



# True Natural Gradient of Collapsed Variational Bayes

Francisco J. R. Ruiz  
Neil D. Lawrence  
James Hensman

University of Sheffield and University Carlos III in Madrid

November 3, 2014

# Outline

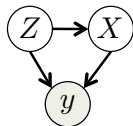
- 1 Review of variational methods
- 2 Our contribution: TNG
- 3 Experiments

## True Natural Gradient of Collapsed Variational Bayes

True Natural Gradient  
of Collapsed Variational Bayes

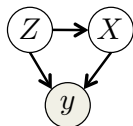
# Variational Bayes

- Notation:
  - $X$  and  $Z$ : Latent variables
  - $y$ : Observations



# Variational Bayes

- Notation:
  - $X$  and  $Z$ : Latent variables
  - $y$ : Observations

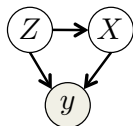


- The posterior is intractable:

$$p(X, Z|y) = \frac{p(y|X, Z)p(X|Z)p(Z)}{p(y)} = \frac{p(y|X, Z)p(X|Z)p(Z)}{\int p(y|X, Z)p(X|Z)p(Z)dXdZ}$$

# Variational Bayes

- Notation:
  - $X$  and  $Z$ : Latent variables
  - $y$ : Observations



- The posterior is intractable:

$$p(X, Z|y) = \frac{p(y|X, Z)p(X|Z)p(Z)}{p(y)} = \frac{p(y|X, Z)p(X|Z)p(Z)}{\int p(y|X, Z)p(X|Z)p(Z)dXdZ}$$

- Assumptions:
  - The model is conditionally conjugate (semi-conjugate)
  - $p(y|X, Z)$ ,  $p(X|Z)$  and  $p(Z)$  in the exponential family

# Variational Bayes

- Models within this class:
  - Bayesian mixture models
  - Latent Dirichlet allocation (LDA)
  - HMMs
  - Factorial models
  - Probabilistic factor analysis/matrix factorization models
  - Bayesian nonparametric mixture models
  - ...

# Variational Bayes

- VB: Approximate the posterior with  $q(X, Z)$ 
  - Minimize  $\text{KL}(q(X, Z) || p(X, Z|y))$

# Variational Bayes

- VB: Approximate the posterior with  $q(X, Z)$ 
  - Minimize  $\text{KL}(q(X, Z) || p(X, Z|y))$
  - Equivalent to maximize  $\mathcal{L}$

$$\mathcal{L} \triangleq \mathbb{E}_q [\log p(y, X, Z)] - \mathbb{E}_q [\log q(X, Z)] \leq p(y)$$

# Variational Bayes

- VB: Approximate the posterior with  $q(X, Z)$ 
  - Minimize  $\text{KL}(q(X, Z) || p(X, Z|y))$
  - Equivalent to maximize  $\mathcal{L}$

$$\mathcal{L} \triangleq \mathbb{E}_q [\log p(y, X, Z)] - \mathbb{E}_q [\log q(X, Z)] \leq p(y)$$

- VBEM: Independence assumption:

$$q(X, Z) = q(X)q(Z)$$

# Variational Bayes

- VB: Approximate the posterior with  $q(X, Z)$ 
  - Minimize  $\text{KL}(q(X, Z) || p(X, Z|y))$
  - Equivalent to maximize  $\mathcal{L}$

$$\mathcal{L} \triangleq \mathbb{E}_q [\log p(y, X, Z)] - \mathbb{E}_q [\log q(X, Z)] \leq p(y)$$

- VBEM: Independence assumption:

$$q(X, Z) = q(X)q(Z)$$

- Two-step algorithm:
  - 1 Optimize  $q(Z)$ , holding  $q(X)$  fixed
  - 2 Optimize  $q(X)$ , holding  $q(Z)$  fixed

# Variational Bayes

- VB: Approximate the posterior with  $q(X, Z)$ 
  - Minimize  $\text{KL}(q(X, Z) || p(X, Z|y))$
  - Equivalent to maximize  $\mathcal{L}$

$$\mathcal{L} \triangleq \mathbb{E}_q [\log p(y, X, Z)] - \mathbb{E}_q [\log q(X, Z)] \leq p(y)$$

- VBEM: Independence assumption:

$$q(X, Z) = q(X)q(Z)$$

- Two-step algorithm:
  - ① Optimize  $q(Z)$ , holding  $q(X)$  fixed
  - ② Optimize  $q(X)$ , holding  $q(Z)$  fixed
- Both  $q(X)$  and  $q(Z)$  are in the exponential family
  - Natural parameters:  $\theta_x$  and  $\theta_z$

# Variational Bayes

- The problem is coordinate-wise convex:

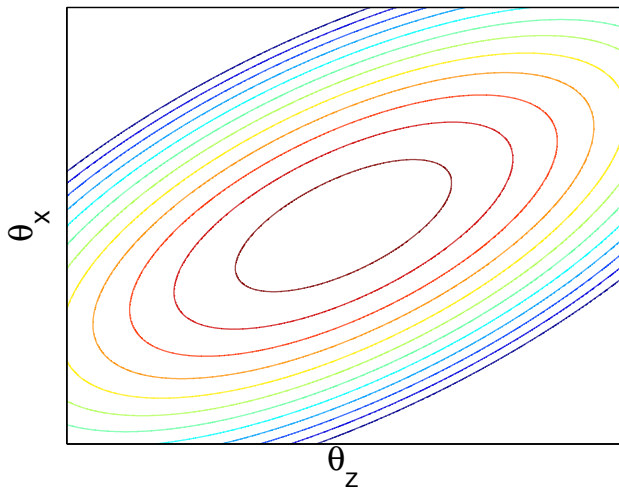
# Variational Bayes

- The problem is coordinate-wise convex:
  - $q^*(Z)$  is unique for a given  $q(X)$
  - $q^*(X)$  is unique for a given  $q(Z)$

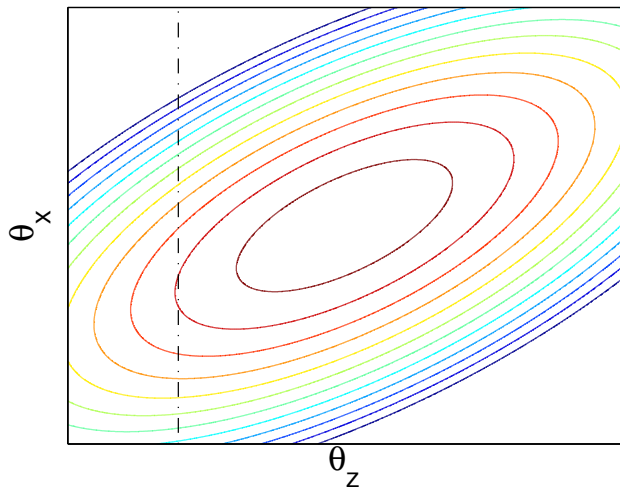
# Variational Bayes

- The problem is coordinate-wise convex:
  - $q^*(Z)$  is unique for a given  $q(X)$ 
    - $\theta_Z^*$  is unique for any given  $\theta_X$
  - $q^*(X)$  is unique for a given  $q(Z)$ 
    - $\theta_X^*$  is unique for any given  $\theta_Z$

