A Divergence Bound for Hybrids of MCMC and Variational Inference and an Application to Langevin Dynamics and SGVI

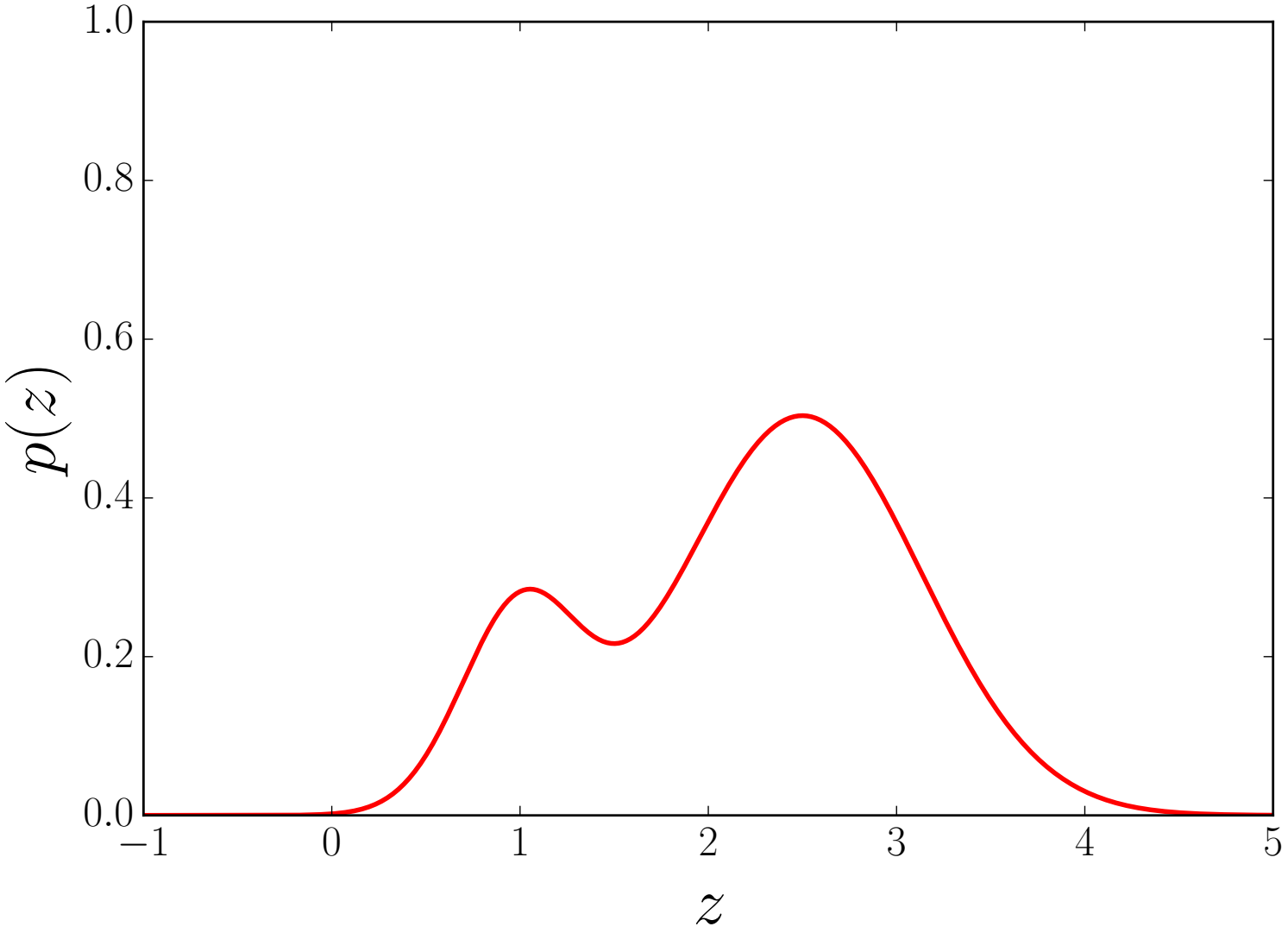
Justin Domke

University of Massachustts, Amherst

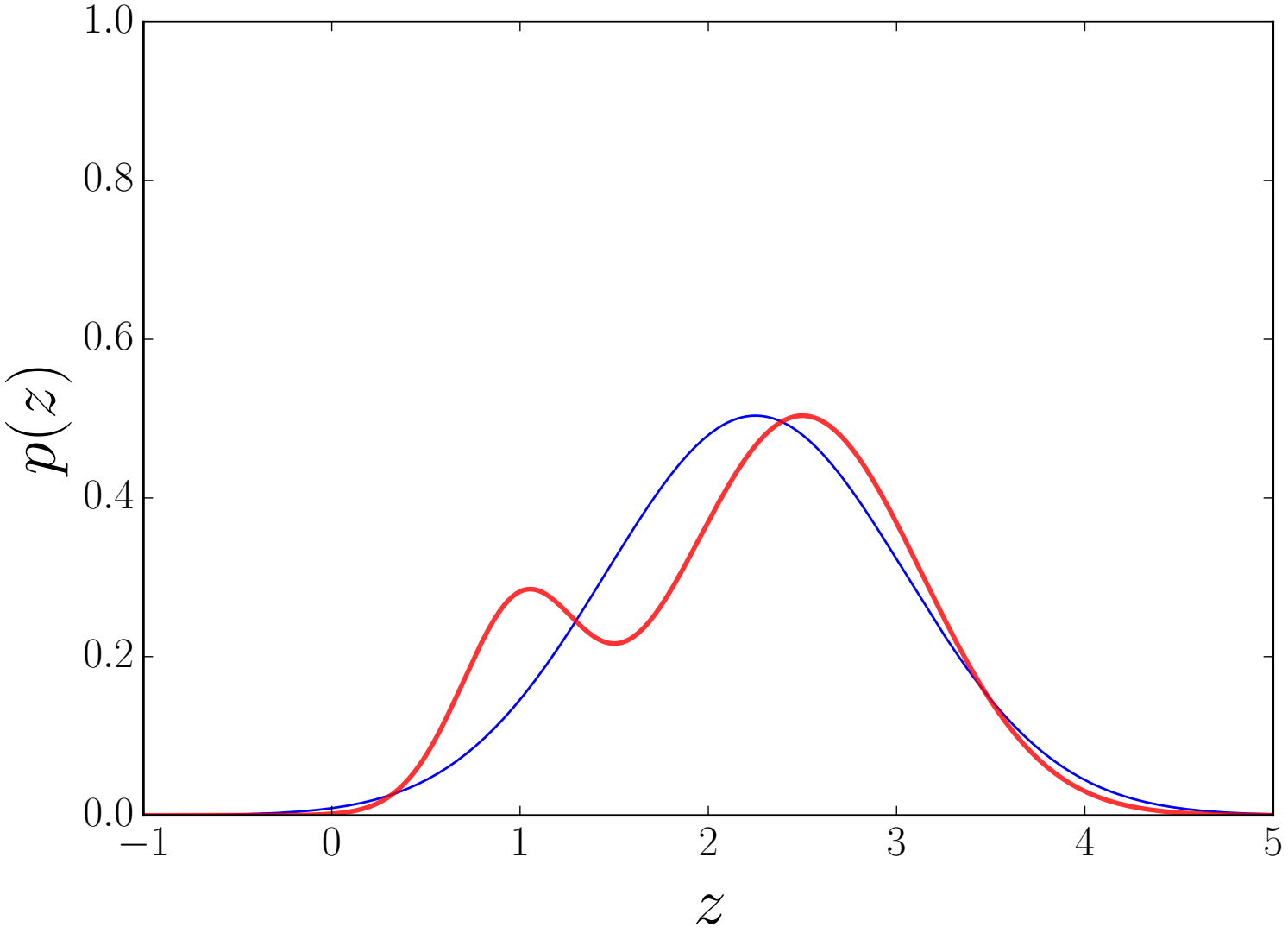
ICML 2017

Outline

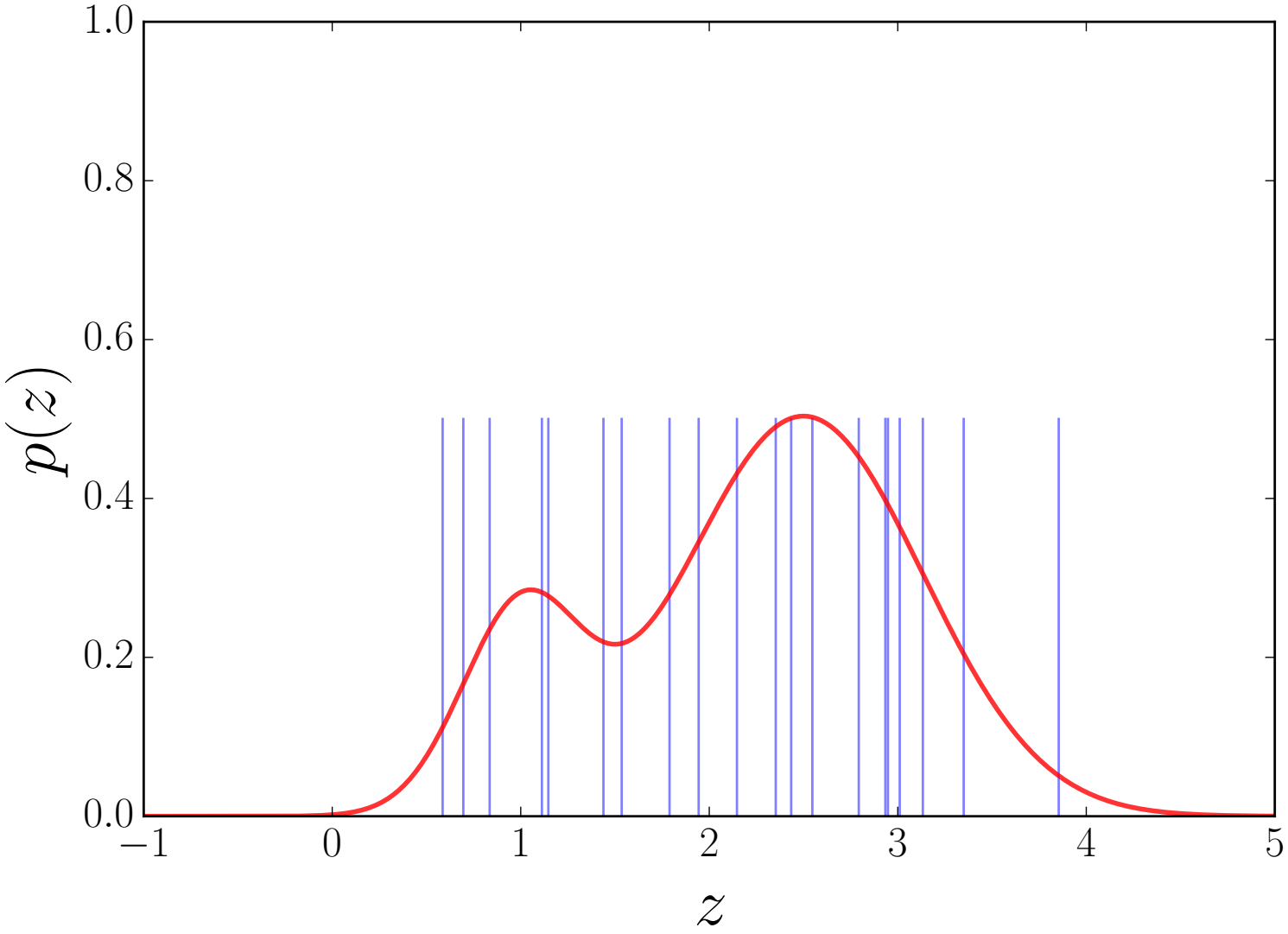
Outline



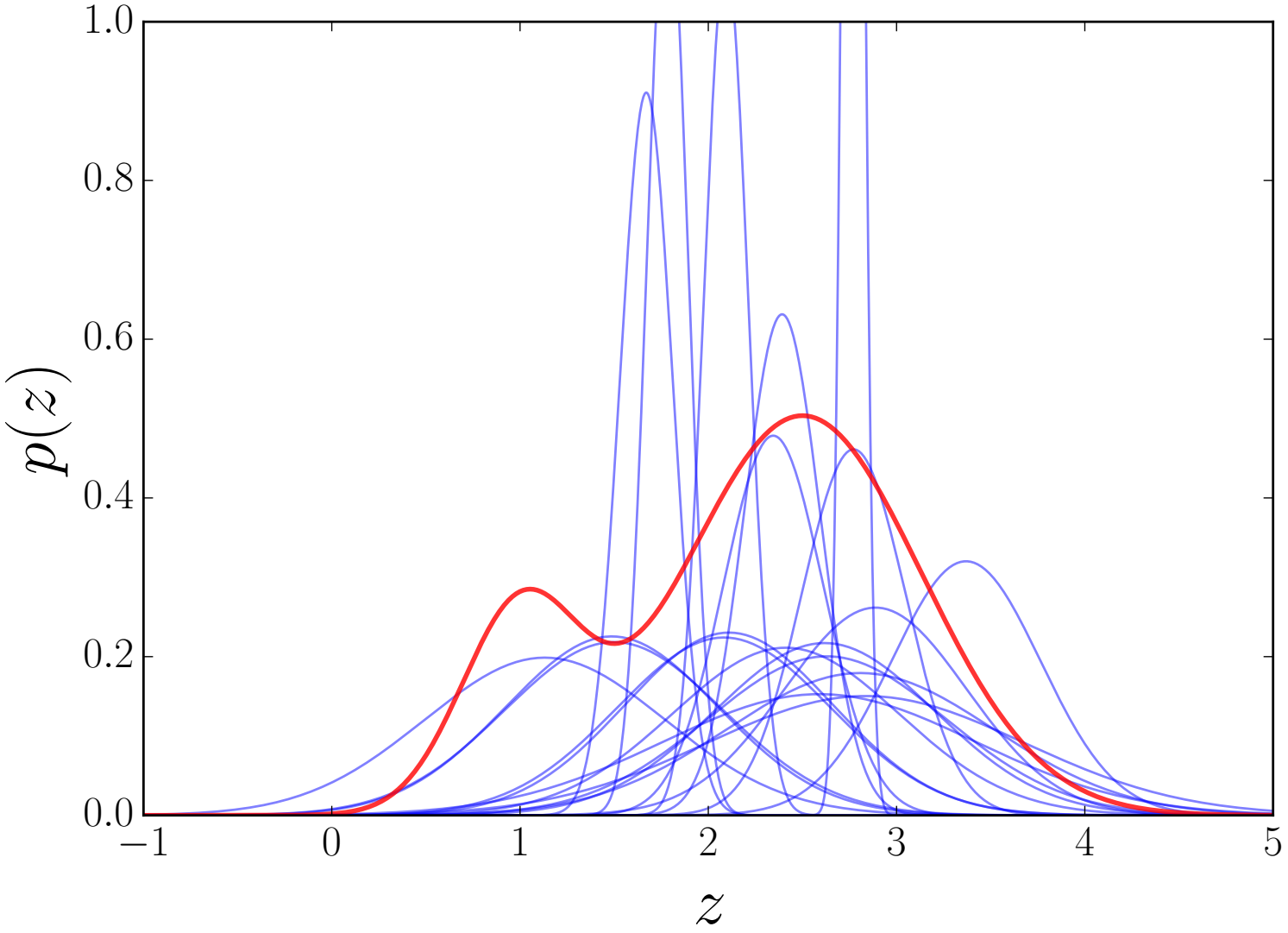
Target distribution: $p(z)$



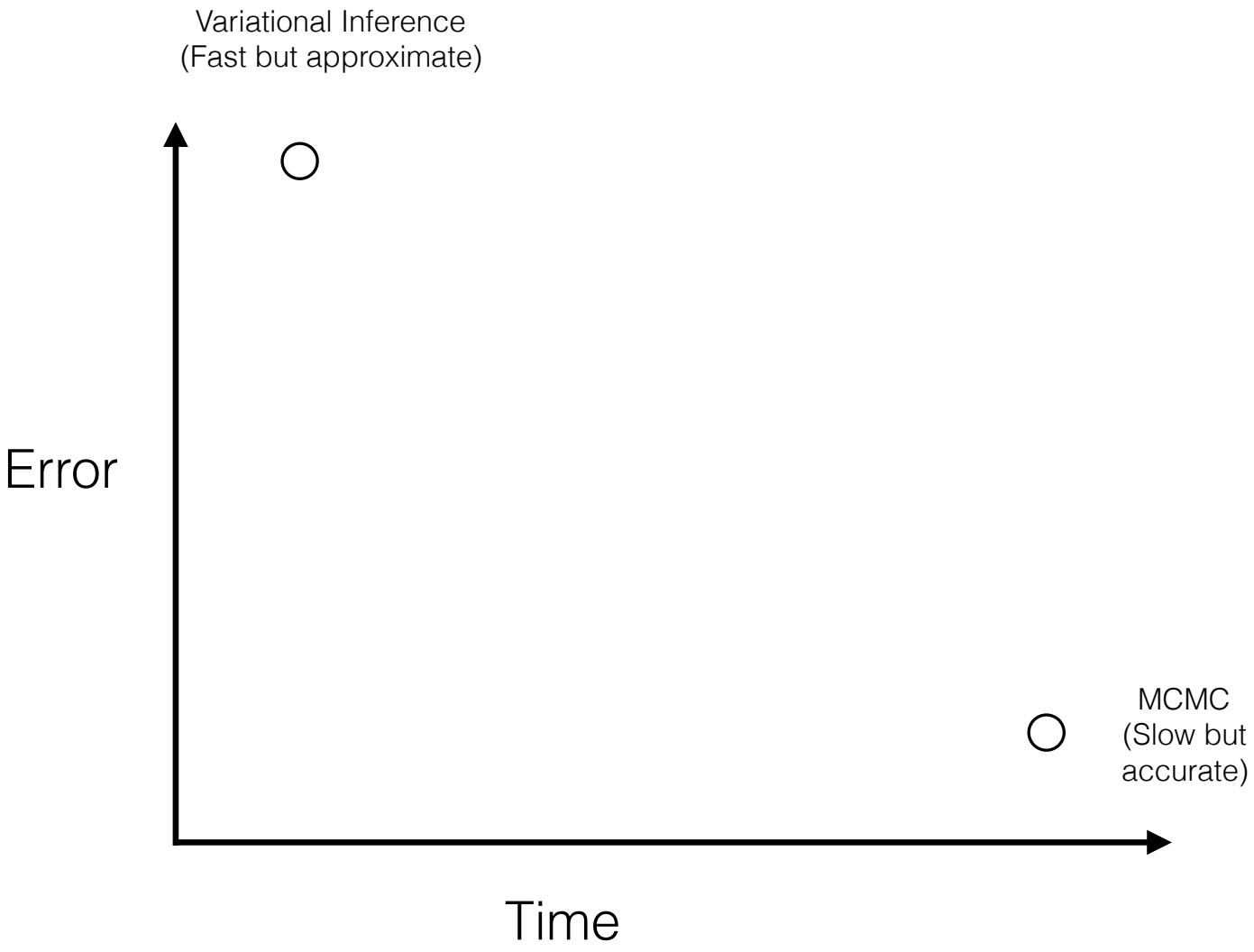
Variational inference (VI): $\min_{w} KL(q(Z|w)||p(Z))$

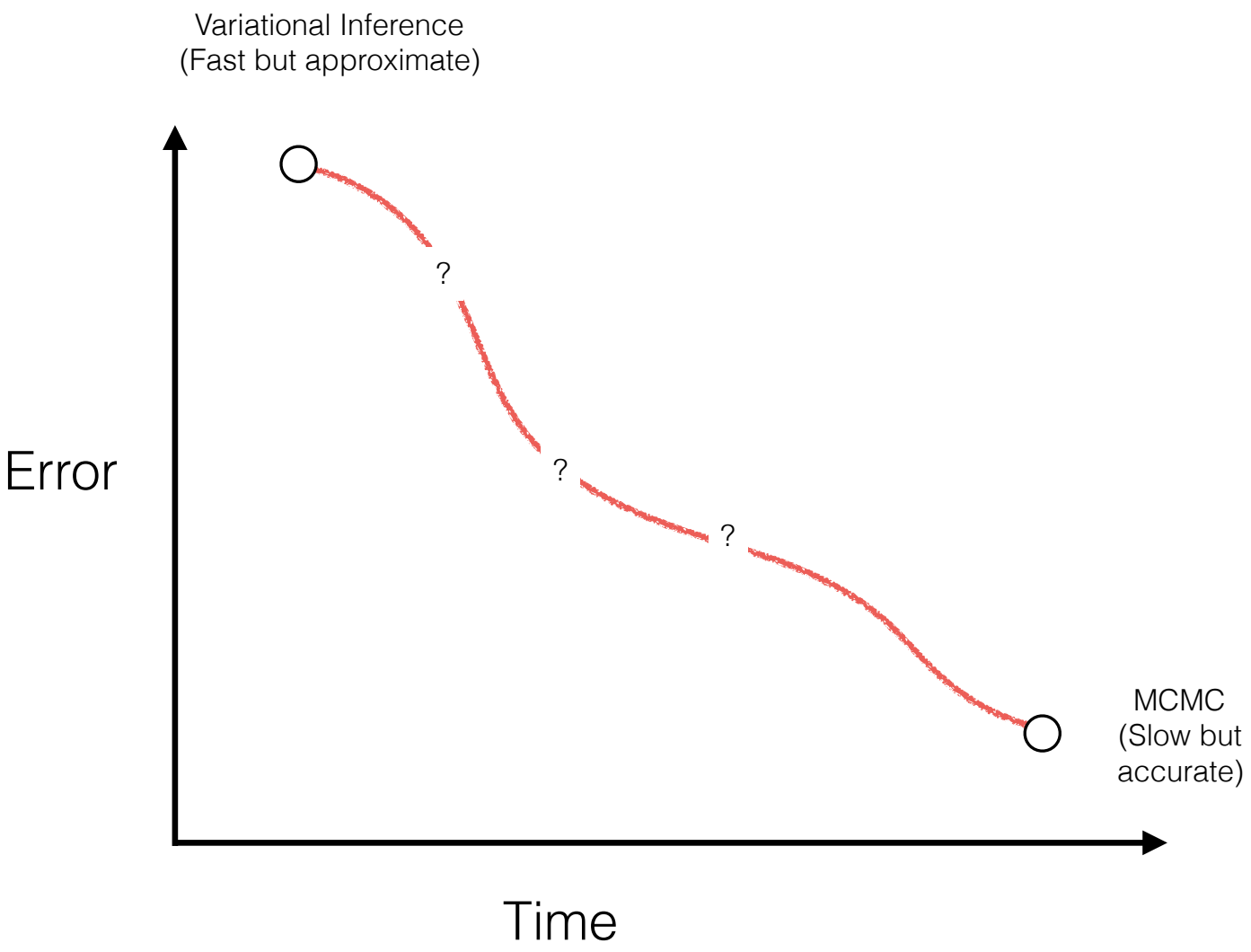


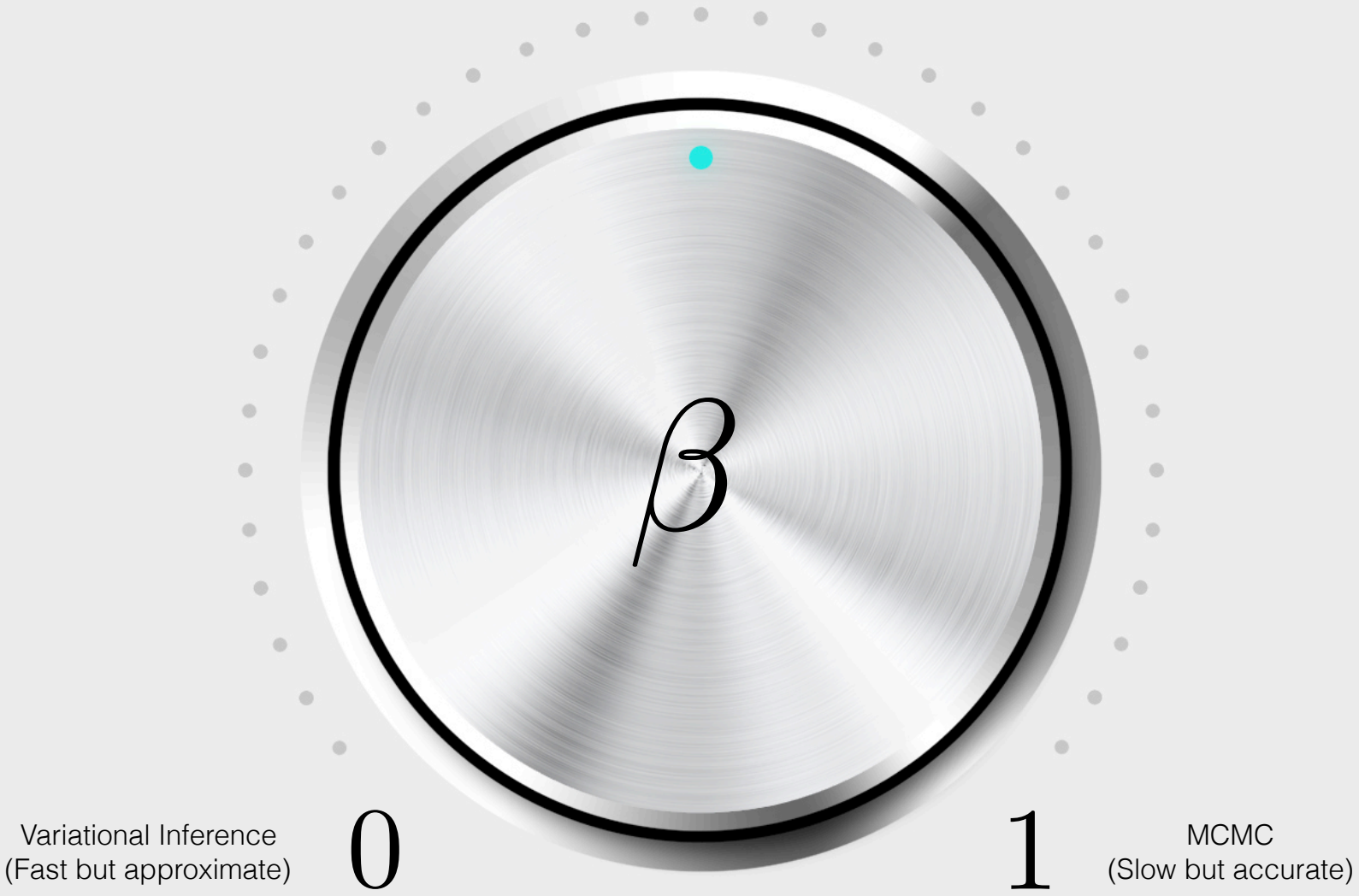
