Annealing Between Distributions by Averaging Moments

Chris J. Maddison

Dept. of Comp. Sci.
University of Toronto

Roger Grosse

CSAIL MIT

Ruslan Salakhutdinov

University of Toronto

Partition Functions

We usually specify distributions up to a normalizing constant,

$$p(\mathbf{y}) = f(\mathbf{y})/\mathcal{Z}$$

	MRFs	Posteriors
$\mathbf{y}$	$\mathbf{x}$	$\boldsymbol{\theta}$
f	$\exp(-E(\mathbf{x}, \boldsymbol{\theta}))$	$p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$
$\mathcal{Z}$	$\mathcal{Z}(\boldsymbol{\theta})$	$p(\mathbf{x})$

Partition Functions

We usually specify distributions up to a normalizing constant,

$$p(\mathbf{y}) = f(\mathbf{y})/\mathcal{Z}$$

	MRFs	Posteriors
$\mathbf{y}$	$\mathbf{x}$	$\boldsymbol{\theta}$
f	$\exp(-E(\mathbf{x}, \boldsymbol{\theta}))$	$p(\mathbf{x}\vert\boldsymbol{\theta})p(\boldsymbol{\theta})$
$\mathcal{Z}$	$\mathcal{Z}(\boldsymbol{\theta})$	$p(\mathbf{x})$

For Markov Random Fields (MRFs)

- **partition function** $\mathcal{Z}(\boldsymbol{\theta}) = \sum_{\mathbf{x}} \exp(-E(\mathbf{x}, \boldsymbol{\theta}))$ is intractable

Goal: Estimate $\log \mathcal{Z}(\boldsymbol{\theta})$.

Estimating Partition Functions

- Variational approximations and bounds on $\log \mathcal{Z}$ (Yedida et al., 2005; Wainwright et al., 2005).
  - We want our models to reflect a highly dependent world, this can hurt variational approaches as we assume more and more independence.
  - This assumption less costly for posterior inference over parameters.
- Sampling methods such as **path sampling** (Gelman and Meng, 1998), **sequential Monte Carlo** (e.g. del Moral et al., 2006), **simple importance sampling**, and **annealed importance sampling** (Neal, 2002).
  - Slow, finicky, and hard to diagnose
  - In principle, can deal with multimodality

Simple Importance Sampling (SIS)

- Two distributions $p_a(\mathbf{x})$ and $p_b(\mathbf{x})$ over $\mathcal{X}$

$f_a(\mathbf{x})/\mathcal{Z}_a$	$f_b(\mathbf{x})/\mathcal{Z}_b$
tractable $\mathcal{Z}$	intractable $\mathcal{Z}$
easy to sample	hard to sample

- Then

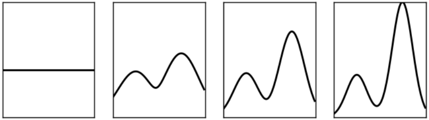
$$\frac{\mathcal{Z}_a}{M}\sum_{i=1}^{M}\frac{f_b(\mathbf{x}^{(i)})}{f_a(\mathbf{x}^{(i)})} \to \int \frac{f_b(\mathbf{x})}{p_a(\mathbf{x})} p_a(\mathbf{x})\, d\mathbf{x} = \mathcal{Z}_b$$

  for $\mathbf{x}^{(i)} \sim p_a(\mathbf{x})$.

- Variance is high (sometimes ∞) if $p_a << p_b$ in some regions.

