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

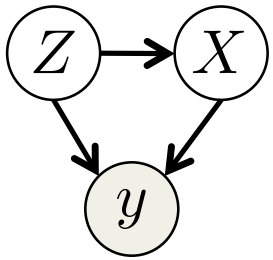
1. Review of variational methods
2. Our contribution: TNG
3. Experiments

True Natural Gradient of Collapsed Variational Bayes

True Natural Gradient
of Collapsed Variational Bayes

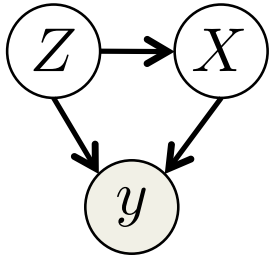
Variational Bayes

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



Variational Bayes

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

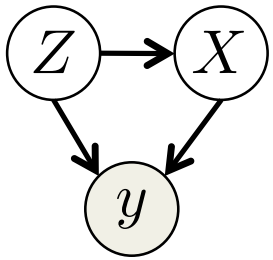


- The posterior is intractable:

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

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 $\mathrm{KL}\big(q(X, Z)||p(X, Z|y)\big)$

Variational Bayes

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

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

Variational Bayes

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

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

- VBEM: Independence assumption:

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

Variational Bayes

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

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

- 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 $\mathrm{KL}\big(q(X, Z)||p(X, Z|y)\big)$
  - Equivalent to maximize $\mathcal{L}$

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

- 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: θ_x and θ_z

Variational Bayes

- The problem is coordinate-wise convex:

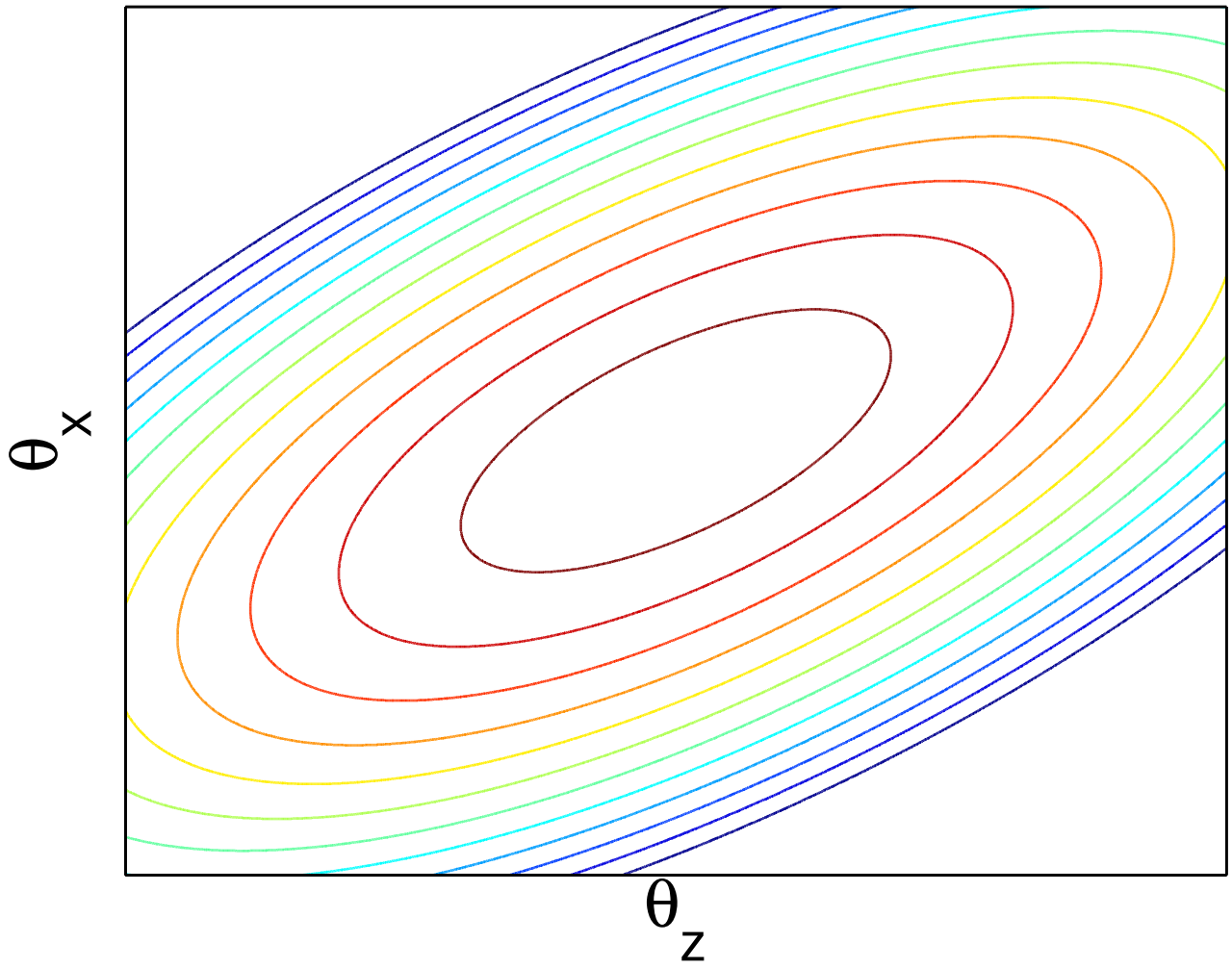
Variational Bayes

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

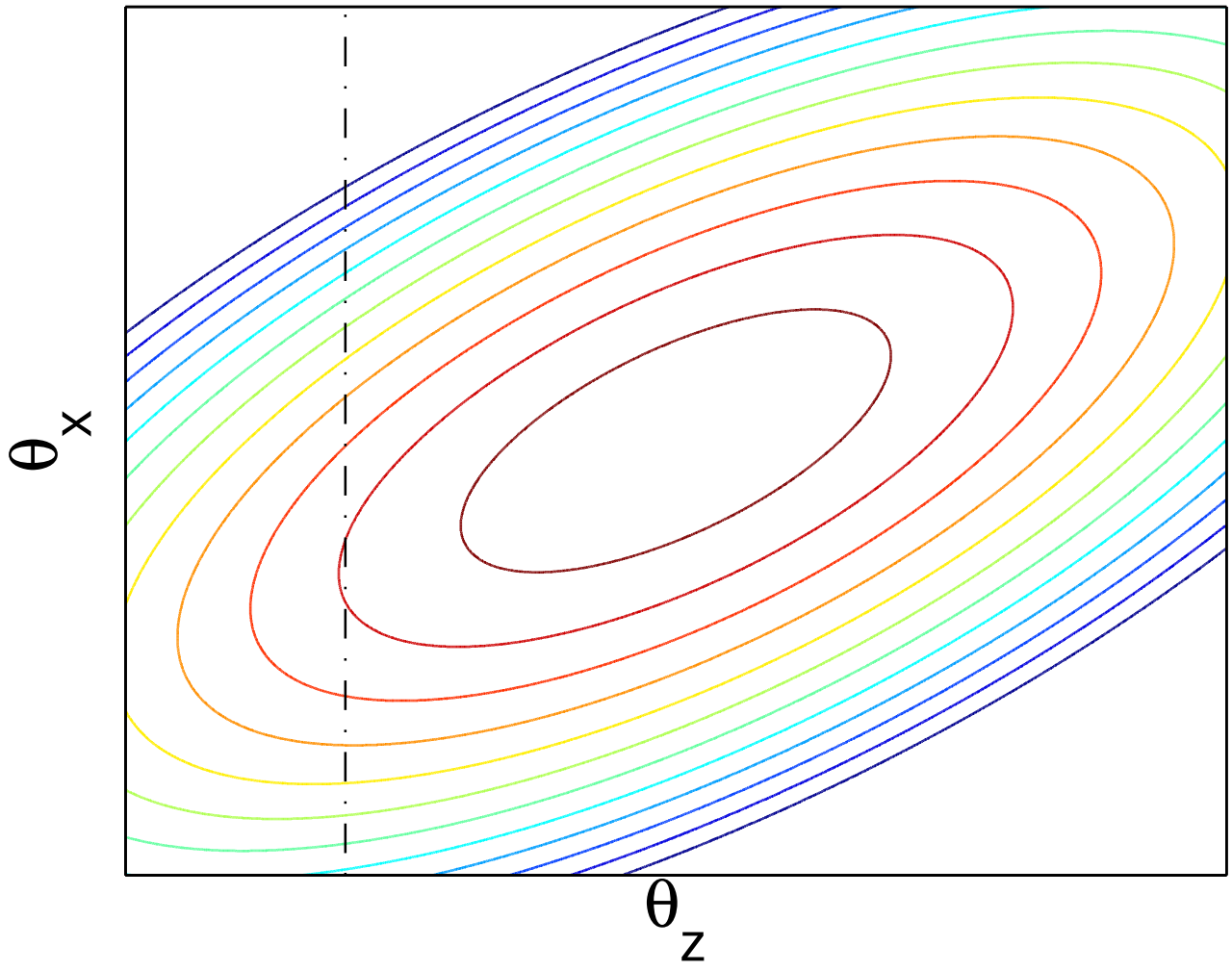
Variational Bayes

- The problem is coordinate-wise convex:
  - $q^\star(Z)$ is unique for a given $q(X)$
    - $\theta_z^\star$ is unique for any given θ_x
  - $q^\star(X)$ is unique for a given $q(Z)$
    - $\theta_x^\star$ is unique for any given θ_z