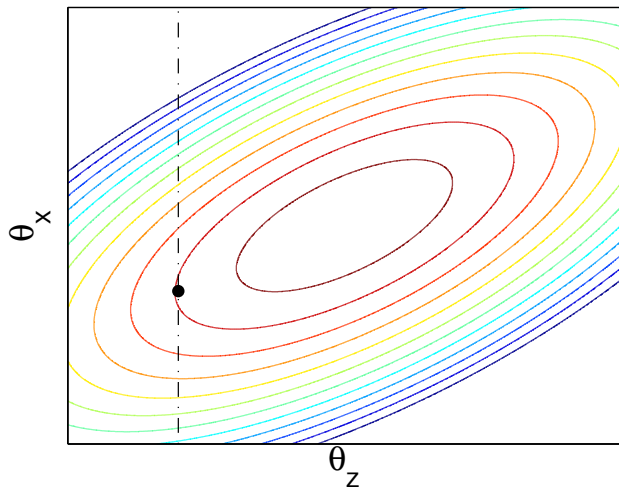
# Variational Bayes



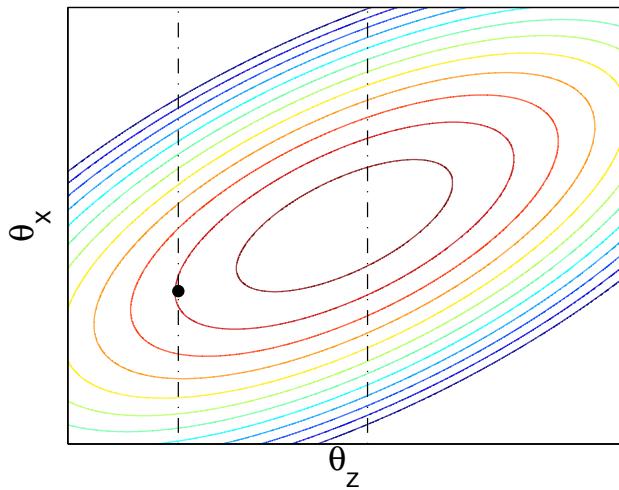
# Variational Bayes



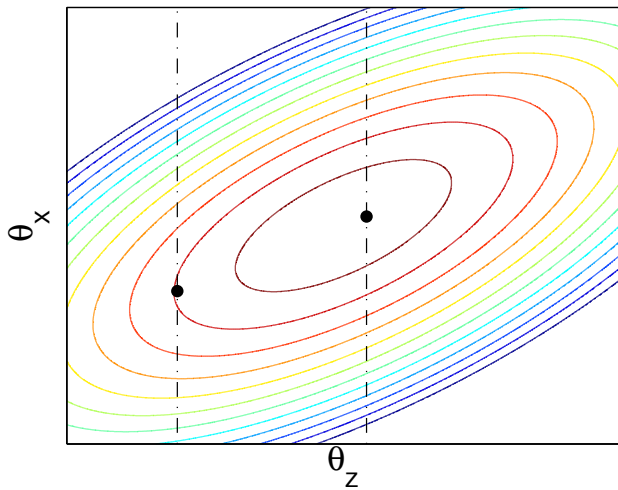
# Variational Bayes



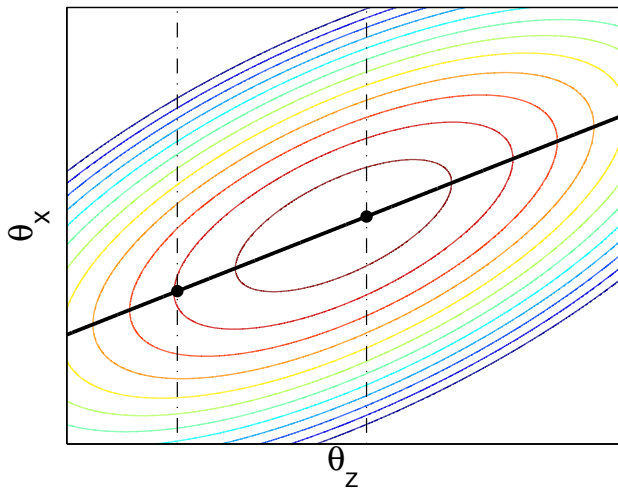
# Variational Bayes



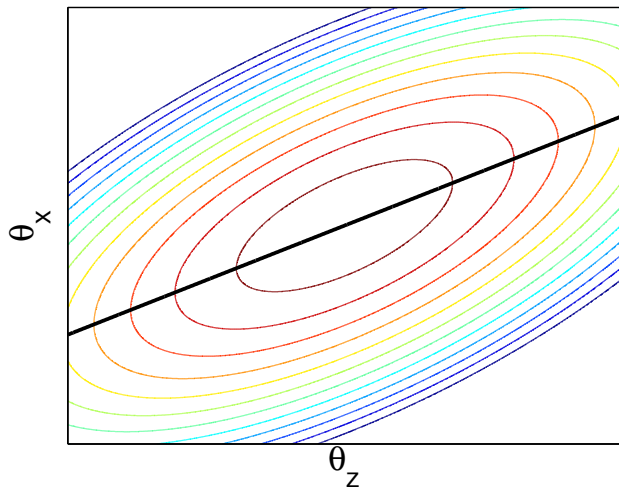
# Variational Bayes



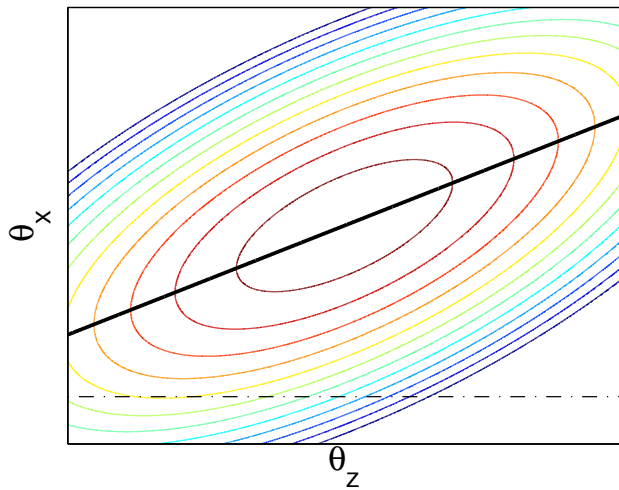
# Variational Bayes



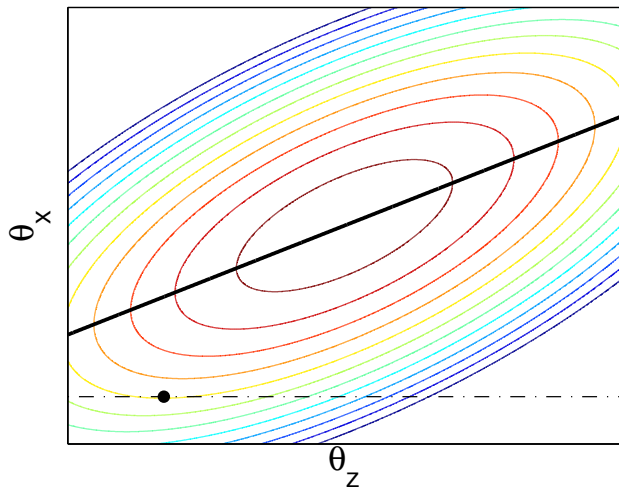
# Variational Bayes



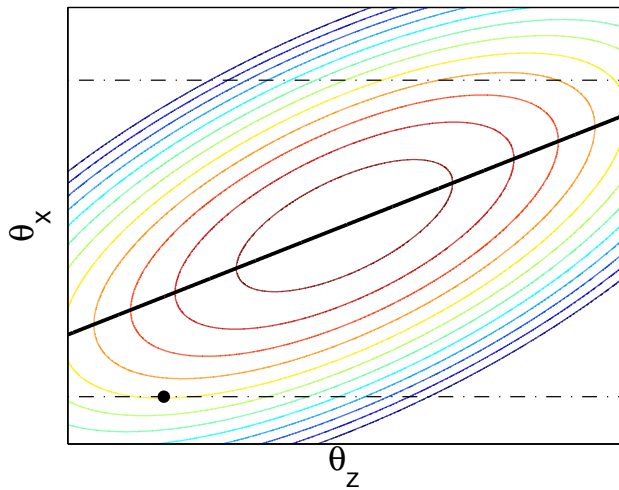
# Variational Bayes



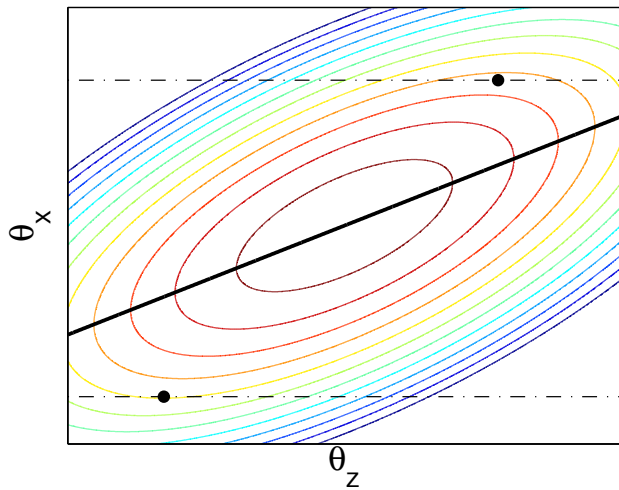
