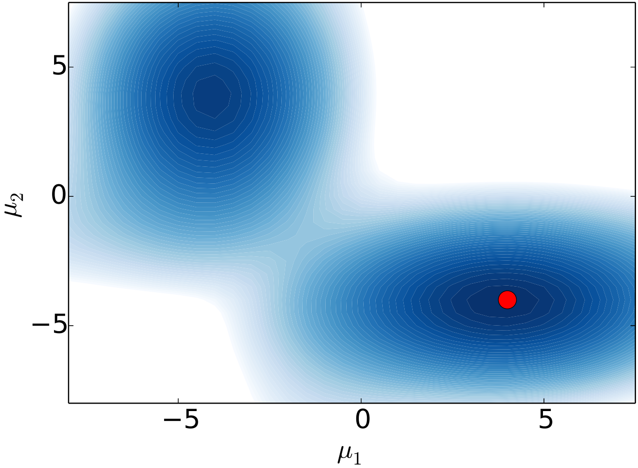
Problem: Local Optima

ELBO for mixture model of two Gaussians



Problem: Local Optima

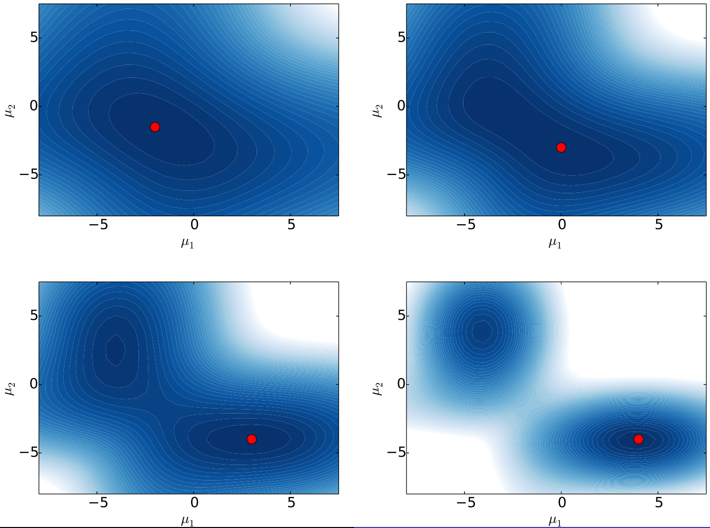
ELBO for mixture model of two Gaussians



Solution is to anneal the likelihood $p(x \mid z)^{\frac{1}{T}}$

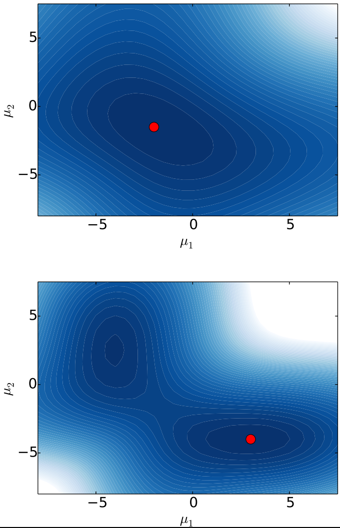
Annealing

We show the variational objective for temperatures (left to right) $T = 20$, $T = 13$, $T = 8.5$ and $T = 1$.



Annealing

Annealing slowly reduces temperature while optimizing the T-ELBO.



Why only anneal the likelihood?

Modern variational inference methods subsample data

Why only anneal the likelihood?

What happens when we anneal the prior and subsample data?

Why only anneal the likelihood?

What happens when we anneal the prior and subsample data?

- Consider Latent Dirichlet Allocation:
- The prior on the topics is Dirichlet with parameter $\alpha = \eta$
- The annealed prior is Dirichlet with parameter $\alpha = \frac{\eta}{T}$
- Given a batch of documents the update for the topics is

$$\lambda_{t+1} = (1 - \rho_t)\lambda_t + \rho_t \left(\alpha + \frac{D}{B} \sum_d \phi_{dw} W_{dn} \right).$$

- When a topic is not assigned a word it is quickly driven to α
- For $T = 10$ and $\eta = .01$, $exp(\Psi(\frac{\eta}{T}) - \Psi(\eta)) \approx 10^{-400}$

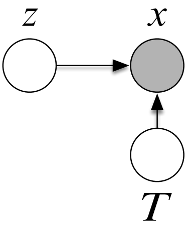
This is the digamma problem or the zero forcing problem.

Multicanonical Methods

- Annealing introduces a temperature sequence T_1 to T_t
- Results are very sensitive to the temperature schedule

Multicanonical Methods

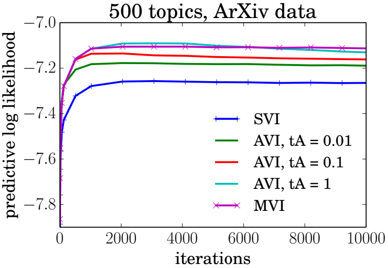
- Annealing introduces a temperature sequence T_1 to T_t
- Results are very sensitive to the temperature schedule
- Solution: Make T part of the model

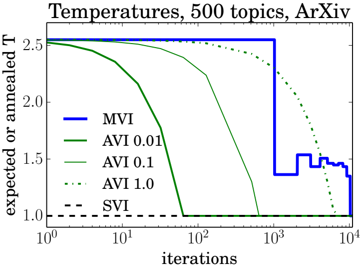


- Place a multinomial prior on T
- Variational update needs $Z(T)$

This trades a sequence of parameters settings for integral computation to renormalize.