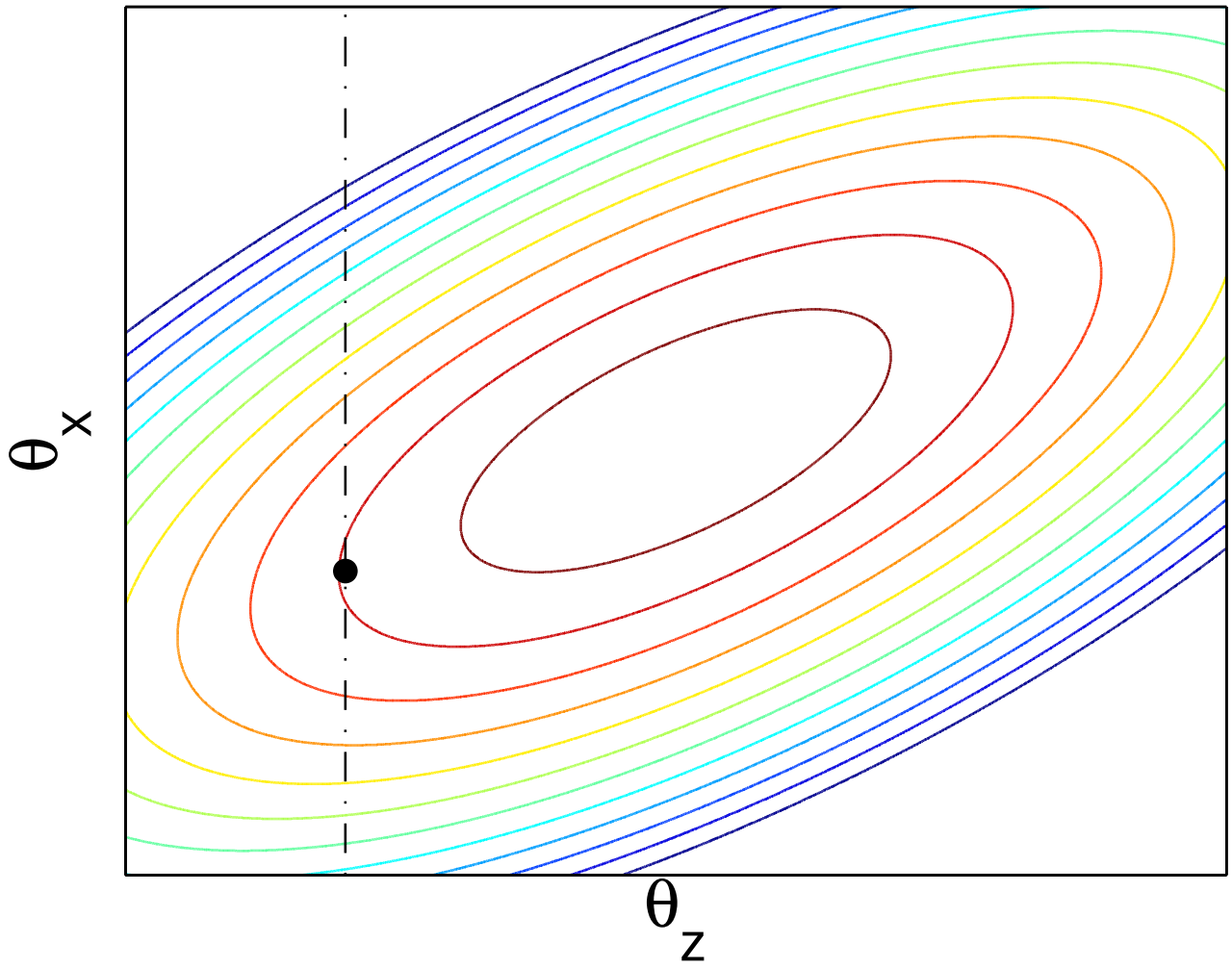
Variational Bayes



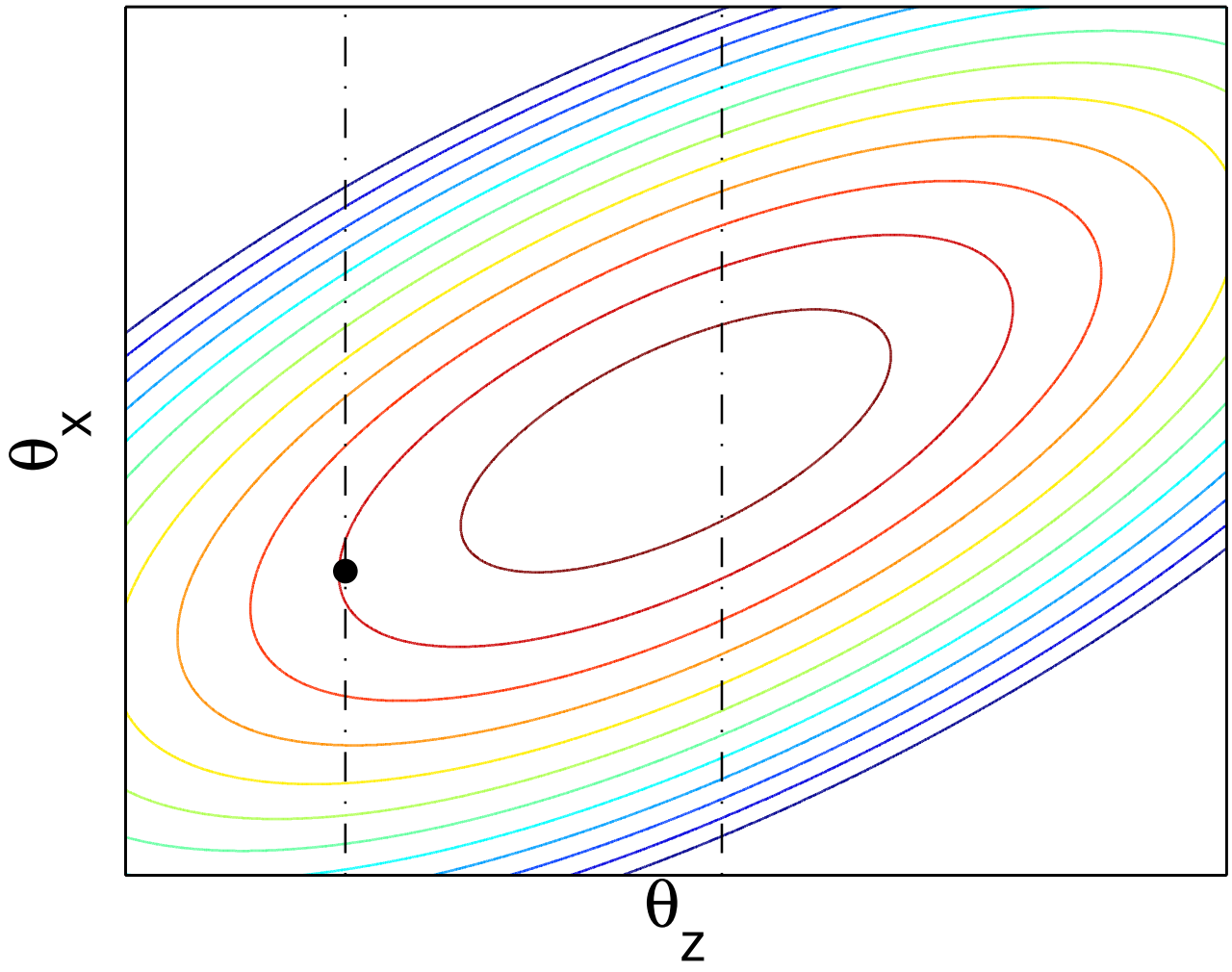
Variational Bayes



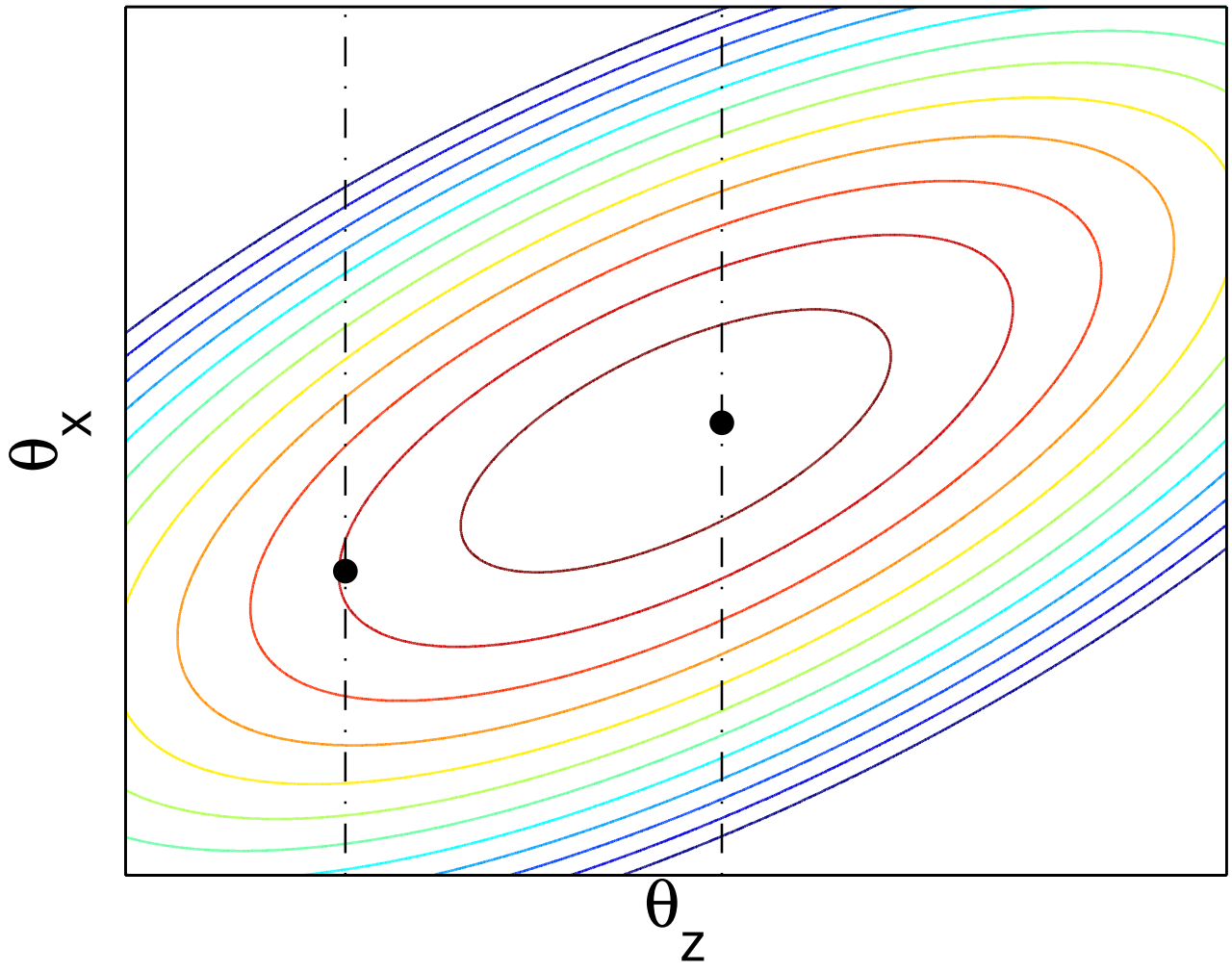
Variational Bayes



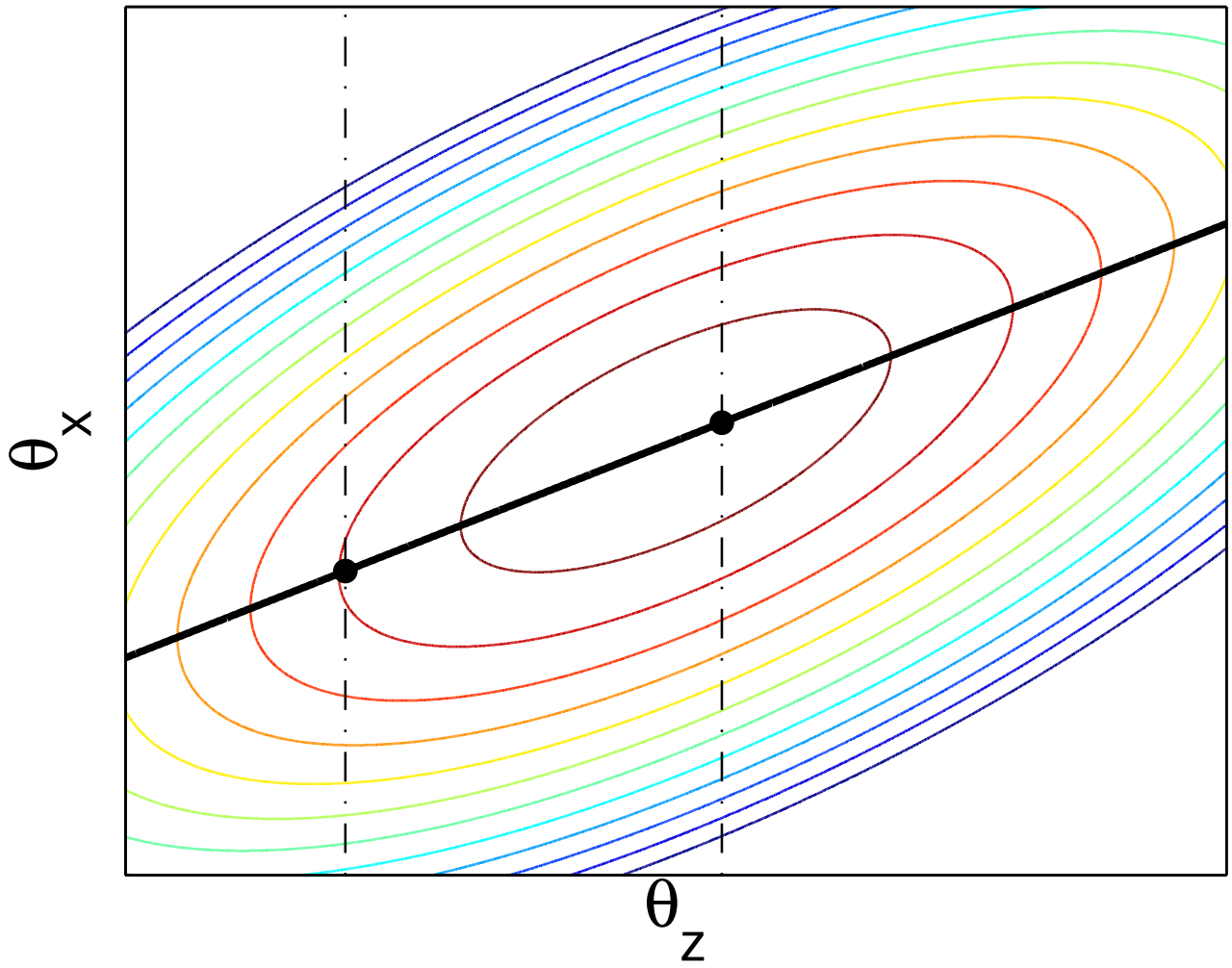
Variational Bayes



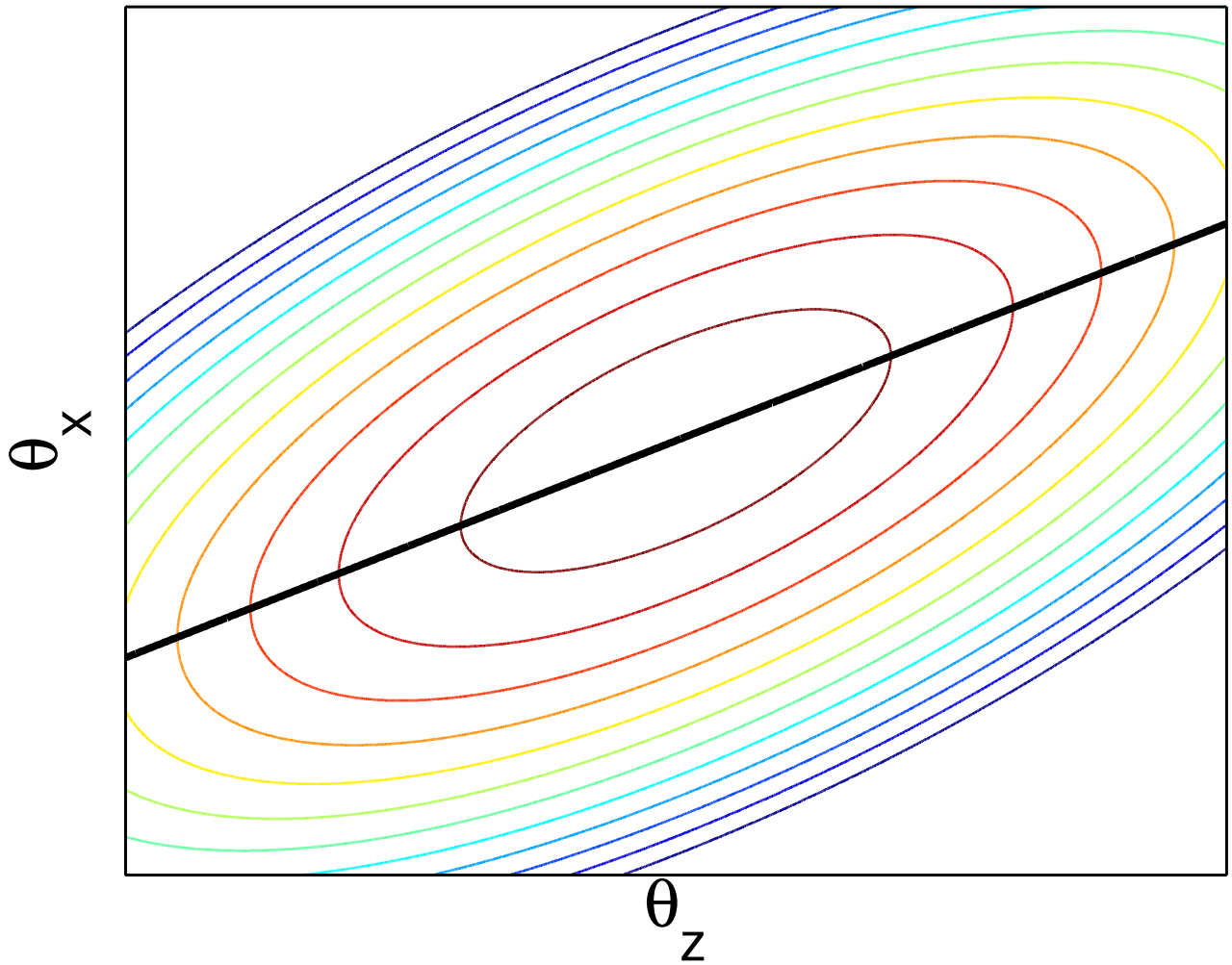
Variational Bayes



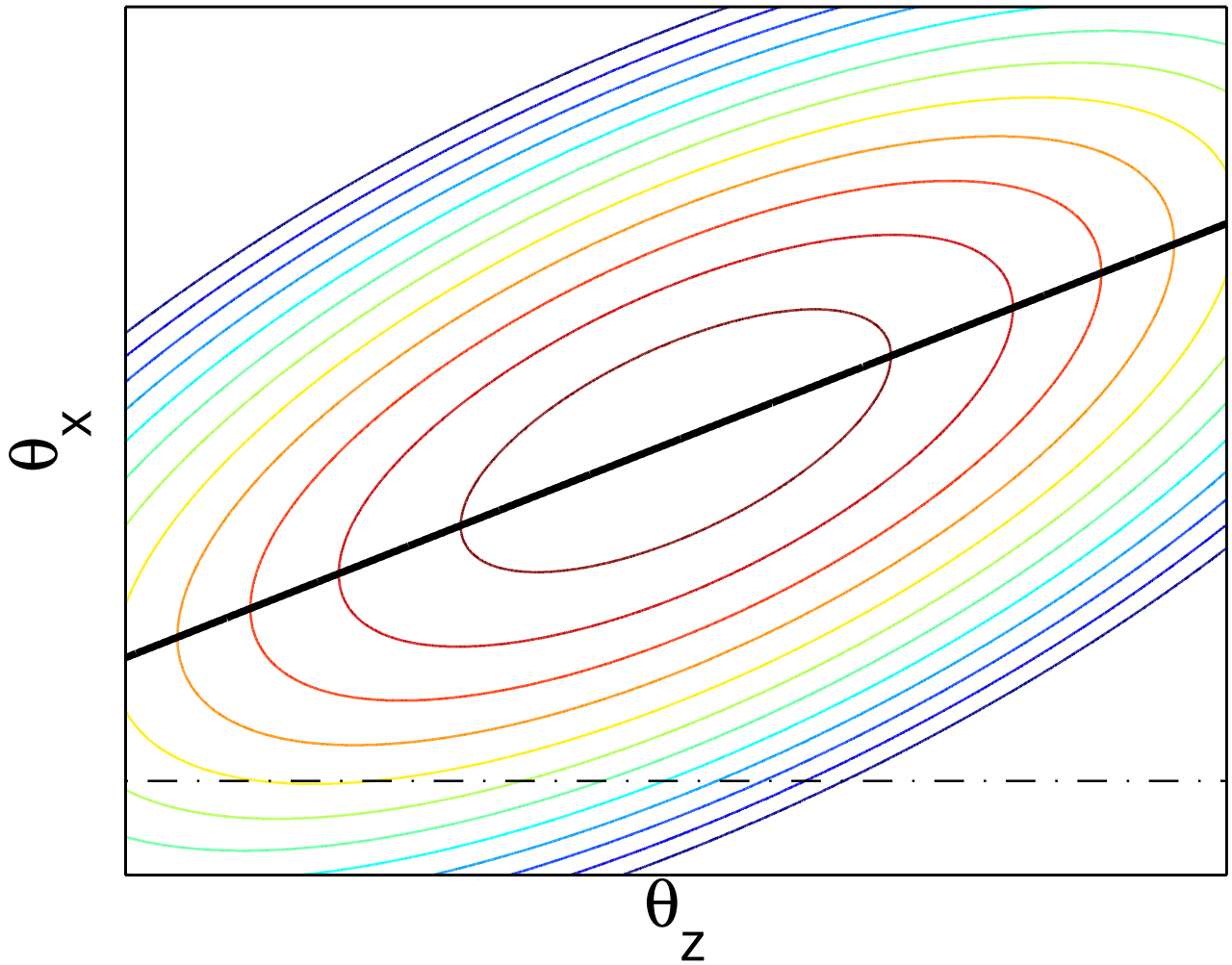
Variational Bayes



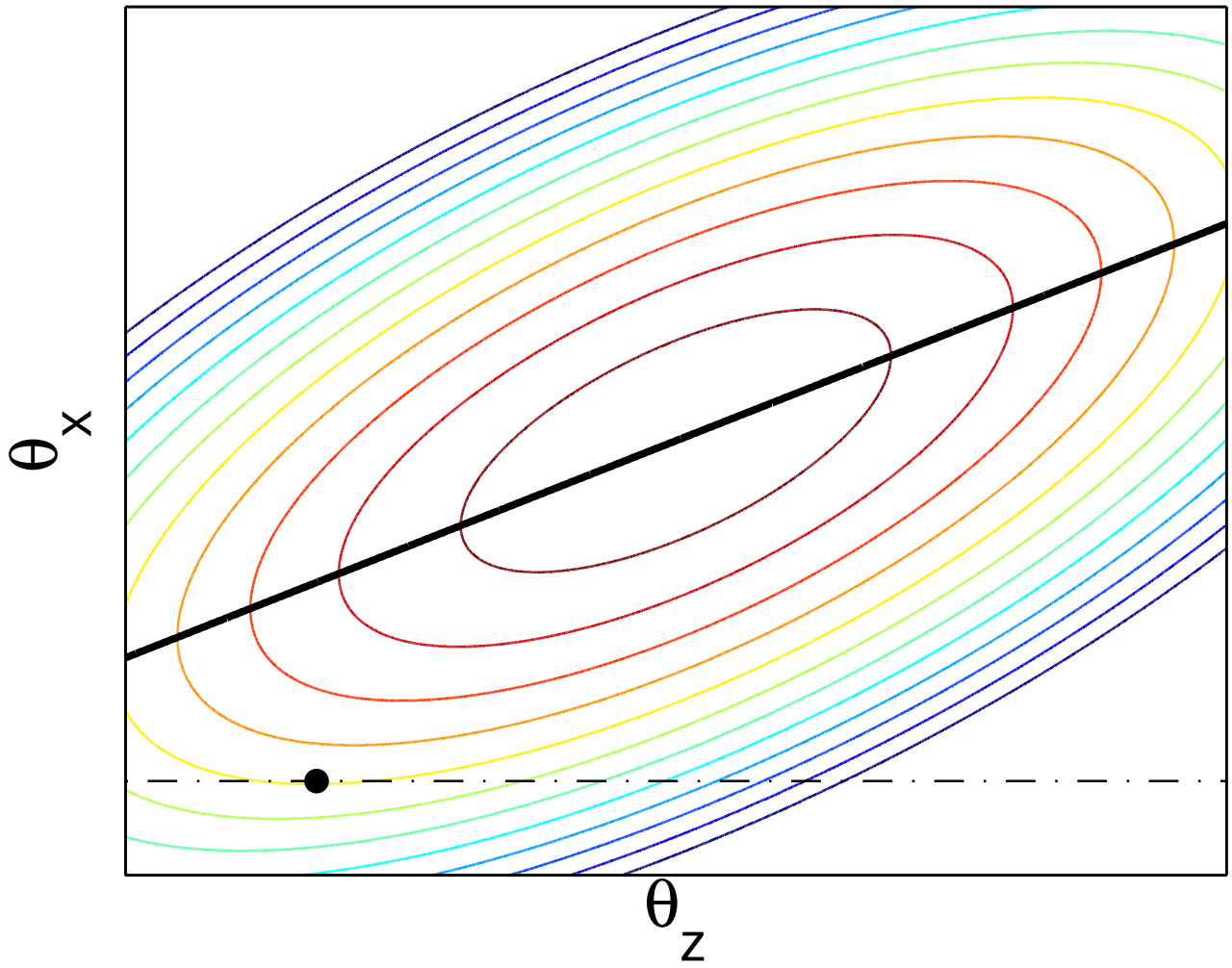
