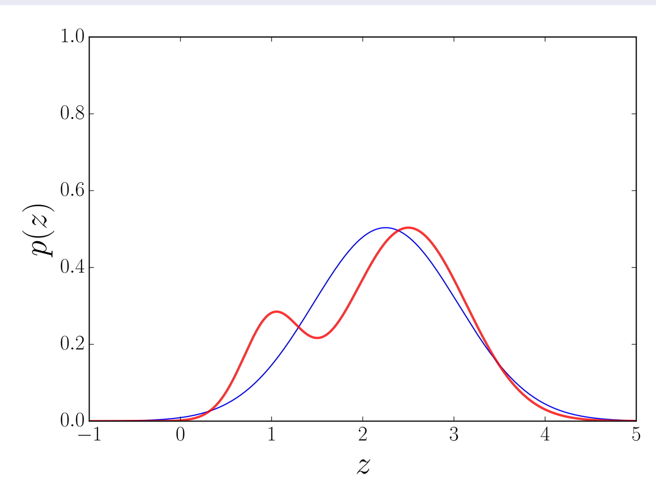
A Divergence Bound for Hybrids of MCMC and VI and an application to Langevin Dynamics and SGVI

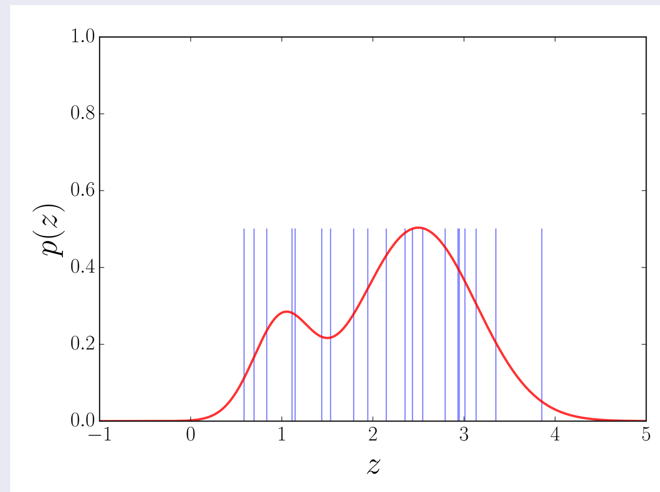
Justin Domke, UMass Amherst

Introduction

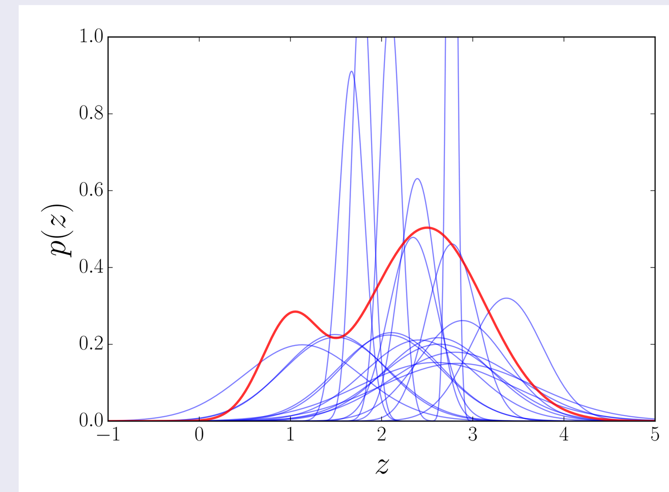
Variational inference (VI): $\min_w KL(q(Z|w)||p(Z))$



Markov chain Monte Carlo (MCMC): Sample from $p(z)$



This paper: Something in the middle



Intuition

VI and MCMC both seek high probability z.

Different **coverage** strategies.

- VI: include entropy $H(w) = -\int_z q(z|w) \log q(z|w)$ in objective.
- MCMC inject randomness.

Idea: Random walk over w. Trade off:

- "How random" the walk is
- "How much" $H(w)$ is favored

Easy to imagine... **but what are we doing**?

Divergence Bounds

Goal: Choose $q(w)$ so $q(z) = \int_w q(w)\, q(z|w) \approx p(z)$.

Impossible: minimize $KL(q(Z)\|p(Z)) = \int_z q(z) \log \frac{q(z)}{p(z)}$.