# Variational Bayes



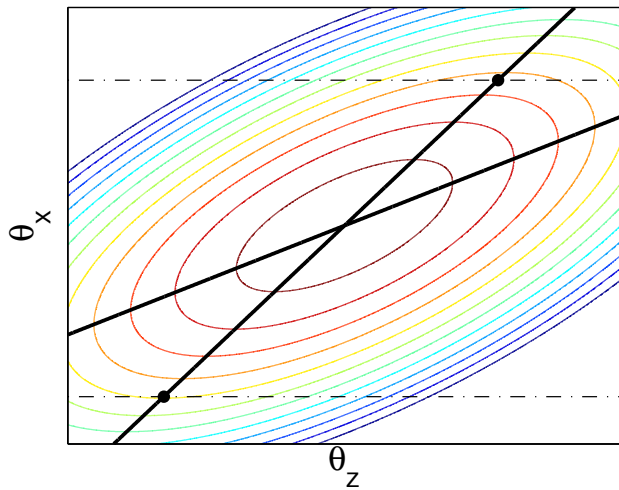
# Variational Bayes



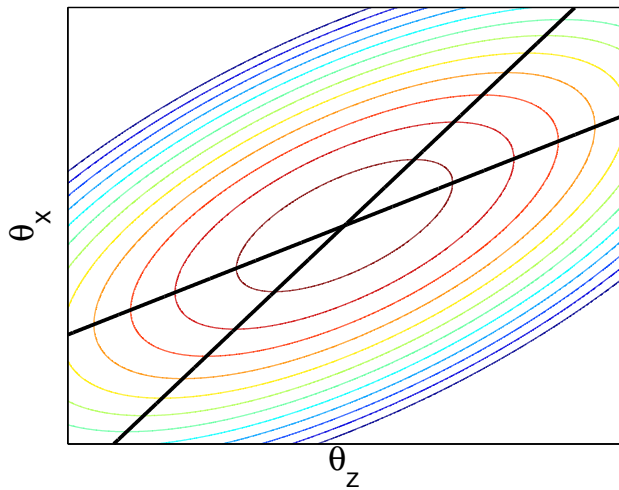
# Variational Bayes



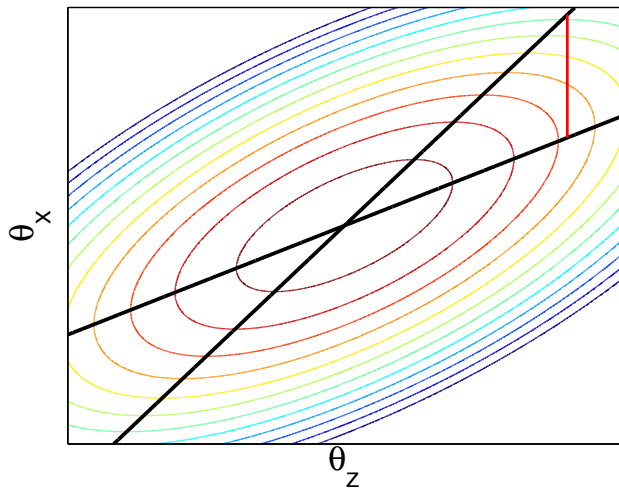
# Variational Bayes



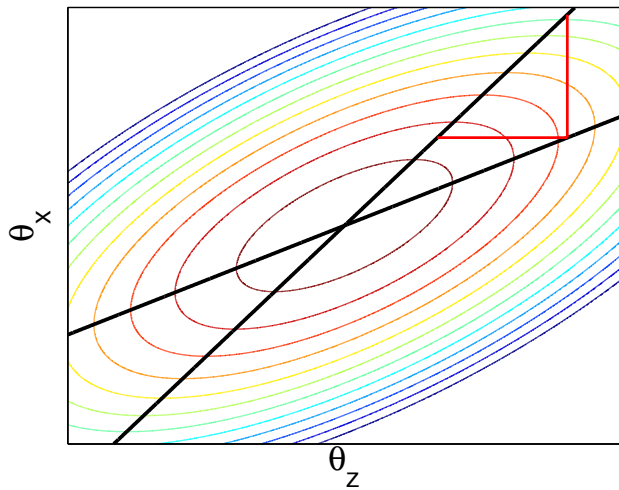
# Variational Bayes



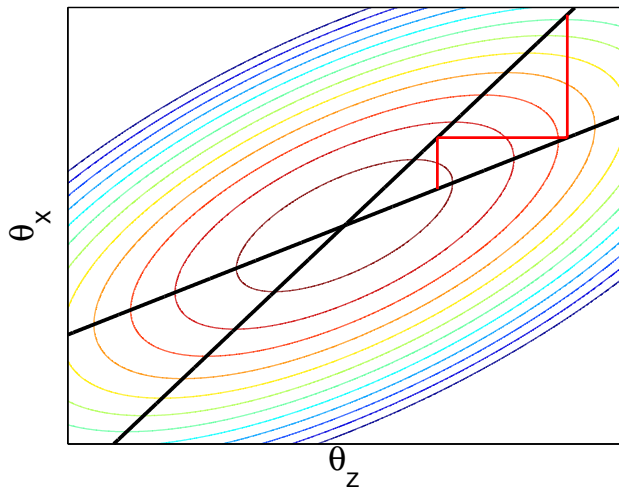
# Variational Bayes



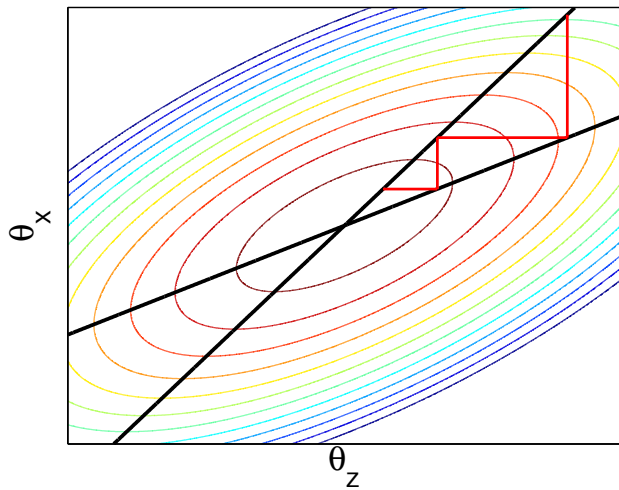
# Variational Bayes



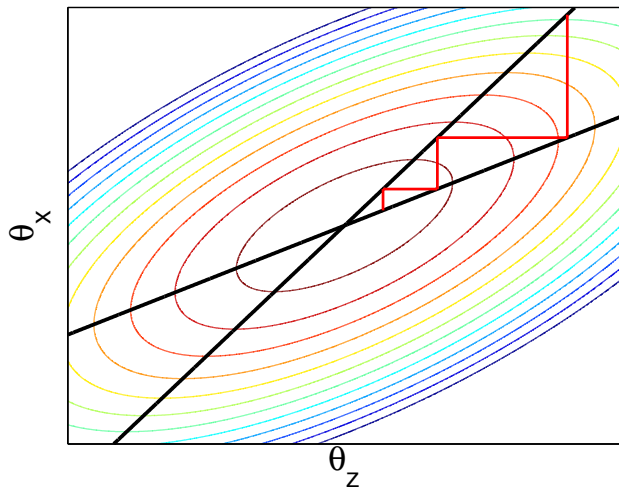
# Variational Bayes



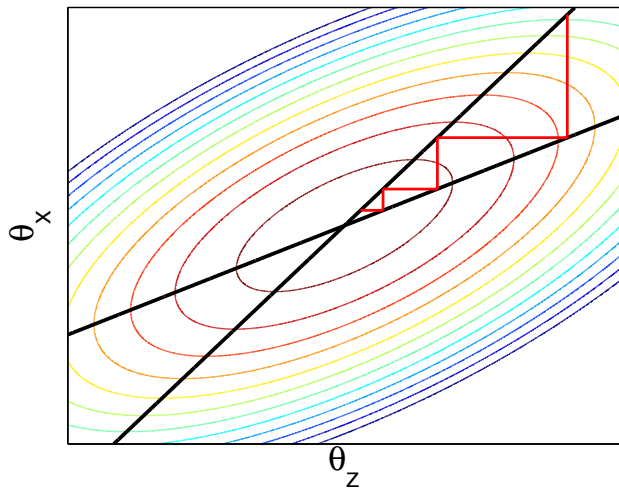