Variational Bayes



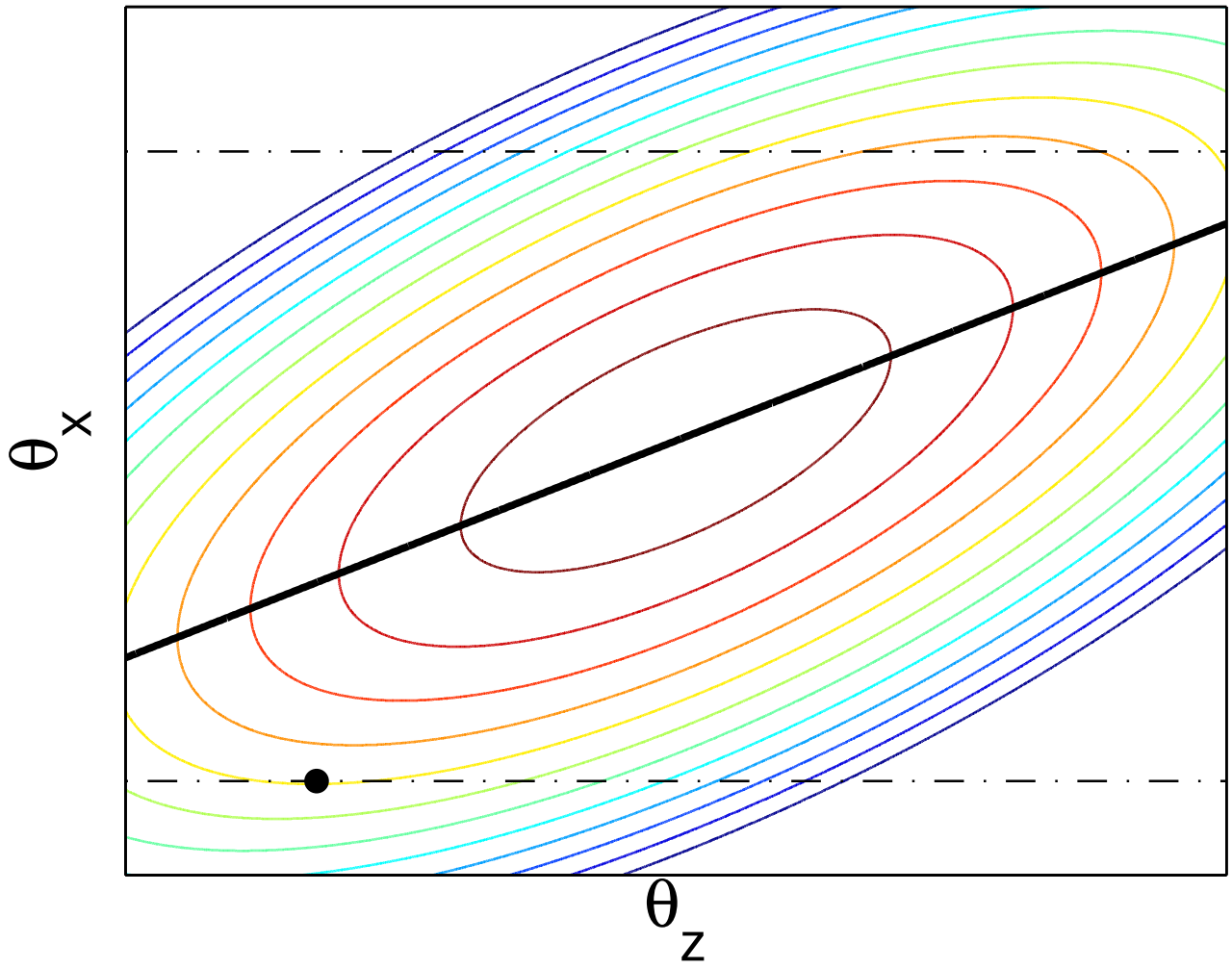
Variational Bayes



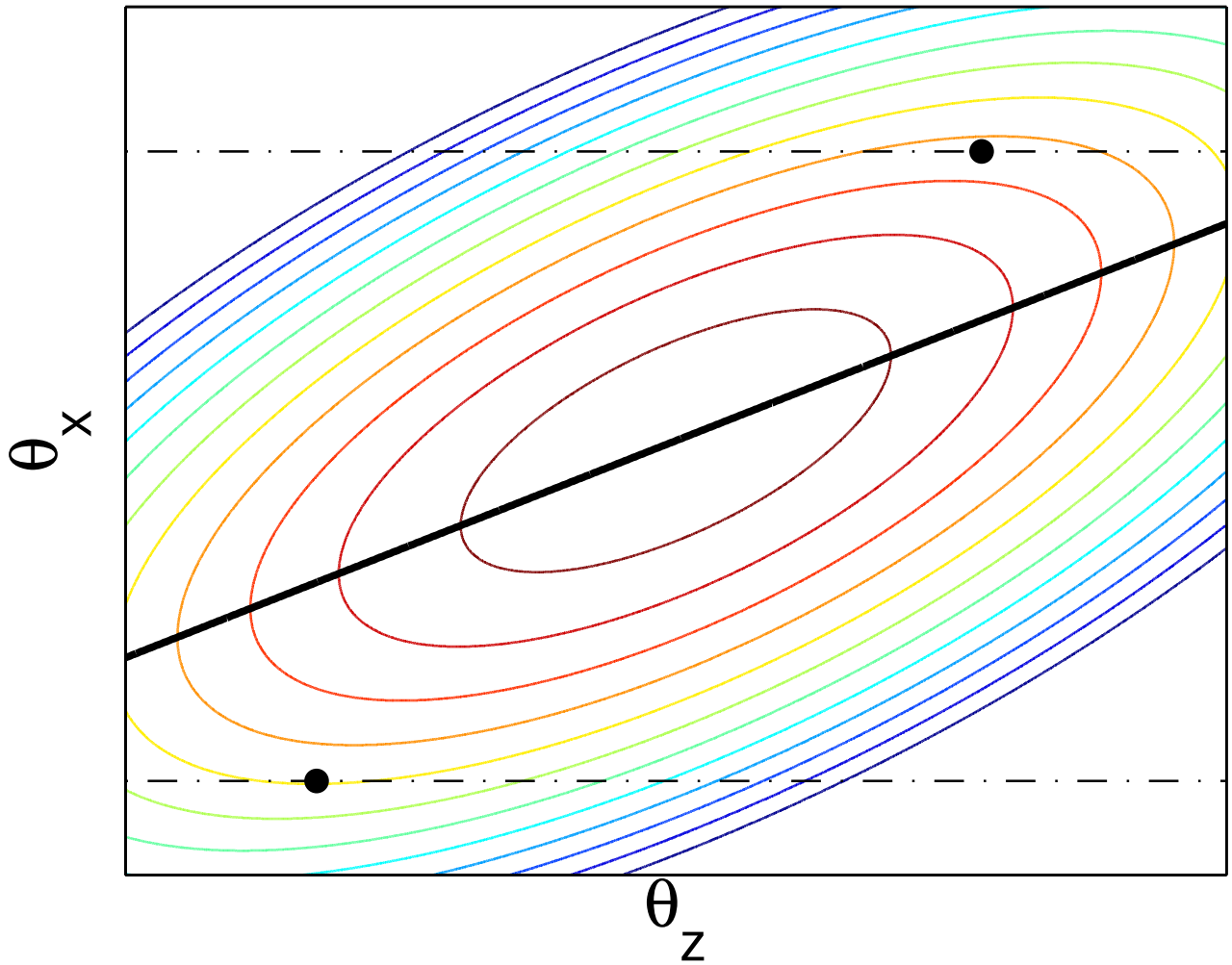
Variational Bayes



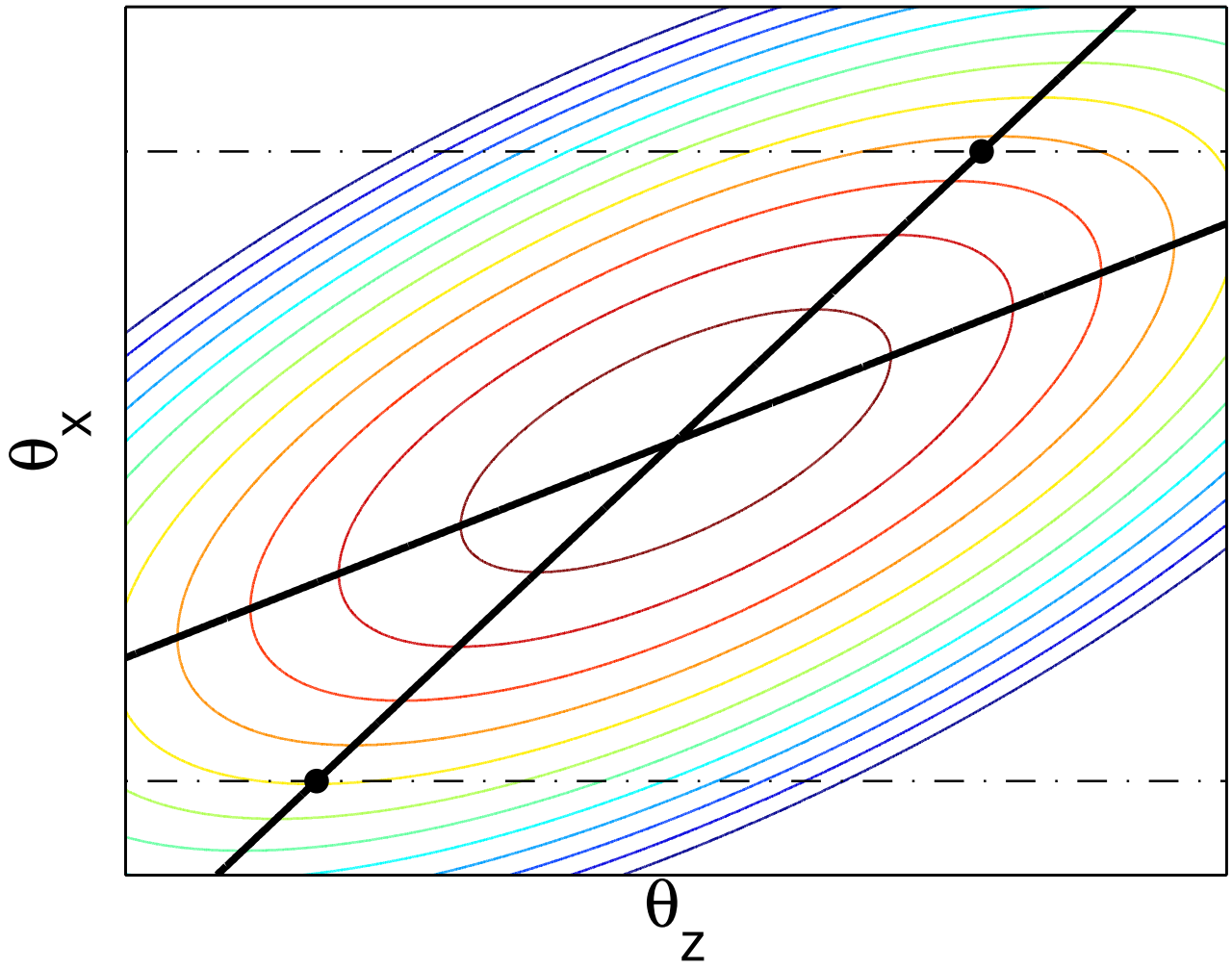
Variational Bayes



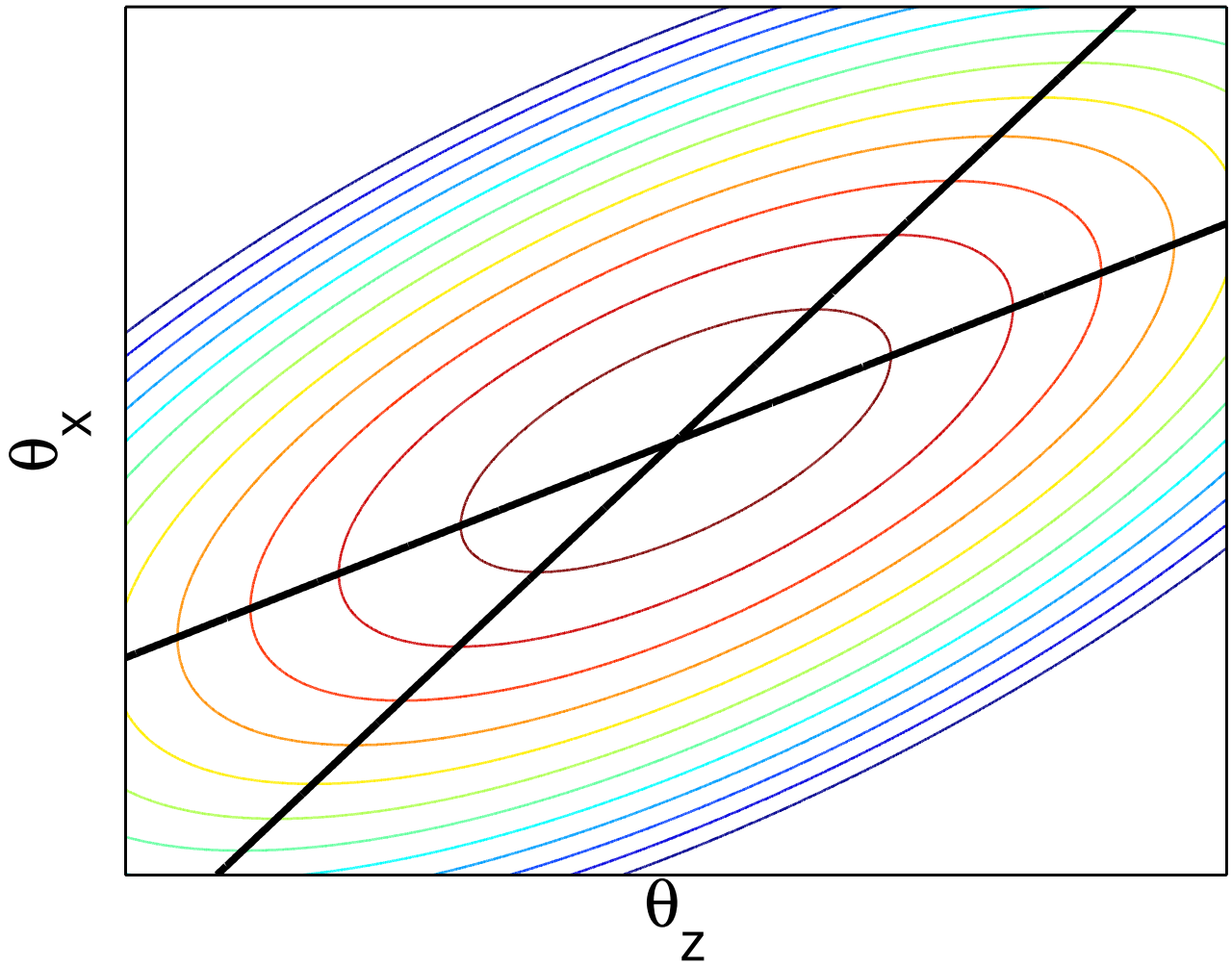
Variational Bayes



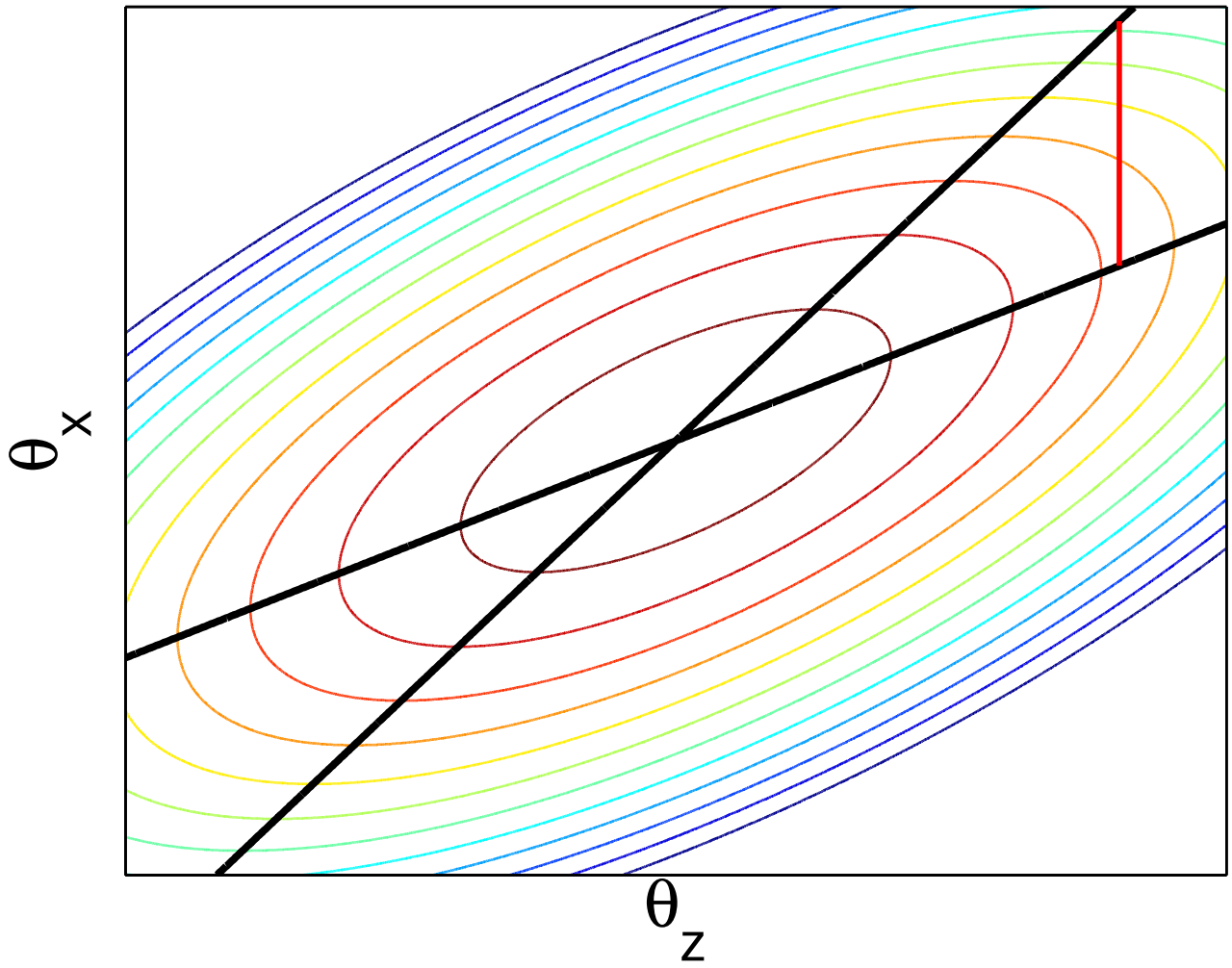
Variational Bayes



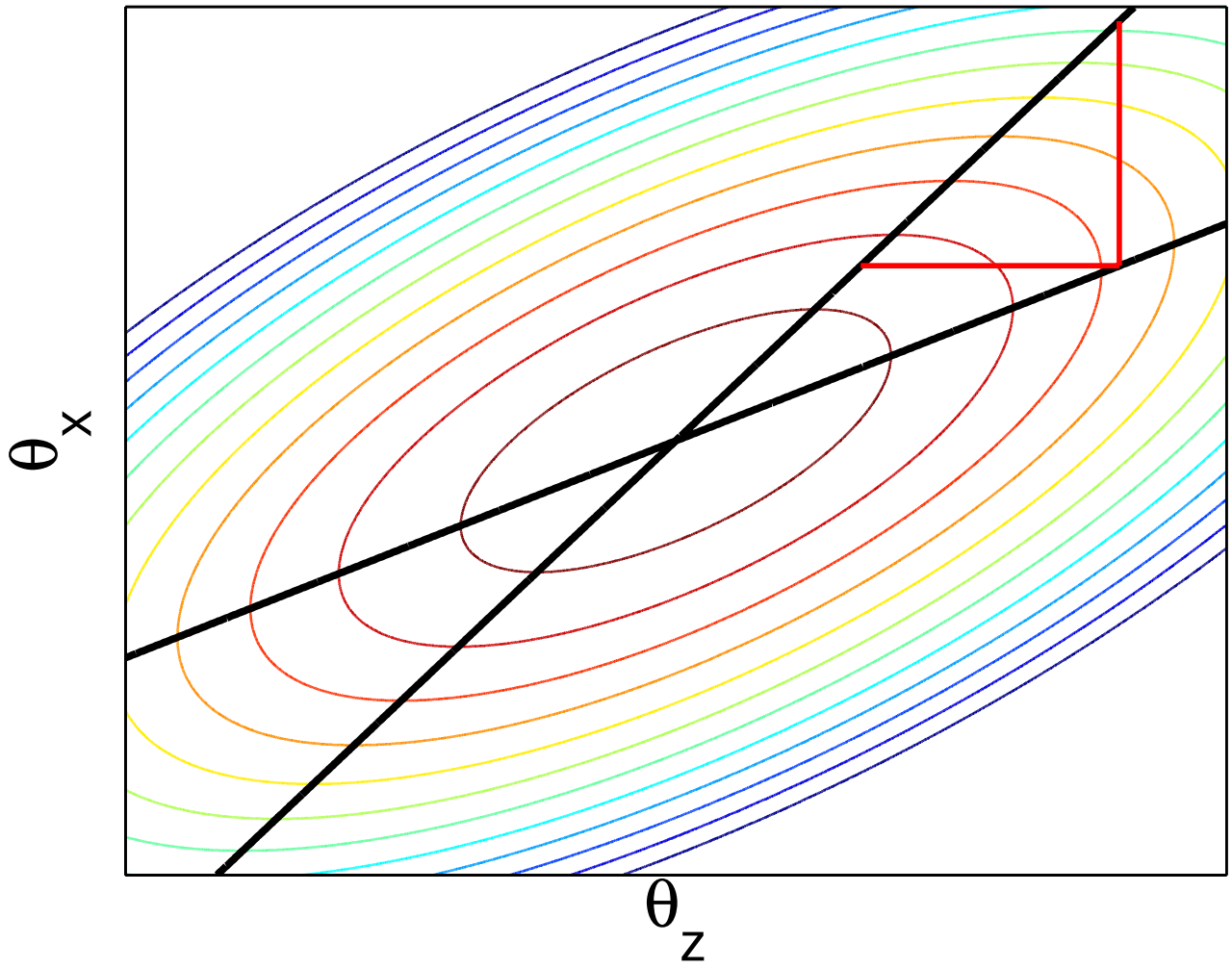