1st bound: (conditional divergence)

$$KL(q(Z)\|p(z)) \leq \int_w q(w) \int_z q(z|w) \log \frac{q(z|w)}{p(z)} = D_0.$$

2nd bound: (joint divergence) "Augment" with $p(w|z)$.

$$KL(q(Z)\|p(z)) \leq \int_w q(w) \int_z q(z|w) \log \frac{q(z|w)}{p(z)p(w|z)} = D_1.$$

Use **convex combination**: $D_\beta = (1-\beta)D_0 + \beta D_1$

Minimizer of the bound

Thm: Choose $p(w|z) = r(w)q(z|w)/r_z$ where $r_z = \int_w r(w)q(z|w)$ is constant. Then, D_β minimized by

$$\begin{aligned} q^*(w) &= \exp(s(w) - A) \\ s(w) &= \log r(w) - \log r_z \\ &\quad + \mathbb{E}_{q(Z|w)}\left[\beta^{-1} \log p(Z) + (1-\beta^{-1}) \log q(Z|w)\right] \\ A &= \log \int_w \exp(s(w)) \end{aligned}$$

Furthermore, the divergence at q^* is $D^*_\beta = -\beta A$.

Algorithms

Langevin (MCMC): $z \leftarrow z + \frac{\epsilon}{2}\nabla_z \log p(z) + \sqrt{\epsilon}\underbrace{\eta}_{\text{noise}}$

(Stochastic) Gradient VI: $w \leftarrow w - \frac{\epsilon}{2}\nabla_w KL(q(Z|w)\|p(Z))$

Hybrid (this paper): (Apply Langevin to q^* and scale ϵ)

$$w \leftarrow w + \frac{\epsilon}{2}\nabla_w \left(-KL(q(Z|w)\|p(Z)) - \beta H(w) + \beta \log r_\beta(w) \right) + \sqrt{\beta\epsilon\eta}$$

Becomes VI when $\beta \to 0$ VI (easy)

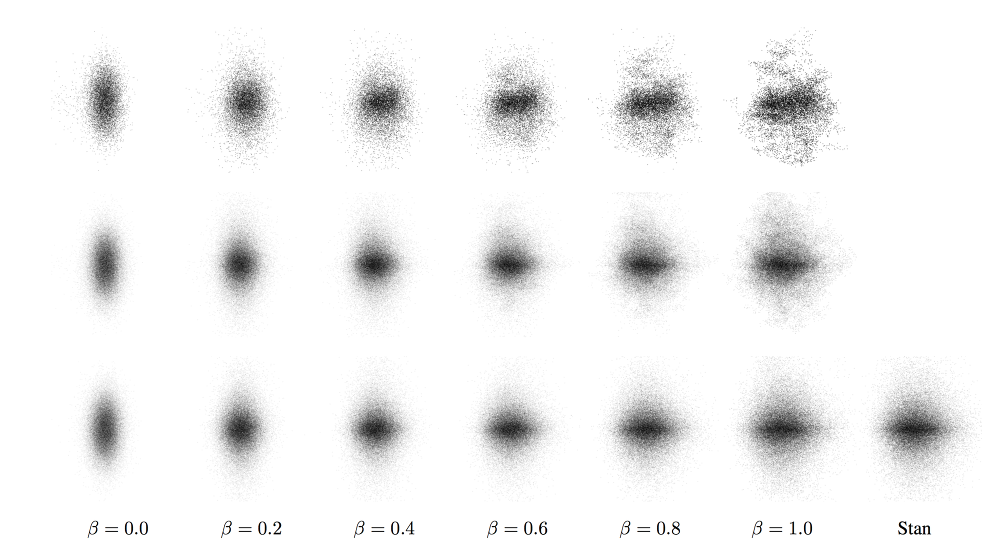
Becomes Langevin (on z) when $\beta \to 1$

- r_β likes w where $q(Z|w)$ concentrates.