# Variational Bayes



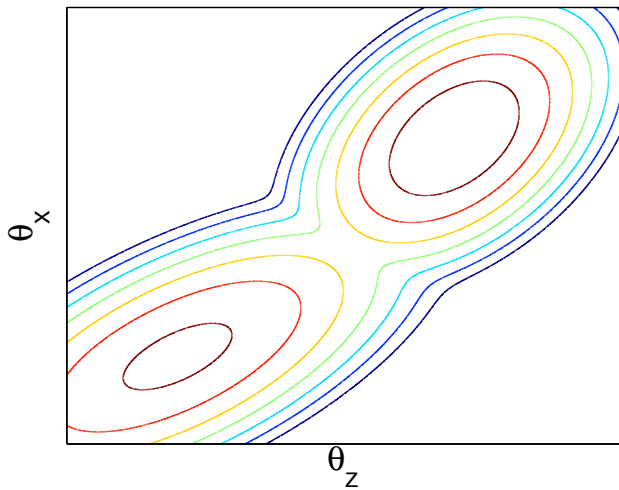
# Variational Bayes



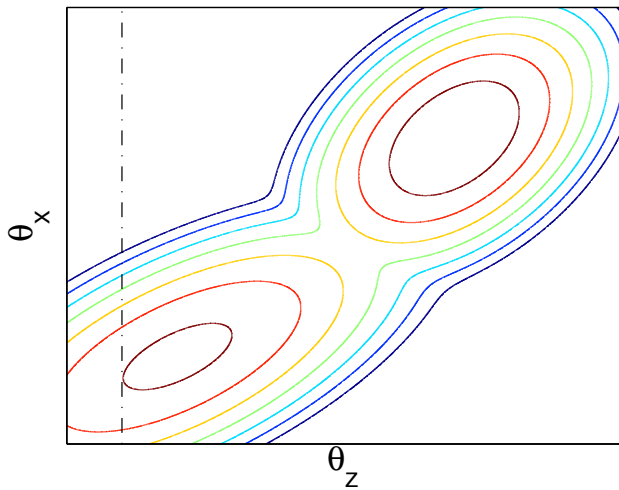
# Variational Bayes



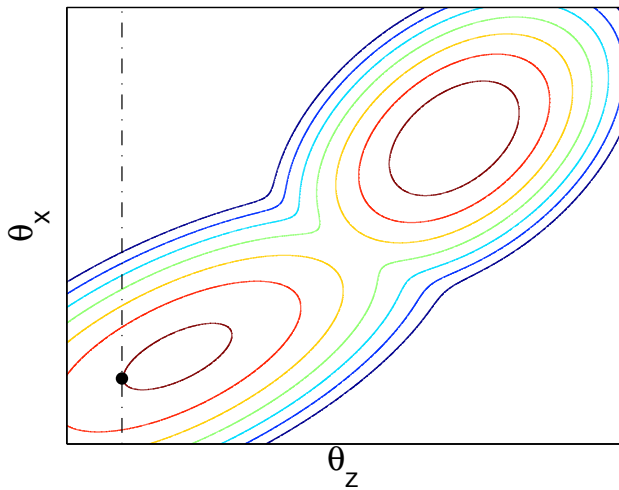
# Variational Bayes



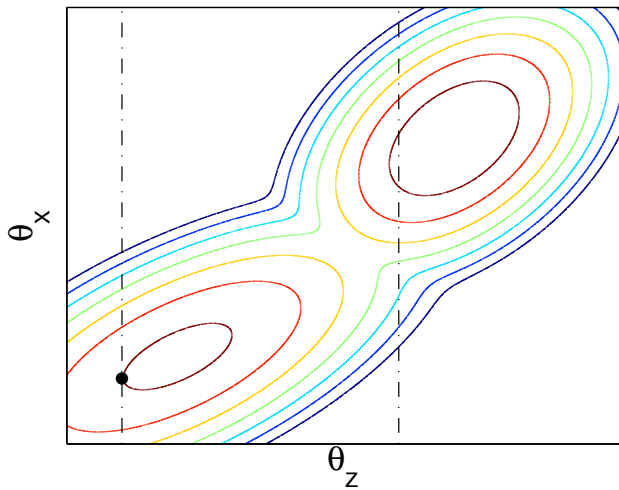
# Variational Bayes



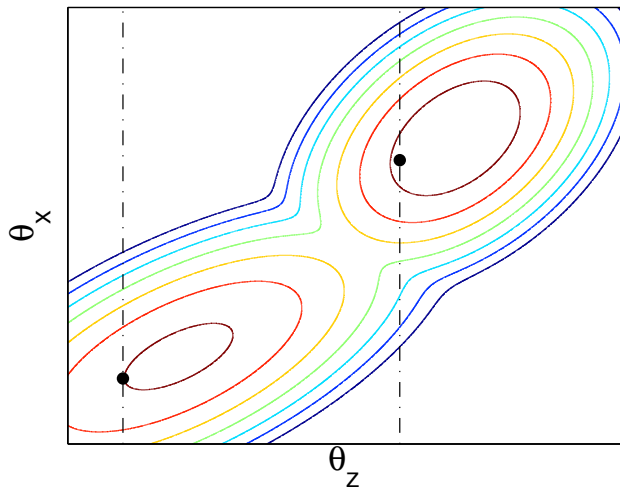
# Variational Bayes



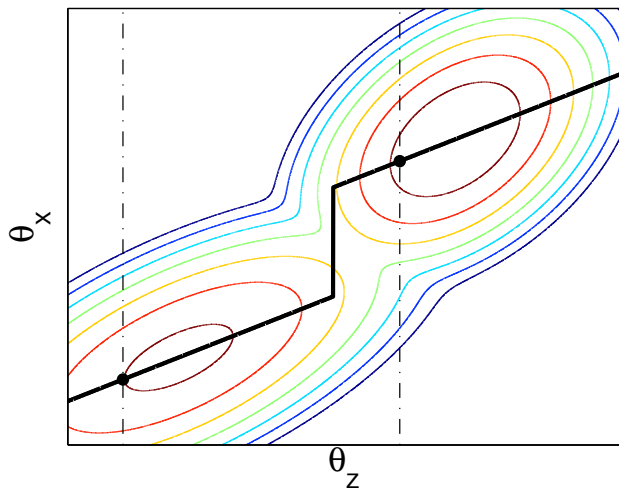
# Variational Bayes



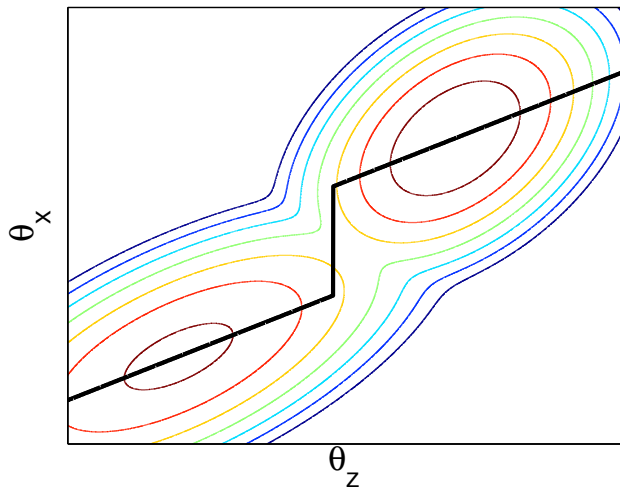
# Variational Bayes



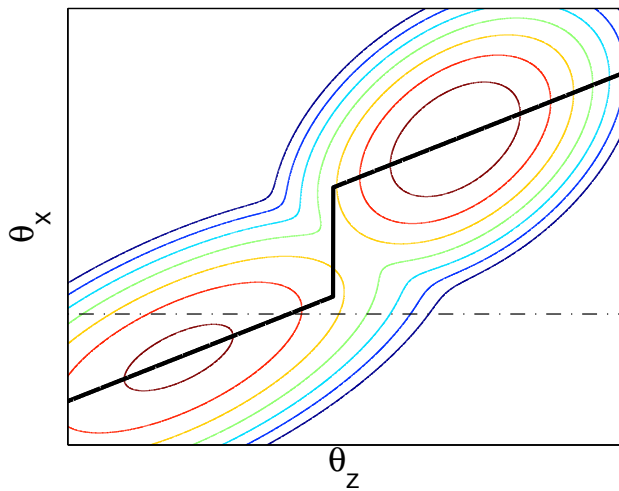