An Intuition

- Move gradually from a hotter p_a to a colder p_b — annealing



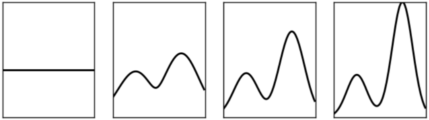
- Reduce variance by chaining importance samplers

$$\frac{\mathcal{Z}_0}{M}\sum_{i=1}^{M}\frac{f_1(\mathbf{x}_0^{(i)})}{f_0(\mathbf{x}_0^{(i)})}\cdots\frac{f_K(\mathbf{x}_{K-1}^{(i)})}{f_{K-1}(\mathbf{x}_{K-1}^{(i)})}\to\mathcal{Z}_0\frac{\mathcal{Z}_1}{\mathcal{Z}_0}\cdots\frac{\mathcal{Z}_K}{\mathcal{Z}_{K-1}}=\mathcal{Z}_K$$

where we independently draw $\mathbf{x}_k \sim p_k(\mathbf{x})$

An Intuition

- Move gradually from a hotter p_a to a colder p_b — annealing



- Reduce variance by chaining importance samplers

$$\frac{\mathcal{Z}_0}{M}\sum_{i=1}^{M}\frac{f_1(\mathbf{x}_0^{(i)})}{f_0(\mathbf{x}_0^{(i)})}\cdots\frac{f_K(\mathbf{x}_{K-1}^{(i)})}{f_{K-1}(\mathbf{x}_{K-1}^{(i)})}\to\mathcal{Z}_0\frac{\mathcal{Z}_1}{\mathcal{Z}_0}\cdots\frac{\mathcal{Z}_K}{\mathcal{Z}_{K-1}}=\mathcal{Z}_K$$

where we independently draw $\mathbf{x}_k \sim p_k(\mathbf{x})$ $\leftarrow$ hard to do

Annealed Importance Sampling (AIS)

We can still do this with certain *dependent* samplers! (Neal, 2002)

Annealed Importance Sampling (AIS)

We can still do this with certain *dependent* samplers! (Neal, 2002)

$$\mathbf{x}_k \quad \sim \quad \underbrace{T_k\left(\mathbf{x} \,\middle|\, \mathbf{x}_{k-1}\right)}_{\text{leaves } p_k(\mathbf{x}) \text{ invariant}}$$

Annealed Importance Sampling (AIS)

We can still do this with certain *dependent* samplers! (Neal, 2002)

$$\mathbf{x}_k \sim \underbrace{T_k\left(\mathbf{x} \mid \mathbf{x}_{k-1}\right)}_{\text{leaves } p_k(\mathbf{x}) \text{ invariant}}$$

For chain i

$\mathbf{x}_0 \sim p_0$

$\mathbf{x}_0$

$$w^{(i)} = \frac{f_1(\mathbf{x}_0^{(i)})}{f_0(\mathbf{x}_0^{(i)})}$$

Annealed Importance Sampling (AIS)

We can still do this with certain *dependent* samplers! (Neal, 2002)

$$\mathbf{x}_k \sim \underbrace{T_k\left(\mathbf{x} \,|\, \mathbf{x}_{k-1}\right)}_{\text{leaves } p_k(\mathbf{x}) \text{ invariant}}$$

For chain i

$\mathbf{x}_0 \sim p_0$

$$\begin{tikzcd} \mathbf{x}_0 \arrow[r] & \mathbf{x}_1 \end{tikzcd}$$

$\mathbf{x}_1 \sim T_1\left(\mathbf{x} \,|\, \mathbf{x}_0\right)$

$$w^{(i)} = \frac{f_1(\mathbf{x}_0^{(i)})}{f_0(\mathbf{x}_0^{(i)})} \frac{f_2(\mathbf{x}_1^{(i)})}{f_1(\mathbf{x}_1^{(i)})}$$

Annealed Importance Sampling (AIS)

We can still do this with certain *dependent* samplers! (Neal, 2002)

$$\mathbf{x}_k \quad \sim \quad \underbrace{T_k\left(\mathbf{x} \,|\, \mathbf{x}_{k-1}\right)}_{\text{leaves } p_k(\mathbf{x}) \text{ invariant}}$$