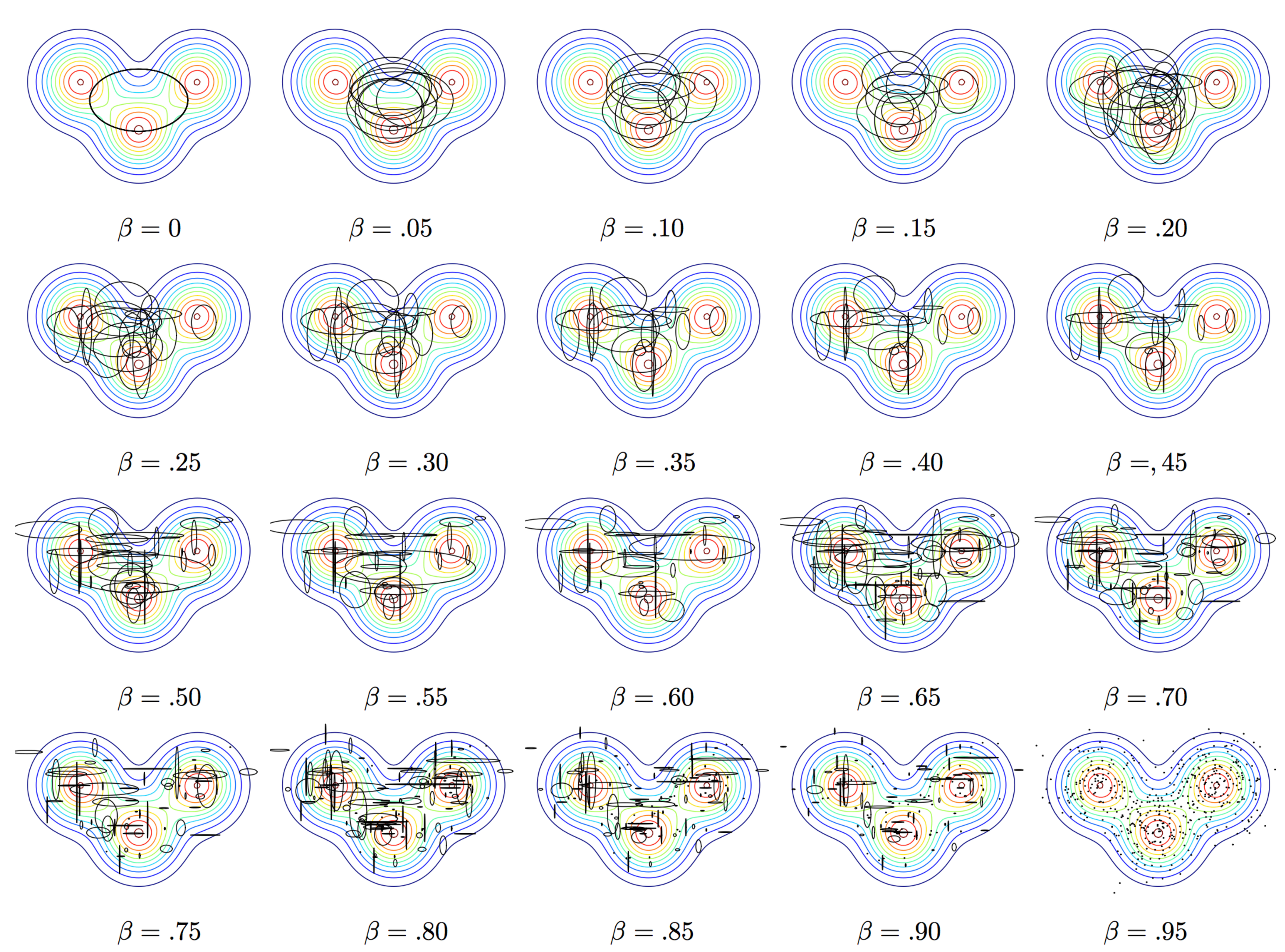
Algorithmic details

- Use a **diagonal Gaussian** for $q(z|w)$, with $w = (\mu, \nu)$, $\nu_i = \log_{10} \sigma_i$.
- To estimate gradient, use standard tricks from SGVI:
  - For Bayesian inference, estimate $\log p(z)$ using subsampling.
  - Reparameterization trick: $\nabla_w \mathbb{E}_{q(Z|w)}[\log p(Z)] \to \mathbb{E}_R[\nabla_w \log p(z_{R,w})]$, then sample R and apply autodiff.
  - Use closed form for entropy $H(w) = -\mathbb{E}_{q(Z|w)}[\log q(Z|w]$.
- Use (improper) $r_\beta(w) \propto \prod_i \mathcal{N}(\nu_i|u_\beta, 1)$. Numerically optimize u_β to minimize D^*_β when $p(z)$ is a standard Gaussian.