Variational Bayes



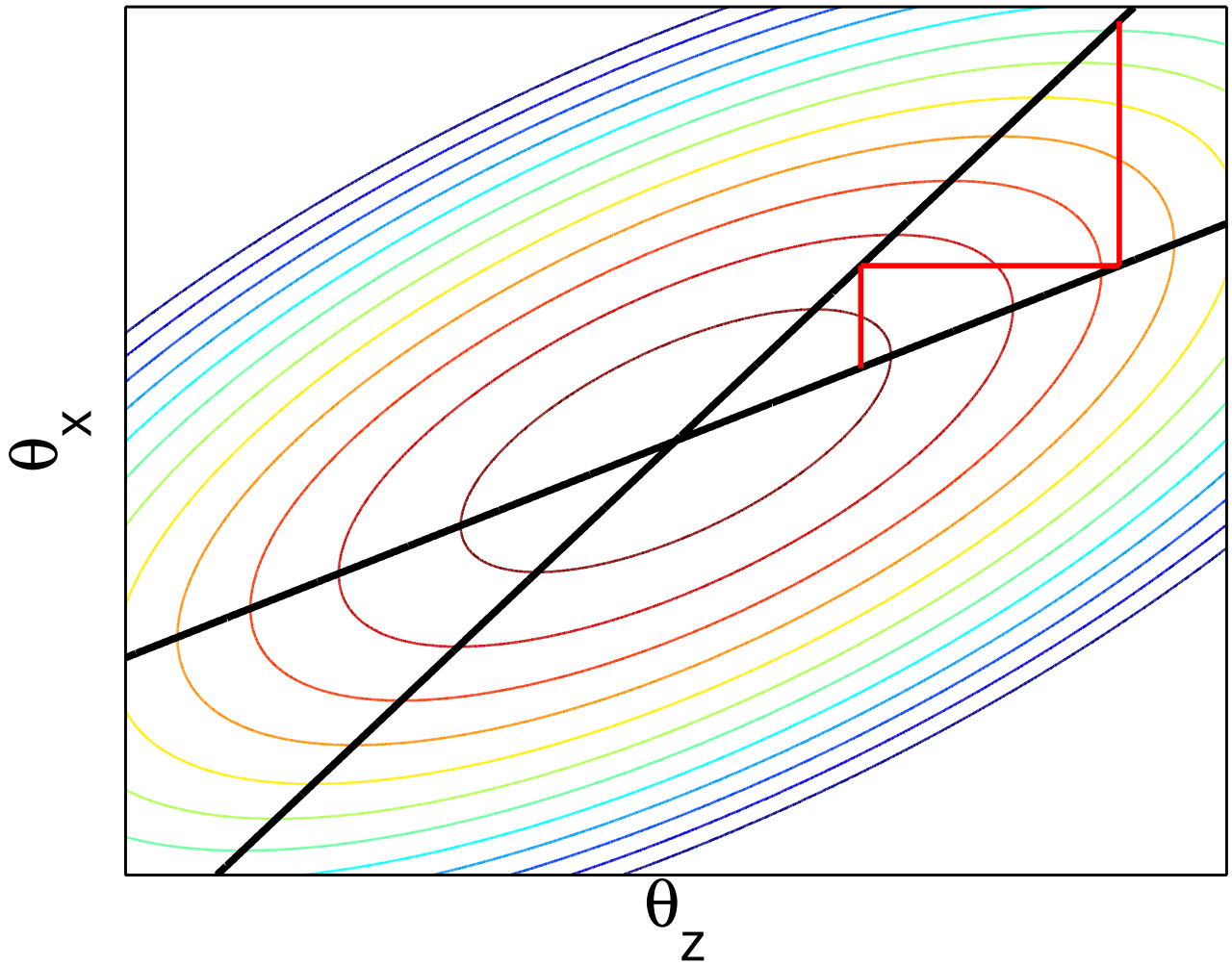
Variational Bayes



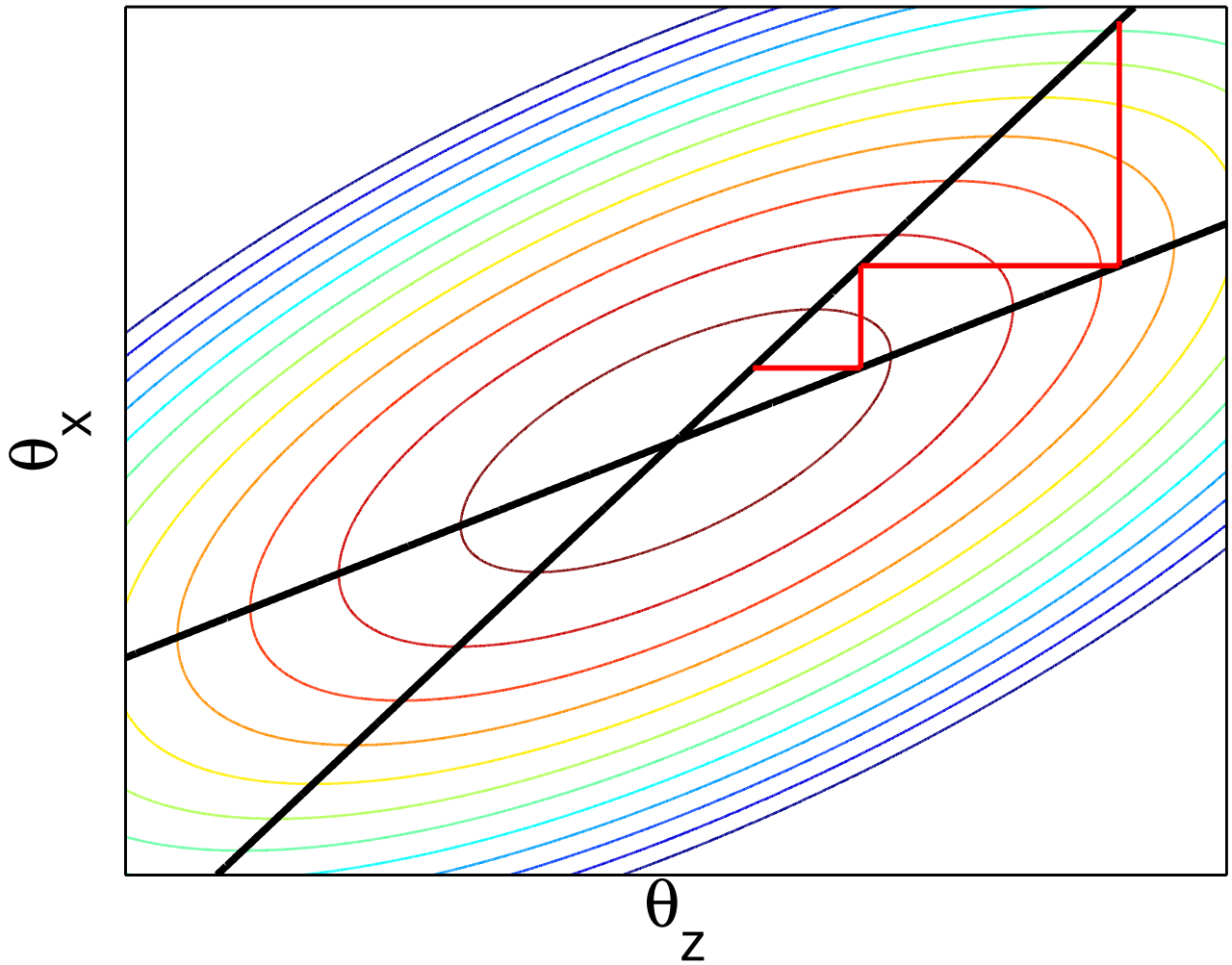
Variational Bayes



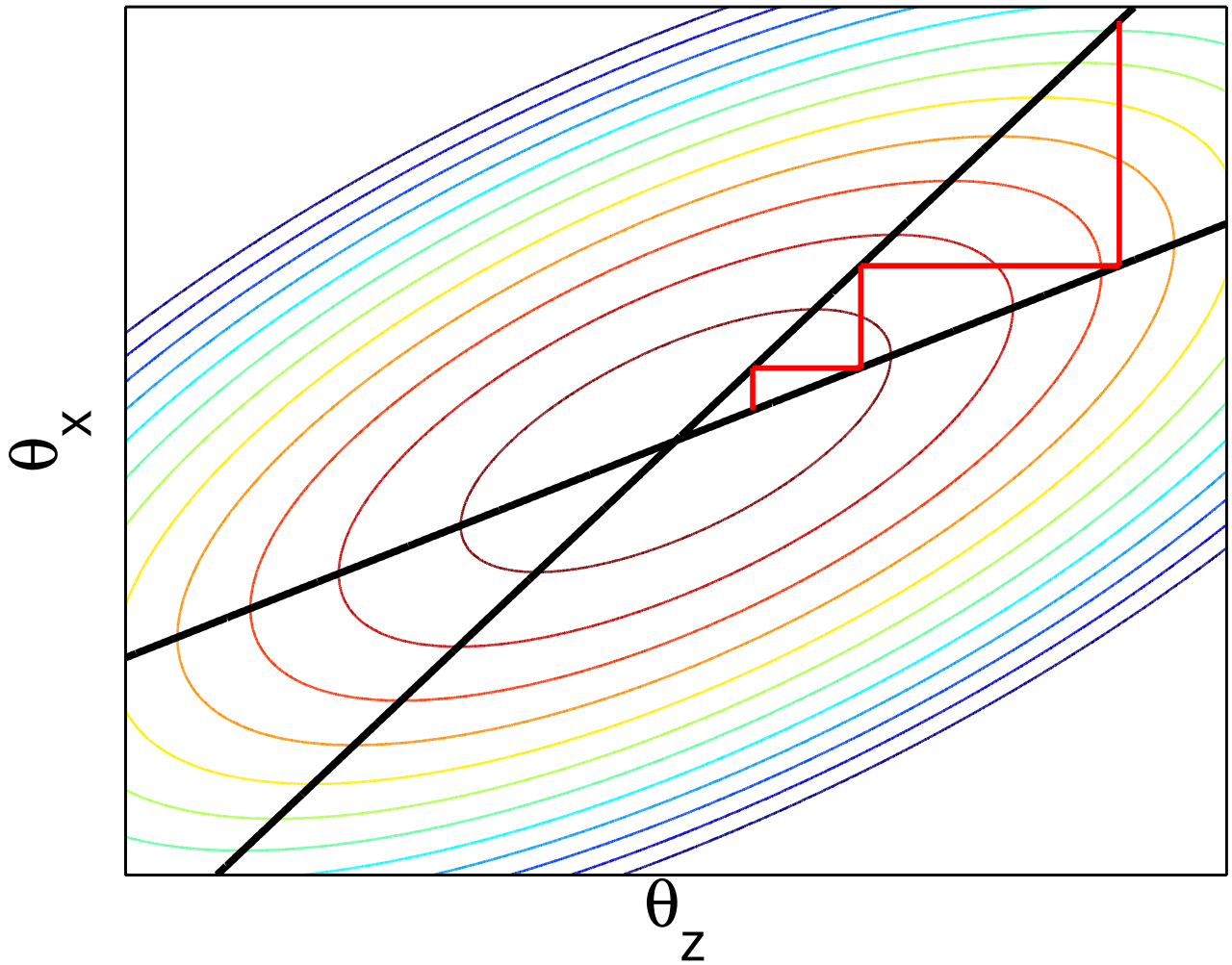
Variational Bayes



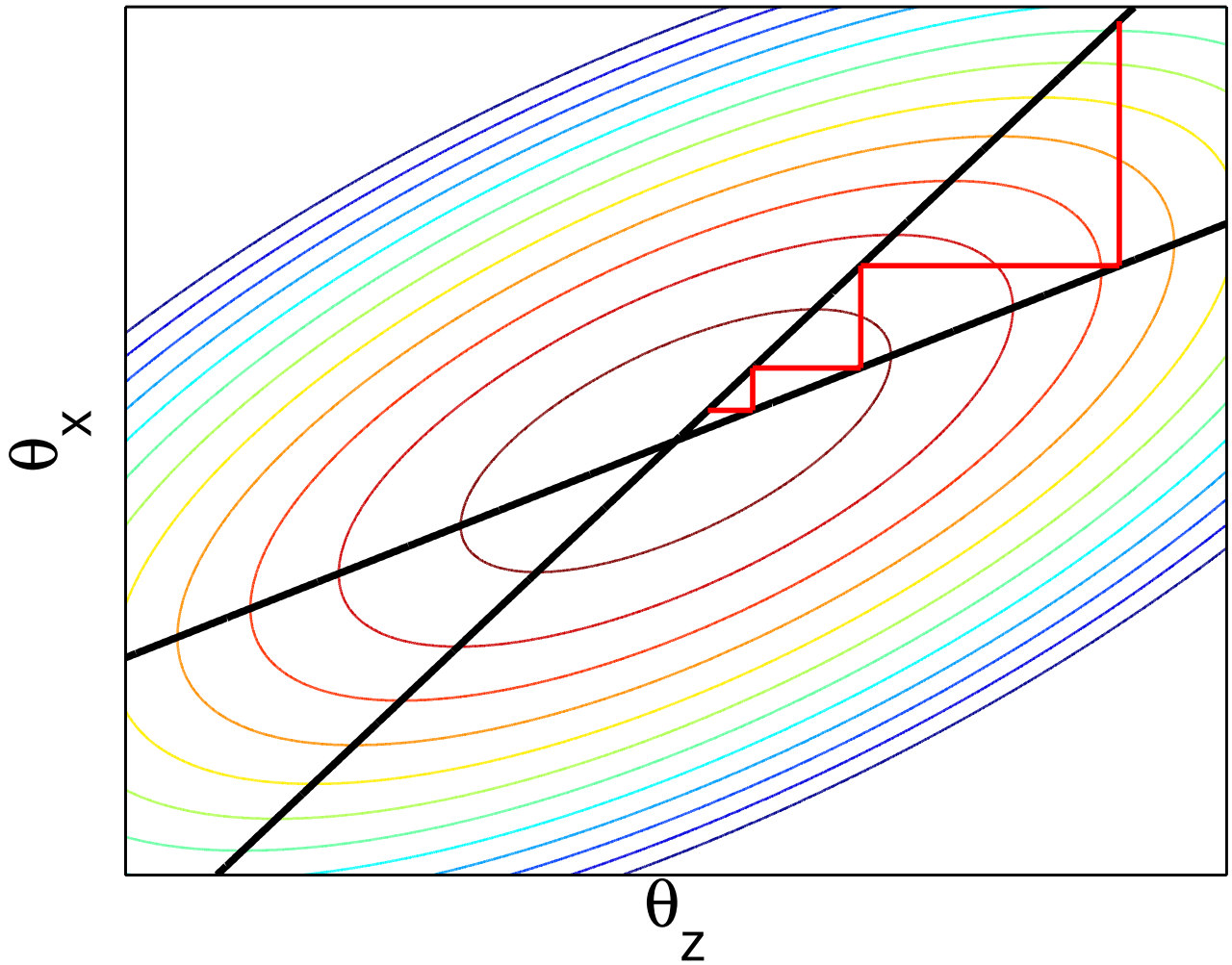
Variational Bayes



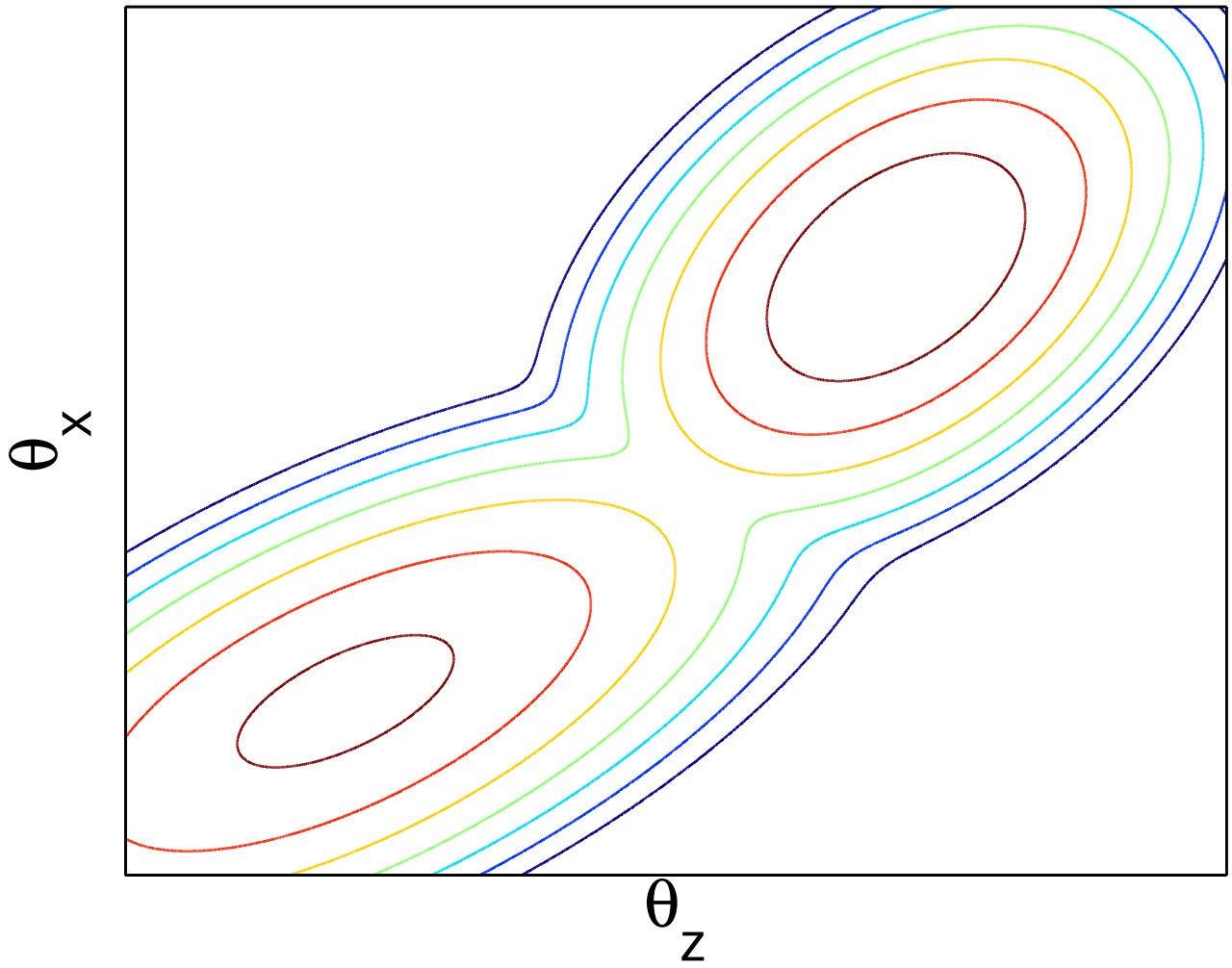
Variational Bayes



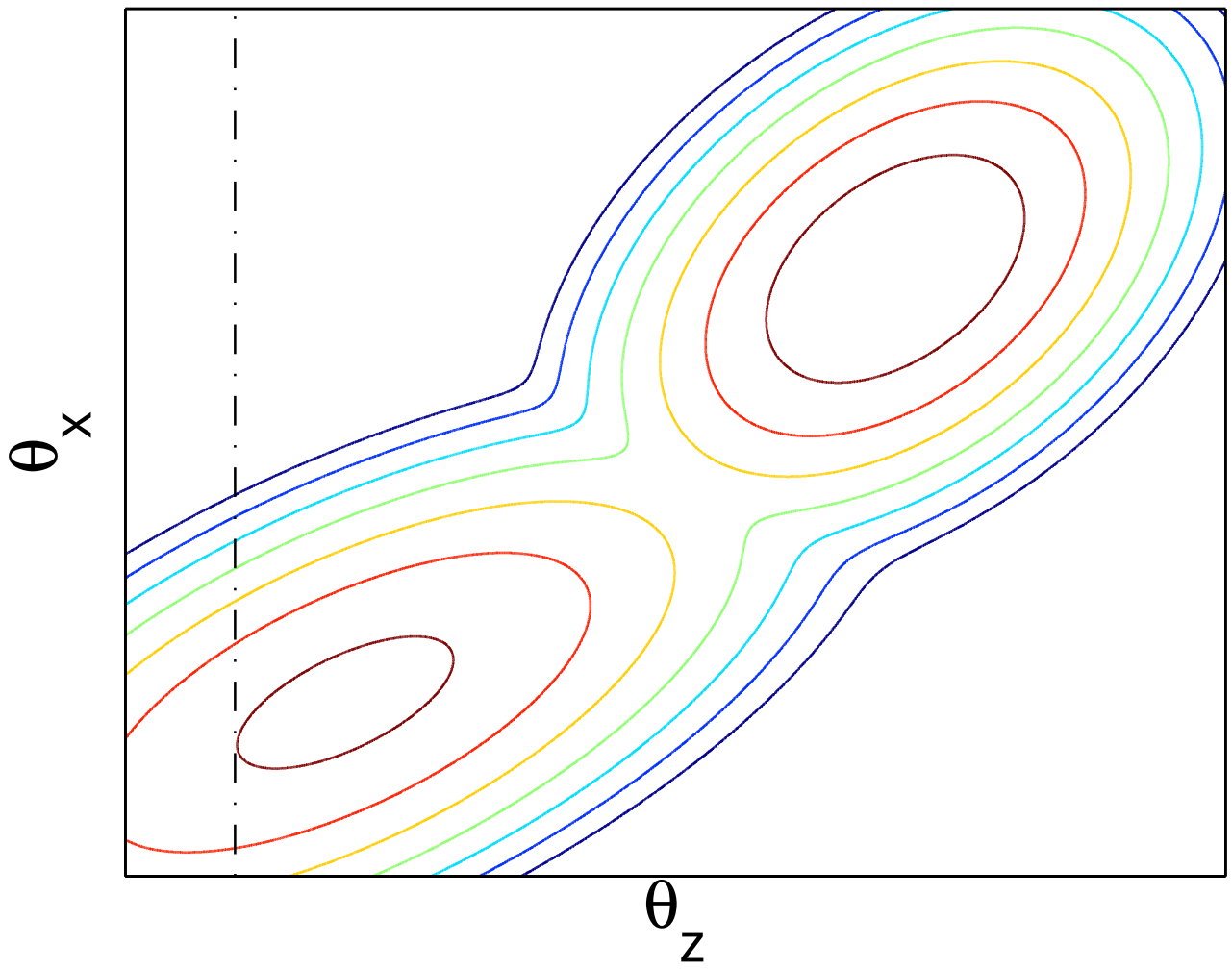
Variational Bayes