For chain i

$$\begin{tikzcd}
\mathbf{x}_0 \sim p_0 & & \\
\mathbf{x}_0 \arrow[r, "\mathbf{x}_1 \sim T_1\left(\mathbf{x} \,|\, \mathbf{x}_0\right)"'] & \mathbf{x}_1 \arrow[r, "\mathbf{x}_2 \sim T_2\left(\mathbf{x} \,|\, \mathbf{x}_1\right)"] & \mathbf{x}_2
\end{tikzcd}$$

$$w^{(i)} = \frac{f_1(\mathbf{x}_0^{(i)})}{f_0(\mathbf{x}_0^{(i)})} \frac{f_2(\mathbf{x}_1^{(i)})}{f_1(\mathbf{x}_1^{(i)})} \frac{f_3(\mathbf{x}_2^{(i)})}{f_2(\mathbf{x}_2^{(i)})}$$

Annealed Importance Sampling (AIS)

We can still do this with certain *dependent* samplers! (Neal, 2002)

$$\mathbf{x}_k \sim \underbrace{T_k\left(\mathbf{x} \mid \mathbf{x}_{k-1}\right)}_{\text{leaves } p_k(\mathbf{x}) \text{ invariant}}$$

For chain i

$\mathbf{x}_0 \sim p_0$

$$\begin{tikzcd} \mathbf{x}_0 \arrow[r, "\mathbf{x}_1 \sim T_1(\mathbf{x} \mid \mathbf{x}_0)"'] & \mathbf{x}_1 \arrow[r, "\mathbf{x}_2 \sim T_2(\mathbf{x} \mid \mathbf{x}_1)"] & \mathbf{x}_2 \arrow[r] & \cdots \arrow[r, "\mathbf{x}_{K-1} \sim T_{K-1}(\mathbf{x} \mid \mathbf{x}_{K-2})"'] & \mathbf{x}_{K-1} \end{tikzcd}$$

$$w^{(i)} = \frac{f_1(\mathbf{x}_0^{(i)})}{f_0(\mathbf{x}_0^{(i)})} \frac{f_2(\mathbf{x}_1^{(i)})}{f_1(\mathbf{x}_1^{(i)})} \frac{f_3(\mathbf{x}_2^{(i)})}{f_2(\mathbf{x}_2^{(i)})} \cdots \frac{f_K(\mathbf{x}_{K-1}^{(i)})}{f_{K-1}(\mathbf{x}_{K-1}^{(i)})}$$

Annealed Importance Sampling (AIS)

We can still do this with certain *dependent* samplers! (Neal, 2002)

$$\mathbf{x}_k \sim \underbrace{T_k\left(\mathbf{x} \,|\, \mathbf{x}_{k-1}\right)}_{\text{leaves } p_k(\mathbf{x}) \text{ invariant}}$$

For chain i

$\mathbf{x}_0 \sim p_0$ $\qquad$ $\mathbf{x}_2 \sim T_2\left(\mathbf{x} \,|\, \mathbf{x}_1\right)$

$$\mathbf{x}_0 \longrightarrow \mathbf{x}_1 \longrightarrow \mathbf{x}_2 \longrightarrow \cdots \longrightarrow \mathbf{x}_{K-1}$$

$\mathbf{x}_1 \sim T_1\left(\mathbf{x} \,|\, \mathbf{x}_0\right)$ $\qquad$ $\mathbf{x}_{K-1} \sim T_{K-1}\left(\mathbf{x} \,|\, \mathbf{x}_{K-2}\right)$

$$\frac{\mathcal{Z}_0}{M} \sum_{i=1}^{M} w^{(i)} \rightarrow \mathcal{Z}_K$$

Intuition: SIS on an extended state space, remarkably *unbiased!*

Analyzing AIS Paths

- Virtually the only scheme used is **geometric averages**,

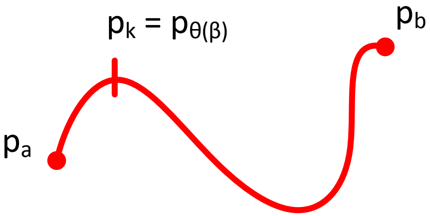
$$f_k(\mathbf{x}) = {f_a}^{1-\beta_k}(\mathbf{x}){f_b}^{\beta_k}(\mathbf{x})$$