Makov chain Monte Carlo (MCMC): Sample from $p(z)$



This paper: A hybrid interpolating between VI and MCMC


Variational Inference
(Fast but approximate)
Error
MCMC
(Slow but
accurate)
Time


Variational Inference
(Fast but approximate)
?
?
?
Error
MCMC
(Slow but
accurate)
Time


β
Variational Inference
(Fast but approximate)
0
1
MCMC
(Slow but accurate)

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**?

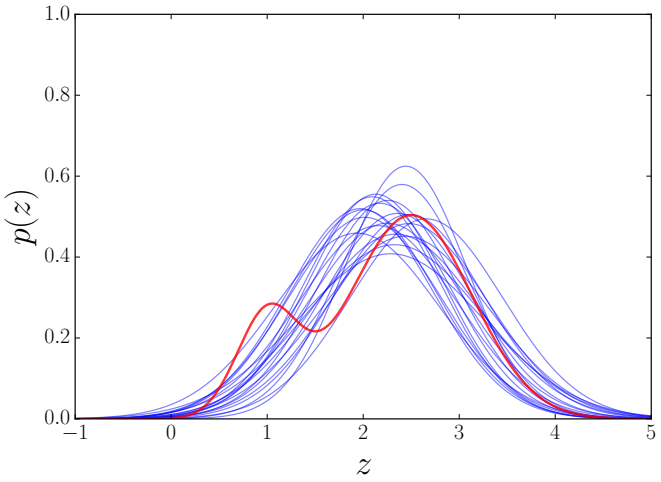
Intuition: VI and MCMC both seek high probability z.

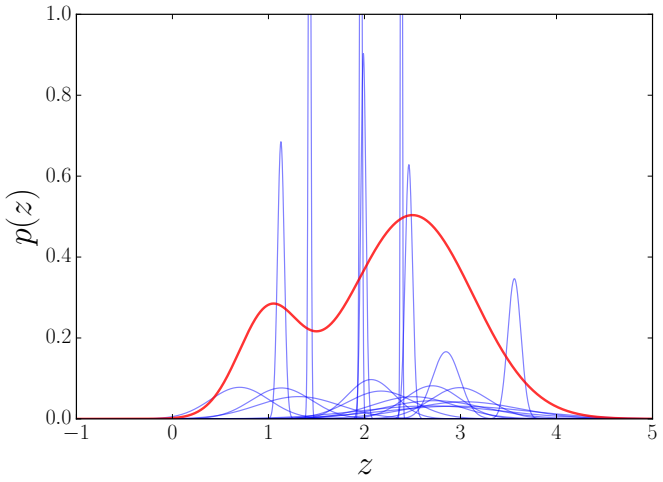
Different **coverage** strategies.