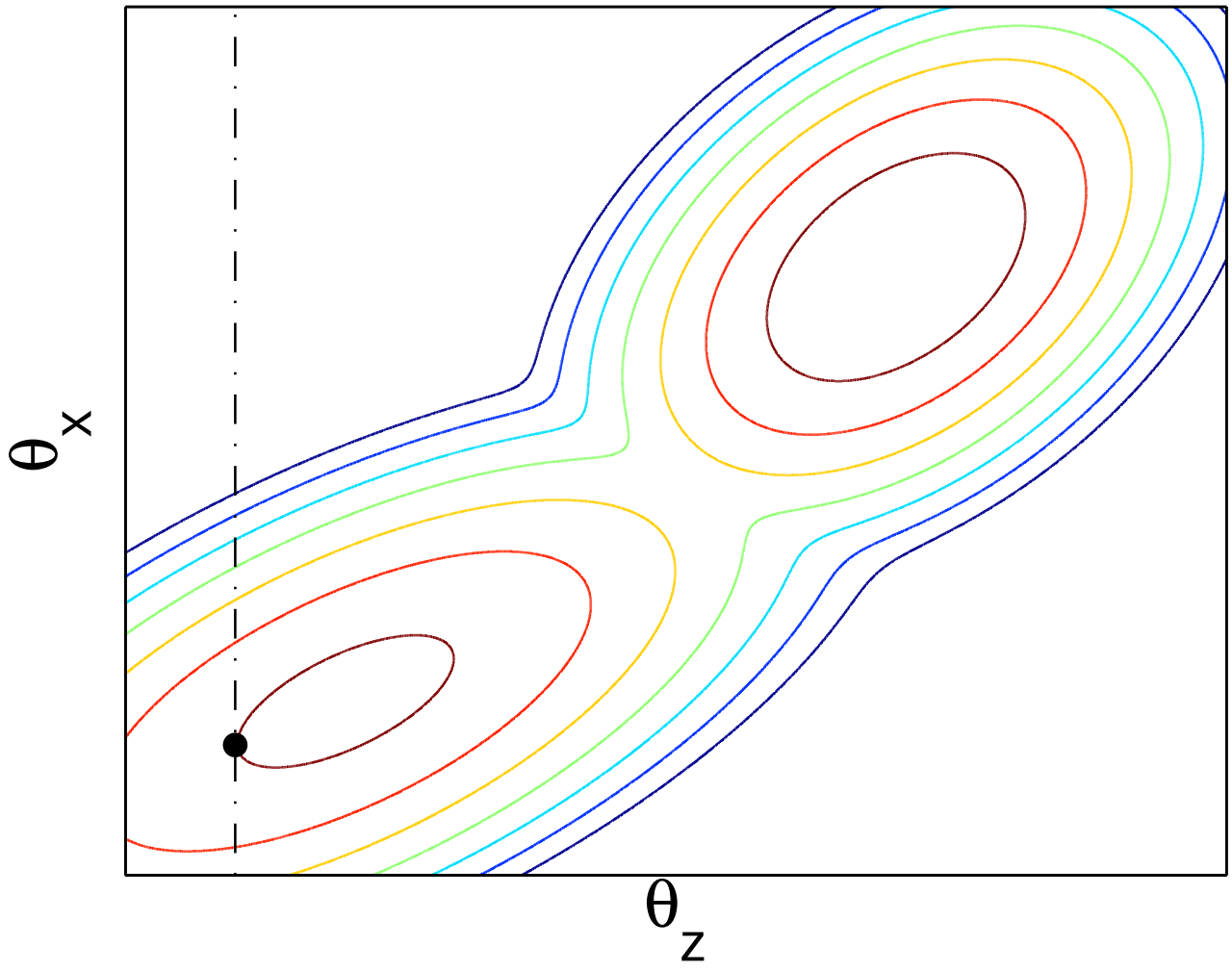
Variational Bayes



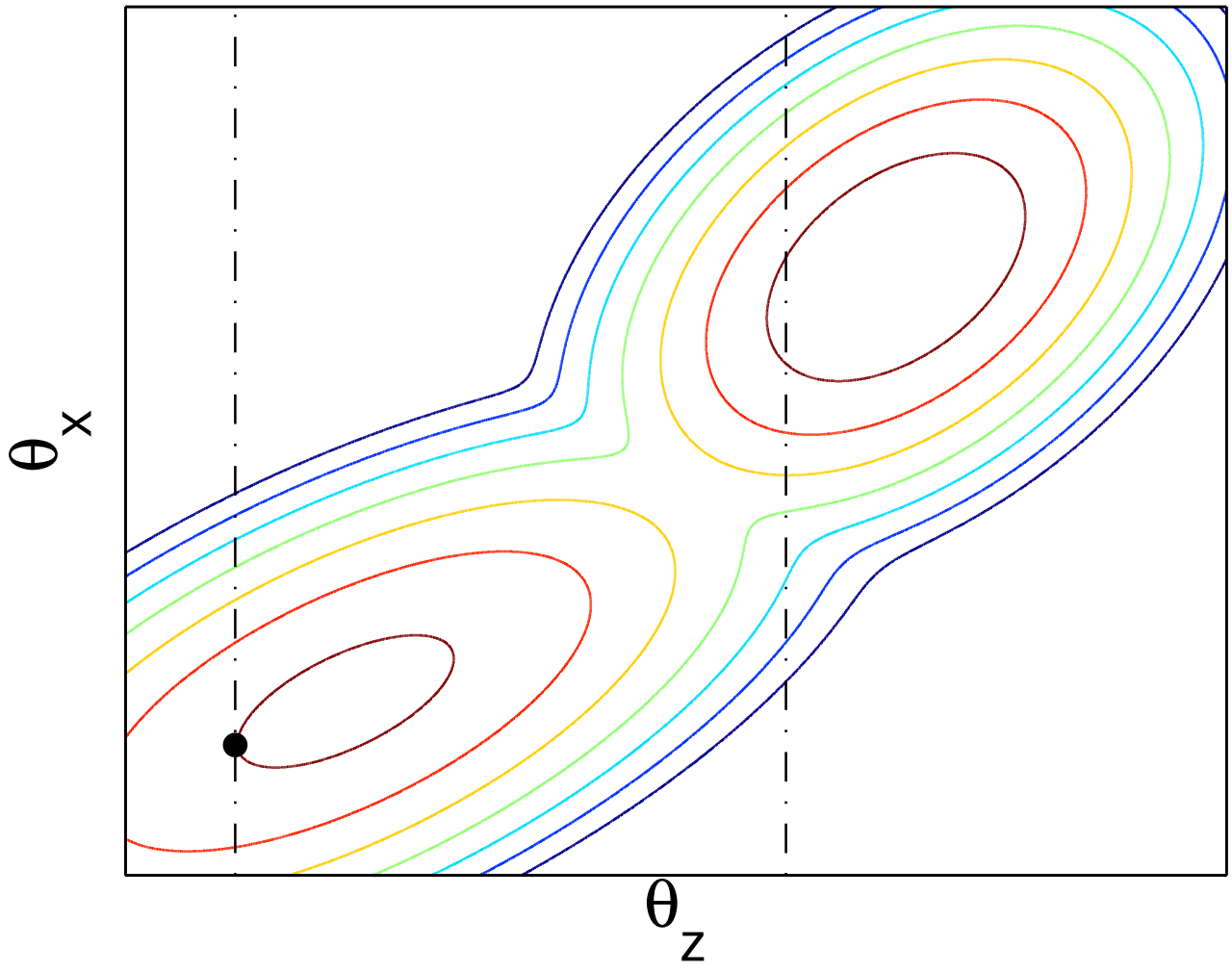
Variational Bayes



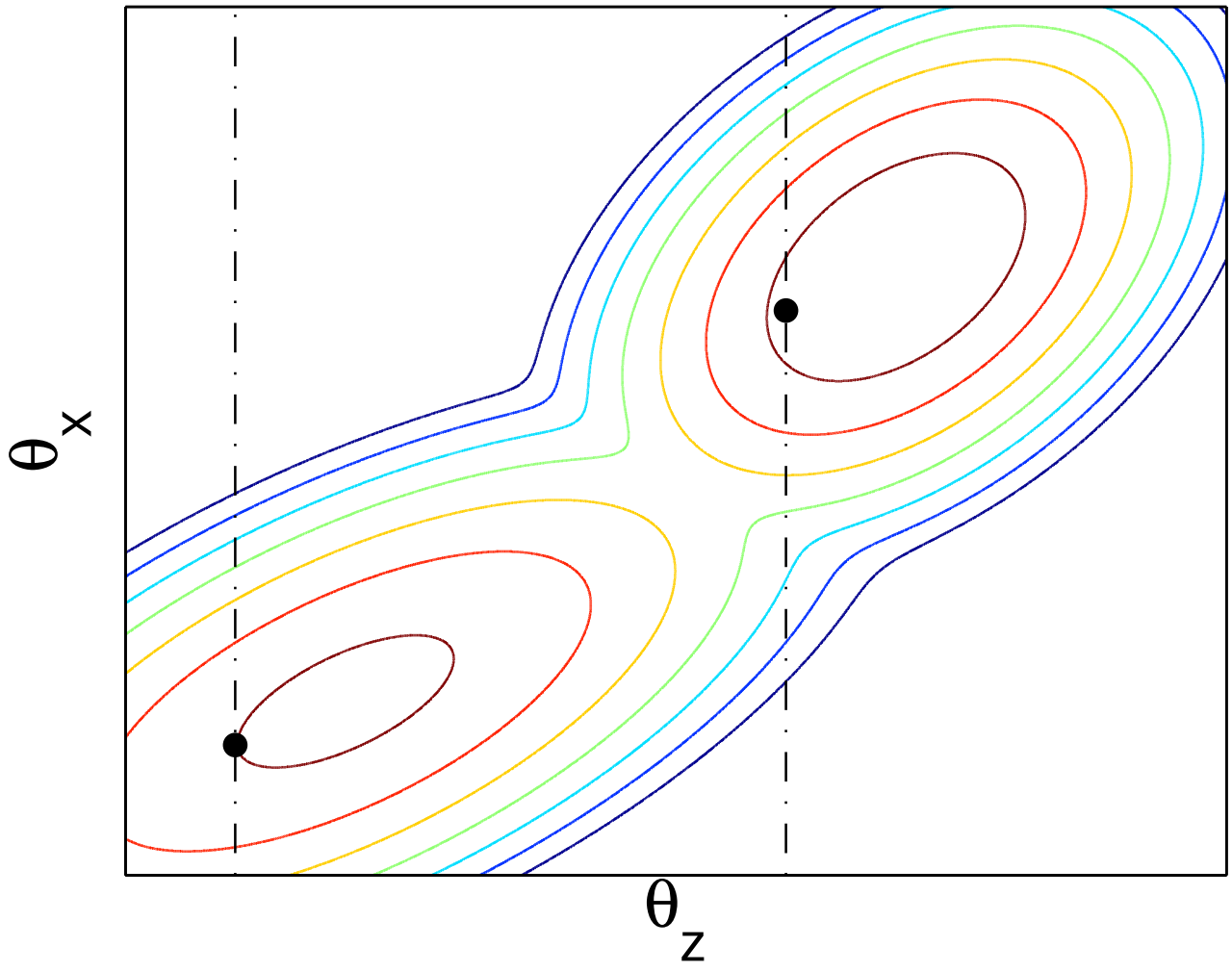
Variational Bayes



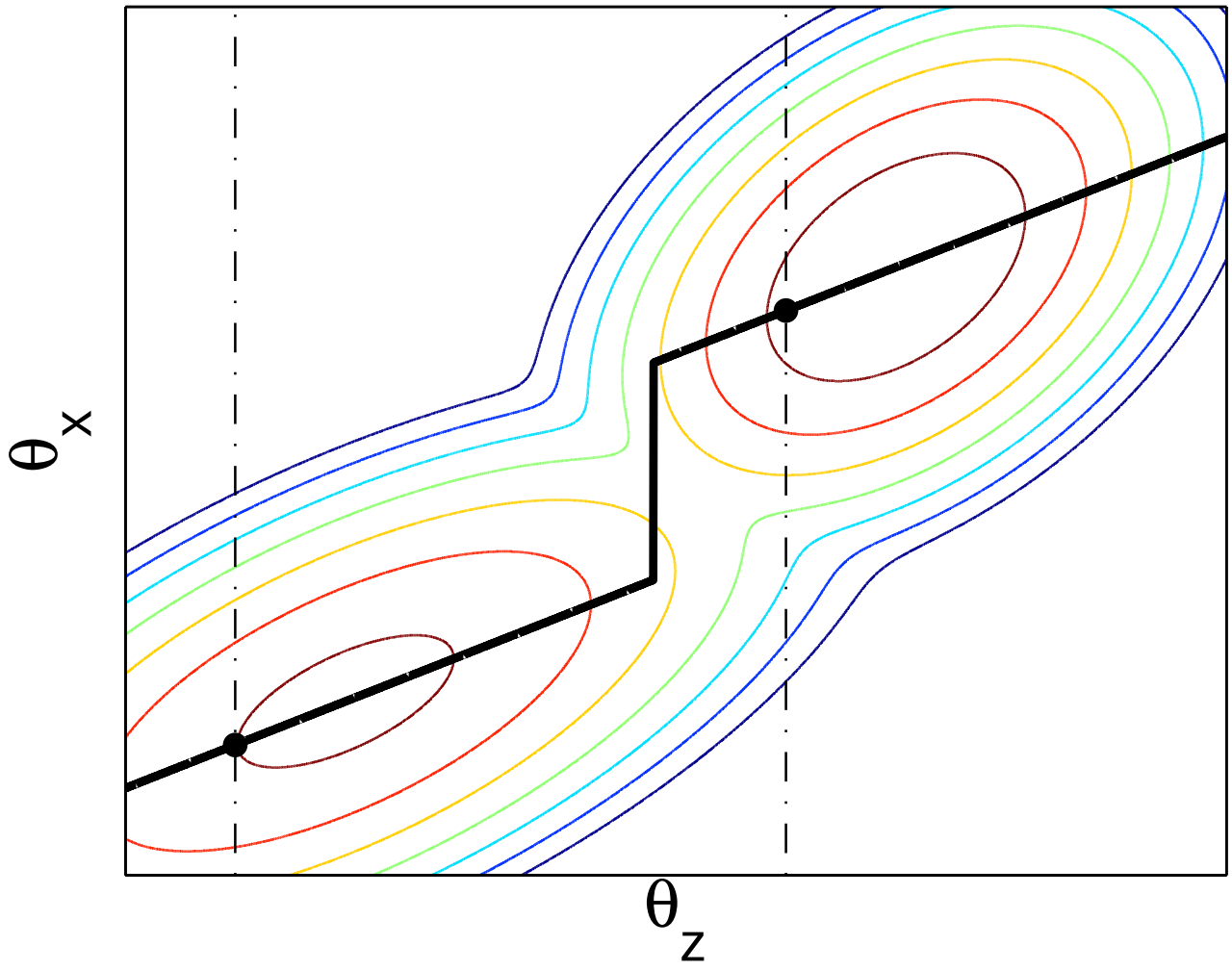
Variational Bayes



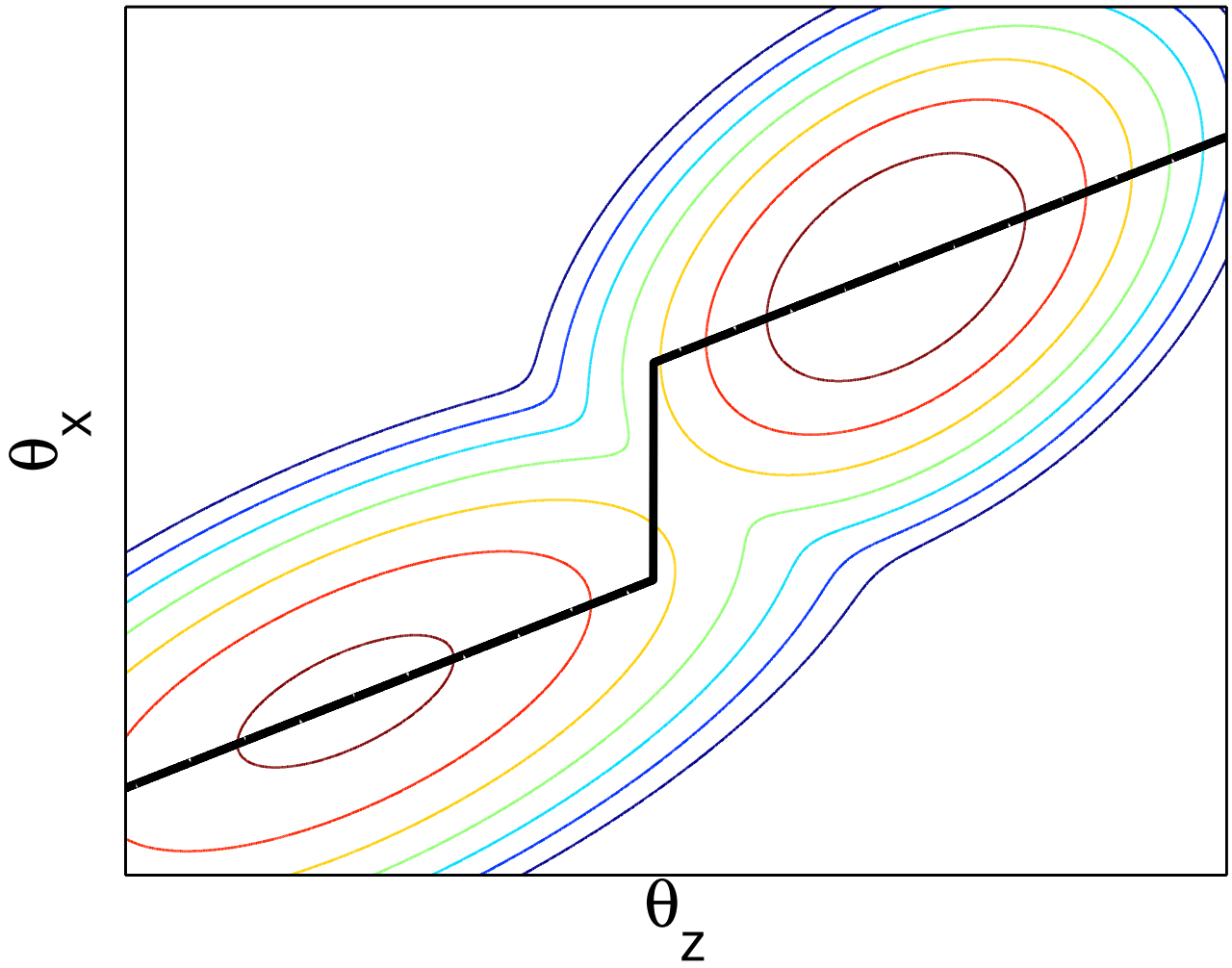
Variational Bayes



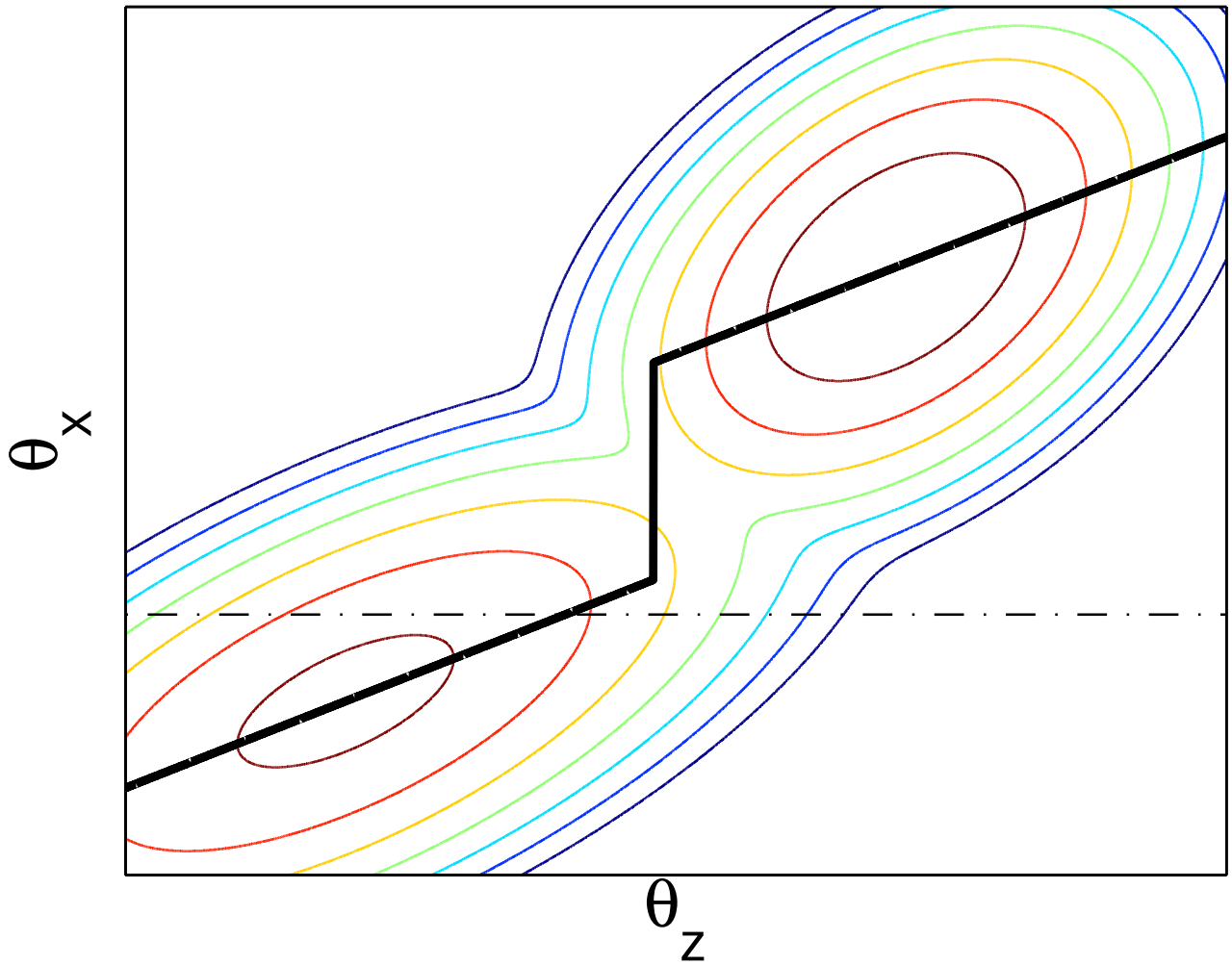
Variational Bayes



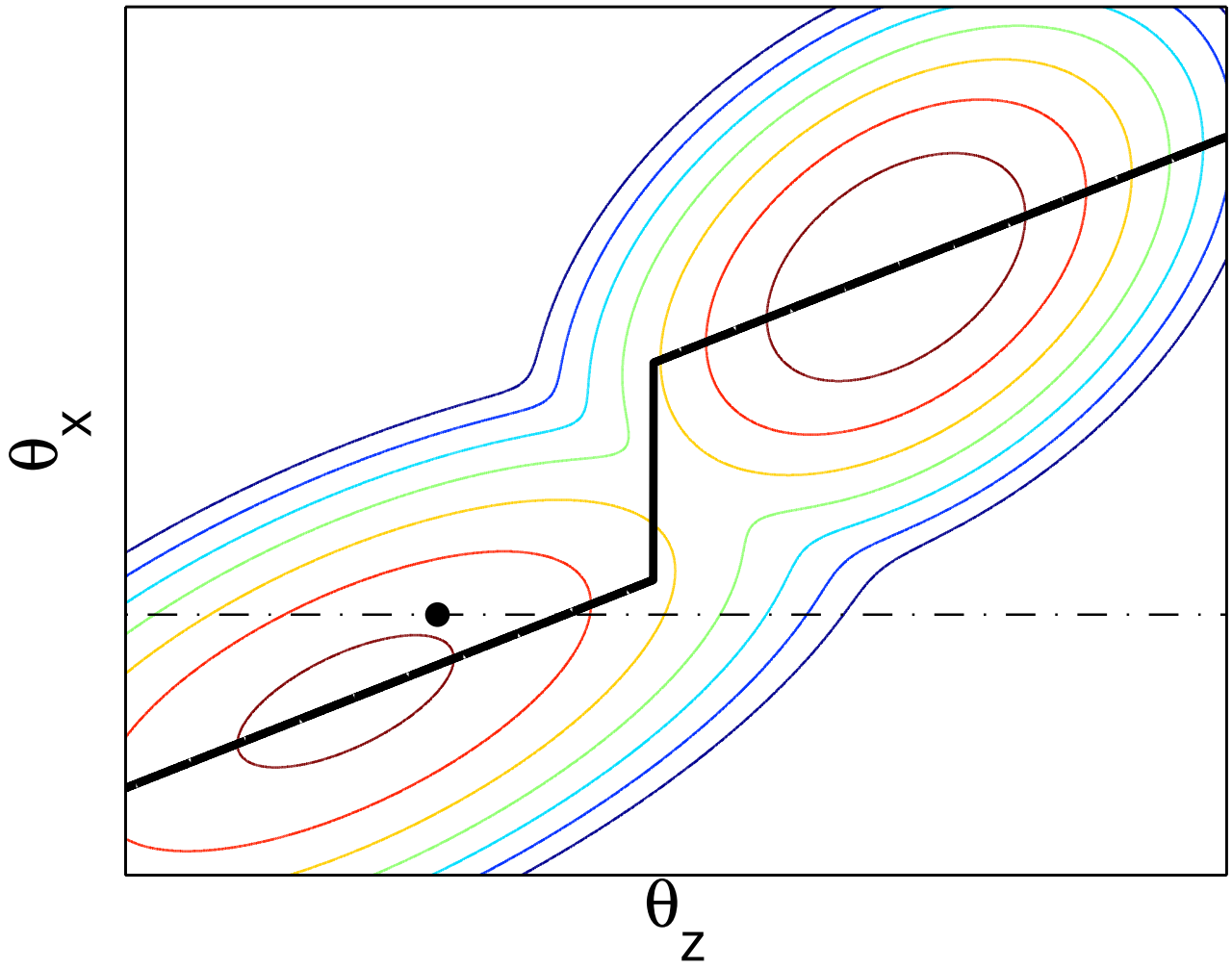