Analyzing AIS Paths

- Virtually the only scheme used is **geometric averages**,

$$f_k(\mathbf{x}) = {f_a}^{1-\beta_k}(\mathbf{x}){f_b}^{\beta_k}(\mathbf{x})$$

- Let $\mathcal{P}$ be a family of distributions parameterized by $\boldsymbol{\theta}$
- Define a **path** $\gamma : [0,1] \to \mathcal{P}$ and a **schedule** of points $0 = \beta_0 < \beta_1 < \ldots < \beta_K = 1$



Analyzing AIS Paths

Assume **perfect transitions**

$$T_k\left(\mathbf{x} \mid \mathbf{x}_{k-1}\right) = p_k(\mathbf{x})$$

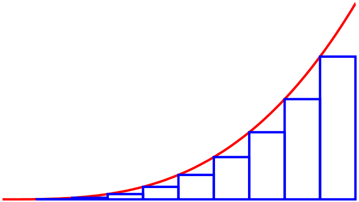
then $\mathbb{E}\left[\log w^{(i)}\right] = \sum_{k=1}^{K} \mathbb{E}_{p_k}\left[\log f_k(\mathbf{x}) - \log f_{k-1}(\mathbf{x})\right]$

Analyzing AIS Paths

Assume **perfect transitions**

$$T_k\left(\mathbf{x} \,|\, \mathbf{x}_{k-1}\right) = p_k(\mathbf{x})$$

then $\mathbb{E}\left[\log w^{(i)}\right] = \sum_{k=1}^{K} \underbrace{\mathbb{E}_{p_k}\left[\log f_k(\mathbf{x}) - \log f_{k-1}(\mathbf{x})\right]}_{\text{finite difference approximation}}$



$$\text{assump. + math} \xrightarrow{k\to\infty} \int_0^1 \frac{d \log \mathcal{Z}(\boldsymbol{\theta}(\beta))}{d\beta}\, d\beta = \log \frac{\mathcal{Z}_K}{\mathcal{Z}_0}$$

Intuition: $\log w^{(i)}$s accumulate finite differences of $\log \mathcal{Z}$

Analyzing AIS Paths

What is the error for a fixed number of intermediate distributions?

Analyzing AIS Paths

What is the error for a fixed number of intermediate distributions?

- $\mathbb{E}\left[\log w^{(i)}\right]$ is biased, unlike $\mathbb{E}\left[w^{(i)}\right]$

$$
\begin{aligned}
\mathbb{E}\left[\log w^{(i)}\right] &= \sum_{k=1}^{K} \mathbb{E}_{p_k}\left[\log f_k(\mathbf{x}) - \log f_{k-1}(\mathbf{x})\right] \\
&= \log \frac{\mathcal{Z}_K}{\mathcal{Z}_0} - \underbrace{\sum_{k=1}^{K} \mathrm{D}_{\mathrm{KL}}(p_{k-1} \| p_k)}_{\text{bias} \,=\, \delta(\log w^{(i)})}
\end{aligned}
$$

Analyzing AIS Paths

What is the error for a fixed number of intermediate distributions?

- $\mathbb{E}\left[\log w^{(i)}\right]$ is biased, unlike $\mathbb{E}\left[w^{(i)}\right]$

$$\mathbb{E}\left[\log w^{(i)}\right] = \sum_{k=1}^{K} \mathbb{E}_{p_k}\left[\log f_k(\mathbf{x}) - \log f_{k-1}(\mathbf{x})\right]$$

$$= \log \frac{\mathcal{Z}_K}{\mathcal{Z}_0} - \underbrace{\sum_{k=1}^{K} \mathrm{D}_{\mathrm{KL}}(p_{k-1} \| p_k)}_{\text{bias} = \delta(\log w^{(i)})}$$

- With perfect transitions the following are monotonic in the sum of KL divergences

$$\delta(\log w^{(i)}) \quad \mathrm{var}\left[\log w^{(i)}\right] \quad \mathrm{var}\left[w^{(i)}\right]$$

Goal: Minimize the sum of KL divergences.