- 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?**

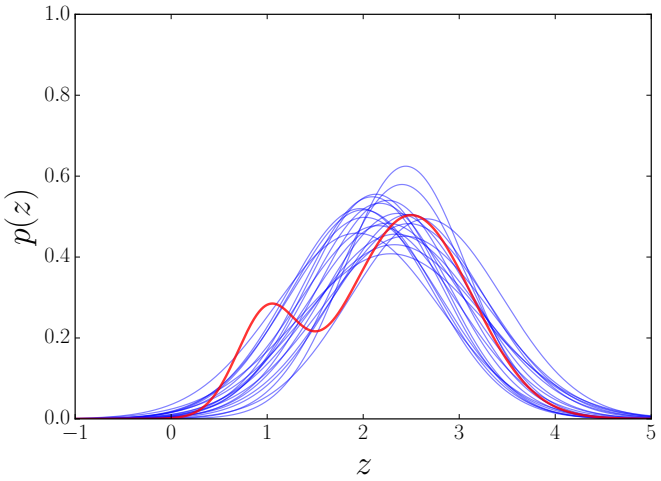
Intuition: VI and MCMC both seek high probability z.

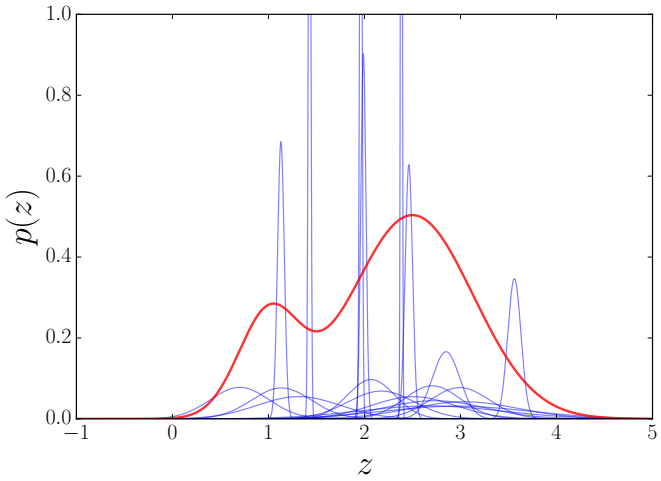
Different **coverage** strategies.

- 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**?

First contribution: Divergence bound and it's minimizer

Define $q(z) = \int_w \underbrace{q(w)}_{\text{distr. over w}} \underbrace{q(z|w)}_{\text{variational family}}$.

Impossible: Directly choose $q(w)$ so that $KL(q(z)\|p(z))$ is small.

Instead: upper-bound: $KL(q(z)\|p(z)) \leq D_\beta$.

Possible: Derive $q^*(w)$ that minimizes D_β.

First contribution: Divergence bound and it's minimizer

Define $q(z) = \int_w \underbrace{q(w)}_{\text{distr. over w}} \underbrace{q(z|w)}_{\text{variational family}}$.

Impossible: Directly choose $q(w)$ so that $KL(q(z)\|p(z))$ is small.

Instead: upper-bound: $KL(q(z)\|p(z)) \leq D_\beta$.

Possible: Derive $q^*(w)$ that minimizes D_β.

First contribution: Divergence bound and it's minimizer

Define $q(z) = \int_w \underbrace{q(w)}_{\text{distr. over w}} \underbrace{q(z|w)}_{\text{variational family}}$.

Impossible: Directly choose $q(w)$ so that $KL(q(z)\|p(z))$ is small.

Instead: upper-bound: $KL(q(z)\|p(z)) \leq D_\beta$.

Possible: Derive $q^*(w)$ that minimizes D_β.

First contribution: Divergence bound and it's minimizer

Define $q(z) = \int_w \underbrace{q(w)}_{\text{distr. over w}} \underbrace{q(z|w)}_{\text{variational family}}$.

Impossible: Directly choose $q(w)$ so that $KL(q(z)\|p(z))$ is small.

Instead: upper-bound: $KL(q(z)\|p(z)) \leq D_\beta$.

Possible: Derive $q^*(w)$ that minimizes D_β.

Second contribution: How to sample from q^*?

Langevin (MCMC): Gradient ascent (in z) with injected noise on

$$\log p(z).$$

(Stochastic) Gradient VI: Gradient ascent on

$$-KL(q(Z|w)\|p(Z)).$$

Hybrid: Gradient ascent (in w) with (varying) injected noise on

$$-KL(q(Z|w)\|p(Z)) - \beta H(w).$$

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

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

Second contribution: How to sample from q^*?

Langevin (MCMC): Gradient ascent (in z) with injected noise on

$$\log p(z).$$

(Stochastic) Gradient VI: Gradient ascent on

$$-KL(q(Z|w)\|p(Z)).$$

Hybrid: Gradient ascent (in w) with (varying) injected noise on

$$-KL(q(Z|w)\|p(Z)) - \beta H(w).$$

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

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

Second contribution: How to sample from q^*?

Langevin (MCMC): Gradient ascent (in z) with injected noise on

$$\log p(z).$$

(Stochastic) Gradient VI: Gradient ascent on

$$-KL(q(Z|w)\|p(Z)).$$

Hybrid: Gradient ascent (in w) with (varying) injected noise on

$$-KL(q(Z|w)\|p(Z)) - \beta H(w).$$

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

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

Outline

Divergence Bounds

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

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

1st bound: (conditional divergence)

$$KL(q(Z)\|p(z)) \leq KL(q(Z|W)\|p(Z)) = D_0.$$

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

$$KL(q(Z)\|p(z)) \leq KL(q(Z,W)\|p(Z,W)) = D_1.$$

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

Divergence Bounds

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

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

1st bound: (conditional divergence)

$$KL(q(Z)\|p(z)) \leq KL(q(Z|W)\|p(Z)) = D_0.$$

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

$$KL(q(Z)\|p(z)) \leq KL(q(Z,W)\|p(Z,W)) = D_1.$$

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

Divergence Bounds

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

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

1st bound: (conditional divergence)

$$KL(q(Z)\|p(z)) \leq KL(q(Z|W)\|p(Z)) = D_0.$$

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

$$KL(q(Z)\|p(z)) \leq KL(q(Z,W)\|p(Z,W)) = D_1.$$

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

Divergence Bounds

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

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

1st bound: (conditional divergence)

$$KL(q(Z)\|p(z)) \leq KL(q(Z|W)\|p(Z)) = D_0.$$

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

$$KL(q(Z)\|p(z)) \leq KL(q(Z,W)\|p(Z,W)) = D_1.$$

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

Minimizer of the bound

Thm: Choose[1] $p(w|z) \propto r(w)q(z|w)$. Then, D_β is minimized by

$$\log q^*(w) = \log r(w) + \mathbb{E}_{q(Z|w)}\left[\beta^{-1}\log p(Z) + (1-\beta^{-1})\log q(Z|w)\right] + \mathrm{C}.$$

- Resembles a divergence (similar to VI).
- But defines a distribution (similar to MCMC).

[1] Assume $\int_w r(w)q(z|w)$ is constant.

Outline

Algorithm

To get final algorithm: Apply Langevin dynamics to $q^*(w)$ (and scale step-size with β). Leads to

$$w \leftarrow w + \frac{\varepsilon}{2}\left(\nabla_w \mathbb{E}[\log p(Z) + (\beta - 1)\log q(Z|w)] + \beta \log r_\beta(w)\right) + \sqrt{\varepsilon\beta}\,\eta.$$

Borrow "local reparamaterization trick" from VI to estimate grad.

Algorithm

To get final algorithm: Apply Langevin dynamics to $q^*(w)$ (and scale step-size with β). Leads to

$$w \leftarrow w + \frac{\varepsilon}{2}\left(\nabla_w \mathbb{E}[\log p(Z) + (\beta - 1)\log q(Z|w)] + \beta \log r_\beta(w)\right) + \sqrt{\varepsilon\beta}\,\eta.$$

Borrow "local reparamaterization trick" from VI to estimate grad.

Algorithm

To get final algorithm: Apply Langevin dynamics to $q^*(w)$ (and scale step-size with β). Leads to

$$w \leftarrow w + \frac{\varepsilon}{2}\left(\nabla_w \mathbb{E}[\log p(Z) + (\cancel{\beta} - 1)\log q(Z|w)] + \cancel{\beta \log r_\beta(w)}\right) + \cancel{\sqrt{\varepsilon\beta}\,\eta}.$$

Borrow "local reparamaterization trick" from VI to estimate grad.

Reductions:

- If $\beta \to 0$, becomes VI:

$$w \leftarrow w + \frac{\varepsilon}{2}\nabla_w \mathbb{E}[\log p(Z) - \log q(Z|w)]$$

Algorithm

To get final algorithm: Apply Langevin dynamics to $q^*(w)$ (and scale step-size with β). Leads to

$$w \leftarrow w + \frac{\varepsilon}{2}\left(\nabla_w \mathbb{E}[\log p(Z) + \cancel{(\beta-1)\log q(Z|w)}] + \beta \log r_\beta(w)\right) + \sqrt{\varepsilon\beta}\,\eta.$$

Borrow "local reparamaterization trick" from VI to estimate grad.

Reductions:

- If $\beta \to 1$, becomes:

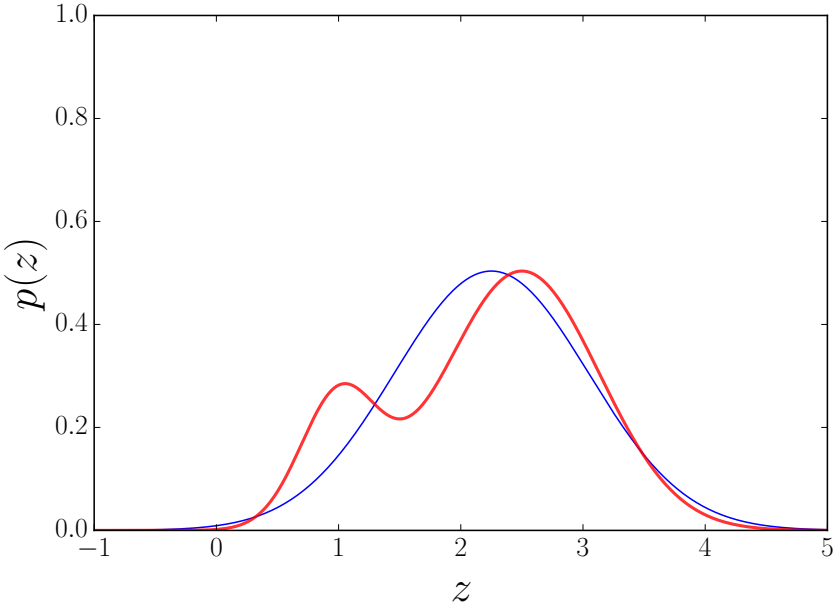
$$w \leftarrow w + \frac{\varepsilon}{2}\left(\nabla_w \mathbb{E}[\log p(Z)] + \log r_\beta(w)\right) + \sqrt{\varepsilon}\,\eta$$

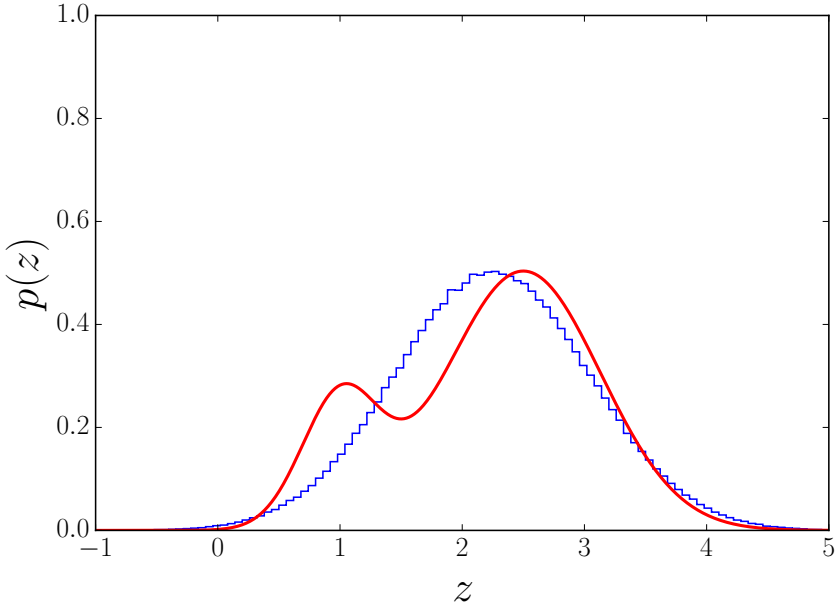
  But when $\beta \to 1$, $r_\beta(w)$ makes $q(z|w)$ concentrate around a single point. Thus, equivalent to Langevin dynamics on z.

