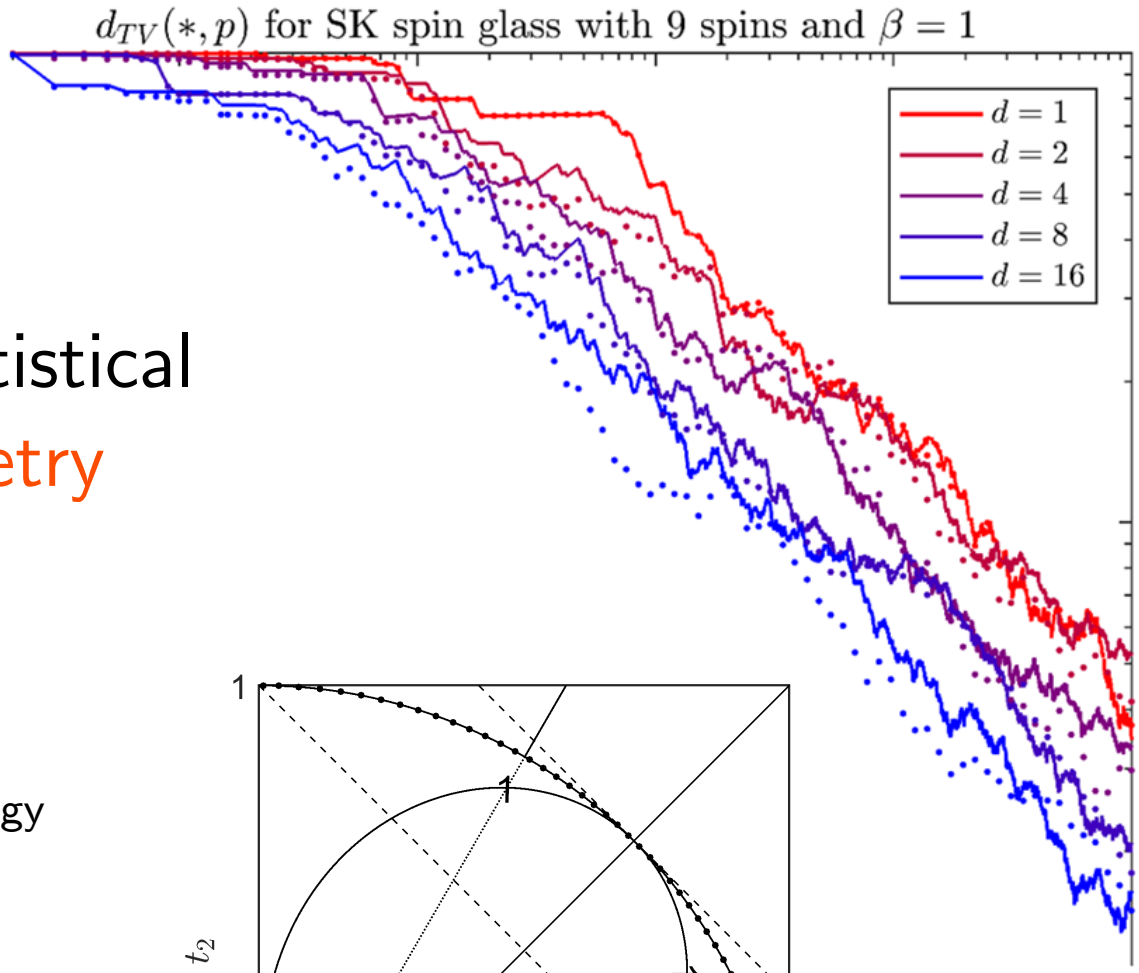
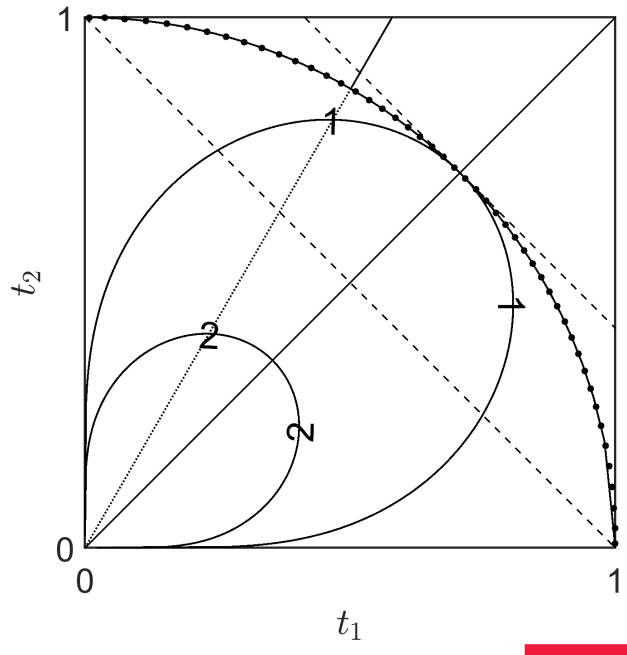


Sampling and statistical physics via symmetry

Steve Huntsman

FAST Labs / Cyber Technology

27 July 2020



BAE SYSTEMS

MCMC builds $X_t \sim p$ in the limit of large t

- Sampling from $p_j := \exp(-\beta E_j)/Z$ is generally hard
- MCMC: build a nice Markov chain with invariant measure p

MCMC builds $X_t \sim p$ in the limit of large t

- Sampling from $p_j := \exp(-\beta E_j)/Z$ is generally hard
- MCMC: build a nice Markov chain with invariant measure p
- $\mathbb{E}_p f(X) = \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{t} f(X_j)$
  - Even though X_j are correlated!
- The price: slow convergence (and no good way to track it)

MCMC builds $X_t \sim p$ in the limit of large t

- Sampling from $p_j := \exp(-\beta E_j)/Z$ is generally hard
- MCMC: build a nice Markov chain with invariant measure p
- $\mathbb{E}_p f(X) = \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{t} f(X_j)$
  - Even though X_j are correlated!
- The price: slow convergence (and no good way to track it)
- Typically decompose into *proposal* and *acceptance* steps
  - Proposal probability: $q_{jk} := \mathbb{P}(X' = k | X_t = j)$
  - Acceptance probability: $\alpha_{jk} := \mathbb{P}(X_{t+1} = k | X' = k, X_t = j)$
  - Chain transition matrix: $P_{jk} := \mathbb{P}(X_{t+1} = k | X_t = j) = q_{jk}\alpha_{jk}$
- Most attention in the literature is devoted to proposals, but we will only be concerned with acceptances