Analyzing AIS Paths

Approach: Approximate the sum of KL with a functional.

Let γ be a path and β_k a linearly spaced schedule. Then

$$K \sum_{k=1}^{K} \mathrm{D_{KL}}(p_{k-1} \| p_k) \xrightarrow{K \to \infty} \mathcal{F}(\gamma) \equiv \frac{1}{2} \int_0^1 \underbrace{\dot{\boldsymbol{\theta}}(\beta)^T \mathbf{G}_{\boldsymbol{\theta}}(\beta) \dot{\boldsymbol{\theta}}(\beta)}_{\text{metric on manifold defined by Fisher Inform.}} d\beta,$$

where $\mathbf{G}_{\boldsymbol{\theta}} = \mathrm{cov}_{p_{\boldsymbol{\theta}}}(\nabla_{\boldsymbol{\theta}} \log p_\theta(\mathbf{x}))$ is the Fisher Information and $\dot{\boldsymbol{\theta}}(\beta) = d\boldsymbol{\theta}(\beta)/d\beta$.

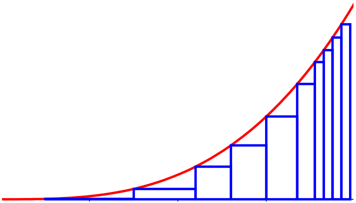
- Ties to information geometry.
- Analogous functional for path sampling (Gelman and Meng, 1998).

Analyzing AIS Paths

Under the optimal schedule, the value of the functional is $\mathcal{F}(\gamma) = \ell(\gamma)^2/2$ where

$$\ell(\gamma) = \int_0^1 \sqrt{\dot{\boldsymbol{\theta}}(\beta)^T \mathbf{G}_{\boldsymbol{\theta}}(\beta) \dot{\boldsymbol{\theta}}(\beta)} d\beta$$

is the path length on the Riemannian manifold defined by $\mathbf{G}_{\boldsymbol{\theta}}(\beta)$.



Intuition: Spend more time on segments with high curvature.

Paths for Exponential Family Distributions

- Let us restrict ourselves to the exponential family

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{g}(\mathbf{x}))$$

- $\boldsymbol{\eta}$ or $\mathbf{s} = \mathbb{E}[\mathbf{g}(\mathbf{x})]$ completely specifies a distribution.
    - One-to-one correspondence between $\boldsymbol{\eta}$ and $\mathbf{s}$

Paths for Exponential Family Distributions

- Let us restrict ourselves to the exponential family

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{g}(\mathbf{x}))$$

- $\boldsymbol{\eta}$ or $\mathbf{s} = \mathbb{E}[\mathbf{g}(\mathbf{x})]$ completely specifies a distribution.
    - One-to-one correspondence between $\boldsymbol{\eta}$ and $\mathbf{s}$

Old Geom. Averaged Path

$\gamma_{GA}(\beta)$ is the distribution with

$$\boldsymbol{\eta}(\beta) =$$

$$(1-\beta)\boldsymbol{\eta}(0) + \beta\boldsymbol{\eta}(1)$$

Paths for Exponential Family Distributions

- Let us restrict ourselves to the exponential family

$$p(\mathbf{x}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\mathbf{x}) \exp(\boldsymbol{\eta}^T \mathbf{g}(\mathbf{x}))$$

- $\boldsymbol{\eta}$ or $\mathbf{s} = \mathbb{E}[\mathbf{g}(\mathbf{x})]$ completely specifies a distribution.
  - One-to-one correspondence between $\boldsymbol{\eta}$ and $\mathbf{s}$