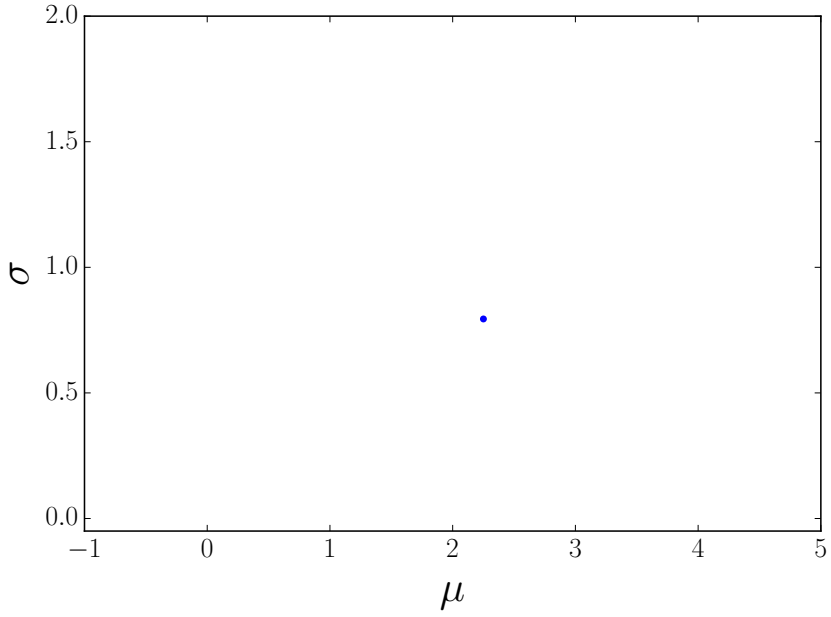
Outline

Outline

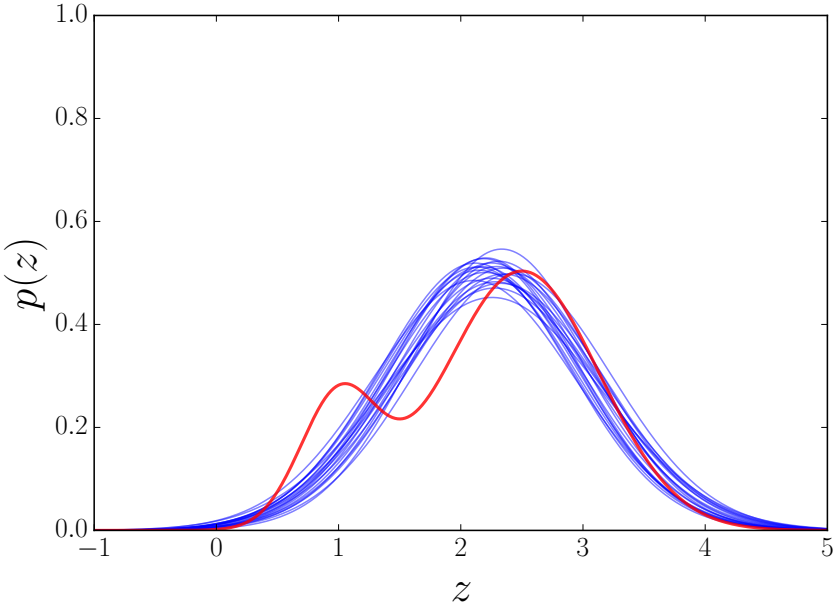
$\beta = 0$

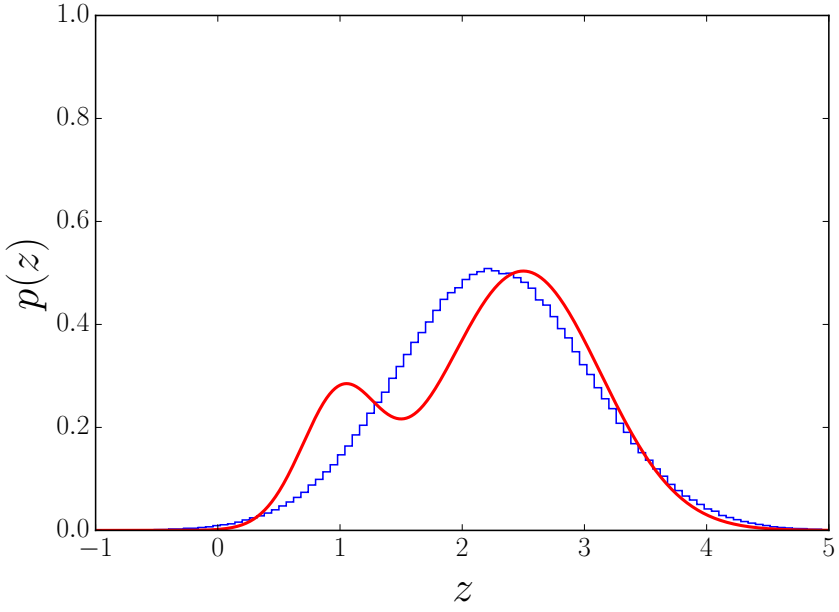

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z

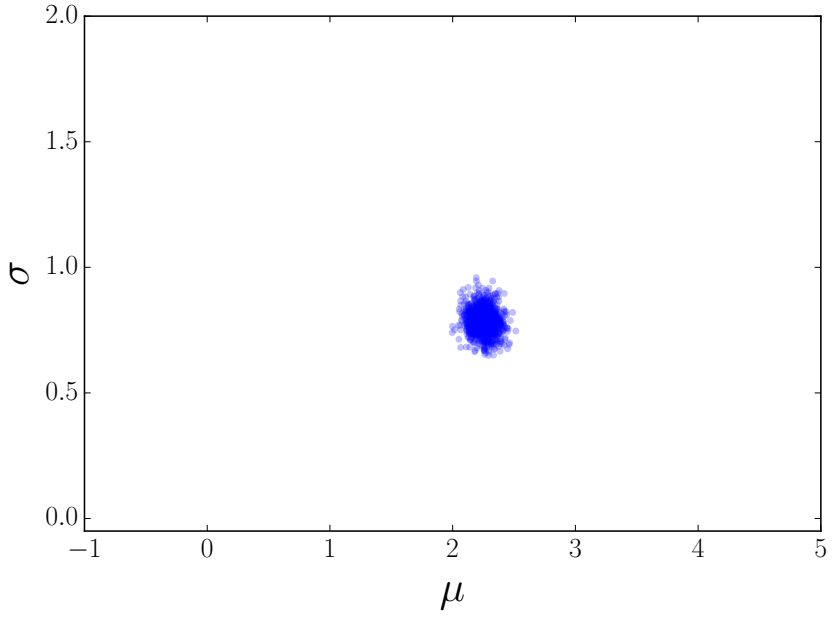

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z


2.0
1.5
1.0
0.5
0.0
σ
−1
0
1
2
3
4
5
μ

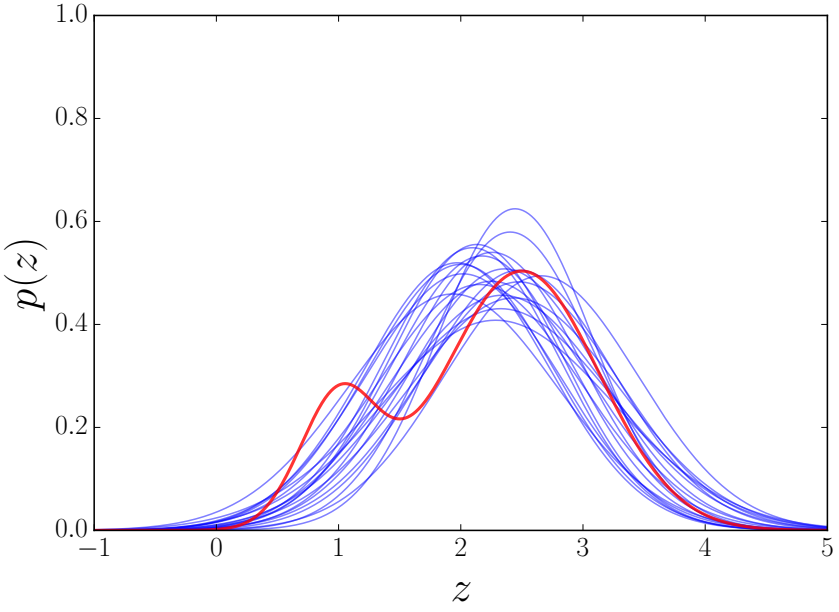
$\beta = 0.01$

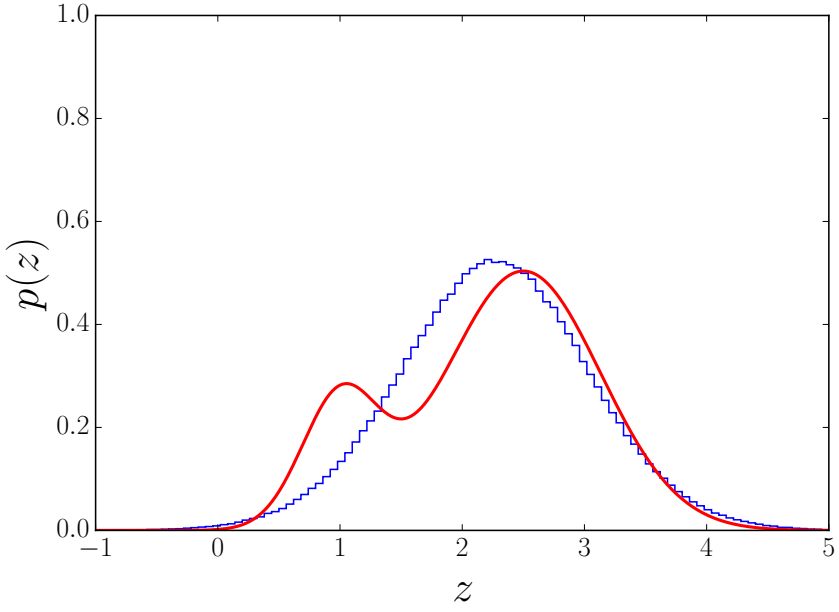

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z

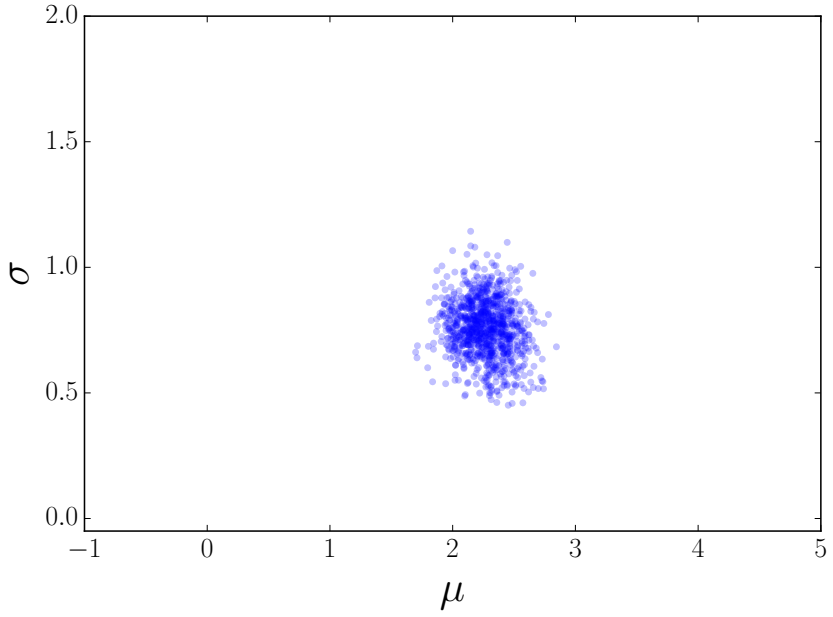

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z


2.0
1.5
1.0
0.5
0.0
σ
−1
0
1
2
3
4
5
μ

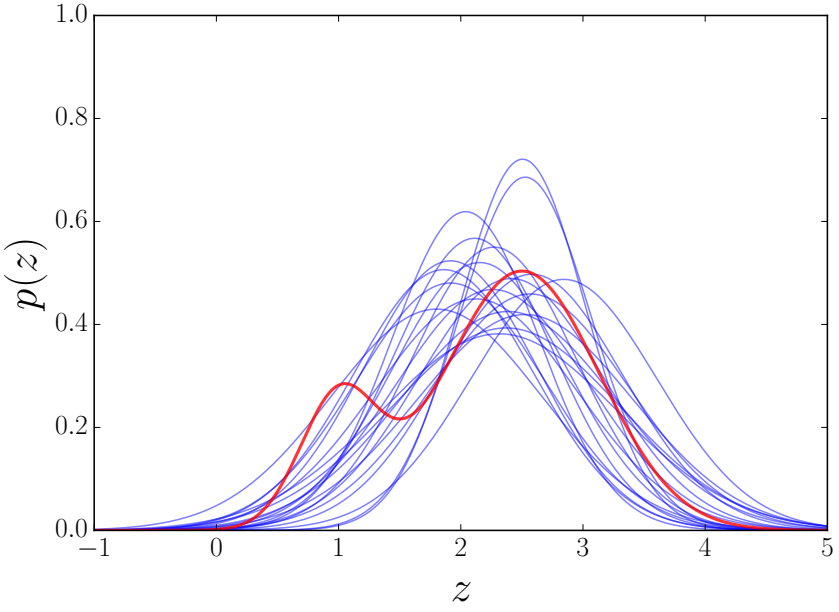
$\beta = 0.05$

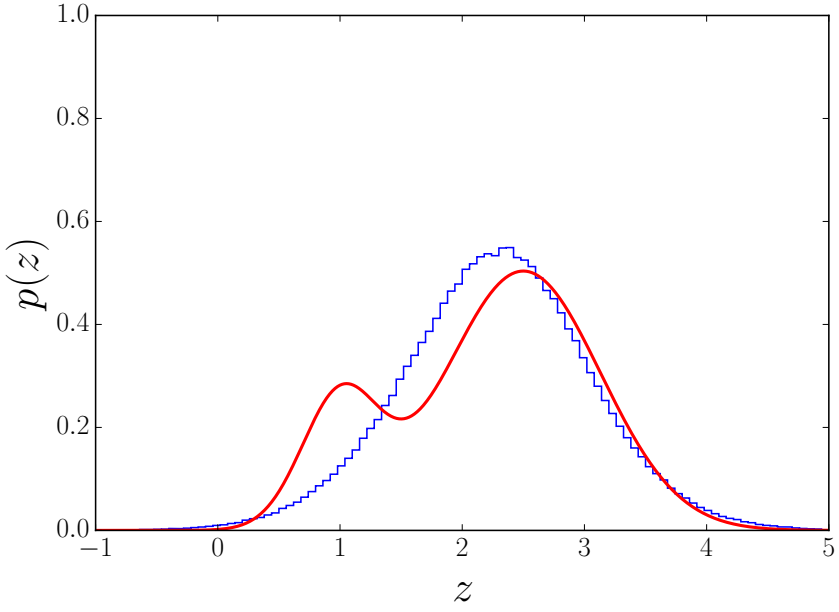

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z

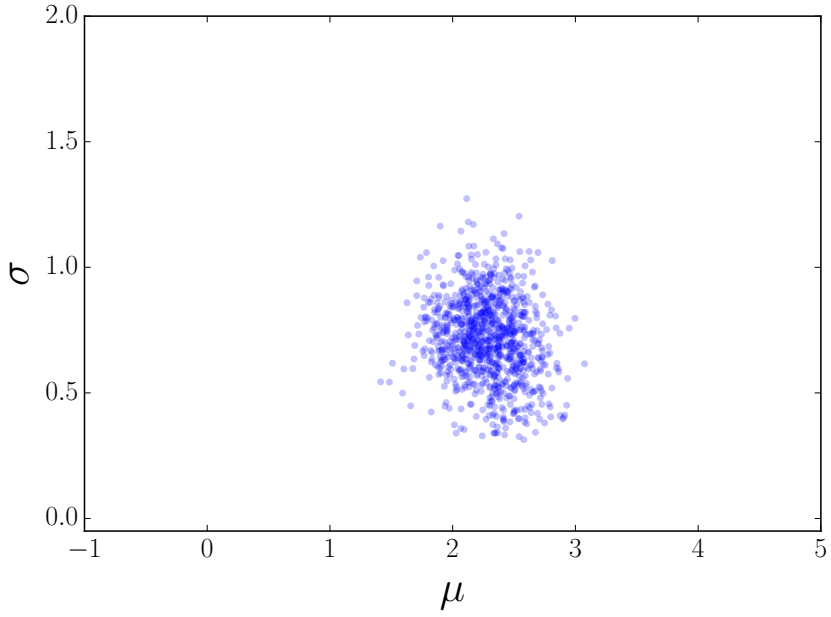

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z