The typical MCMC algorithm has a simple form

Input: Runtime T and $P_{jk} = q_{jk}\alpha_{jk}$ with $pP = p$
Initialize $t = 0$ and X_0
repeat
  for each state k **do**
    Propose k with probability q_{jk}
  end for
  Accept $X_{t+1} = k$ with probability α_{jk}
  Set $t = t + 1$
until $t = T$
Output: $\{X_t\}_{t=0}^{T} \sim p^{\times(T+1)}$ (approximately)

The Hastings algorithm is an acceptance specification

- Hastings: accept proposal with probability $\alpha_{jk} = s_{jk}/(1 + t_{jk})$
  - $t_{jk} := p_j q_{jk} / p_k q_{kj}$
- Requirements on s:
  - $s_{jk} = s_{kj}$
  - $s_{jk} \leq 1 + \min(t_{jk}, t_{kj})$
- $s_{jk} = 1$: Barker sampler
- $s_{jk} = 1 + \min(t_{jk}, t_{kj})$: Metropolis-Hastings sampler

Lie groups and algebras might seem unrelated . . .

- A *Lie group* is a group and manifold with smooth group ops
- Classical real matrix examples sit inside $GL(n, \mathbb{R})$:
  - E.g., $O(n) := \{A \in GL(n, \mathbb{R}) : AA^T = I\}$
  - "Special" $(\det = 1)$ subgroups, e.g., $SL(n, \mathbb{R})$, $SO(n)$, . . .

Lie groups and algebras might seem unrelated . . .

- A *Lie group* is a group and manifold with smooth group ops
- Classical real matrix examples sit inside $GL(n, \mathbb{R})$:
  - E.g., $O(n) := \{A \in GL(n, \mathbb{R}) : AA^T = I\}$
  - "Special" $(\det = 1)$ subgroups, e.g., $SL(n, \mathbb{R})$, $SO(n)$, . . .
- Tangent space at the identity of a Lie group is a *Lie algebra*
  - Echoes Lie group structure via bilinear *bracket* satisfying the *Jacobi identity* $[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0$

Lie groups and algebras might seem unrelated . . .

- A *Lie group* is a group and manifold with smooth group ops
- Classical real matrix examples sit inside $GL(n, \mathbb{R})$:
  - E.g., $O(n) := \{A \in GL(n, \mathbb{R}) : AA^T = I\}$
  - "Special" ($\det = 1$) subgroups, e.g., $SL(n, \mathbb{R})$, $SO(n)$, . . .
- Tangent space at the identity of a Lie group is a *Lie algebra*
  - Echoes Lie group structure via bilinear *bracket* satisfying the *Jacobi identity* $[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0$
- Ado's theorem: a real finite-dimensional Lie group is isomorphic to a subgroup of $GL(n, \mathbb{R})$
  - Corresponding Lie algebra isomorphic to a subalgebra of $M_n(\mathbb{R})$
  - Bracket is the matrix commutator: $[X, Y] := XY - YX$
  - Matrix exponential gives a map from the Lie algebra to the corresponding Lie group that respects both structures

. . . but note that a measure generates a Lie group

- The *stochastic group* $STO(n) := \{P \in GL(n, \mathbb{R}) : P1 = 1\}$ has a Lie algebra with basis indexed by $(j, k) \in [n] \times [n-1]$:

$$e_{(j,k)} := e_j(e_k^T - e_n^T)$$

. . . but note that a measure generates a Lie group

- The *stochastic group* $STO(n) := \{P \in GL(n, \mathbb{R}) : P1 = 1\}$ has a Lie algebra with basis indexed by $(j, k) \in [n] \times [n-1]$:

$$e_{(j,k)} := e_j(e_k^T - e_n^T)$$

- The *group generated by* p, $\langle p \rangle := \{P \in STO(n) : pP = p\}$, has a Lie algebra with basis indexed by $(j, k) \in [n-1]^2$:

$$\begin{aligned} e^{(p)}_{(j,k)} &:= e_{(j,k)} - r_j e_{(n,k)} \\ &= (e_j - r_j e_n)(e_k^T - e_n^T) \end{aligned}$$

where $r_j := p_j / p_n$

. . . but note that a measure generates a Lie group

- The *stochastic group* $STO(n) := \{P \in GL(n, \mathbb{R}) : P1 = 1\}$ has a Lie algebra with basis indexed by $(j, k) \in [n] \times [n-1]$:

$$e_{(j,k)} := e_j(e_k^T - e_n^T)$$

- The *group generated by* p, $\langle p \rangle := \{P \in STO(n) : pP = p\}$, has a Lie algebra with basis indexed by $(j, k) \in [n-1]^2$:

$$\begin{aligned} e^{(p)}_{(j,k)} &:= e_{(j,k)} - r_j e_{(n,k)} \\ &= (e_j - r_j e_n)(e_k^T - e_n^T) \end{aligned}$$

  where $r_j := p_j/p_n$

- If $p_j \equiv \mathcal{L}_j/Z$ then $r_j = \mathcal{L}_j/\mathcal{L}_n$ does not depend on Z: this is the fundamental reason why MCMC works

Our chosen basis yields convenient formulae

- For $i \in \mathbb{Z}_+$

$$\left(e^{(p)}_{(j,k)}\right)^i = (\delta_{jk} + r_j)^{i-1}\, e^{(p)}_{(j,k)}$$

Our chosen basis yields convenient formulae

- For $i \in \mathbb{Z}_+$

$$\left(e^{(p)}_{(j,k)}\right)^i = (\delta_{jk} + r_j)^{i-1}\, e^{(p)}_{(j,k)}$$

- It follows that

$$\begin{aligned}\exp te^{(p)}_{(j,k)} &= I + \frac{e^{t(\delta_{jk}+r_j)} - 1}{\delta_{jk} + r_j} e^{(p)}_{(j,k)} \\ &:= I + f^{(p)}_{(j,k)}(-t) \cdot e^{(p)}_{(j,k)}\end{aligned}$$

- The case $j = k$ will be particularly important

We consider special monoids in $STO(n)$ and $\langle p \rangle$

- Define the *monoids* (i.e., semigroups with identity)

$$STO^+(n) := \{P \in M(n, \mathbb{R}) : P1 = 1 \text{ and } P \geq 0\}$$

where $P \geq 0$ is interpreted per entry, and

$$\langle p \rangle^+ := \{P \in STO^+(n) : pP = p\}$$

We consider special monoids in $STO(n)$ and $\langle p \rangle$

- Define the *monoids* (i.e., semigroups with identity)

$$STO^+(n) := \{P \in M(n, \mathbb{R}) : P1 = 1 \text{ and } P \geq 0\}$$

where $P \geq 0$ is interpreted per entry, and

$$\langle p \rangle^+ := \{P \in STO^+(n) : pP = p\}$$

- $STO^+(n) \not\subset STO(n)$ and $\langle p \rangle^+ \not\subset \langle p \rangle$
  - The LHSs have noninvertible elements; the RHSs have matrices with negative entries
- $STO^+(n)$ and $\langle p \rangle^+$ are bounded convex polytopes that respectively embody *bona fide* transition and candidate MCMC matrices