# Variational Bayes



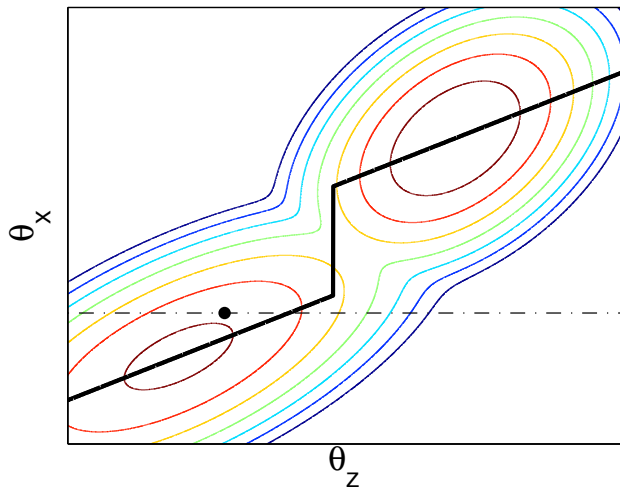
# Variational Bayes



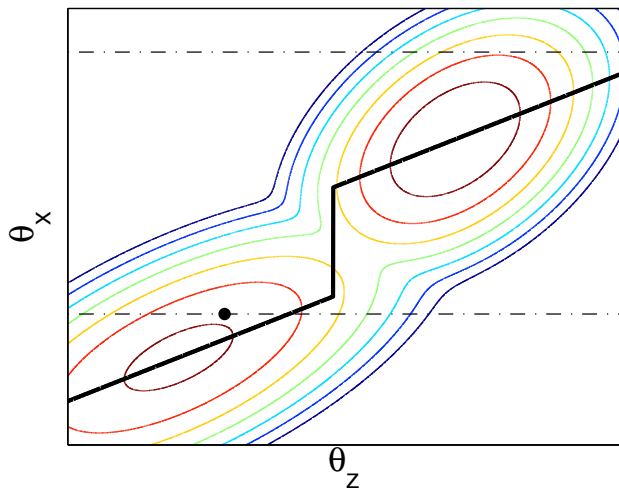
# Variational Bayes



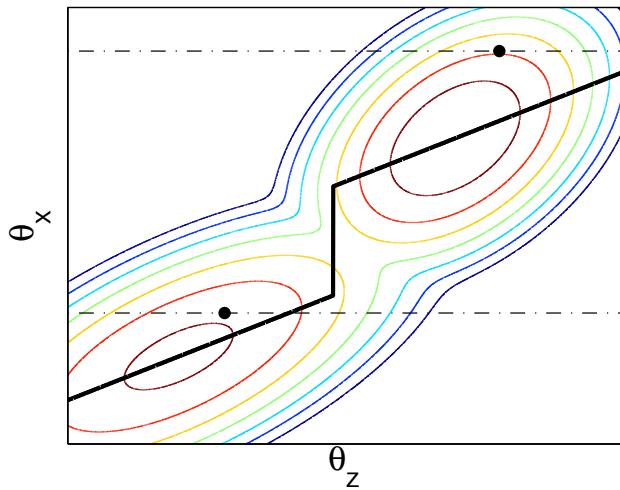
# Variational Bayes



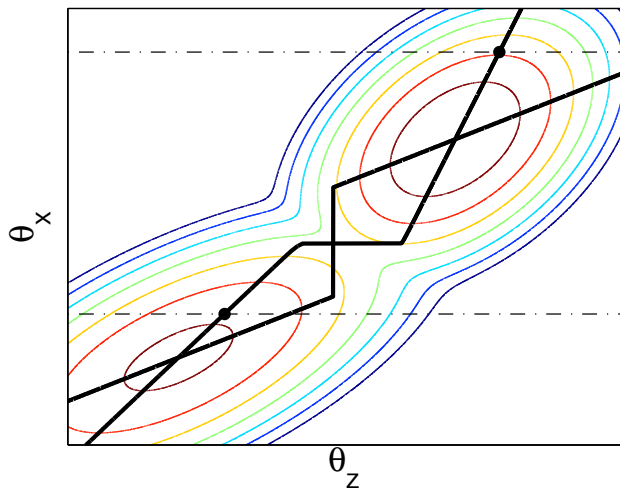
# Variational Bayes



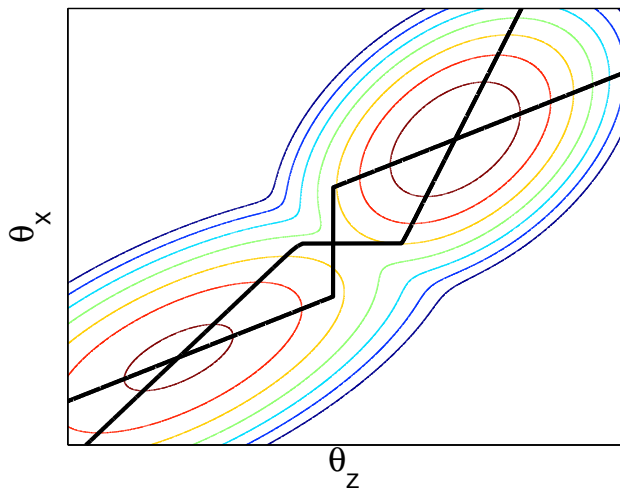
# Variational Bayes



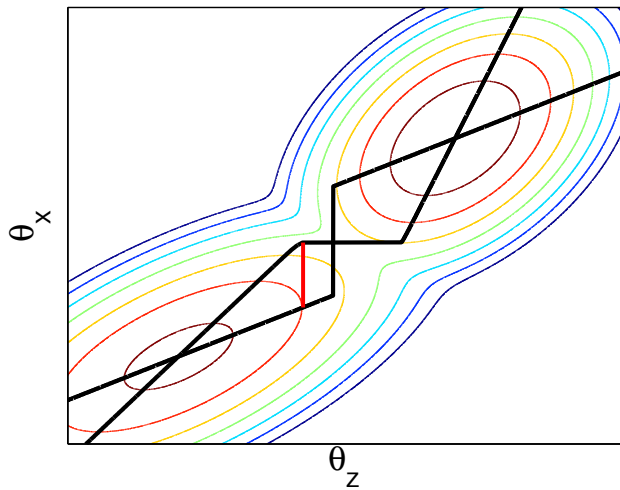
# Variational Bayes



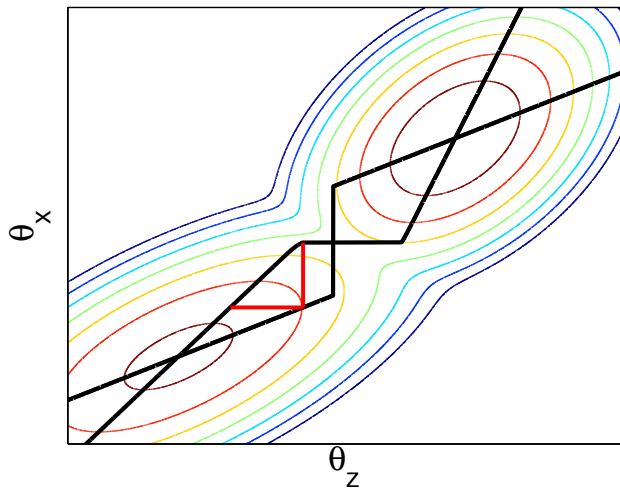
# Variational Bayes



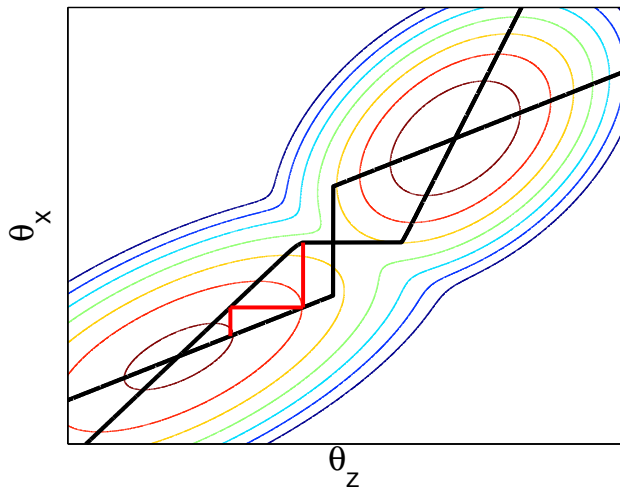