2.0
1.5
1.0
0.5
0.0
σ
−1
0
1
2
3
4
5
μ

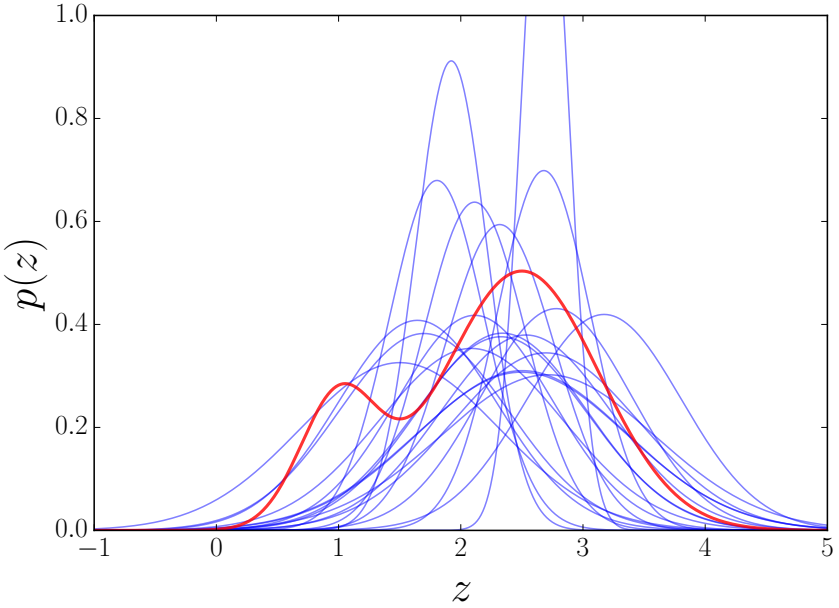
$\beta = 0.10$

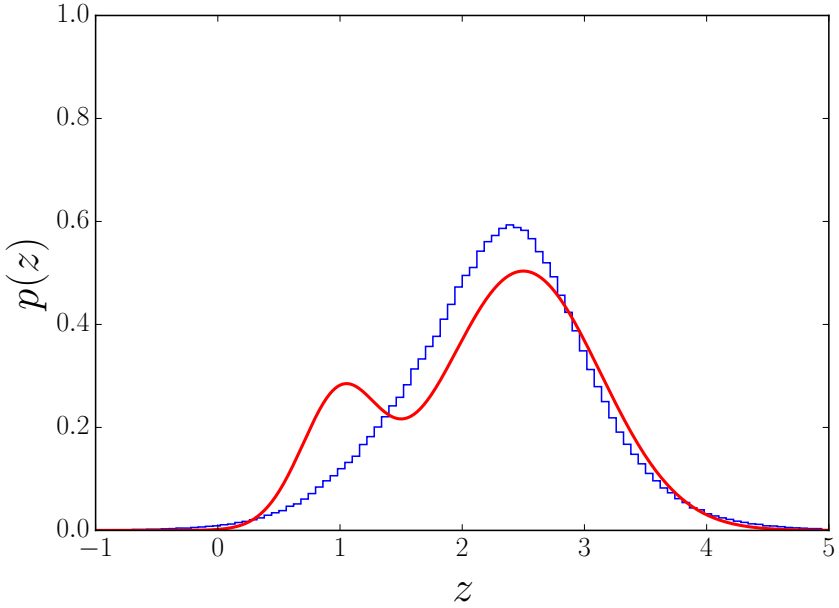

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z

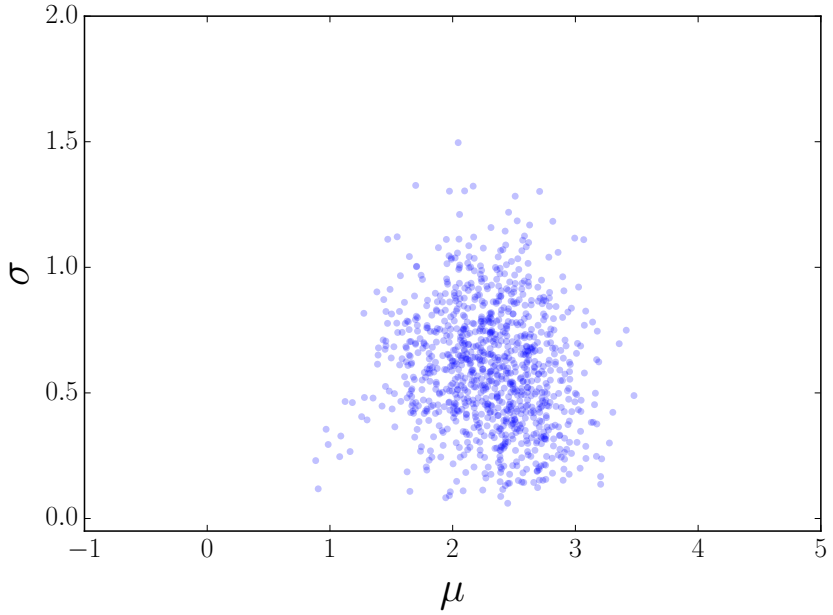

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z


2.0
1.5
1.0
0.5
0.0
σ
−1
0
1
2
3
4
5
μ

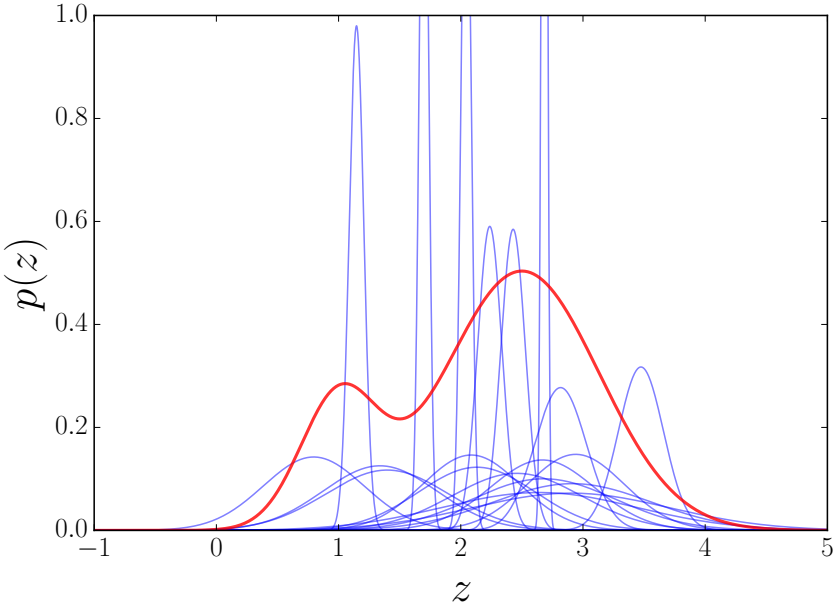
$\beta = 0.25$

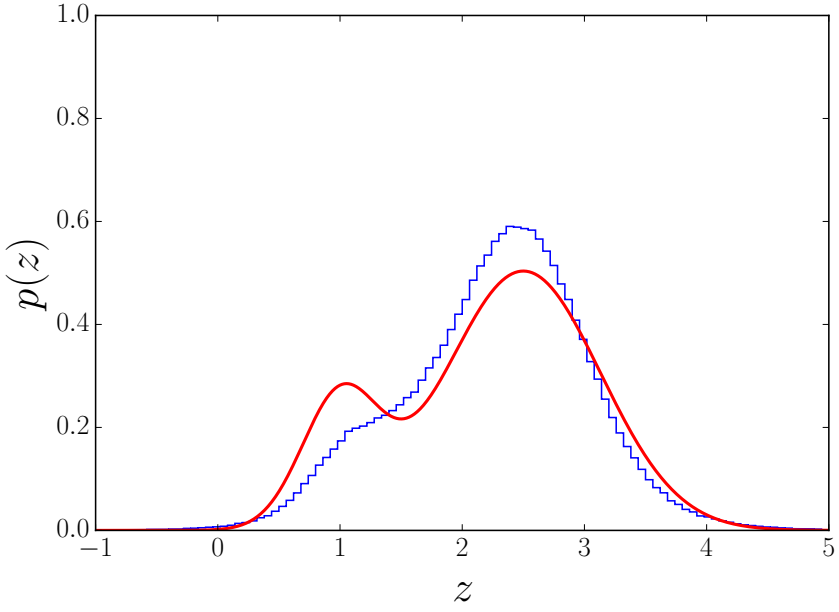

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z

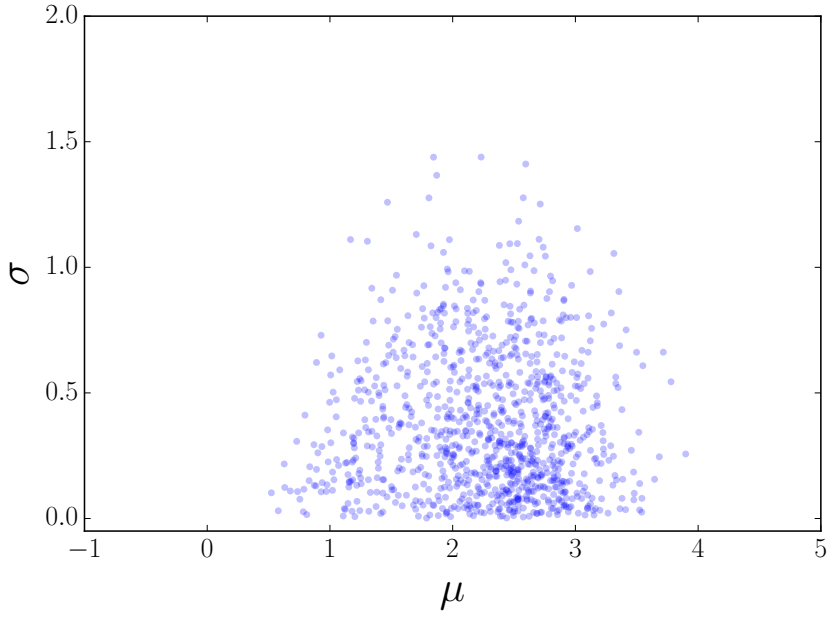

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z


2.0
1.5
1.0
0.5
0.0
σ
−1
0
1
2
3
4
5
μ

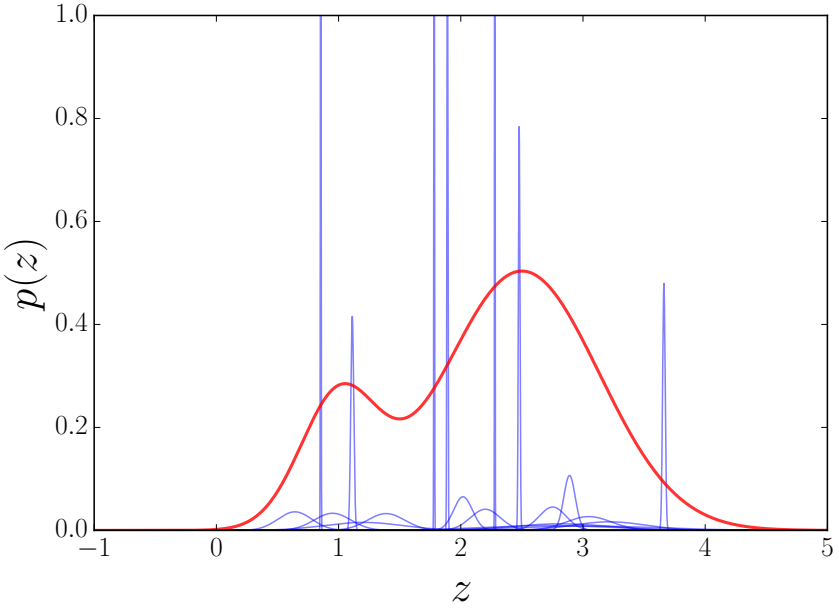
$\beta = 0.50$

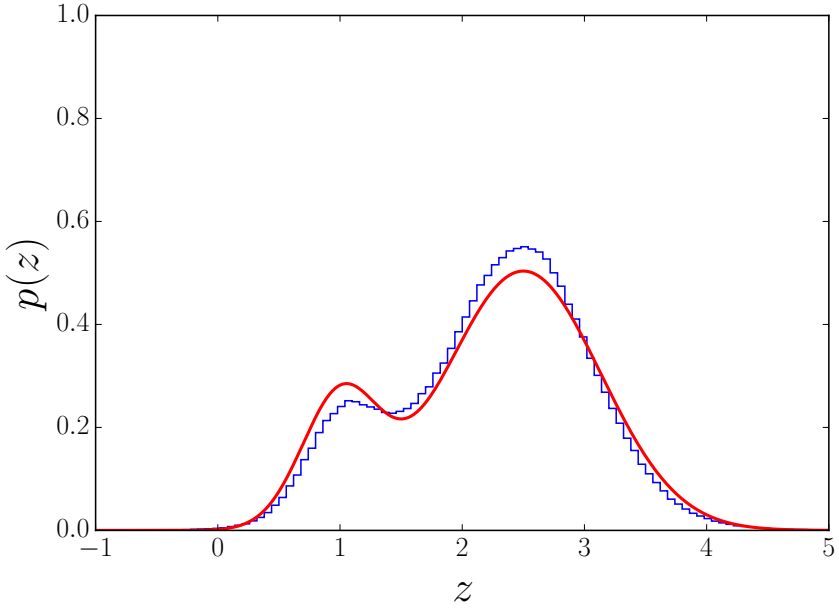

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z

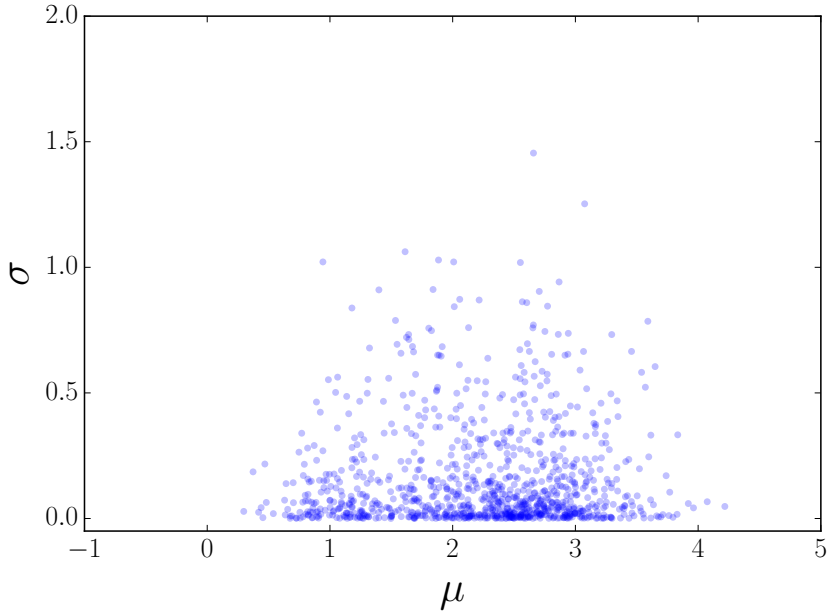

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z


2.0
1.5
1.0
0.5