We can construct a nice element of $\langle p \rangle^+$

Lemma If $t_j \geq 0$, then $\exp\left(-\sum_j t_j e^{(p)}_{(j,j)}\right) \in \langle p \rangle^+$

Proof $-\sum_j t_j e^{(p)}_{(j,j)}$ is a continuous-time Markov generator matrix. $\square$

We can construct a nice element of $\langle p \rangle^+$

Lemma If $t_j \geq 0$, then $\exp\left(-\sum_j t_j e^{(p)}_{(j,j)}\right) \in \langle p \rangle^+$

Proof $-\sum_j t_j e^{(p)}_{(j,j)}$ is a continuous-time Markov generator matrix. □

- In particular, for $t \geq 0$ we get a closed-form element of $\langle p \rangle^+$:

$$\exp\left(-te^{(p)}_{(j,j)}\right) = I + f^{(p)}_{(j,j)}(t) \cdot e^{(p)}_{(j,j)}$$

- No obvious useful generalization of this expression
  - Closed form for $\exp\left(-t_{(j,k)} e^{(p)}_{(j,k)} - t_{(\ell,m)} e^{(p)}_{(\ell,m)}\right)$ runs many pages or has some manifestly negative entries (but wouldn't count these out *a priori*)

We can construct a nice element of $\langle p \rangle^+$

Lemma If $t_j \geq 0$, then $\exp\left(-\sum_j t_j e^{(p)}_{(j,j)}\right) \in \langle p \rangle^+$

Proof $-\sum_j t_j e^{(p)}_{(j,j)}$ is a continuous-time Markov generator matrix. □

- In particular, for $t \geq 0$ we get a closed-form element of $\langle p \rangle^+$:

$$\exp\left(-t e^{(p)}_{(j,j)}\right) = I + f^{(p)}_{(j,j)}(t) \cdot e^{(p)}_{(j,j)}$$

- No obvious useful generalization of this expression
  - Closed form for $\exp\left(-t_{(j,k)} e^{(p)}_{(j,k)} - t_{(\ell,m)} e^{(p)}_{(\ell,m)}\right)$ runs many pages or has some manifestly negative entries (but wouldn't count these out *a priori*)
- But this is enough to recover classical MCMC samplers!

We recover classical MCMC samplers

- Relabel current state as n; undo after applying matrix in $\langle p \rangle^+$
  - I.e., transition $n \to j$ is generic

We recover classical MCMC samplers

- Relabel current state as n; undo after applying matrix in $\langle p \rangle^+$
  - I.e., transition $n \to j$ is generic
- $P = \exp\left(-te^{(p)}_{(j,j)}\right) = I + f^{(p)}_{(j,j)}(t) \cdot e^{(p)}_{(j,j)} \Rightarrow P_{nj}(t) = -f^{(p)}_{(j,j)}(t) r_j$

We recover classical MCMC samplers

- Relabel current state as n; undo after applying matrix in $\langle p \rangle^+$
  - I.e., transition $n \to j$ is generic
- $P = \exp\left(-te^{(p)}_{(j,j)}\right) = I + f^{(p)}_{(j,j)}(t) \cdot e^{(p)}_{(j,j)} \Rightarrow P_{nj}(t) = -f^{(p)}_{(j,j)}(t) r_j$
- Maximize $P_{nj}(t)$ at $t = \infty$: $P_{nj}(\infty) = r_j/(1 + r_j)$
  - $\mathcal{B}^{(p)} := P(\infty)$ corresponds to the Barker sampler

We recover classical MCMC samplers

- Relabel current state as n; undo after applying matrix in $\langle p \rangle^+$
  - I.e., transition $n \to j$ is generic
- $P = \exp\left(-t e^{(p)}_{(j,j)}\right) = I + f^{(p)}_{(j,j)}(t) \cdot e^{(p)}_{(j,j)} \Rightarrow P_{nj}(t) = -f^{(p)}_{(j,j)}(t) r_j$
- Maximize $P_{nj}(t)$ at $t = \infty$: $P_{nj}(\infty) = r_j/(1 + r_j)$
  - $\mathcal{B}^{(p)} := P(\infty)$ corresponds to the Barker sampler
- But we can almost trivially do better by optimizing over the entire line segment in $\langle p \rangle^+$ that I and $\mathcal{B}^{(p)}$ belong to
  - $I - \tau e^{(p)}_{(j,j)} \in \langle p \rangle^+$ iff $0 \leq \tau \leq \min(1, r_j^{-1})$
  - Taking the upper limit for τ yields the Metropolis sampler:

$$\mathcal{M}^{(p)} := I - \min(1, r_j^{-1}) \cdot e^{(p)}_{(j,j)}; \quad \left(\mathcal{M}^{(p)}\right)_{nj} = \min(1, r_j)$$

Barker sampler

Input: Runtime T and and oracle for r
Initialize $t = 0$ and X_0
repeat
    Relabel states so that $X_t = n$
    Propose $j \in [n-1]$
    Accept $X_{t+1} = j$ with probability $\left(\mathcal{B}^{(p)}\right)_{nj} = r_j/(1+r_j)$
    Undo relabeling; set $t = t+1$
until $t = T$
Output: $\{X_t\}_{t=0}^{T} \sim p^{\times(T+1)}$ (approximately)

Metropolis sampler

Input: Runtime T and and oracle for r
Initialize $t = 0$ and X_0
repeat
    Relabel states so that $X_t = n$
    Propose $j \in [n-1]$
    Accept $X_{t+1} = j$ with probability $\left(\mathcal{M}^{(p)}\right)_{nj} = \min(1, r_j)$
    Undo relabeling; set $t = t + 1$
until $t = T$
Output: $\{X_t\}_{t=0}^{T} \sim p^{\times(T+1)}$ (approximately)

What if we are willing to sacrifice some proposal sparsity?

- Barker/Metropolis samplers are the simplest MCMC methods
  - Simplicity derives from functional form and sparsity of corresponding matrices in $\langle p \rangle^+$
- What if we propose more than one state at a time?
  - Anticipates ensemble/multiple-try MCMC methods

What if we are willing to sacrifice some proposal sparsity?