Variational Bayes



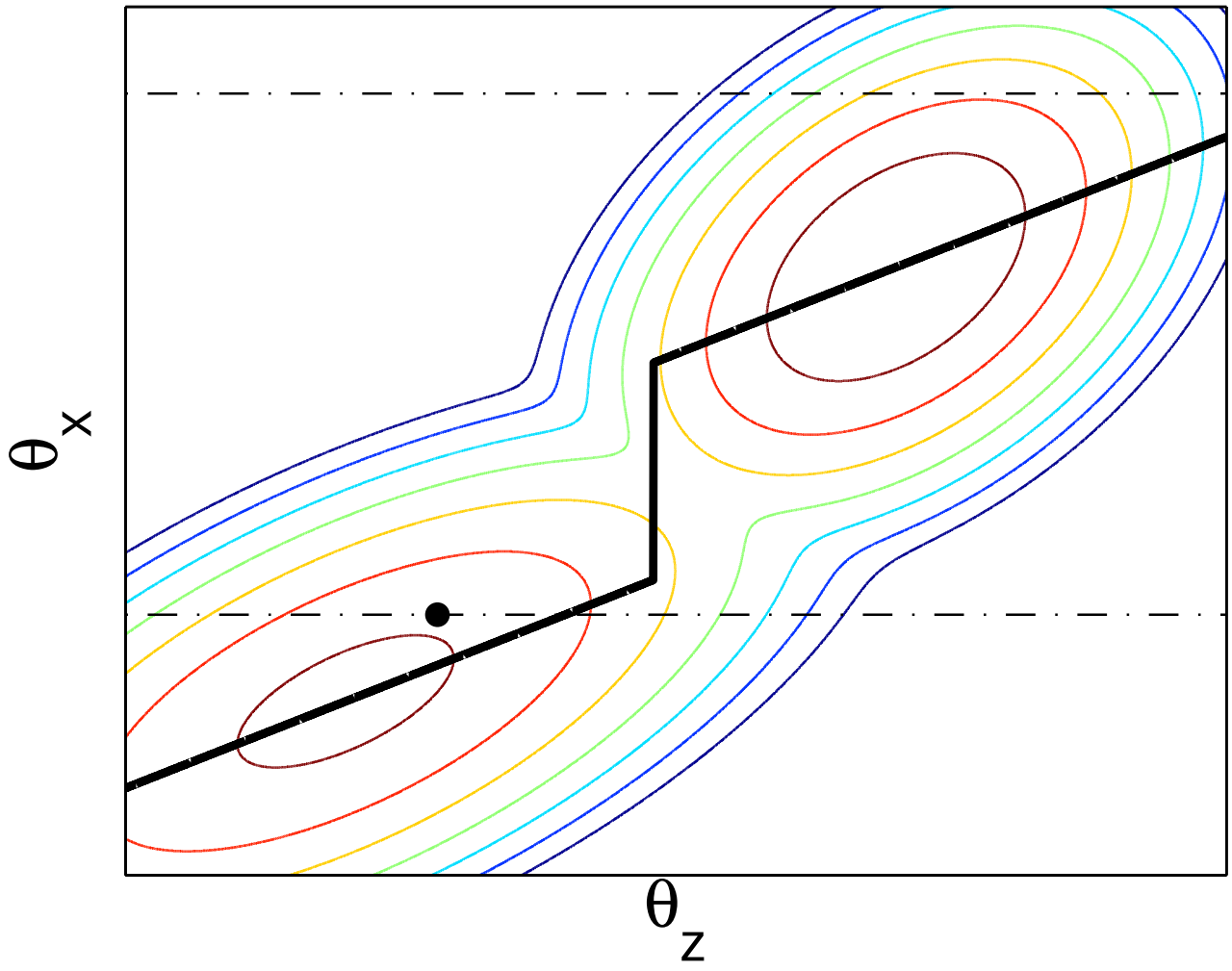
Variational Bayes



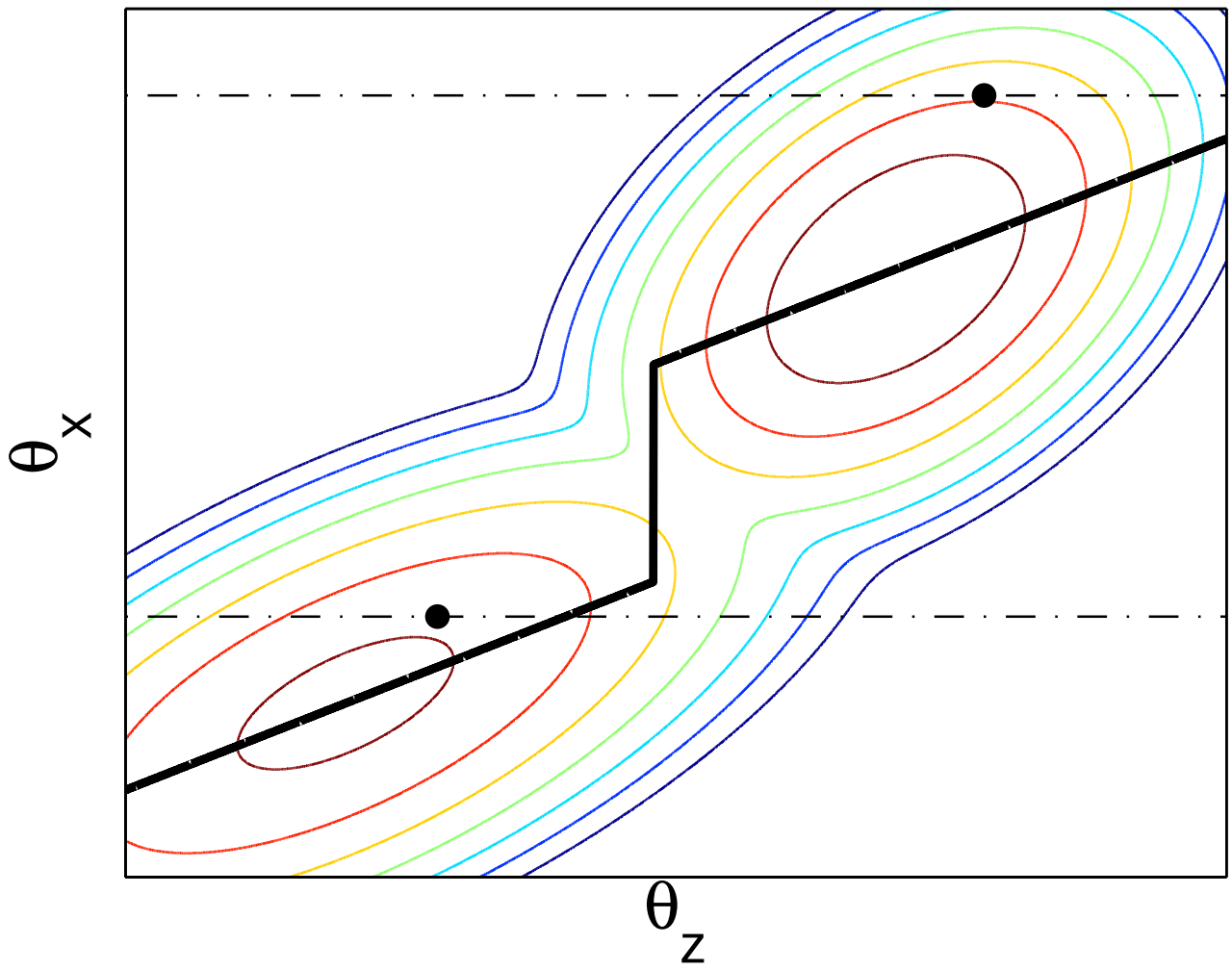
Variational Bayes



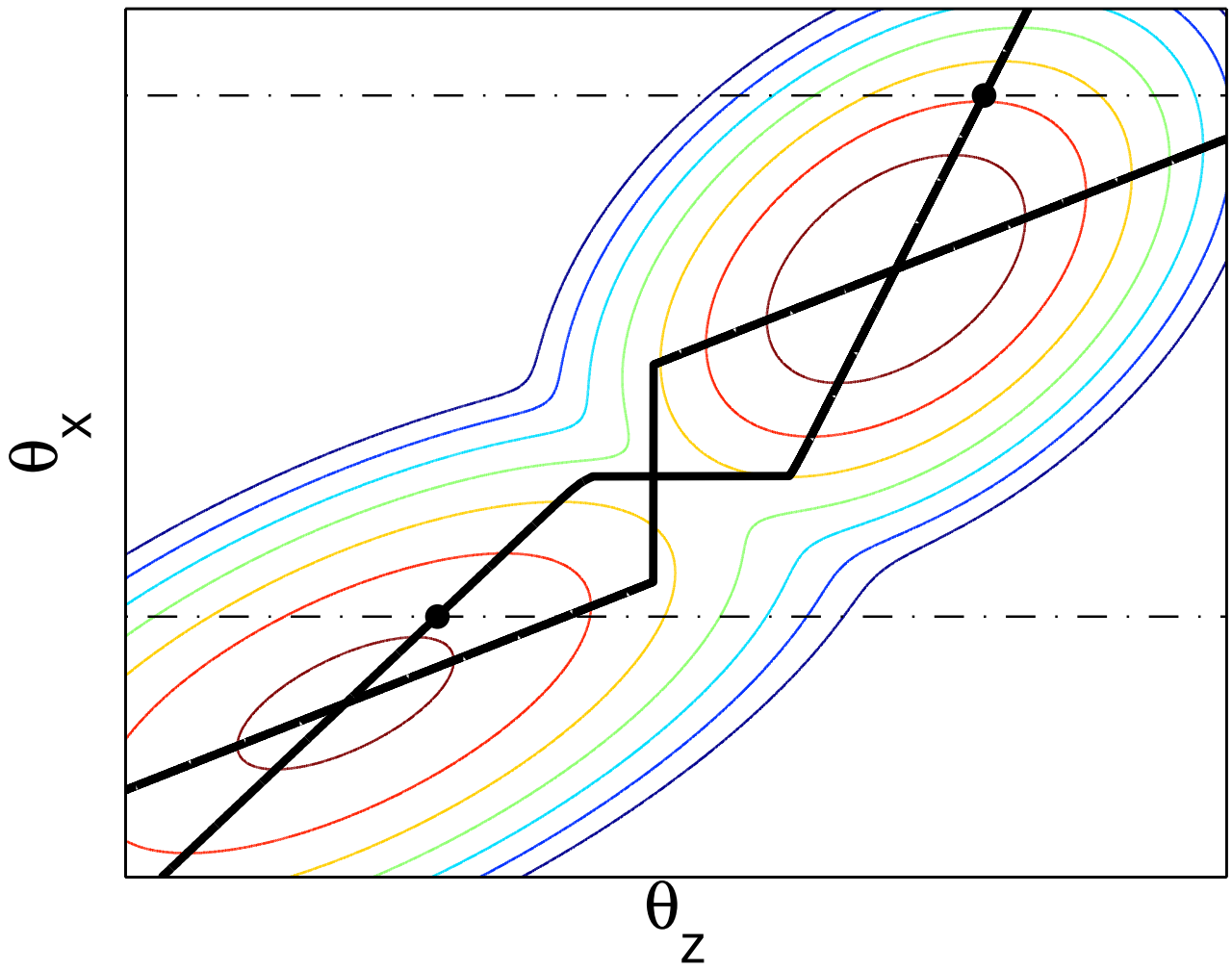
Variational Bayes



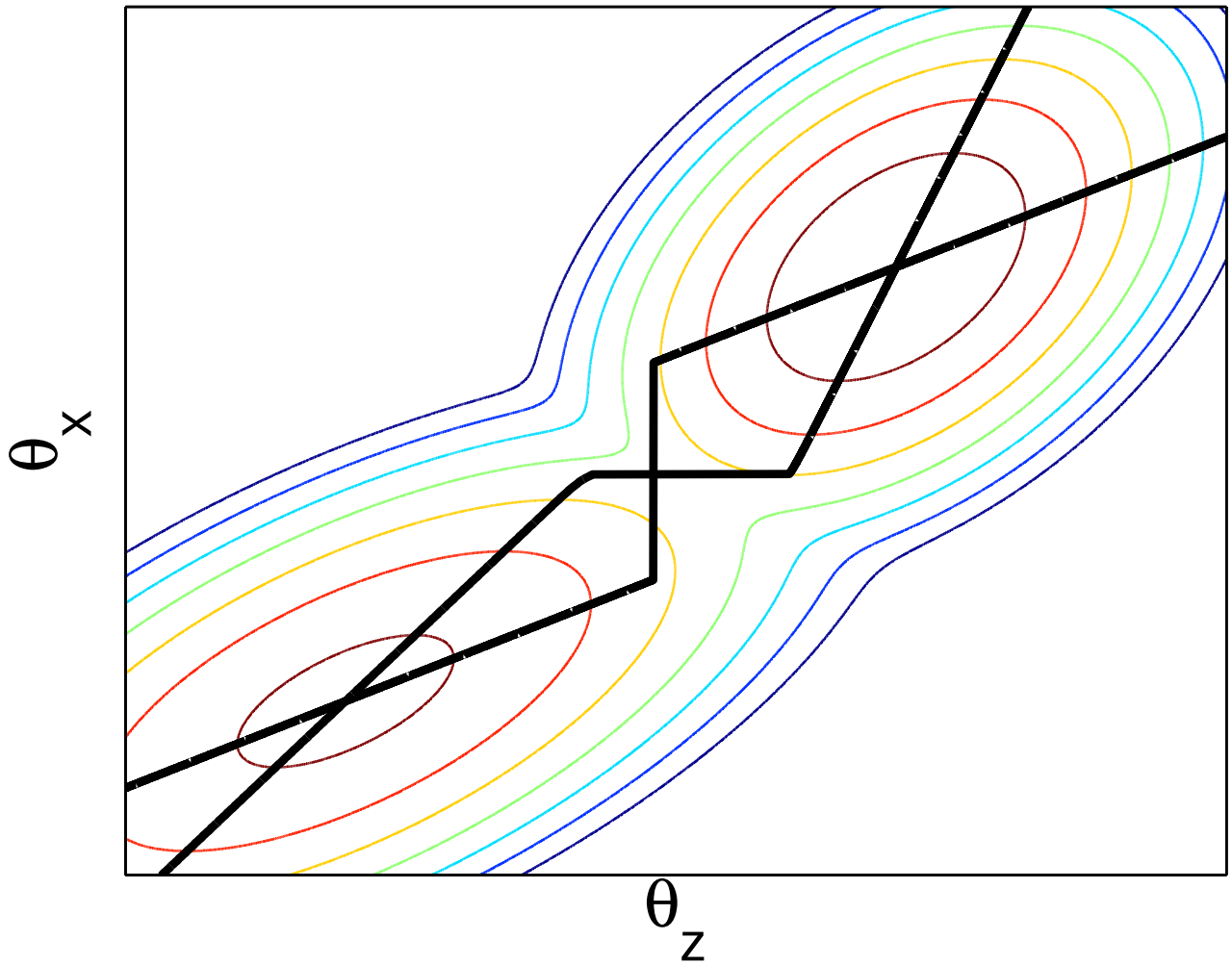
Variational Bayes



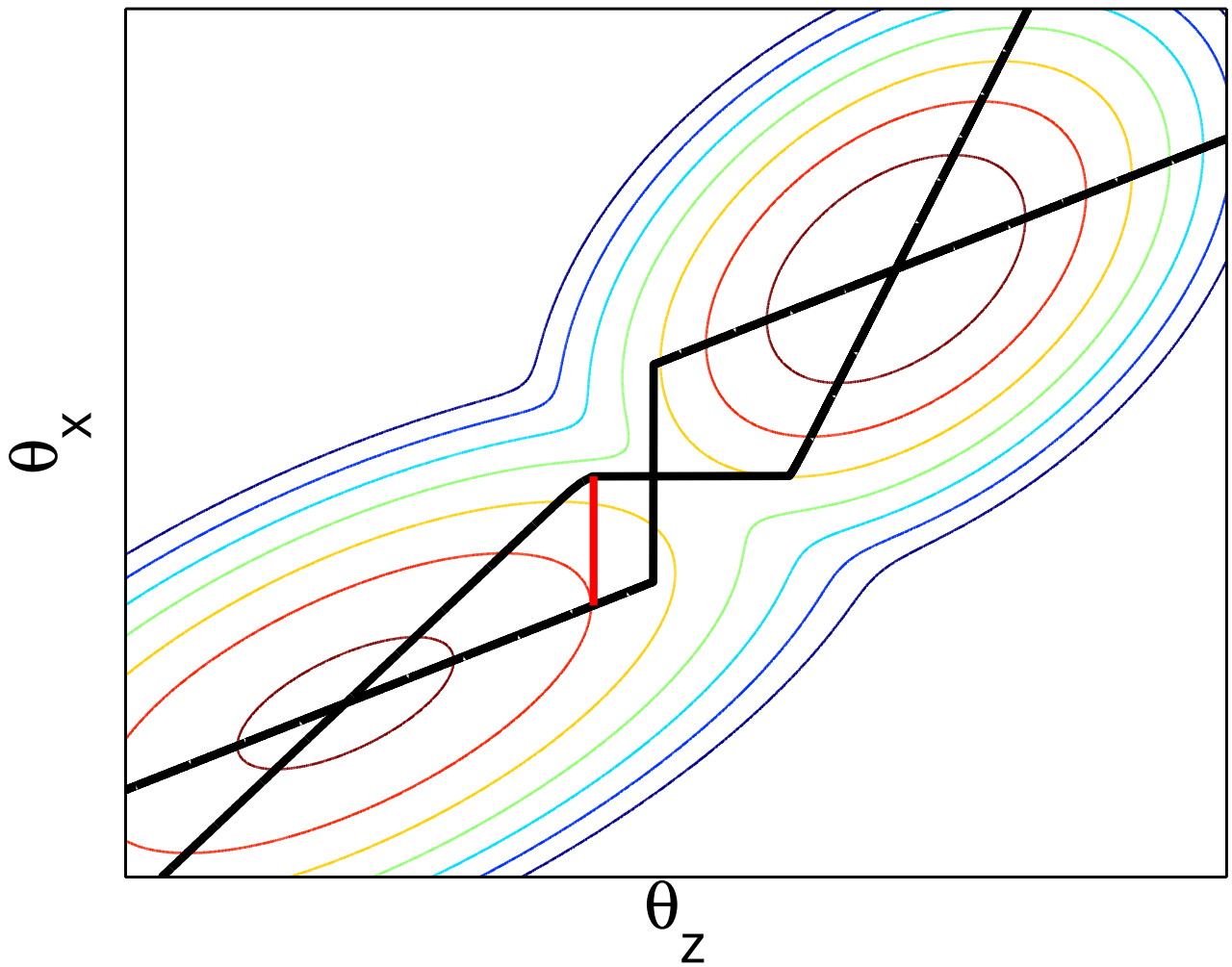
Variational Bayes



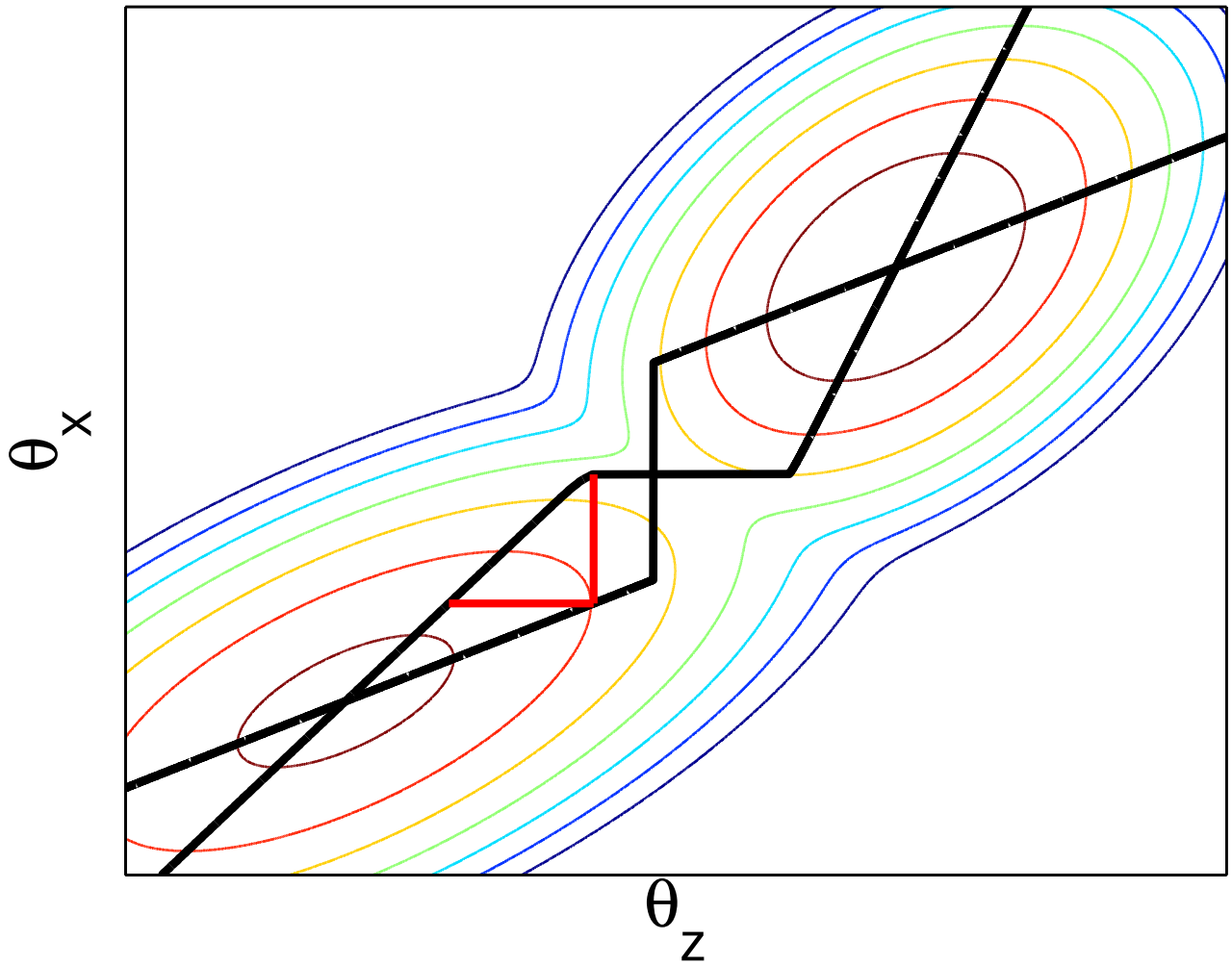