# Variational Bayes



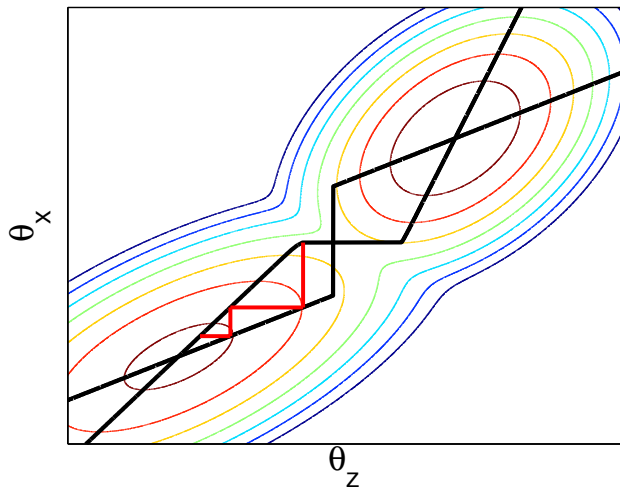
# Variational Bayes



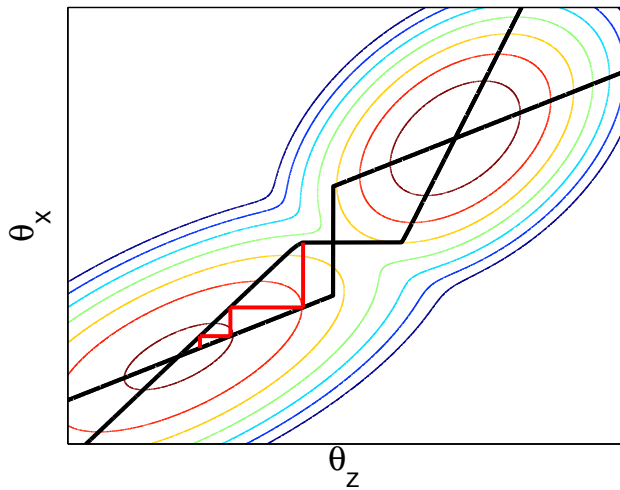
# Variational Bayes



# Variational Bayes



# Variational Bayes



## True Natural Gradient of Collapsed Variational Bayes

# Collapsed Variational Bayes

- We can obtain the optimal  $\theta_x^*$  for any fixed  $\theta_z \dots$

# Collapsed Variational Bayes

- We can obtain the optimal  $\theta_x^*$  for any fixed  $\theta_z \dots$
- $\dots$  Hence, there is no reason to keep both variables