- Barker/Metropolis samplers are the simplest MCMC methods
  - Simplicity derives from functional form and sparsity of corresponding matrices in $\langle p \rangle^+$
- What if we propose more than one state at a time?
  - Anticipates ensemble/multiple-try MCMC methods
- Natural to expect better convergence/higher complexity
  - Impractical and degenerate limiting case is the matrix $1p$
  - Practical starting case is Barker/Metropolis
- Key consideration is how (or if) we can readily construct suitable elements of $\langle p \rangle^+$

Let's do some algebra aimed at building elements of $\langle p \rangle^+$

- Define $r := (r_1, \ldots, r_{n-1}, 1)$ and $r^- := (r_1, \ldots, r_{n-1})$
- For $\mathcal{J} := \{j_1, \ldots, j_d\} \subseteq [n-1]$ and $\alpha \in M_{n-1}(\mathbb{R})$, define
  - $(\alpha_{(\mathcal{J})})_{uv} := \alpha_{j_u j_v}$
  - $\alpha^{(p)}_{(\mathcal{J})} := \sum_{u,v=1}^{d} \alpha_{j_u j_v} e^{(p)}_{(j_u, j_v)} \in \mathfrak{lie}(\langle p \rangle)$
  - $r_{(\mathcal{J})} := (r_{j_1}, \ldots, r_{j_d})$

Let's do some algebra aimed at building elements of $\langle p \rangle^+$

- Define $r := (r_1, \ldots, r_{n-1}, 1)$ and $r^- := (r_1, \ldots, r_{n-1})$
- For $\mathcal{J} := \{j_1, \ldots, j_d\} \subseteq [n-1]$ and $\alpha \in M_{n-1}(\mathbb{R})$, define
  - $(\alpha_{(\mathcal{J})})_{uv} := \alpha_{j_u j_v}$
  - $\alpha^{(p)}_{(\mathcal{J})} := \sum_{u,v=1}^{d} \alpha_{j_u j_v} e^{(p)}_{(j_u, j_v)} \in \mathfrak{lie}(\langle p \rangle)$
  - $r_{(\mathcal{J})} := (r_{j_1}, \ldots, r_{j_d})$

Lemma Let $\mathcal{J} := \{j_1, \ldots, j_d\} \subseteq [n-1]$. If $\gamma^{(p)}_{(\mathcal{J})} = \alpha^{(p)}_{(\mathcal{J})} \beta^{(p)}_{(\mathcal{J})}$, then

$$\gamma_{(\mathcal{J})} = \alpha_{(\mathcal{J})} (I + 1 r_{(\mathcal{J})}) \beta_{(\mathcal{J})}$$

- This is a notational mess but the lemma is worth it
  - $d = 2$ case takes about a page of algebra to check otherwise
- Using this lemma, we can readily construct an analytically convenient matrix in $\mathfrak{lie}(\langle p \rangle)$ ...

Theorem: we can build a Barker matrix

Let $\mathcal{J} := \{j_1, \ldots, j_d\} \subseteq [n-1]$, $\omega \in \mathbb{R}$ and

$$A^{(p;\omega)}_{(\mathcal{J})} := \omega \sum_{u,v} \left(\delta_{j_u j_v} - \frac{1}{1 + r_{(\mathcal{J})} 1} r_{j_v} \right) e^{(p)}_{(j_u, j_v)} = \left(\omega (I + 1 r_{(\mathcal{J})})^{-1}\right)^{(p)}_{(\mathcal{J})}.$$

(We pick this matrix precisely because we can exponentiate it in closed form easily using the preceding lemma.) Then

$$\exp t A^{(p;\omega)}_{(\mathcal{J})} = I + \frac{e^{\omega t} - 1}{\omega} A^{(p;\omega)}_{(\mathcal{J})}.$$

Moreover, $\exp\left(-t A^{(p;\omega)}_{(\mathcal{J})}\right) \in \langle p \rangle^+ \cap GL(n, \mathbb{R})$ if $t \geq 0$. So the *Barker matrix*

$$\mathcal{B}^{(p)}_{(\mathcal{J})} := I - \omega^{-1} A^{(p;\omega)}_{(\mathcal{J})}$$

is in $\langle p \rangle^+$, and does not depend on ω.

BAE SYSTEMS

Lemma: we can build a Metropolis matrix

Let Δ denote the map that takes a matrix to the vector of its diagonal entries, and indicate the boundary of a nice subset of Euclidean space using ∂.

The *Metropolis matrix*

$$\mathcal{M}^{(p)}_{(\mathcal{J})} := I - \frac{1}{\max \Delta \left(A^{(p;\omega)}_{(\mathcal{J})} \right)} A^{(p;\omega)}_{(\mathcal{J})}$$

is in $\partial \langle p \rangle^+$ and does not depend on ω.

Example: $p = (1, 2, 3, 4, 10)/20$ and $\mathcal{J} = \{1, 2, 3\}$

$$A^{(p;\omega)}_{(\mathcal{J})} = \frac{\omega}{16}\begin{pmatrix} 15 & -2 & -3 & 0 & -10 \\ -1 & 14 & -3 & 0 & -10 \\ -1 & -2 & 13 & 0 & -10 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -2 & -3 & 0 & 6 \end{pmatrix}$$

For $\omega = 1$ and $t = -\log 2$,

$$\exp\left(\log 2 \cdot A^{(p;1)}_{(\mathcal{J})}\right) = \frac{1}{32}\begin{pmatrix} 17 & 2 & 3 & 0 & 10 \\ 1 & 18 & 3 & 0 & 10 \\ 1 & 2 & 19 & 0 & 10 \\ 0 & 0 & 0 & 32 & 0 \\ 1 & 2 & 3 & 0 & 26 \end{pmatrix}$$

Finally,

$$\mathcal{B}^{(p)}_{(\mathcal{J})} = \frac{1}{16}\begin{pmatrix} 1 & 2 & 3 & 0 & 10 \\ 1 & 2 & 3 & 0 & 10 \\ 1 & 2 & 3 & 0 & 10 \\ 0 & 0 & 0 & 16 & 0 \\ 1 & 2 & 3 & 0 & 10 \end{pmatrix}; \quad \mathcal{M}^{(p)}_{(\mathcal{J})} = \frac{1}{15}\begin{pmatrix} 0 & 2 & 3 & 0 & 10 \\ 1 & 1 & 3 & 0 & 10 \\ 1 & 2 & 2 & 0 & 10 \\ 0 & 0 & 0 & 15 & 0 \\ 1 & 2 & 3 & 0 & 9 \end{pmatrix}$$