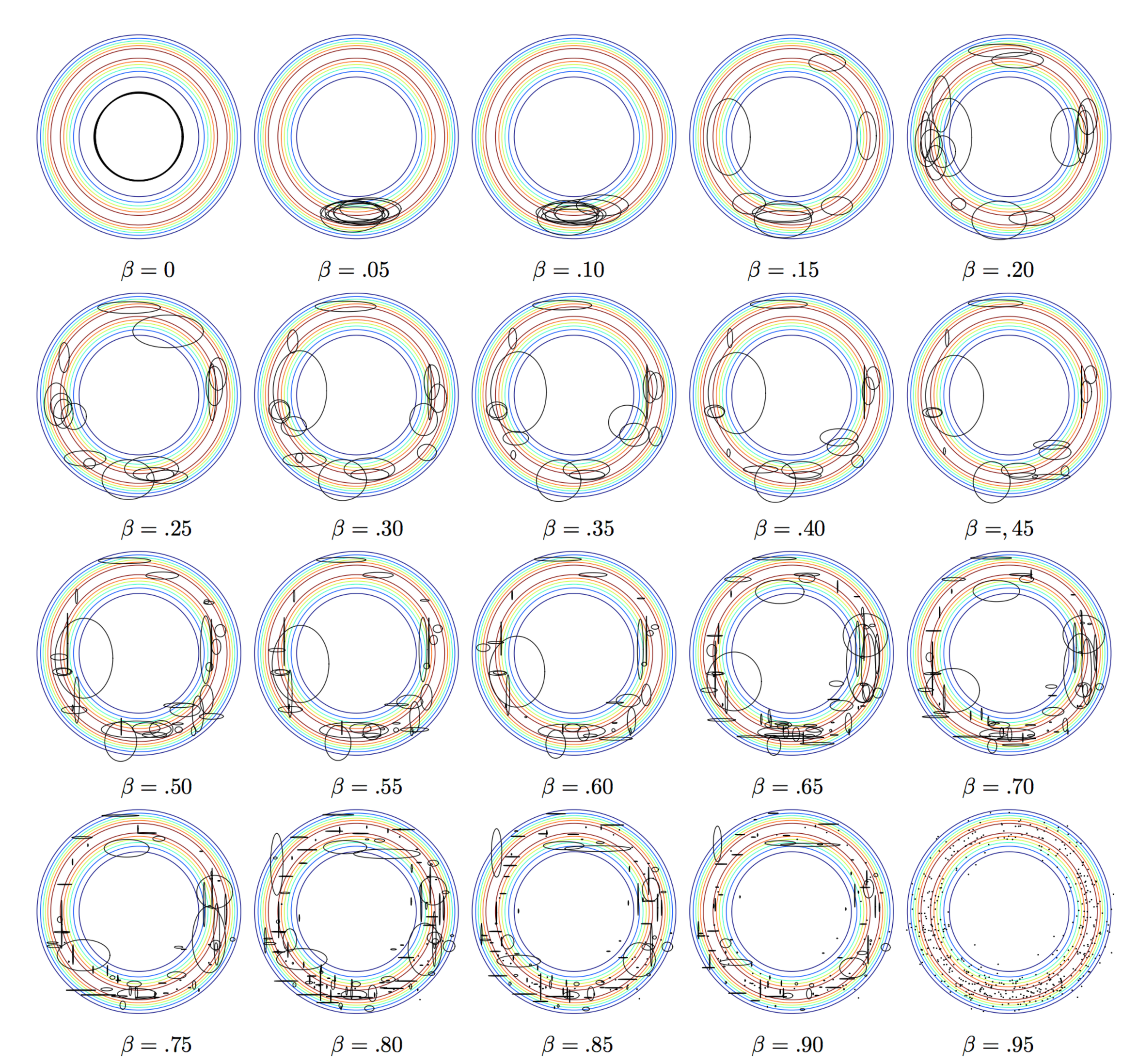
Ionosphere, 10^4 / 10^5 / 10^6 iterations