Variational Bayes



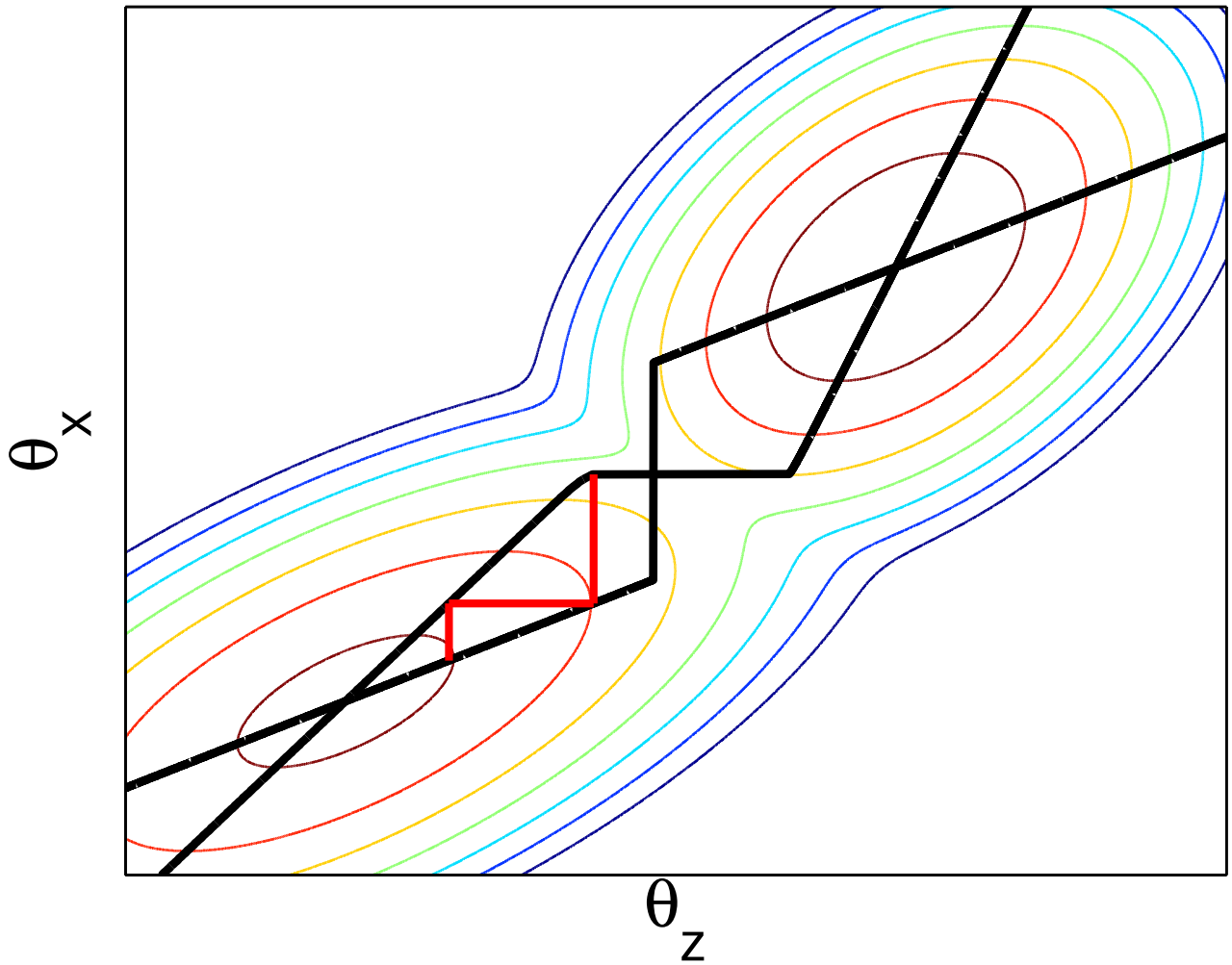
Variational Bayes



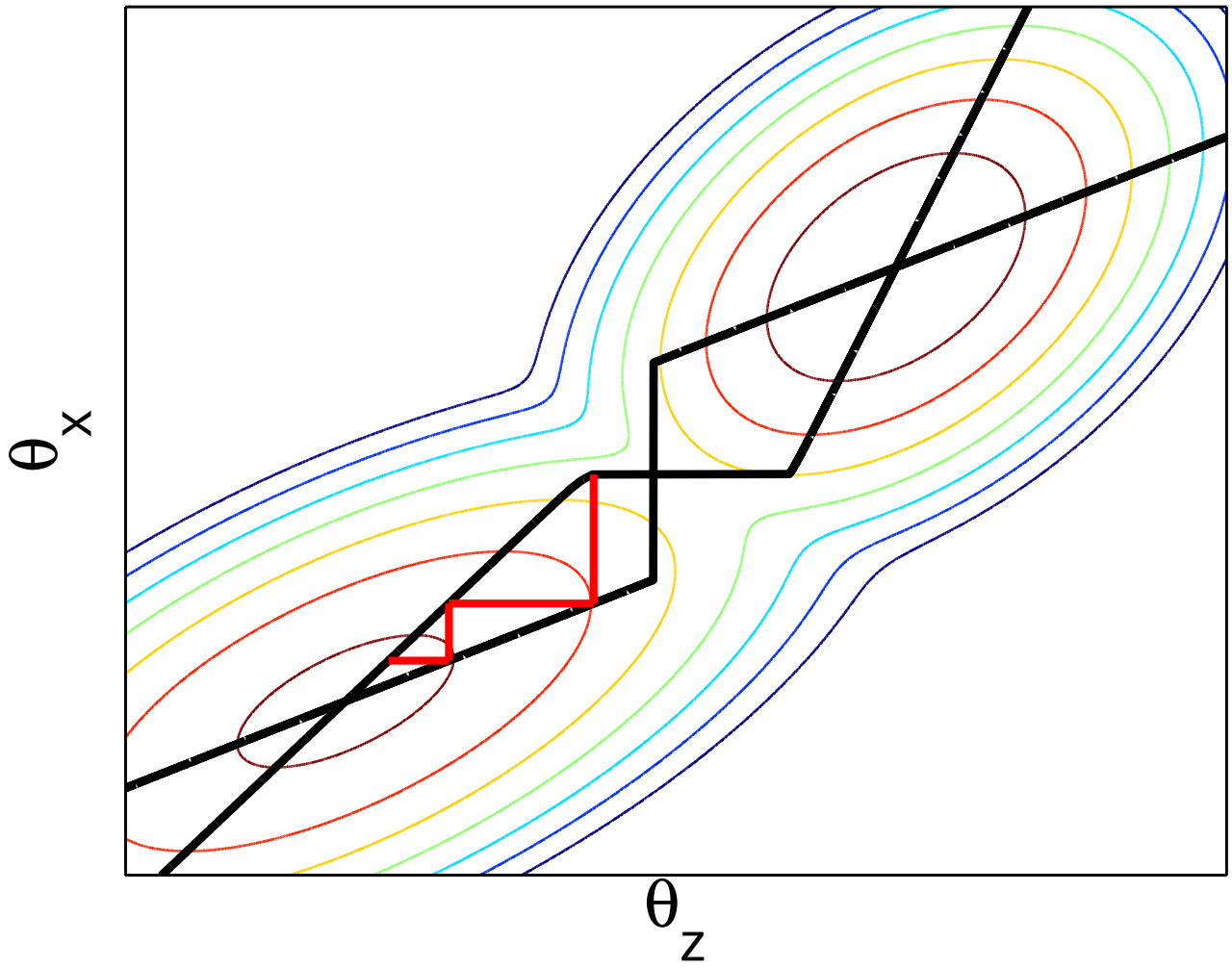
Variational Bayes



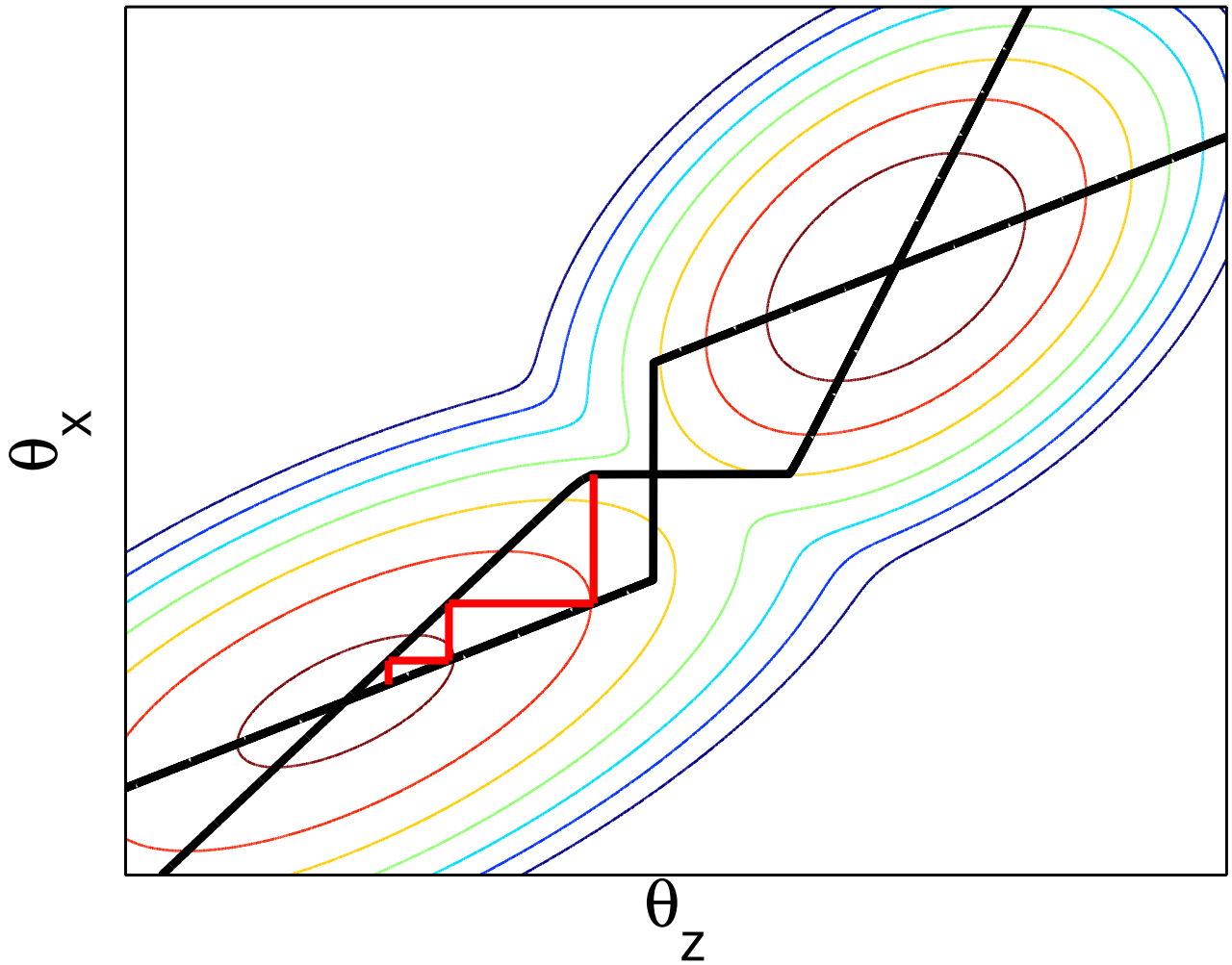
Variational Bayes



Variational Bayes



Variational Bayes



True Natural Gradient
of Collapsed Variational Bayes

Collapsed Variational Bayes

- We can obtain the optimal $\theta_x^\star$ for any fixed θ_z ...