0.0
σ
−1
0
1
2
3
4
5
μ

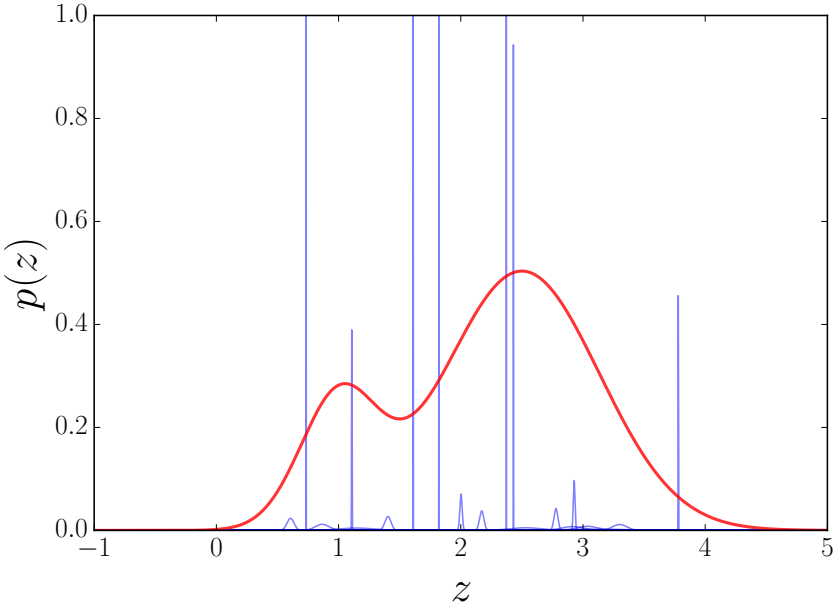
$\beta = 0.75$

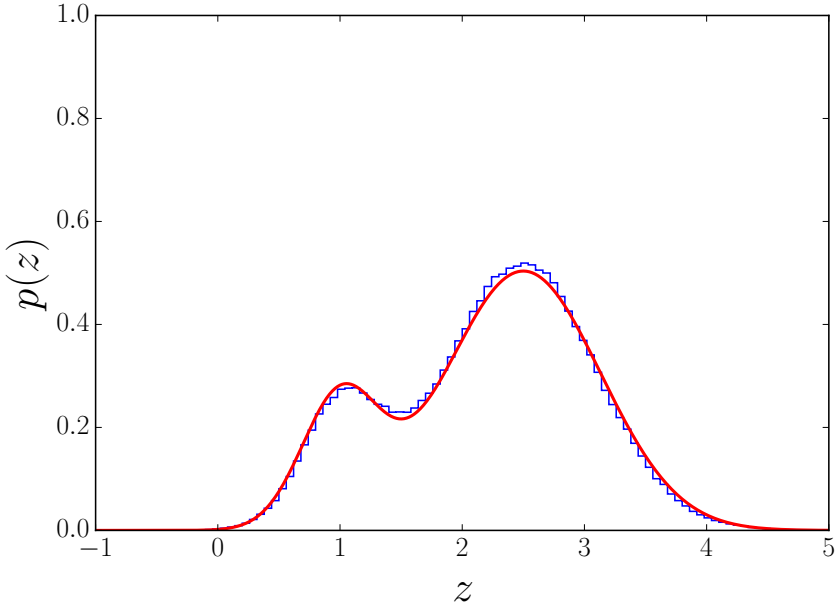

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z

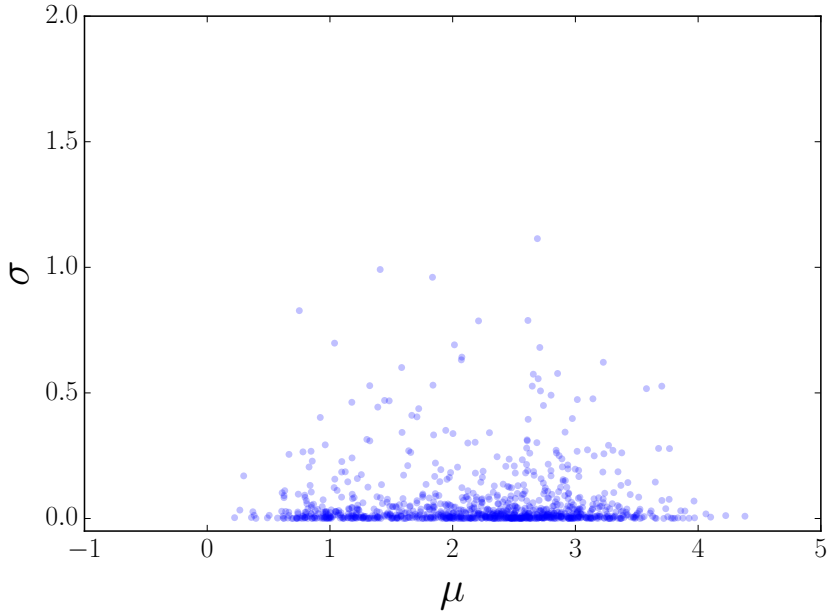

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z


2.0
1.5
1.0
0.5
0.0
σ
−1
0
1
2
3
4
5
μ

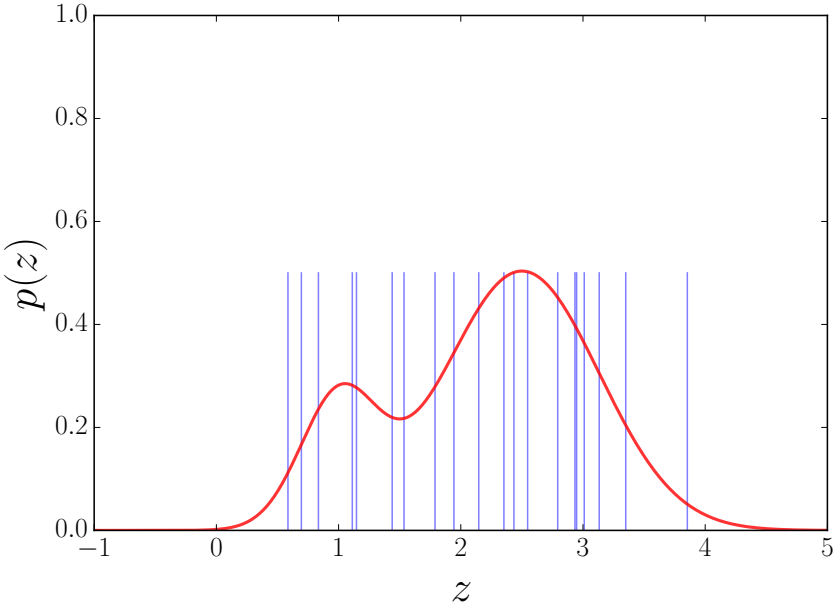
$\beta = 0.90$

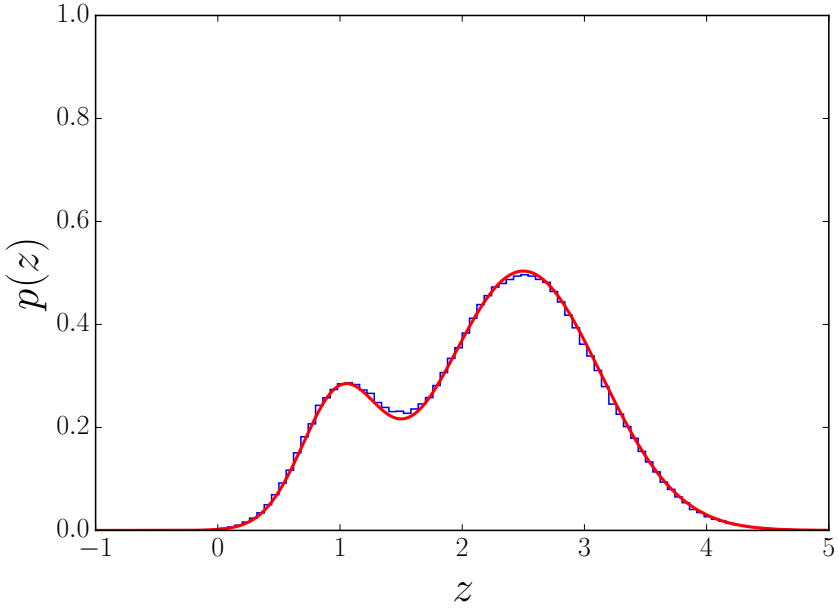

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z

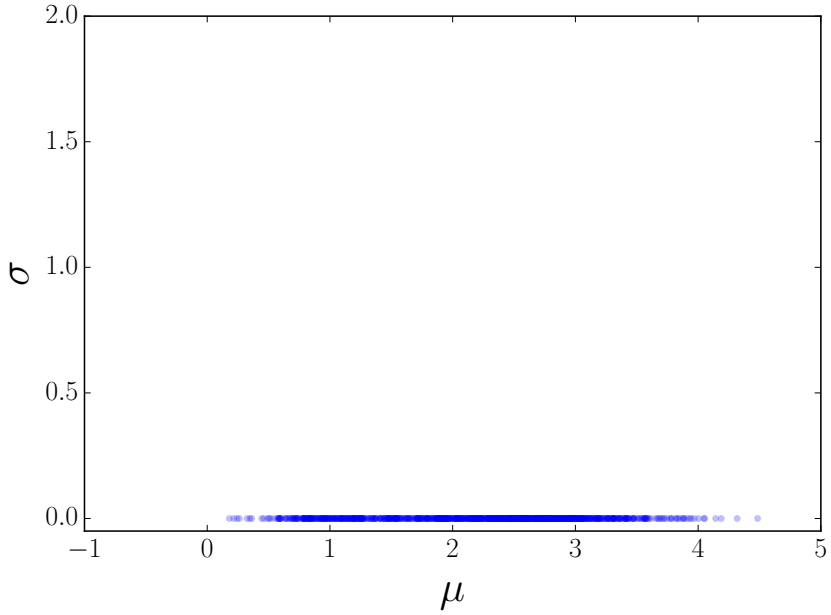

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z


2.0
1.5
1.0
0.5
0.0
σ
−1
0
1
2
3
4
5
μ

$\beta = 1.00$


1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z

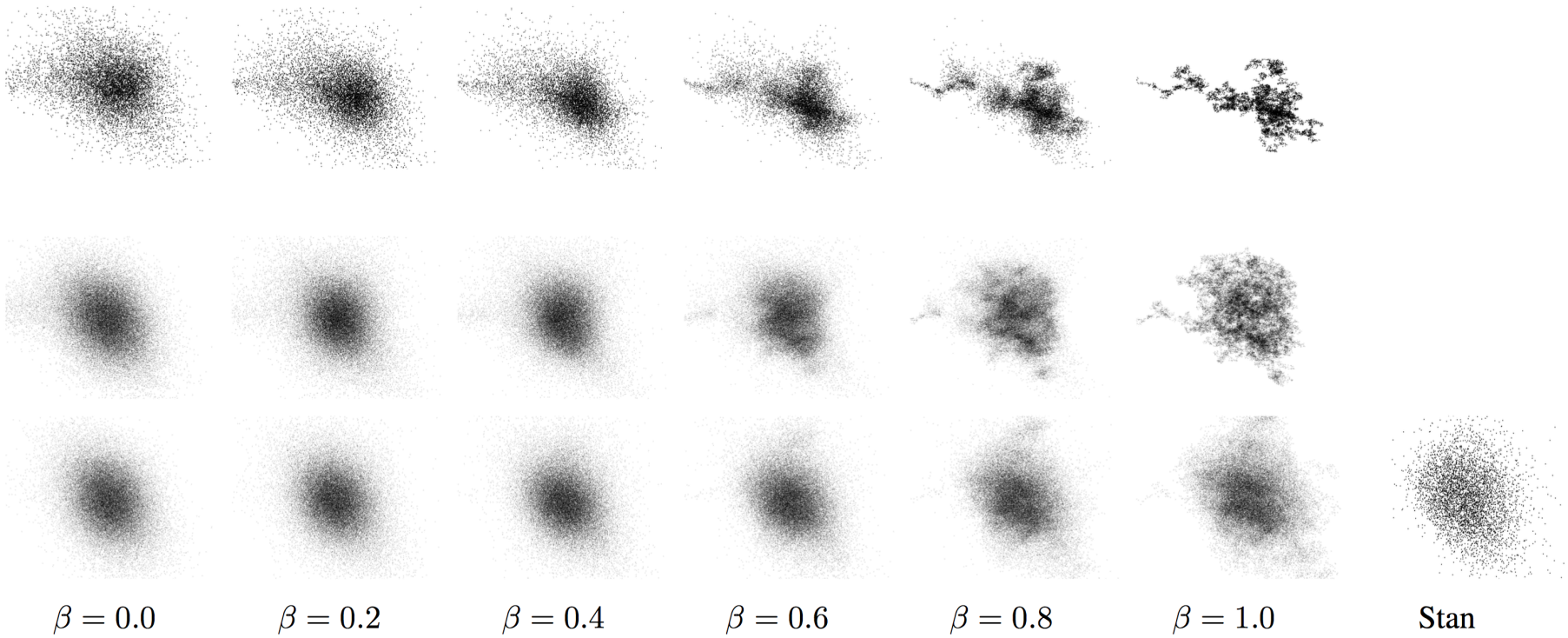

1.0
0.8
0.6
0.4
0.2
0.0
$p(z)$
−1
0
1
2
3
4
5
z


2.0
1.5
1.0
0.5
0.0
σ
−1
0
1
2
3
4
5
μ

Outline

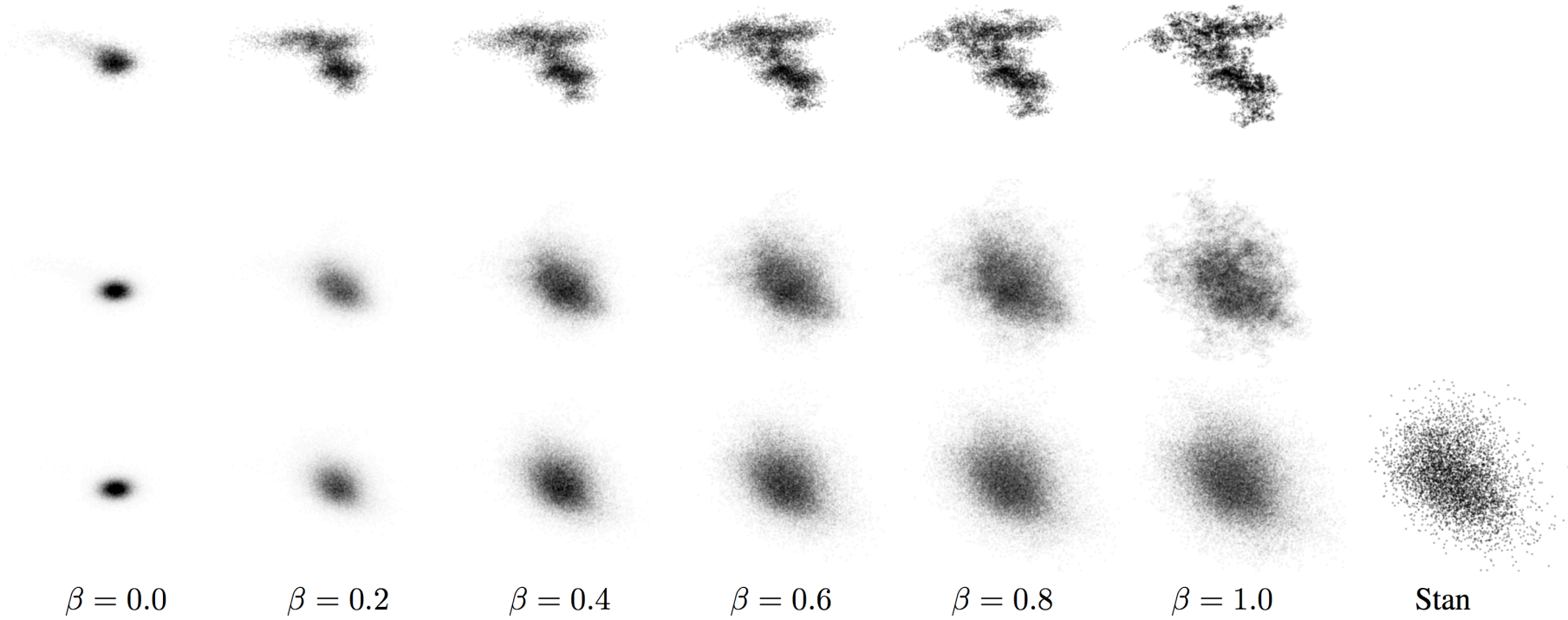
a1a (1st 2 PCA components)