Old Geom. Averaged Path

$\gamma_{GA}(\beta)$ is the distribution with $\boldsymbol{\eta}(\beta) =$

$$(1-\beta)\boldsymbol{\eta}(0) + \beta\boldsymbol{\eta}(1)$$

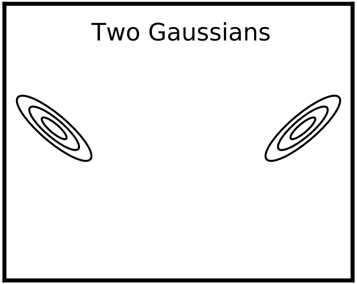
New Moment Averaged Path

$\gamma_{MA}(\beta)$ is the distribution with $\mathbf{s}(\beta) =$

$$(1-\beta)\mathbf{s}(0) + \beta\mathbf{s}(1)$$

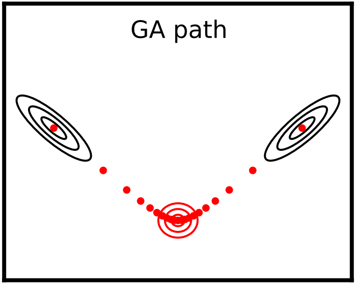
The Picture for Gaussians

Often unintuitive ...



The Picture for Gaussians

Often unintuitive ...

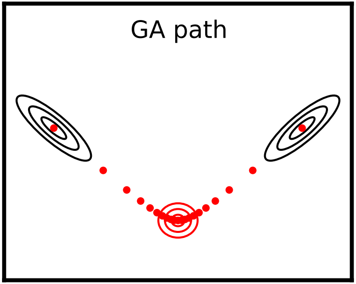


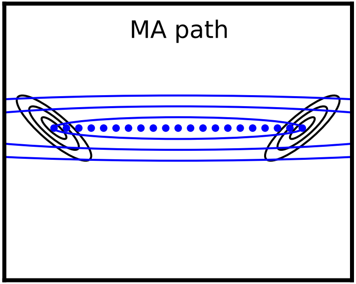
γ_{GA} places mass only where both pdfs agree

"veto" effects

The Picture for Gaussians

Often unintuitive ...





γ_{GA} places mass only where both pdfs agree "veto" effects

γ_{MA} interpolate means and covariances then stretch covariance

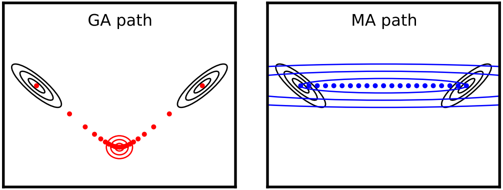
Path Properties: Variational Interpretation

- For geometric averages, the intermediate distribution minimizes a weighted sum of KLs

$$p_\beta^{(GA)} = \arg\min_p \, (1-\beta)\mathrm{D}_{\mathrm{KL}}(p \| p_a) + \beta \mathrm{D}_{\mathrm{KL}}(p \| p_b)$$

- For moment averages, the same but of the reverse KLs

$$p_\beta^{(MA)} = \arg\min_p \, (1-\beta)\mathrm{D}_{\mathrm{KL}}(p_a \| p) + \beta \mathrm{D}_{\mathrm{KL}}(p_b \| p)$$



Path Properties: Cost Functional

- For the exponential family we can find $\mathcal{F}(\gamma)$
  - **Important** this assumes **linear** schedules
- Both γ_{GA} and γ_{MA} have the same functional!

$$\mathcal{F}(\gamma_{GA}) = \mathcal{F}(\gamma_{MA}) = \frac{1}{2}(\mathbf{s}(1) - \mathbf{s}(0))^T(\boldsymbol{\eta}(1) - \boldsymbol{\eta}(0))$$

- If we partition a schedule by distributions p_j into piecewise linear schedules we can optimally allocate distributions from a total budget $= \sum K_j$