Example: $p = (1, 2, 3, 4, 10)/20$ and $\mathcal{J} = \{1, 2, 3\}$

$$A^{(p;\omega)}_{(\mathcal{J})} = \frac{\omega}{16} \begin{pmatrix} 15 & -2 & -3 & 0 & -10 \\ -1 & 14 & -3 & 0 & -10 \\ -1 & -2 & 13 & 0 & -10 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & -2 & -3 & 0 & 6 \end{pmatrix}$$

For $\omega = 2$ and $t = -\log 2$,

$$\exp\left(\log 2 \cdot A^{(p;2)}_{(\mathcal{J})}\right) = \frac{1}{64} \begin{pmatrix} 19 & 6 & 9 & 0 & 30 \\ 3 & 22 & 9 & 0 & 30 \\ 3 & 6 & 25 & 0 & 30 \\ 0 & 0 & 0 & 64 & 0 \\ 3 & 6 & 9 & 0 & 46 \end{pmatrix}$$

Finally,

$$\mathcal{B}^{(p)}_{(\mathcal{J})} = \frac{1}{16} \begin{pmatrix} 1 & 2 & 3 & 0 & 10 \\ 1 & 2 & 3 & 0 & 10 \\ 1 & 2 & 3 & 0 & 10 \\ 0 & 0 & 0 & 16 & 0 \\ 1 & 2 & 3 & 0 & 10 \end{pmatrix}; \quad \mathcal{M}^{(p)}_{(\mathcal{J})} = \frac{1}{15} \begin{pmatrix} 0 & 2 & 3 & 0 & 10 \\ 1 & 1 & 3 & 0 & 10 \\ 1 & 2 & 2 & 0 & 10 \\ 0 & 0 & 0 & 15 & 0 \\ 1 & 2 & 3 & 0 & 9 \end{pmatrix}$$

Algebra yields higher-order Barker sampler

- Key idea: let $n \to j \in \mathcal{J}$ correspond to a generic transition
  - We do not specify or constrain a proposal that produces $\mathcal{J}$
- Get matrix entries

$$\frac{1}{\omega}\left(A^{(p;\omega)}_{(\mathcal{J})}\right)_{j_u j_u} = 1 - \frac{r_{j_u}}{1 + r_{(\mathcal{J})}1}$$

$$\frac{1}{\omega}\left(A^{(p;\omega)}_{(\mathcal{J})}\right)_{n j_u} = -\frac{r_{j_u}}{1 + r_{(\mathcal{J})}1}$$

$$\frac{1}{\omega}\left(A^{(p;\omega)}_{(\mathcal{J})}\right)_{nn} = \frac{r_{(\mathcal{J})}1}{1 + r_{(\mathcal{J})}1}$$

- This yields the *higher-order Barker sampler* (HOBS):

$$\left(\mathcal{B}^{(p)}_{(\mathcal{J})}\right)_{n j_u} = \frac{r_{j_u}}{1 + r_{(\mathcal{J})}1}; \quad \left(\mathcal{B}^{(p)}_{(\mathcal{J})}\right)_{nn} = \frac{1}{1 + r_{(\mathcal{J})}1}$$

BAE SYSTEMS

Algebra yields higher-order Metropolis sampler

- Meanwhile

$$\frac{1}{\omega}\max\Delta\left(A^{(p;\omega)}_{(\mathcal{J})}\right)=\frac{1+r_{(\mathcal{J})}1-\min\{1,\min r_{(\mathcal{J})}\}}{1+r_{(\mathcal{J})}1}$$

- This yields the higher-order Metropolis sampler (HOMS):

$$\left(\mathcal{M}^{(p)}_{(\mathcal{J})}\right)_{nj_u}=\frac{r_{j_u}}{1+r_{(\mathcal{J})}1-\min\{1,\min r_{(\mathcal{J})}\}}$$

$$\left(\mathcal{M}^{(p)}_{(\mathcal{J})}\right)_{nn}=1-\frac{r_{(\mathcal{J})}1}{1+r_{(\mathcal{J})}1-\min\{1,\min r_{(\mathcal{J})}\}}$$

Higher-order Barker sampler (HOBS)

Input: Runtime T and and oracle for r
Initialize $t = 0$ and X_0
repeat
  Relabel states so that $X_t = n$
  Propose $\mathcal{J} = \{j_1, \ldots, j_d\} \subseteq [n-1]$
  Accept $X_{t+1} = j_u$ with probability $\left(\mathcal{B}^{(p)}_{(\mathcal{J})}\right)_{nj_u}$
  Undo relabeling; set $t = t + 1$
until $t = T$
Output: $\{X_t\}_{t=0}^{T} \sim p^{\times(T+1)}$ (approximately)

Ensemble MCMC algorithm of (Neal, 2011) as in (Martino, 2018)

Higher-order Metropolis sampler (HOMS)

Input: Runtime T and and oracle for r
Initialize $t = 0$ and X_0
repeat
  Relabel states so that $X_t = n$
  Propose $\mathcal{J} = \{j_1, \ldots, j_d\} \subseteq [n-1]$
  Accept $X_{t+1} = j_u$ with probability $\left(\mathcal{M}^{(p)}_{(\mathcal{J})}\right)_{nj_u}$
  Undo relabeling; set $t = t + 1$
until $t = T$
Output: $\{X_t\}_{t=0}^{T} \sim p^{\times(T+1)}$ (approximately)

Slight specialization of construction in (Delmas & Jourdain, 2009)