Results





Similar results on factorial mixtures

Questions

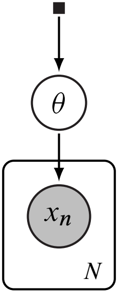
- Is this procedure always better? Adding latent variables can introduce new optima?
- More generally, what are automated model transforms that preserve model semantics while improving computation?

Problem: Automation

A lot of variational inference methods are black box, but what happens if we try to develop them in a programming framework?

Specifying a Model in Stan

$\alpha = 1.5, \sigma = 1$



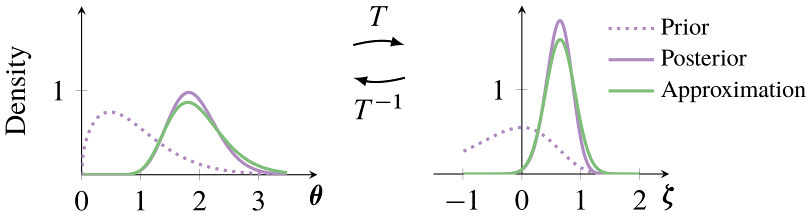

```
data {
  int N;    // number of observations
  int x[N]; // discrete-valued observations
}
parameters {
  // latent variable, must be positive
  real<lower=0> theta;
}
model {
  // non-conjugate prior for latent variable
  theta ~ weibull(1.5, 1);

  // likelihood
  for (n in 1:N)
    x[n] ~ poisson(theta);
}
```

Automatic Differentiation Variational Inference (ADVI)

What does it work for? Differentiable models where the posterior has same support as the prior

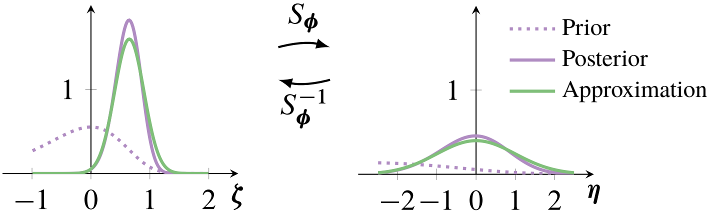
How does it work?



(a) Latent variable space **(b)** Real coordinate space

Posit a factorized normal approximation on this space

How does it work?



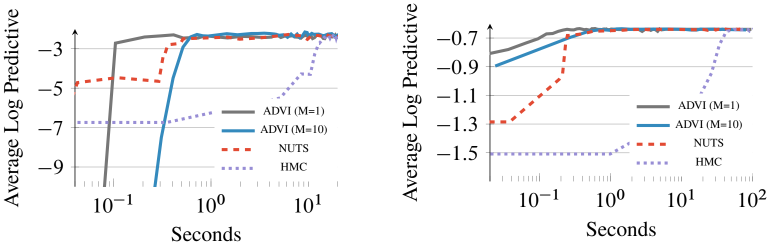
(a) Real coordinate space **(b)** Standardized space

How does it work?

Use Monte Carlo estimate reparameterization gradient to optimize the ELBO

$$\nabla_\lambda \mathcal{L}(\lambda) = \mathbb{E}_{s(\epsilon)}[\nabla_z[\log p(x, z)]\nabla_\lambda z(\epsilon)] + \nabla_\lambda H[q]$$

ADVI: Does it work?



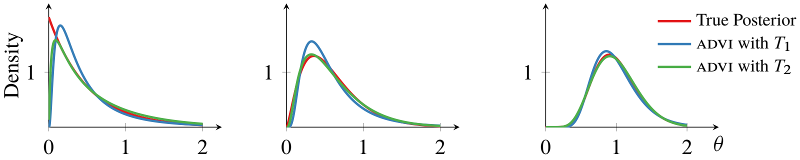
(a) Linear Regression with ARD

(b) Hierarchical Logistic Regression

The Role of Transforms

There exist multiple maps from the constrained to the unconstrained space.

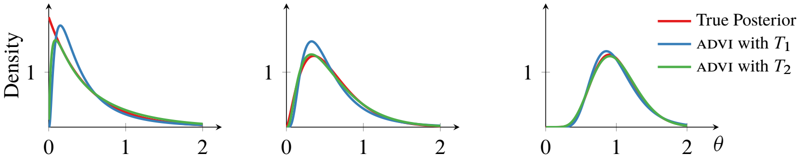
- For example from: $\mathbb{R}_+ \to \mathbb{R}$
- $T1 : \log(x)$ and $T2 : \log(\exp(x) - 1)$



(a) Gamma(1, 2) **(b)** Gamma(2.5, 4.2) **(c)** Gamma(10, 10)

There exist multiple maps from the constrained to the unconstrained space.