Top 10^4 / middle 10^5 / bottom 10^6 iterations



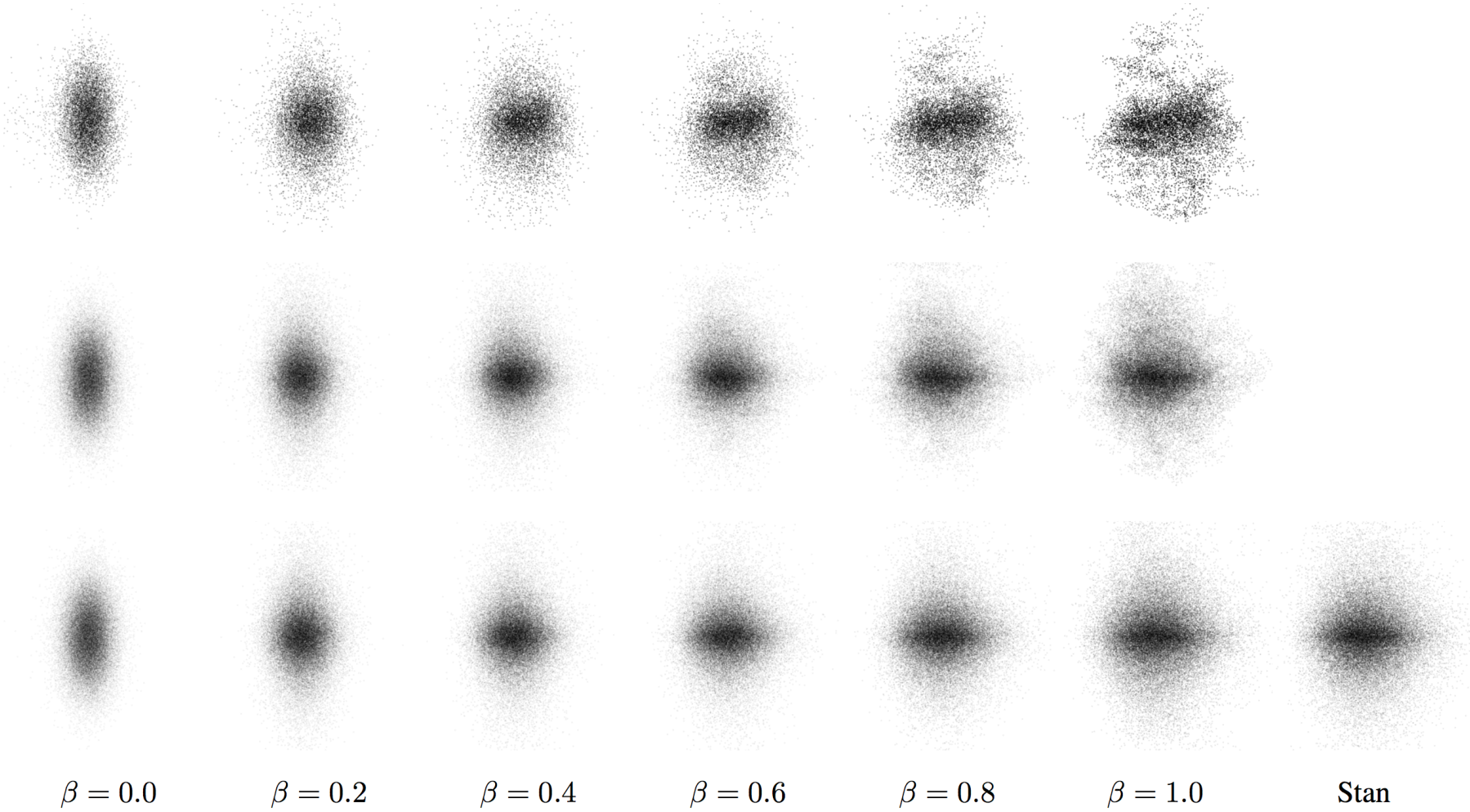
australian (1st 2 PCA components)

Top 10^4 / middle 10^5 / bottom 10^6 iterations



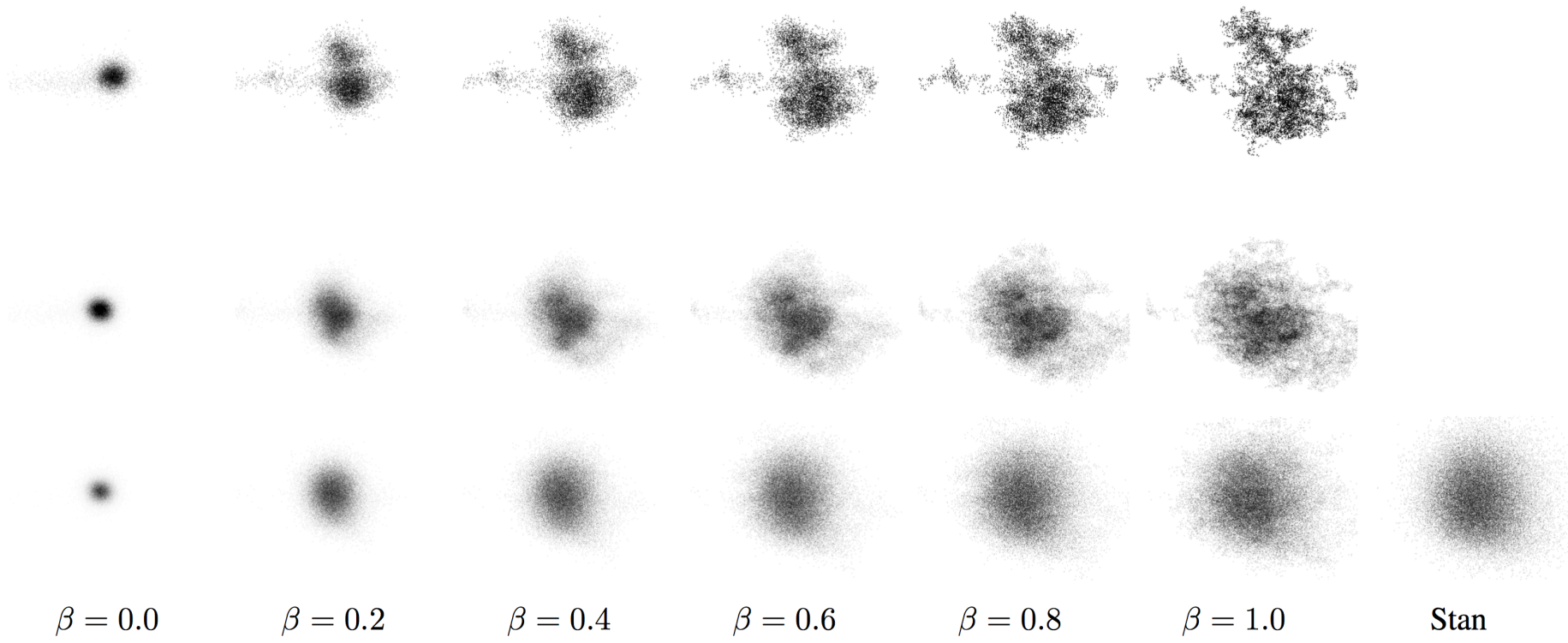
ionosphere (1st 2 PCA components)

Top 10^4 / middle 10^5 / bottom 10^6 iterations

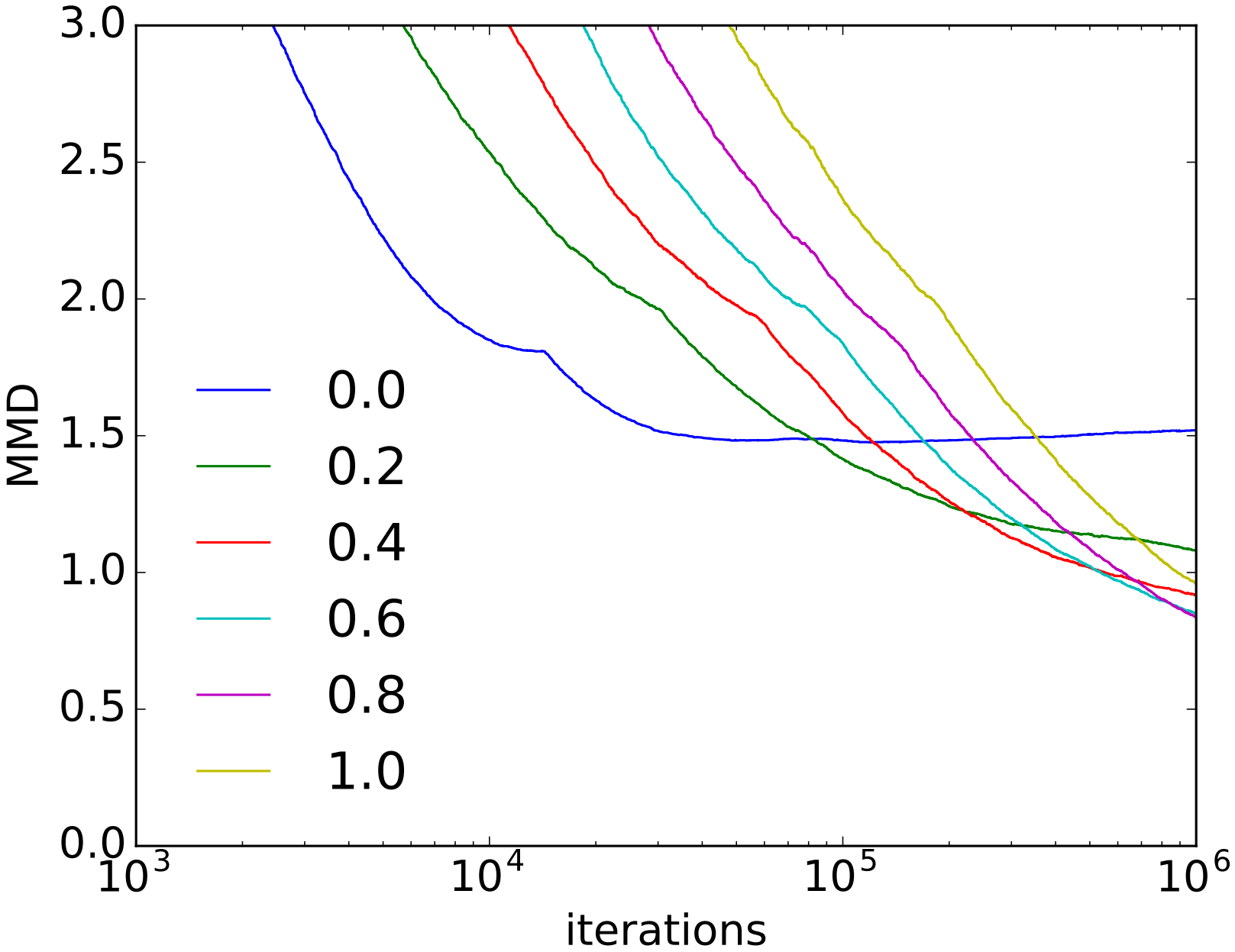


sonar (1st 2 PCA components)

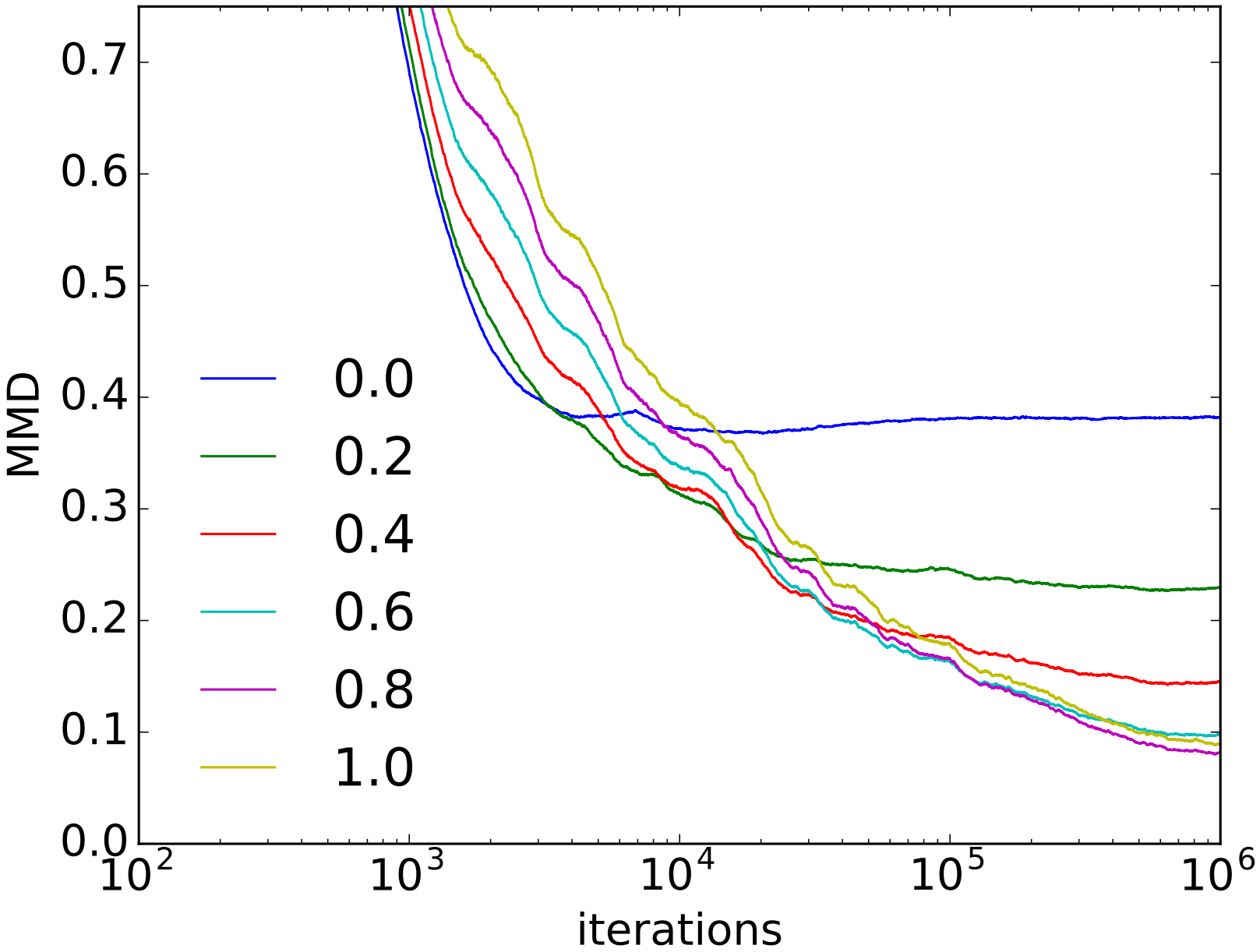
Top 10^4 / middle 10^5 / bottom 10^6 iterations