Look at behavior on a Sherrington-Kirkpatrick spin glass

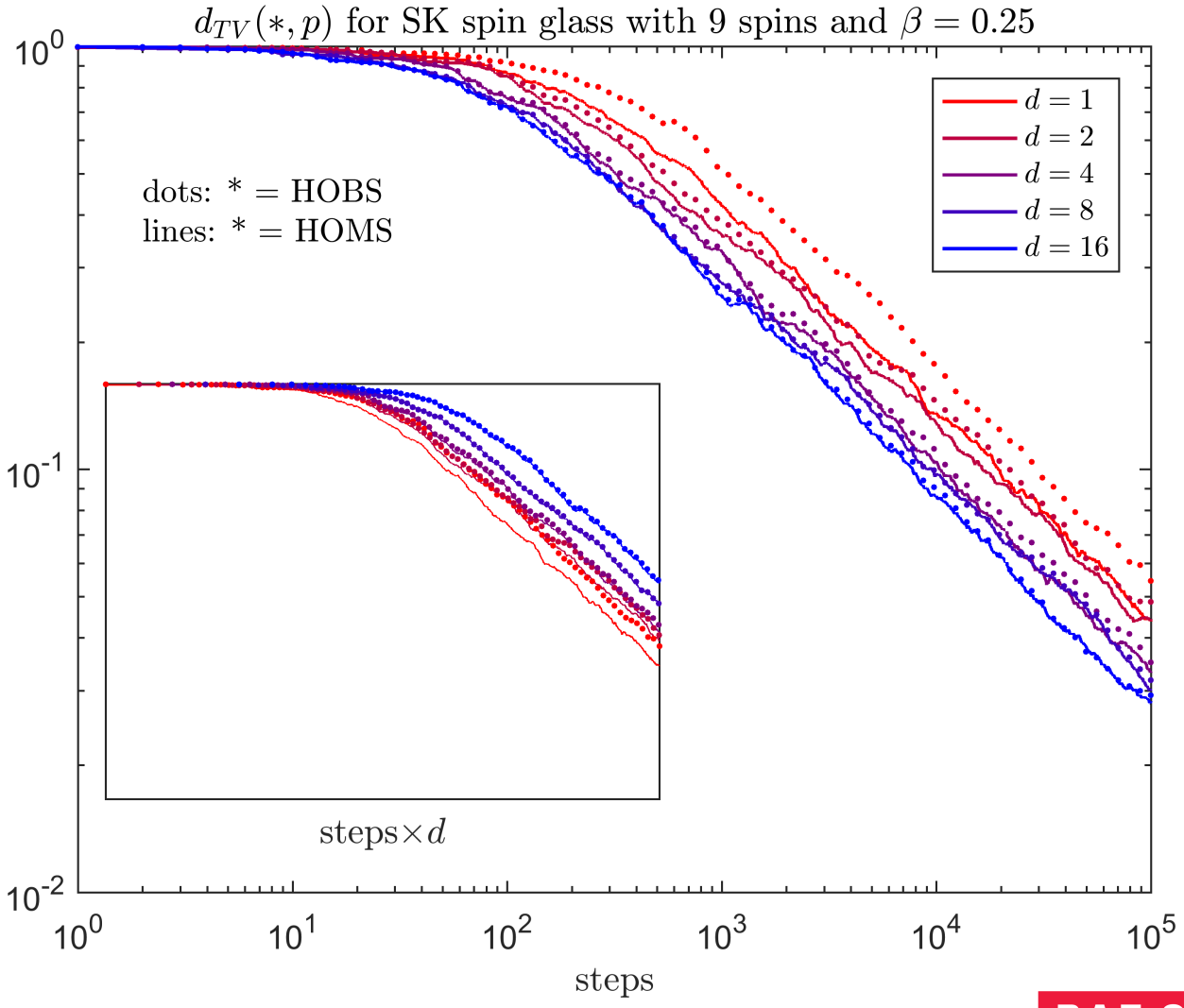
- Sherrington-Kirkpatrick spin glass at inverse temperature β is

$$p(s) := Z^{-1} \exp\left(-\frac{\beta}{\sqrt{N}} \sum_{jk} J_{jk} s_j s_k\right)$$

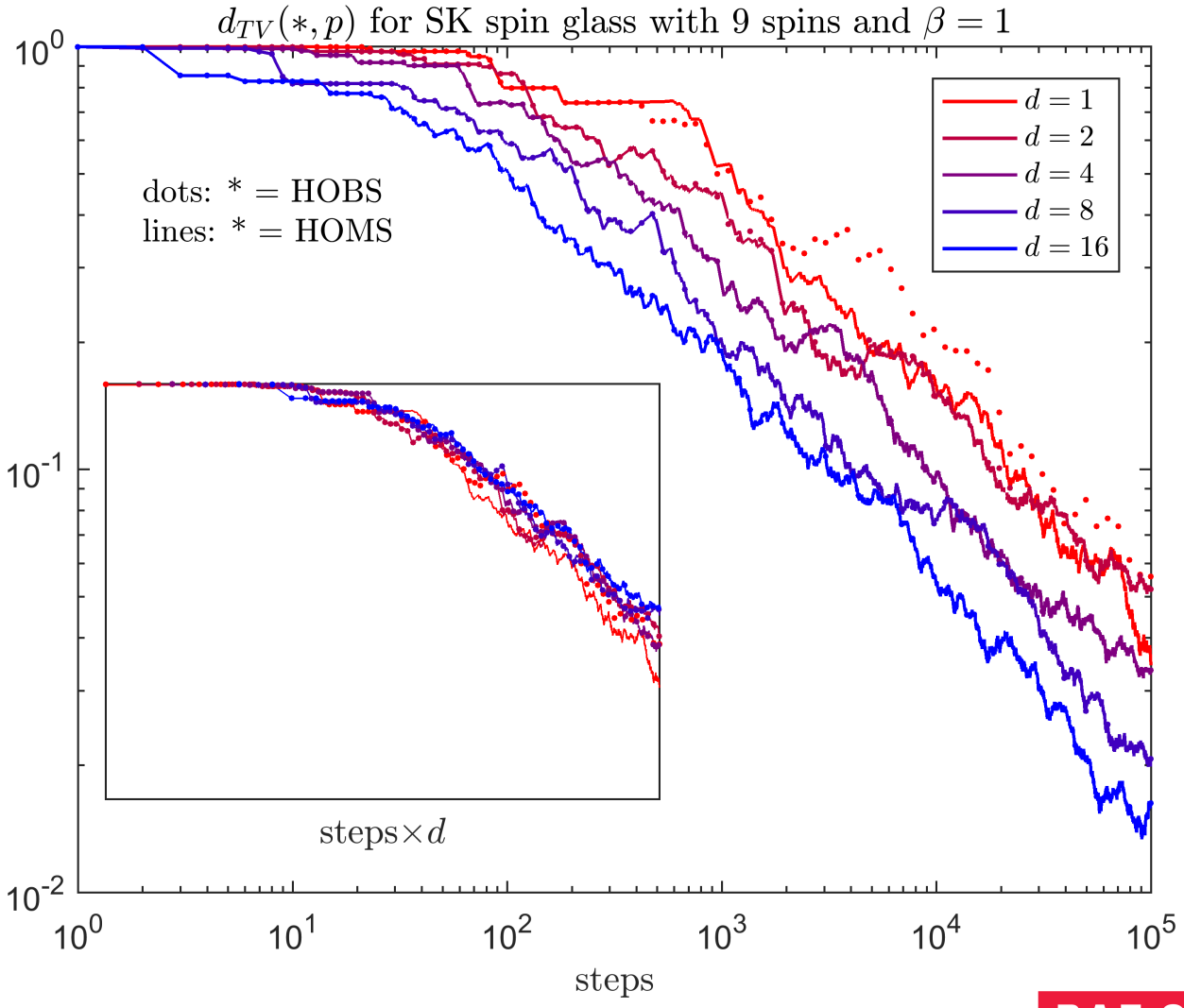
where $s \in \{\pm 1\}^N$; $J_{jk} \sim \mathcal{N}(0,1)$ are IID with $J_{kj} = J_{jk}$