Collapsed Variational Bayes

- We can obtain the optimal $\theta_x^\star$ for any fixed θ_z ...
- ...Hence, there is no reason to keep both variables

Collapsed Variational Bayes

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

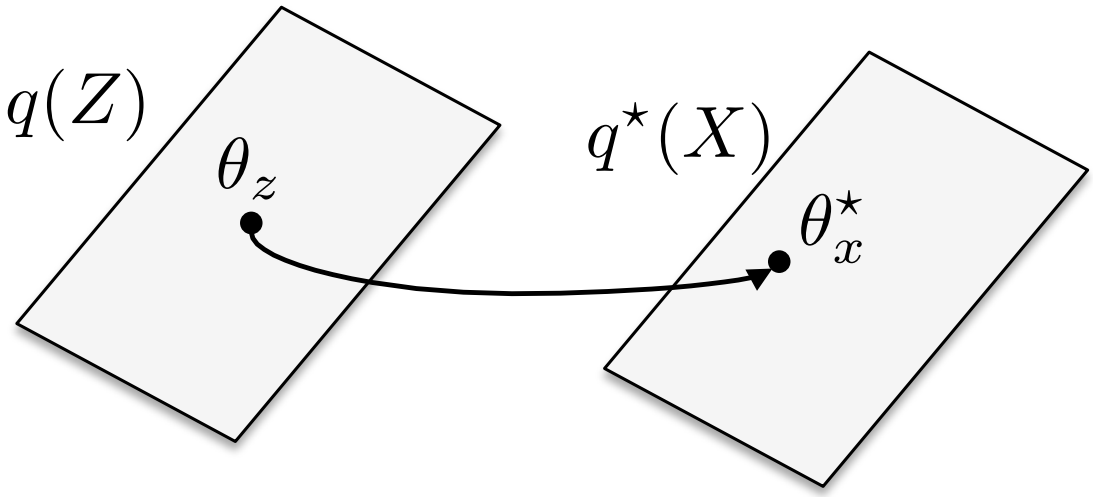
Collapsed Variational Bayes

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

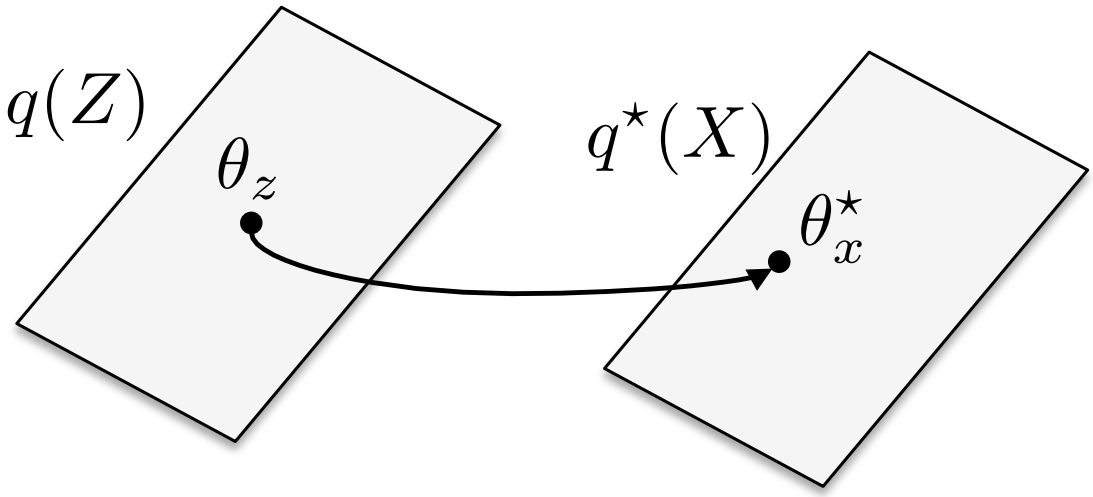
Collapsed Variational Bayes

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

Collapsed Variational Bayes



Collapsed Variational Bayes



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

$$\mathcal{L}_{\mathrm{KL}} = \mathcal{L} + \mathrm{KL}\big(q^\star(X)||q(X)\big)$$

True Natural Gradient
of Collapsed Variational Bayes

Natural Gradient

- Gradient:

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

Natural Gradient

- Gradient:

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

- The space of natural parameters is not Euclidean

Natural Gradient

- Gradient:

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

- 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:

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

- 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

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

Natural Gradient

- Gradient:

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

Natural Gradient

- Gradient:

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

- Natural gradient:

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

Natural Gradient

- Gradient:

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

- Natural gradient:

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

- 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:

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

- Natural gradient:

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

- 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)

Our contribution

True Natural Gradient
of Collapsed Variational Bayes

True Natural Gradient

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

True Natural Gradient

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

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

  only captures the information geometry of the space in which θ_z lives

True Natural Gradient

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

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

  only captures the information geometry of the space in which θ_z lives

- We define instead

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

True Natural Gradient

- Alternative expressions for the TNG:

$$\mathbf{F}_{\mathrm{TNG}}(\theta_z) \triangleq -\mathbb{E}_{q(Z)q^\star(X)}\left[\nabla\nabla \log\left(q(Z)q^\star(X)\right)\right]$$

True Natural Gradient

- Alternative expressions for the TNG:

$$\mathbf{F}_{\mathrm{TNG}}(\theta_z) \triangleq -\mathbb{E}_{q(Z)q^\star(X)}\left[\nabla\nabla \log\left(q(Z)q^\star(X)\right)\right]$$

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

True Natural Gradient

- Alternative expressions for the TNG:

$$\mathbf{F}_{\mathrm{TNG}}(\theta_z) \triangleq -\mathbb{E}_{q(Z)q^\star(X)}\left[\nabla\nabla \log\left(q(Z)q^\star(X)\right)\right]$$

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

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

Experiments

- Mixture of Gaussians with K components:
  - Normal inverse-Gamma prior over cluster means μ_k and variances σ_k^2
  - Dirichlet prior over cluster proportions $\pi = [\pi_1, \ldots, \pi_K]$
  - The assignments $\ell_n \sim \pi$
  - Given its assignment ℓ_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 μ_k and variances σ_k^2
  - Dirichlet prior over cluster proportions $\pi = [\pi_1, \ldots, \pi_K]$
  - The assignments $\ell_n \sim \pi$
  - Given its assignment ℓ_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

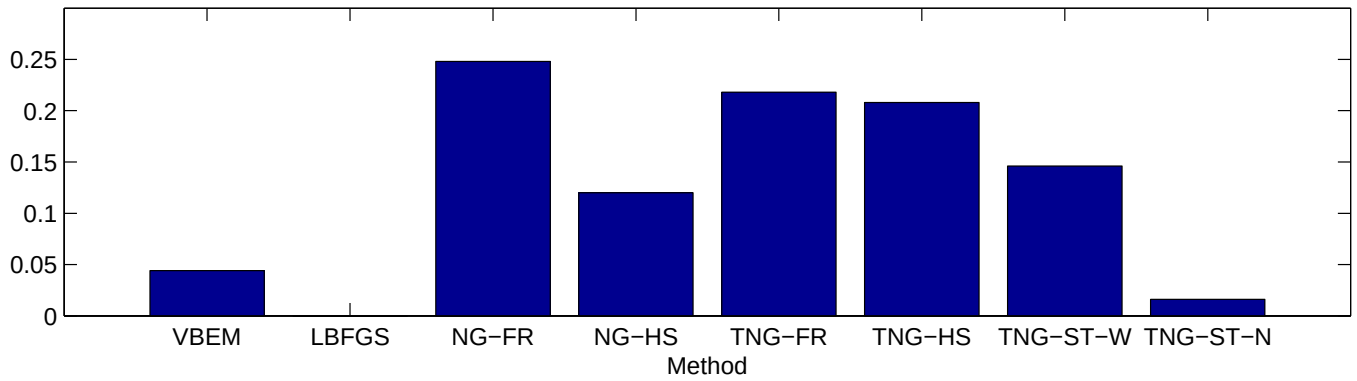
- Mixture of Gaussians with K components:
  - Normal inverse-Gamma prior over cluster means μ_k and variances σ_k^2
  - Dirichlet prior over cluster proportions $\pi = [\pi_1, \ldots, \pi_K]$
  - The assignments $\ell_n \sim \pi$
  - Given its assignment ℓ_n, each observation is $y_n|\ell_n \sim \mathcal{N}(y_n|\mu_{\ell_n},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

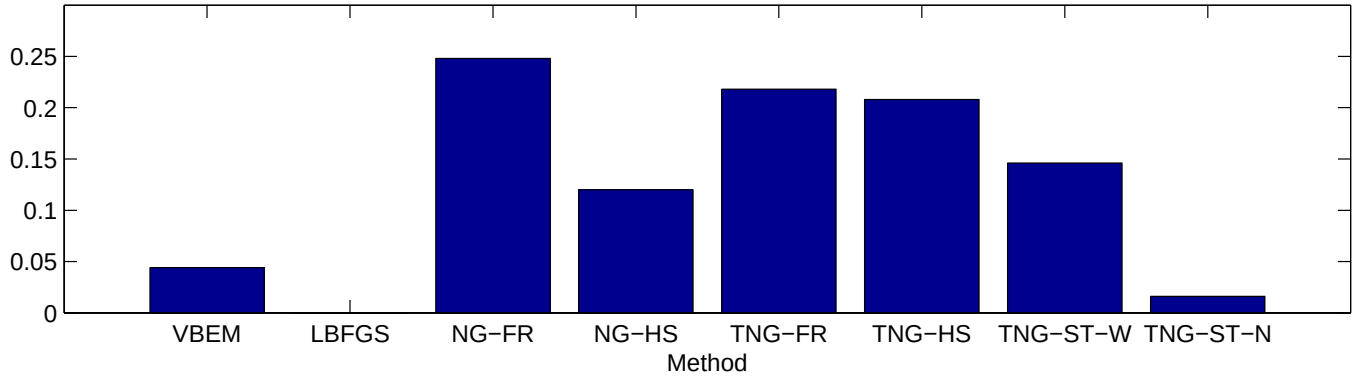
Experiments

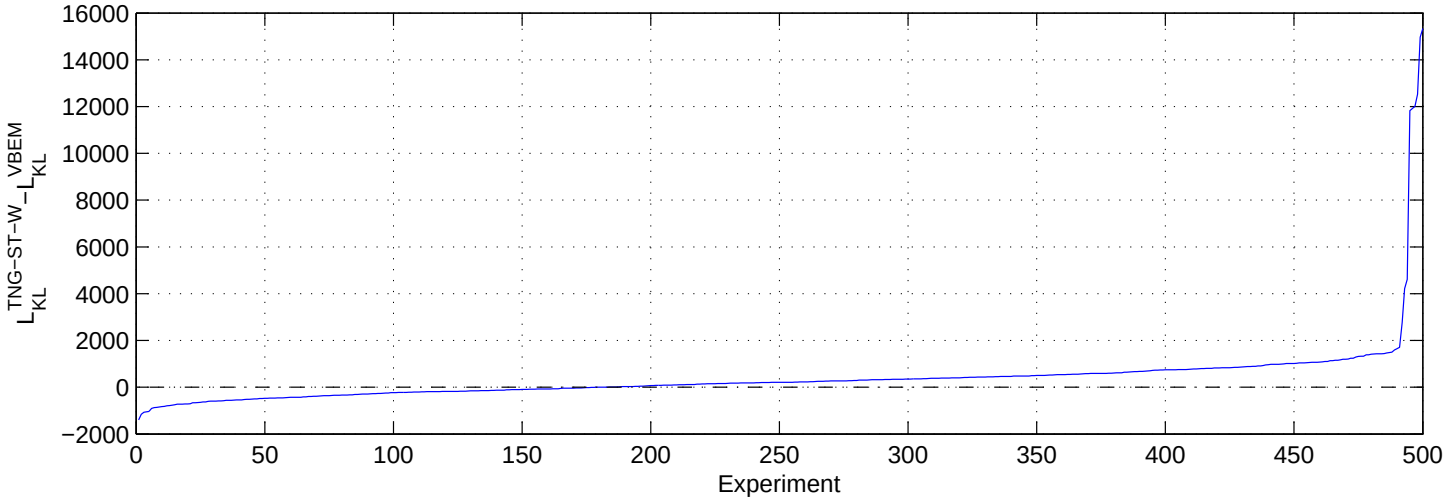
- 500 initializations of 8 algorithms



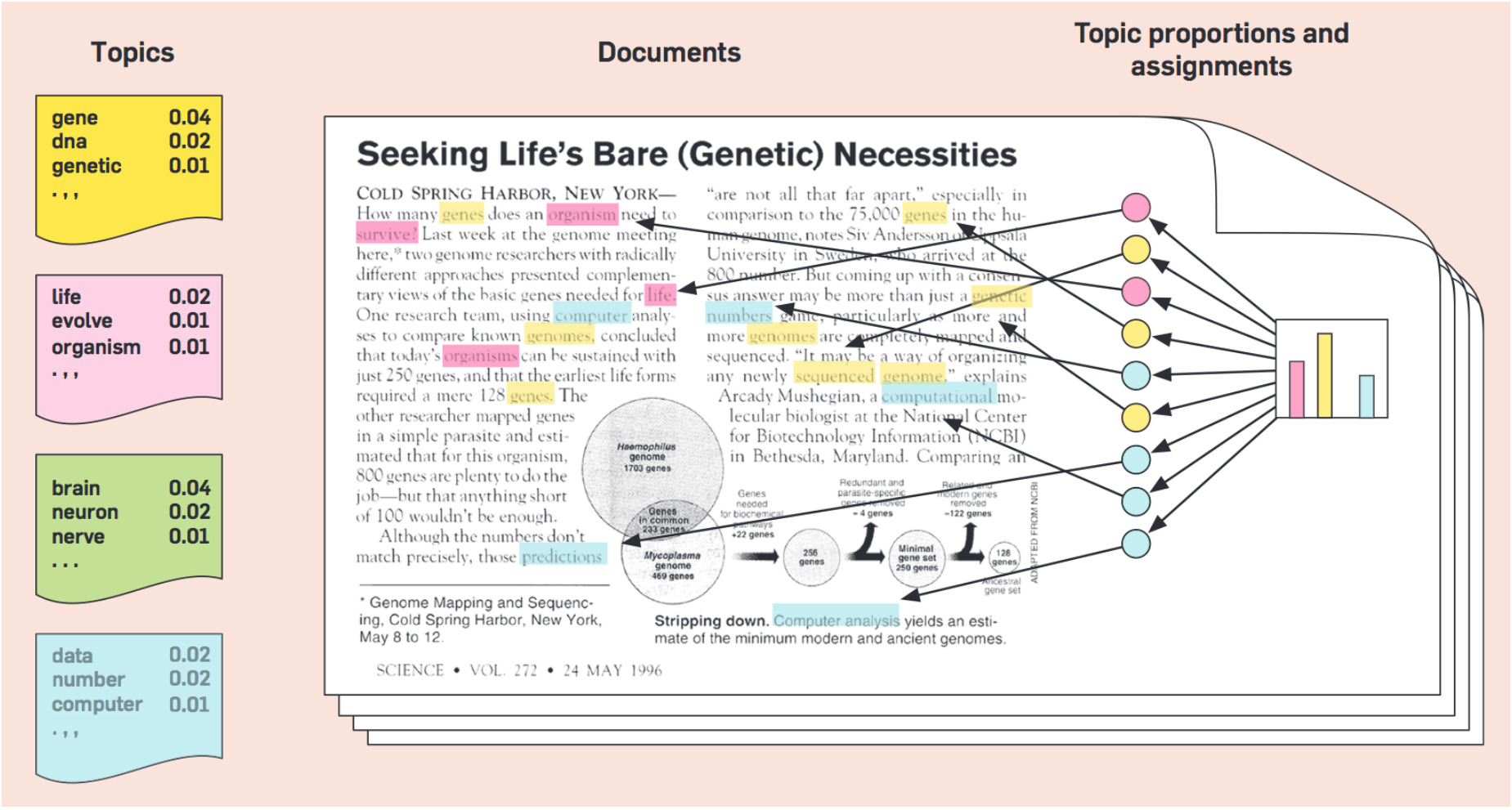
Experiments

- 500 initializations of 8 algorithms





Bonus Slides: Experiments on LDA

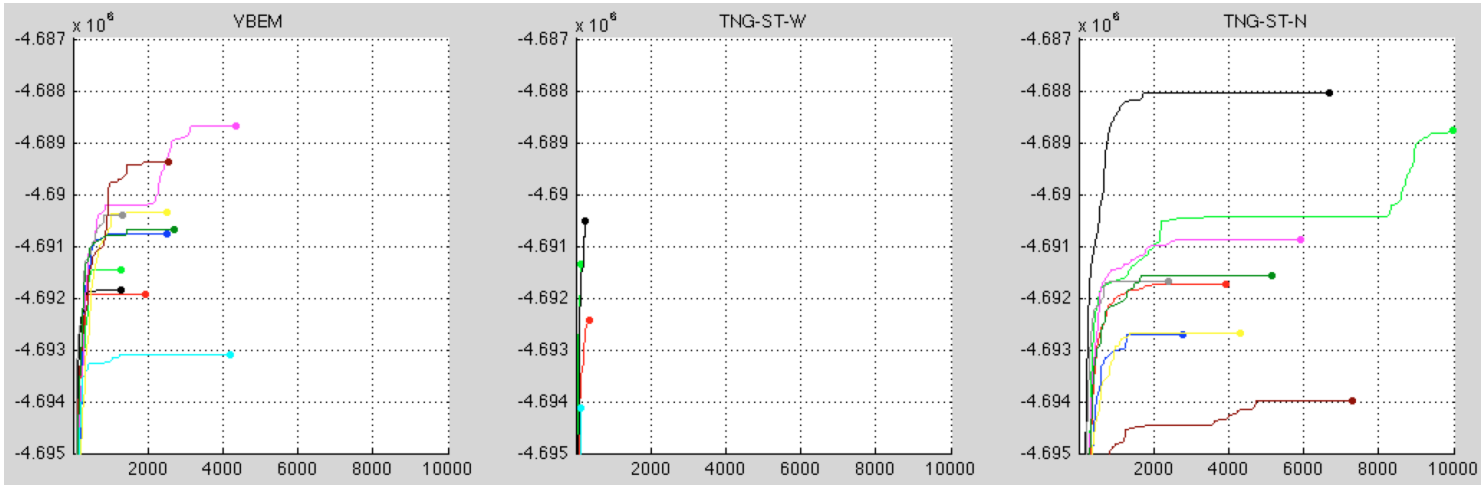


Bonus Slides: Experiments on LDA

- 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