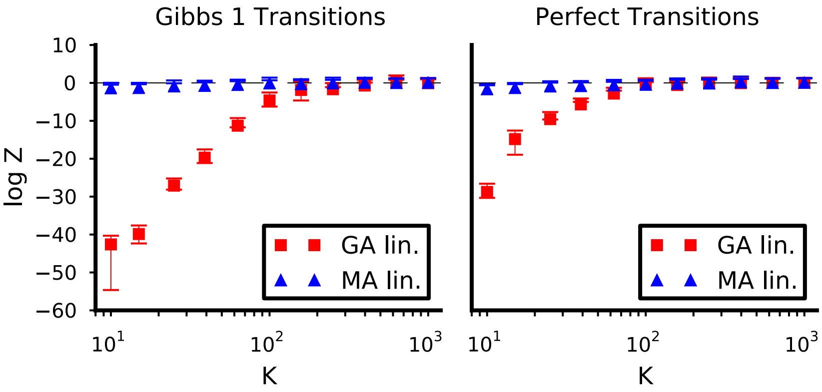
$$K_j \propto \sqrt{(\boldsymbol{\eta}_{j+1} - \boldsymbol{\eta}_j)^T(\mathbf{s}_{j+1} - \mathbf{s}_j)}$$

- Biggest effect is the choice of path, not schedule.

Gaussians

Two Gaussians

$$\mathcal{N}\left(\begin{pmatrix}-10\\0\end{pmatrix}, \begin{pmatrix}1 & -0.85\\-0.85 & 1\end{pmatrix}\right) \text{ and } \mathcal{N}\left(\begin{pmatrix}10\\0\end{pmatrix}, \begin{pmatrix}1 & 0.85\\0.85 & 1\end{pmatrix}\right)$$



Gaussians

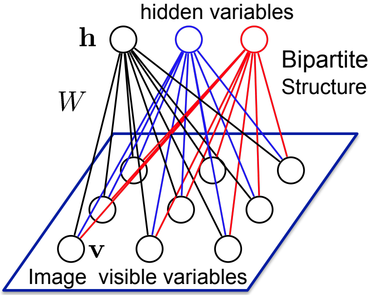
Number of intermediate distributions required to anneal between two Gaussians with means μ_1 and μ_2 and variance σ as a function of $d = (\mu_2 - \mu 1)/\sigma$

GA, linear schedule	$\mathcal{O}(d^2)$
MA, linear schedule	$\mathcal{O}(d^2)$
GA, optimal schedule	$\mathcal{O}(d^2)$
MA, optimal schedule	$\mathcal{O}((\log d)^2)$
Optimal path (Gelman and Meng, 1998)	$\mathcal{O}((\log d)^2)$

- MA within constant factor of optimal *under its optimal scheduling* — no general proof yet.
- Mixing issues dominate performance in practice — where MA shines.

Restricted Boltzmann Machines (RBMs)

- RBMs are Markov Random Fields of coupled **visible and hidden** binary variables $\mathbf{x} = (\mathbf{v}, \mathbf{h}) \in \{0,1\}^D \times \{0,1\}^F$ with a special **bipartite structure**.



The energy of a joint configuration is $E(\mathbf{v}, \mathbf{h}, \boldsymbol{\theta})$

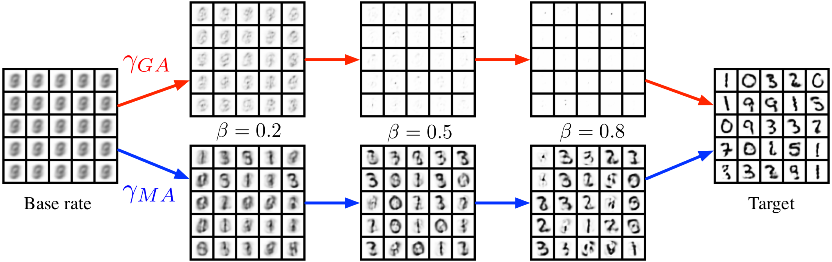
$$-\overbrace{\mathbf{v}^T W \mathbf{h}}^{\text{binary potentials}} - \underbrace{\mathbf{v}^T c}_{\text{unary potential}} - \overbrace{\mathbf{h}^T b}^{\text{unary potential}}$$

where $\boldsymbol{\theta} = (W, c, b)$.

- $\mathcal{Z}(\boldsymbol{\theta}) = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}, \boldsymbol{\theta}))$ is generally intractable, except if we have a small number (< 25) of hidden units *or* few visible units.

Restricted Boltzmann Machines (RBMs)

Two different paths for an RBM trained on the MNIST digit dataset (60,000 B&W 28×28 images of digits).