  - We use the same PRNG initial state for each run
  - β low enough ($1/4$ and 1) so single runs are representative

Look at behavior on a Sherrington-Kirkpatrick spin glass



Look at behavior on a Sherrington-Kirkpatrick spin glass



We can still do better than the preceding algorithms

- $I - \tau^{(p)}_{(\mathcal{J})} \in \langle p \rangle^+$ iff τ satisfies various linear constraints
  - $\tau^{(p)}_{(\mathcal{J})} := \begin{pmatrix} I_{n-1} \\ -r^-_{\mathcal{J}} \end{pmatrix} \tau \begin{pmatrix} I_{n-1} & -1^-_{\mathcal{J}} \end{pmatrix}$
  - τ is a generic parameter matrix
  - See paper for exact/simple definitions of $r^-_{\mathcal{J}}$ and $1^-_{\mathcal{J}}$

We can still do better than the preceding algorithms

- $I - \tau^{(p)}_{(\mathcal{J})} \in \langle p \rangle^+$ iff τ satisfies various linear constraints
  - $\tau^{(p)}_{(\mathcal{J})} := \begin{pmatrix} I_{n-1} \\ -r^-_{\mathcal{J}} \end{pmatrix} \tau \begin{pmatrix} I_{n-1} & -1^-_{\mathcal{J}} \end{pmatrix}$
  - τ is a generic parameter matrix
  - See paper for exact/simple definitions of $r^-_{\mathcal{J}}$ and $1^-_{\mathcal{J}}$
- Optimize via linear program
  - Generic objective $x^T \tau^{(p)}_{(\mathcal{J})} y$ for fixed x, y
- There is a natural choice of x, y that yields an optimal Frobenius norm approximation of (the appropriately sparse submatrix of) the "ultimate" transition matrix $1p$
  - Detailed in paper

Example: $p = (1, 2, 3, 4, 10)/20$ and $\mathcal{J} = \{1, 2, 3\}$

$$\mathcal{B}^{(p)}_{(\mathcal{J})} = \frac{1}{16}\begin{pmatrix} 1 & 2 & 3 & 0 & 10 \\ 1 & 2 & 3 & 0 & 10 \\ 1 & 2 & 3 & 0 & 10 \\ 0 & 0 & 0 & 16 & 0 \\ 1 & 2 & 3 & 0 & 10 \end{pmatrix}$$

$$\mathcal{M}^{(p)}_{(\mathcal{J})} = \frac{1}{15}\begin{pmatrix} 0 & 2 & 3 & 0 & 10 \\ 1 & 1 & 3 & 0 & 10 \\ 1 & 2 & 2 & 0 & 10 \\ 0 & 0 & 0 & 15 & 0 \\ 1 & 2 & 3 & 0 & 9 \end{pmatrix}$$

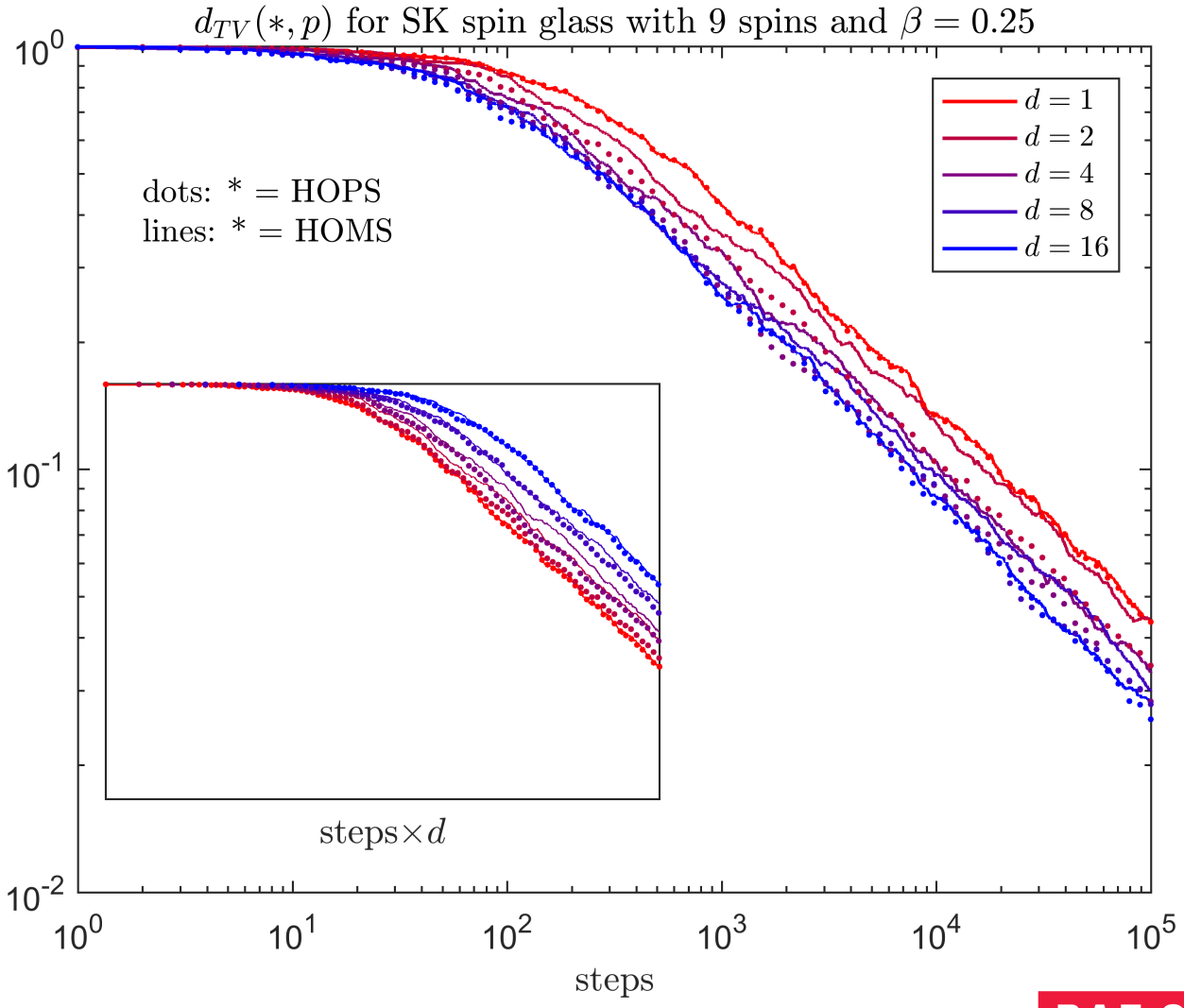
$$\text{opt} = \frac{1}{10}\begin{pmatrix} 0 & 0 & 0 & 0 & 10 \\ 0 & 0 & 0 & 0 & 10 \\ 0 & 0 & 0 & 0 & 10 \\ 0 & 0 & 0 & 10 & 0 \\ 1 & 2 & 3 & 0 & 4 \end{pmatrix}$$