# Collapsed Variational Bayes

- We can obtain the optimal  $\theta_x^*$  for any fixed  $\theta_z \dots$
- $\dots$  Hence, there is no reason to keep both variables
- We can optimize w.r.t.  $\theta_z$

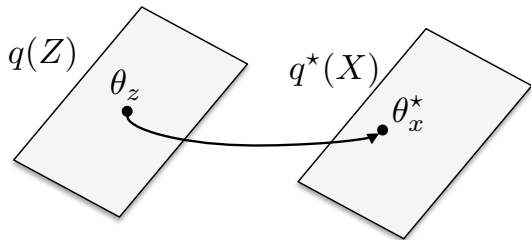
# Collapsed Variational Bayes

- We can obtain the optimal  $\theta_x^*$  for any fixed  $\theta_z \dots$
- $\dots$  Hence, there is no reason to keep both variables
- We can optimize w.r.t.  $\theta_z$
- A change in  $\theta_z$  affects  $q(Z)$  **and** the updated distribution  $q^*(X)$

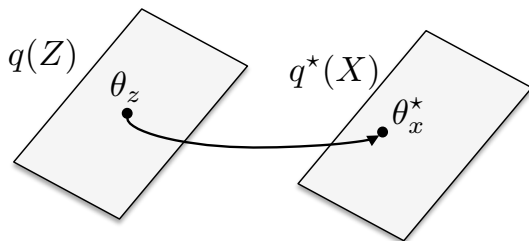
# Collapsed Variational Bayes

- We can obtain the optimal  $\theta_x^*$  for any fixed  $\theta_z \dots$
- $\dots$  Hence, there is no reason to keep both variables
- We can optimize w.r.t.  $\theta_z$
- A change in  $\theta_z$  affects  $q(Z)$  **and** the updated distribution  $q^*(X)$
- Equivalent to moving along one of the lines

# Collapsed Variational Bayes



# Collapsed Variational Bayes



- For a given  $q(Z)$ , the collapsed bound is

$$\mathcal{L}_{\text{KL}} = \mathcal{L} + \text{KL}(q^*(X) || q(X))$$

True **Natural Gradient**  
of Collapsed Variational Bayes

# Natural Gradient

- Gradient:

$$g = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \|d\theta_z\| \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

# Natural Gradient

- Gradient:

$$g = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \|d\theta_z\| \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

- The space of natural parameters is not Euclidean

# Natural Gradient

- Gradient:

$$g = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \|d\theta_z\| \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

- The space of natural parameters is not Euclidean
- $\|\theta_z^{(1)} - \theta_z^{(2)}\|$  is not a metric of the distance between  $q^{(1)}(Z)$  and  $q^{(2)}(Z)$

# Natural Gradient

- Gradient:

$$g = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \|d\theta_z\| \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

- The space of natural parameters is not Euclidean
- $\|\theta_z^{(1)} - \theta_z^{(2)}\|$  is not a metric of the distance between  $q^{(1)}(Z)$  and  $q^{(2)}(Z)$
- A reasonable metric is the symmetrized KL divergence

$$\text{KL}^{\text{sym}}(\theta_z^{(1)}, \theta_z^{(2)}) = \text{KL}(q^{(1)}(Z) \| q^{(2)}(Z)) + \text{KL}(q^{(2)}(Z) \| q^{(1)}(Z))$$

# Natural Gradient

- Gradient:

$$g = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \|d\theta_z\| \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

# Natural Gradient

- Gradient:

$$g = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \|d\theta_z\| \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

- Natural gradient:

$$\tilde{g} = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \text{KL}^{\text{sym}}(\theta_z, \theta_z + d\theta_z) \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

# Natural Gradient

- Gradient:

$$g = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \|d\theta_z\| \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

- Natural gradient:

$$\tilde{g} = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \text{KL}^{\text{sym}}(\theta_z, \theta_z + d\theta_z) \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

- Riemannian metric defined by the information geometry of the parameter space:

$$\tilde{g}(\theta_z) = \mathbf{F}_z^{-1}(\theta_z) \nabla_{\theta_z} \mathcal{L}$$

$$\tilde{g}(\theta_x) = \mathbf{F}_x^{-1}(\theta_x) \nabla_{\theta_x} \mathcal{L}$$

# Natural Gradient

- Gradient:

$$g = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \|d\theta_z\| \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

- Natural gradient:

$$\tilde{g} = \arg \max_{d\theta_z} \mathcal{L}(\theta_z + d\theta_z) \quad \text{s.t. } \text{KL}^{\text{sym}}(\theta_z, \theta_z + d\theta_z) \leq \epsilon \quad \text{as } \epsilon \rightarrow 0$$

- Riemannian metric defined by the information geometry of the parameter space:

$$\tilde{g}(\theta_z) = \mathbf{F}_z^{-1}(\theta_z) \nabla_{\theta_z} \mathcal{L}$$

$$\tilde{g}(\theta_x) = \mathbf{F}_x^{-1}(\theta_x) \nabla_{\theta_x} \mathcal{L}$$

- This works for the **uncollapsed** setting (VBEM)

True Natural Gradient  
of Collapsed Variational Bayes

# True Natural Gradient

- We only deal with one (set of) variables,  $\theta_z$
- Changing  $\theta_z$  affects  $q(Z)$  *and*  $q^*(X)$

# True Natural Gradient

- We only deal with one (set of) variables,  $\theta_z$
- Changing  $\theta_z$  affects  $q(Z)$  and  $q^*(X)$
- The “natural gradient”

$$\mathbf{F}_z(\theta_z) \triangleq -\mathbb{E}_{q(Z)} [\nabla \nabla \log q(Z)]$$
$$\tilde{g}(\theta_z) = \mathbf{F}_z^{-1}(\theta_z) \nabla_{\theta_z} \mathcal{L}_{\text{KL}}$$

only captures the information geometry of the space in which  $\theta_z$  lives

# True Natural Gradient

- We only deal with one (set of) variables,  $\theta_z$
- Changing  $\theta_z$  affects  $q(Z)$  and  $q^*(X)$
- The “natural gradient”

$$\mathbf{F}_z(\theta_z) \triangleq -\mathbb{E}_{q(Z)} [\nabla \nabla \log q(Z)]$$
$$\tilde{g}(\theta_z) = \mathbf{F}_z^{-1}(\theta_z) \nabla_{\theta_z} \mathcal{L}_{\text{KL}}$$

only captures the information geometry of the space in which  $\theta_z$  lives

- We define instead

$$\mathbf{F}_{\text{TNG}}(\theta_z) \triangleq -\mathbb{E}_{q(Z)q^*(X)} [\nabla \nabla \log (q(Z)q^*(X))]$$
$$\tilde{g}_{\text{TNG}}(\theta_z) = \mathbf{F}_{\text{TNG}}^{-1}(\theta_z) \nabla_{\theta_z} \mathcal{L}_{\text{KL}}$$