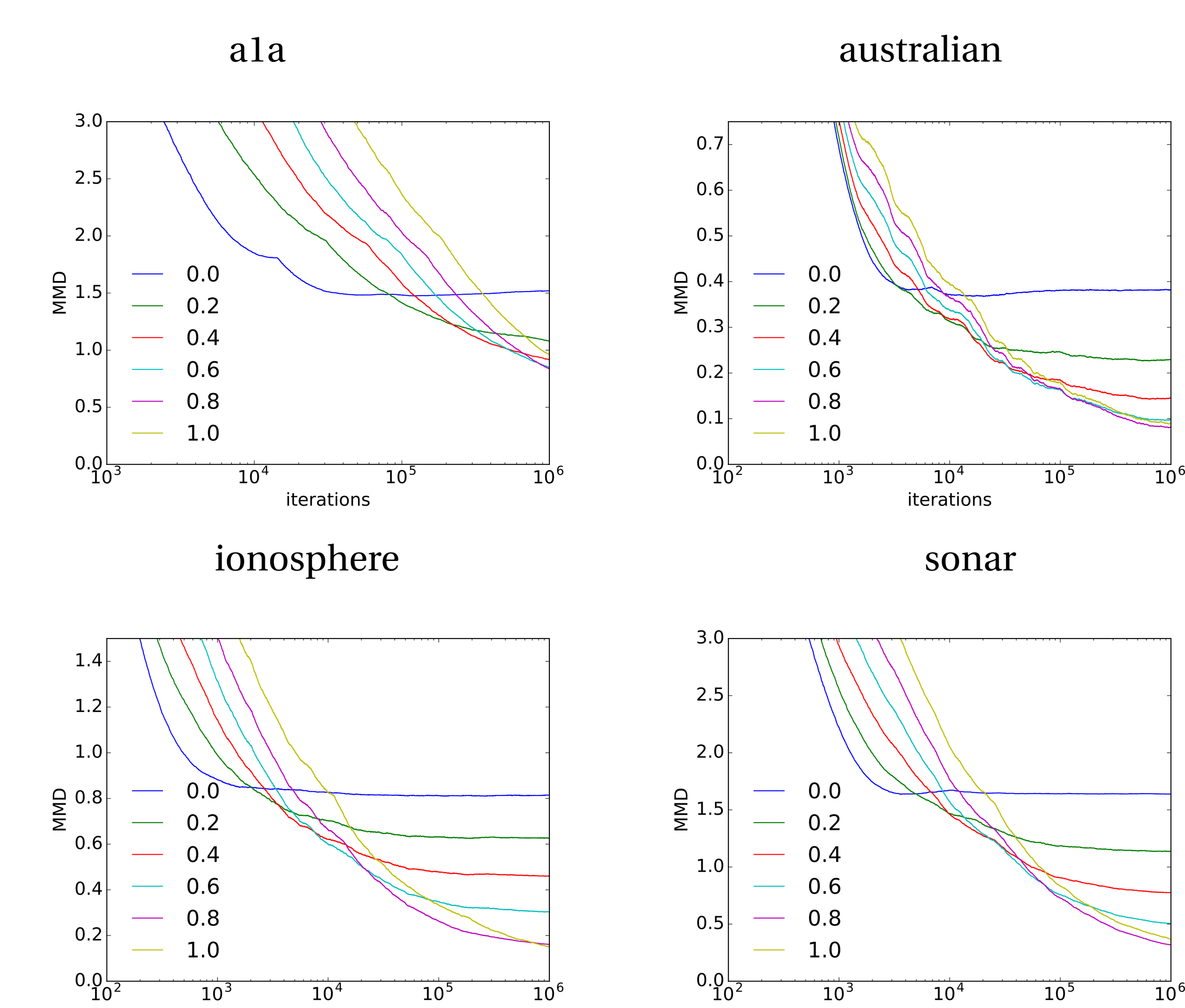
Toy 2-D Example



Toy 2-D Example



Logistic Regression



Toy 1-D Visualization