- For example from: $\mathbb{R}_+ \rightarrow \mathbb{R}$
- $T1 : \log(x)$ and $T2 : \log(\exp(x) - 1)$

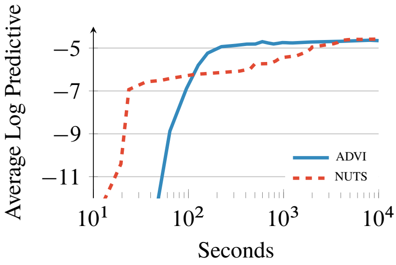


(a) Gamma(1, 2) **(b)** Gamma(2.5, 4.2) **(c)** Gamma(10, 10)

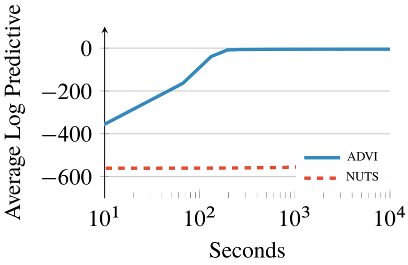
- The optimal transform can be written as $\phi^{-1}(P(z))$

What's the value of automation?

Studying multiple models



(a) Gamma Poisson Predictive Likelihood



(b) Dirichlet Exponential Predictive Likelihood

(c) Gamma Poisson Factors

(d) Dirichlet Exponential Factors

Questions

- Can you learn to initialize from the Stan program?
- Is there a lightweight way to choose hyperparameters?
- Can we expand the class of models to say where the posterior support doesn't match the prior?

Problem: Accuracy

Consider the model

$$y_t \sim \mathcal{N}(0, \exp(h_t/2))$$

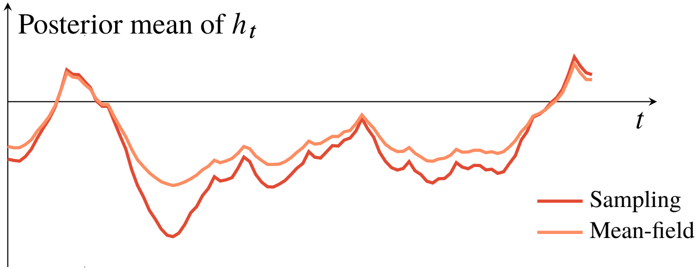
where the volatility itself follows an auto-regressive process

$$h_t \sim \mathcal{N}(\mu + \phi(h_{t-1} - \mu), \sigma) \quad \text{with initialization} \quad h_1 \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{1-\phi^2}}).$$

We posit the following priors for the latent variables

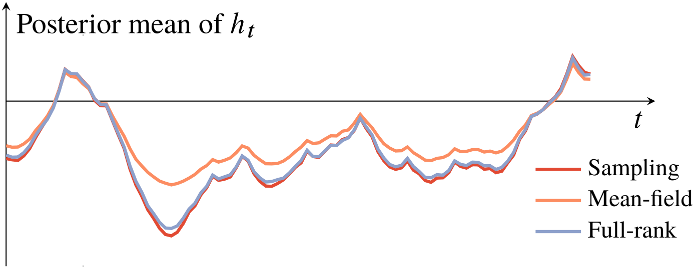
$$\mu \sim \text{Cauchy}(0, 10), \quad \phi \sim \text{Unif}(-1, 1), \quad \text{and} \quad \sigma \sim \text{LogNormal}(0, 10).$$

Mean-Field Variational Bayes



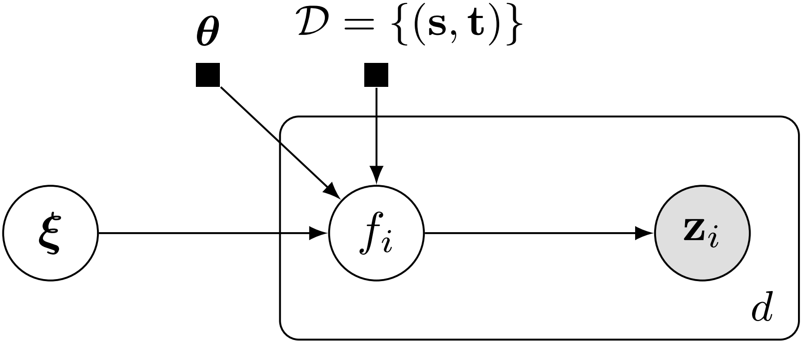
Posit a richer approximation

Instead of a factorized normal, consider a multivariate normal approximation on the unconstrained model.



Fewer iterations are needed with the un-factorized approximation.

Finding good variational distributions is modeling problem



$\xi \sim \text{Normal}(0, I),\ f_i \sim \text{GP}(0, K)|\mathcal{D}_i$

Questions

- Can you choose dependence based on the property of interest of the posterior?
- What are other distances between probability distributions amenable to finding good posterior approximations?

Thanks