# True Natural Gradient

- Alternative expressions for the TNG:

$$\mathbf{F}_{\text{TNG}}(\theta_z) \triangleq -\mathbb{E}_{q(Z)q^*(X)} [\nabla \nabla \log (q(Z)q^*(X))]$$

# True Natural Gradient

- Alternative expressions for the TNG:

$$\mathbf{F}_{\text{TNG}}(\theta_z) \triangleq -\mathbb{E}_{q(Z)q^*(X)} [\nabla \nabla \log (q(Z)q^*(X))]$$

$$\mathbf{F}_{\text{TNG}}(\theta_z) = \nabla \nabla \mathcal{L}_{\text{KL}} + \mathbf{F}_z(\theta_z) - \mathbb{E}_{q^*(X)} [\nabla \nabla \mathcal{L}_1(X)]$$

# True Natural Gradient

- Alternative expressions for the TNG:

$$\mathbf{F}_{\text{TNG}}(\theta_z) \triangleq -\mathbb{E}_{q(Z)q^*(X)} [\nabla \nabla \log (q(Z)q^*(X))]$$

$$\mathbf{F}_{\text{TNG}}(\theta_z) = \nabla \nabla \mathcal{L}_{\text{KL}} + \mathbf{F}_z(\theta_z) - \mathbb{E}_{q^*(X)} [\nabla \nabla \mathcal{L}_1(X)]$$

$$\mathbf{F}_{\text{TNG}}(\theta_z) = \mathbf{F}_z(\theta_z) + \mathbf{B}(\theta_z) \mathbf{F}_x(\theta_x^*) \mathbf{B}^\top(\theta_z)$$

# Experiments

- Mixture of Gaussians with  $K$  components:
  - Normal inverse-Gamma prior over cluster means  $\mu_k$  and variances  $\sigma_k^2$
  - Dirichlet prior over cluster proportions  $\pi = [\pi_1, \dots, \pi_K]$
  - The assignments  $\ell_n \sim \pi$
  - Given its assignment  $\ell_n$ , each observation is  $y_n | \ell_n \sim \mathcal{N}(y_n | \mu_{\ell_n}, \sigma_{\ell_n}^2 \mathbf{I})$

# Experiments

- Mixture of Gaussians with  $K$  components:
  - Normal inverse-Gamma prior over cluster means  $\mu_k$  and variances  $\sigma_k^2$
  - Dirichlet prior over cluster proportions  $\pi = [\pi_1, \dots, \pi_K]$
  - The assignments  $\ell_n \sim \pi$
  - Given its assignment  $\ell_n$ , each observation is  $y_n | \ell_n \sim \mathcal{N}(y_n | \mu_{\ell_n}, \sigma_{\ell_n}^2 \mathbf{I})$
- We choose  $Z = \{\ell_n\}$ ,  $X = \{\mu_k, \sigma_k^2, \pi\}$

# Experiments

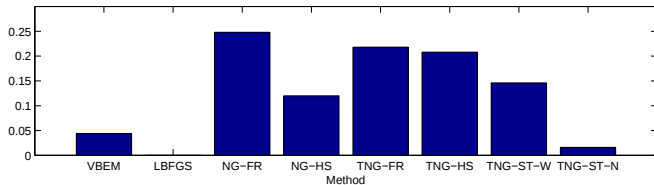
- Mixture of Gaussians with  $K$  components:
  - Normal inverse-Gamma prior over cluster means  $\mu_k$  and variances  $\sigma_k^2$
  - Dirichlet prior over cluster proportions  $\pi = [\pi_1, \dots, \pi_K]$
  - The assignments  $\ell_n \sim \pi$
  - Given its assignment  $\ell_n$ , each observation is  $y_n | \ell_n \sim \mathcal{N}(y_n | \mu_{\ell_n}, \sigma_{\ell_n}^2 \mathbf{I})$
- We choose  $Z = \{\ell_n\}$ ,  $X = \{\mu_k, \sigma_k^2, \pi\}$
- Experiments over the “LIBRAS” dataset
  - Available in the UCI repository
  - $K = 15$  clusters,  $N = 360$  observations,  $D = 90$  dimensions

# Experiments

- 500 initializations of 8 algorithms

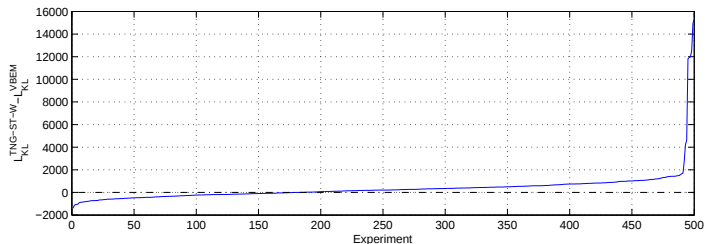
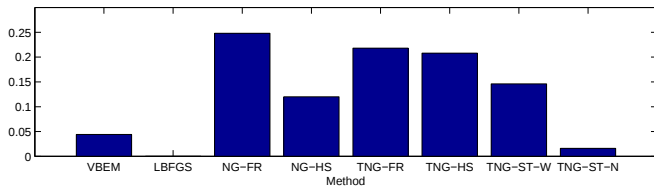
# Experiments

- 500 initializations of 8 algorithms



# Experiments

- 500 initializations of 8 algorithms



# Bonus Slides: Experiments on LDA

## Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

## Documents

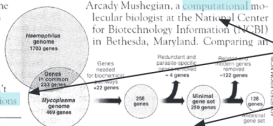
### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Anderson at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic** numbers game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

## Topic proportions and assignments

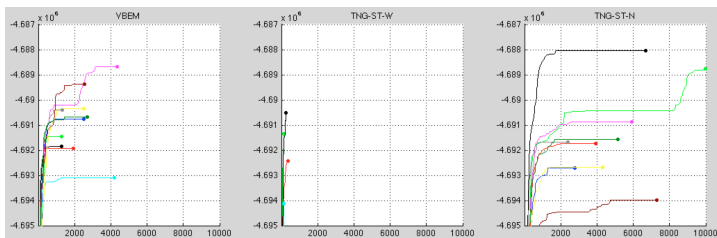


## Bonus Slides: Experiments on LDA

- NIPS 2011 dataset
- 305 documents, 5715 vocabulary words, 612508 words in total
- $K = 150$  topics

# Bonus Slides: Experiments on LDA

- NIPS 2011 dataset
- 305 documents, 5715 vocabulary words, 612508 words in total
- $K = 150$  topics



- Simulations are currently running!

# Bonus Slides: Experiments on LDA

- #1: data, learning, set, algorithm, function, model, number, problem, results
- #2: image, object, images, features, training, classes, model, class, segmentation
- #3: model, neurons, neural, spike, neuron, time, models, parameters, state
- #4: learning, theorem, bound, convex, rate, log, convergence, kernel, algorithm
- #5: matrix, sparse, problem, log, norm, matrices, recovery, rank, lasso
- #6: algorithm, regret, bound, loss, log, setting, problem, bandit, time
- #7: model, latent, gaussian, posterior, noise, bayesian, likelihood, kernel, prior
- #8: policy, state, reward, function, reinforcement, action, learning, optimal, states
- #9: loss, classification, target, classifier, training, prediction, domain, source, feature
- #10: model, process, time, inference, distribution, state, markov, sampling, processes
- ...
- #17: topic, topics, models, model, lda, dirichlet, document, distribution, words

**Thank you for your attention**