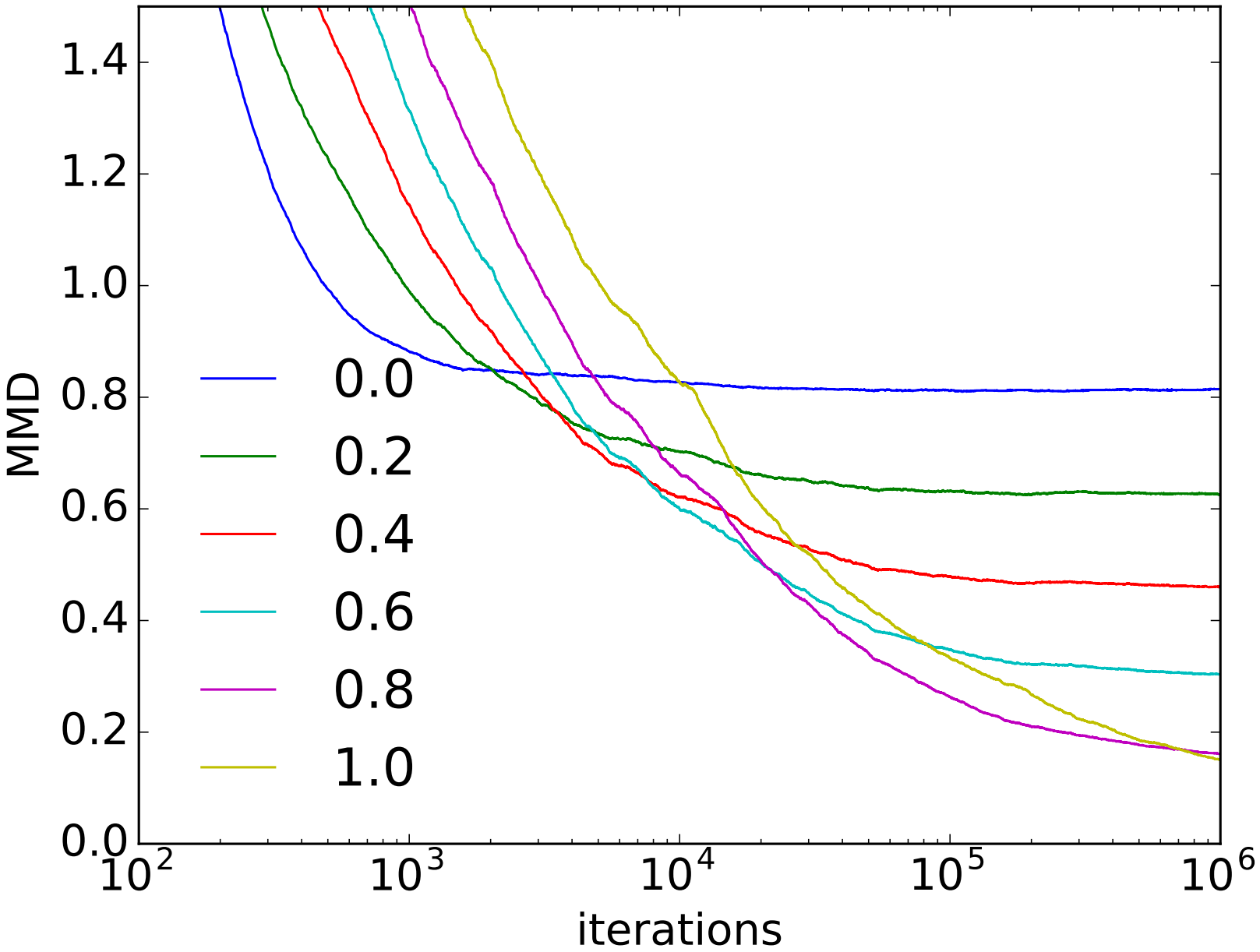
a1a errors vs. time



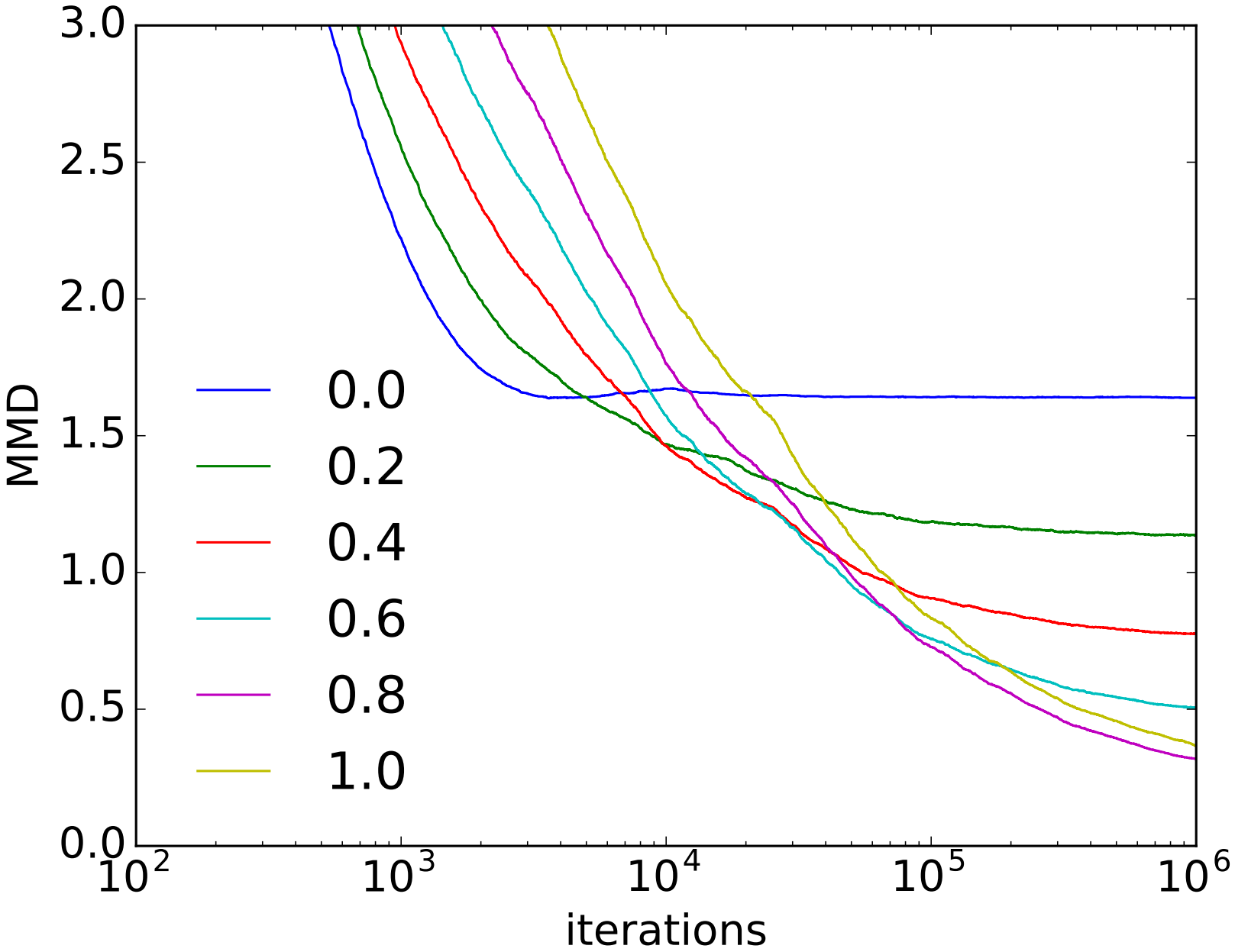
australian errors vs. time



ionosphere errors vs. time



sonar errors vs. time



Outline

Summary:

- Parameterized bound $KL(q(z)\|p(z)) \leq D_\beta$
- Found $q^*(w)$ that minimizes D_β
- Alg. to sample from $q(w)$, simplifies to Langevin and SGVI.
- Experimental evidence beneficial time/accuracy tradeoff.

Questions:

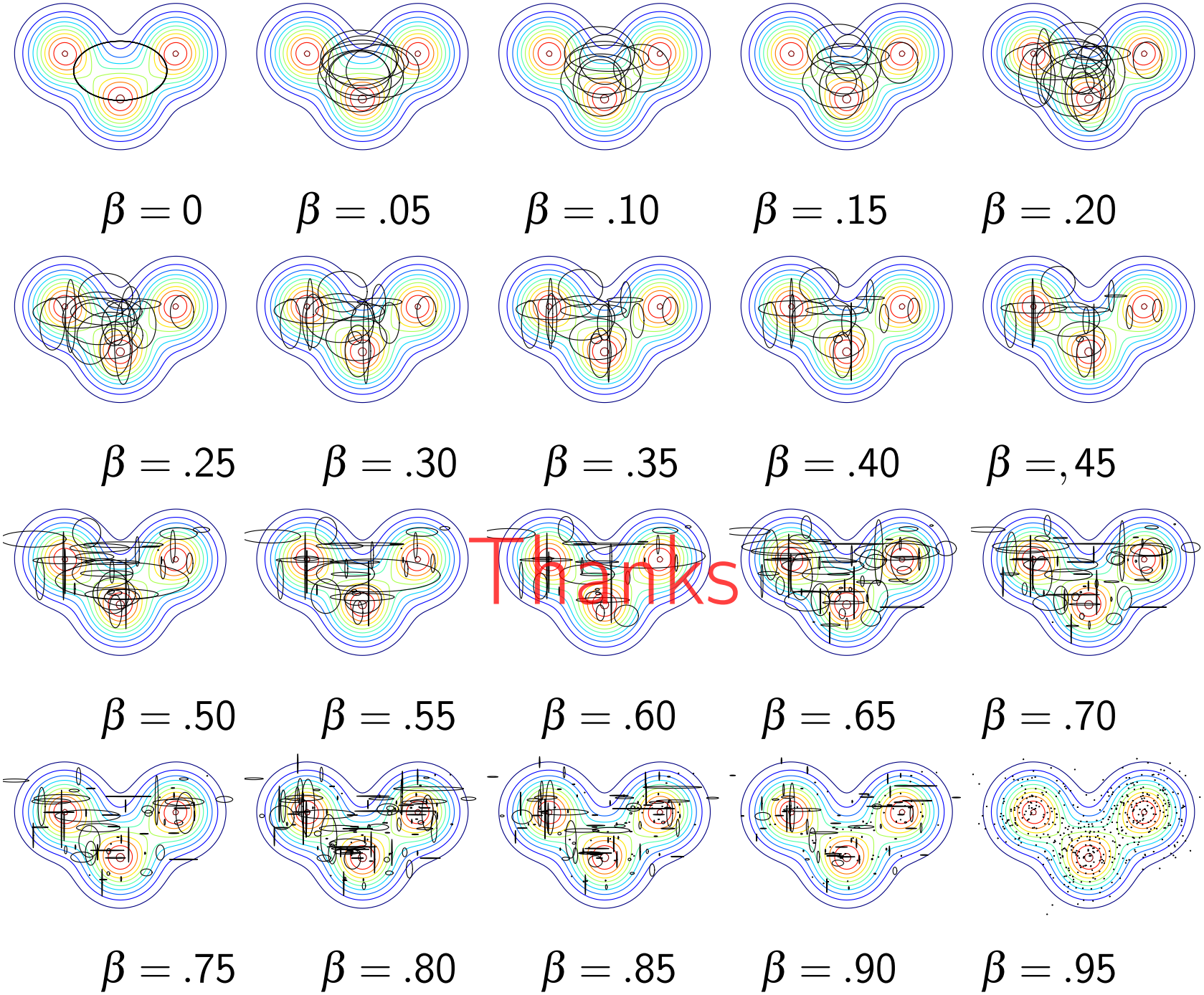
- Theory for time/accuracy tradeoff?
- What is best β for a situation?
- Other algorithmic pairs than Langevin / SGVI?
- Is this the "right" way to do things?

Summary:

- Parameterized bound $KL(q(z)\|p(z)) \leq D_\beta$
- Found $q^*(w)$ that minimizes D_β
- Alg. to sample from $q(w)$, simplifies to Langevin and SGVI.
- Experimental evidence beneficial time/accuracy tradeoff.

Questions:

- Theory for time/accuracy tradeoff?
- What is best β for a situation?
- Other algorithmic pairs than Langevin / SGVI?
- Is this the "right" way to do things?


$\beta = 0$
$\beta = .05$
$\beta = .10$
$\beta = .15$
$\beta = .20$
$\beta = .25$
$\beta = .30$
$\beta = .35$
$\beta = .40$
$\beta =, 45$
$\beta = .50$
$\beta = .55$
$\beta = .60$
$\beta = .65$
$\beta = .70$
$\beta = .75$
$\beta = .80$
$\beta = .85$
$\beta = .90$
$\beta = .95$

Thanks

$p(w|z)$

Choosing $p(w|z)$ very similar to *auxiliary random variables* in VI.
Differences:

- Here, $p(w|z)$ is fixed, numerically optimized at runtime in VI.
- Here, optimal $q^*(w)$ is found (with math), optimal w^* found numerically with VI.