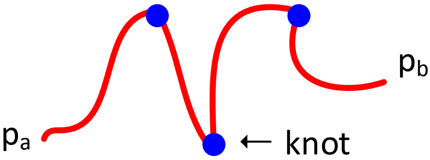
Restricted Boltzmann Machines (RBMs)

- MA path infeasible for most RBMs

$$\overbrace{\mathbb{E}[\mathbf{vh}^T]_\beta}^{\text{solve for natural parameters}} = (1-\beta)\mathbb{E}[\mathbf{vh}^T]_0 + \beta\,\overbrace{\mathbb{E}[\mathbf{vh}^T]_1}^{\text{estimate moments}}$$

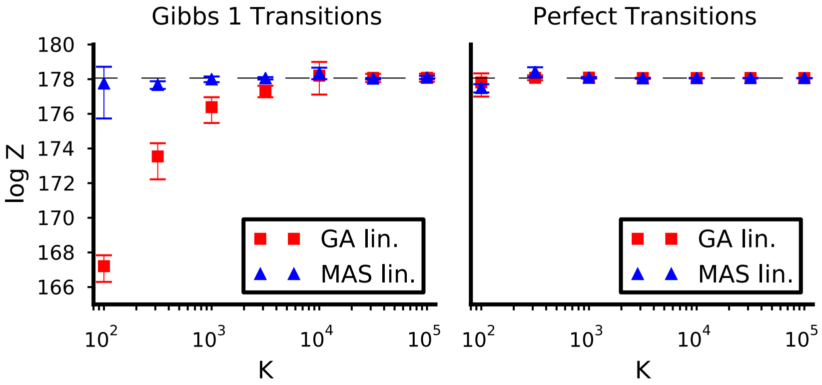
  - Can do approximately for a few intermediate models.
- **Moment Averaged Spline Path** (γ_{MAS} now in blue)

Knots are moment matched, annealing between them with GA



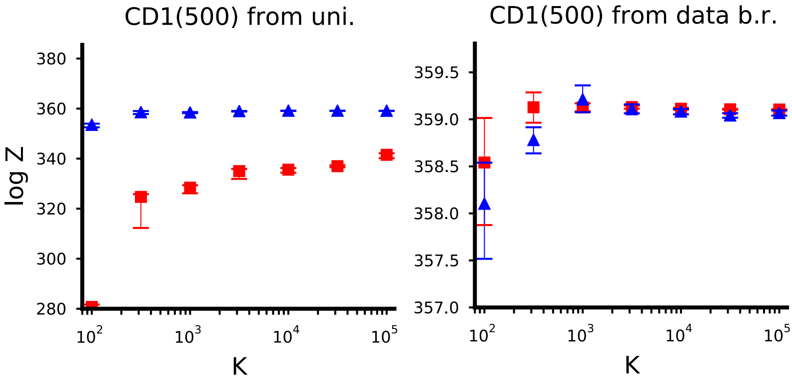
Estimating Partition Functions of RBMs

Estimating partition function of an RBM with 20 hidden units trained on MNIST with PCD



Estimating Partition Functions of RBMs

Estimating partition function of an RBM with 500 hidden units trained on MNIST with CD1.



Under estimating by **20 nats** is difference between a log probability of 130 and 110 on MNIST test set!

overestimate $\rightarrow p(\mathbf{x}) = \exp(-E(\mathbf{x}, \boldsymbol{\theta}))/\mathcal{Z}(\boldsymbol{\theta}) \leftarrow$ underestimate

Conclusions

- Theoretical foundations for studying AIS under prefect mixing
- A new annealing scheme for exponential family distributions with practical approximations
- Improved estimates of partition functions for RBMs
- Ongoing work
  - Diagnostics for AIS
  - Extend this work to models that are harder for AIS
  - MA intermediate distributions for other tempering-based methods such as learning MRFs with MA parallel tempering
  - Using MA path in marginal likelihood estimation for directed models

Thanks!