Higher-order programming sampler (HOPS)

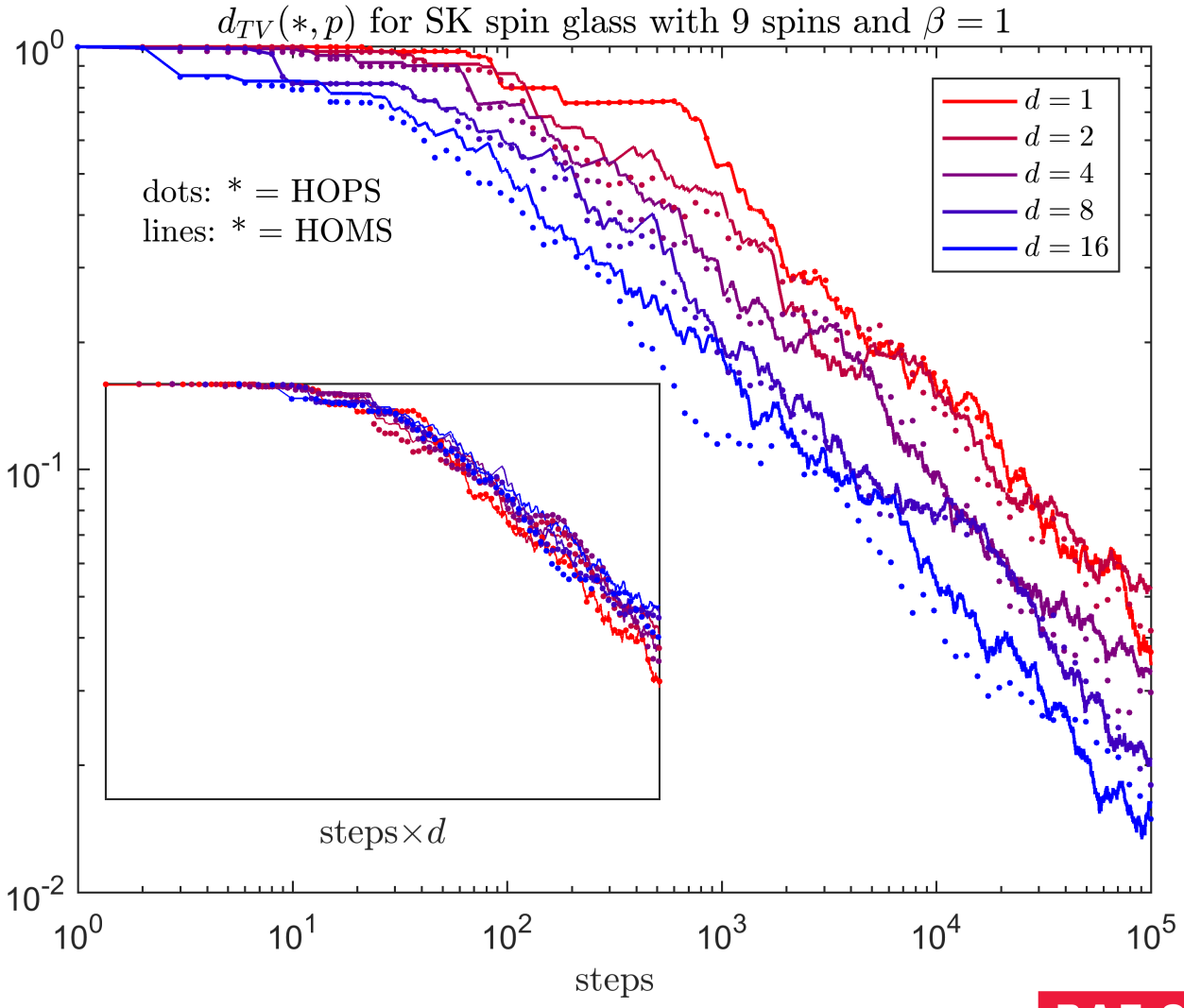
Input: Runtime T and and oracle for r
Initialize $t = 0$ and X_0
repeat
  Relabel states so that $X_t = n$
  Propose $\mathcal{J} = \{j_1, \ldots, j_d\} \subseteq [n-1]$
  Compute optimal τ via linear program
  Set $P = I - \tau^{(p)}_{(\mathcal{J})}$
  Accept $X_{t+1} = j_u$ with probability P_{nj_u}
  Undo relabeling; set $t = t + 1$
until $t = T$
Output: $\{X_t\}_{t=0}^{T} \sim p^{\times(T+1)}$ (approximately)

This algorithm appears to be new

The HOPS outperforms the HOMS



The HOPS outperforms the HOMS



Symmetry unifies MCMC algorithms and gives new ones

- HOPS may be useful for Bayesian inverse problems
- Not tried yet:
  - Continuous variables (would be much more technical)
  - Incorporating proposal mechanism into HOPS objective
  - Generalizing HOPS using convex optimization
  - Determining if HOPS is reversible
- It is possible to produce transiton matrices (even in closed form) with nonnegative nth row but negative entries elsewhere. Not clear if this actually breaks MCMC, though initial experiments in this direction were not encouraging
- Would be nice to sample vertices of $\langle p \rangle^+$, but it's **NP**-hard to sample even approximately uniformly (Khachiyan, 2001)

Part 2: statistical physics from symmetry

We will derive β from data

- There is a unique effective temperature β^{-1} for finite systems consistent both with Gibbs relation in equilibrium and physical scaling requirements
  - Immediately yields an effective energy function
  - Form suggests application to nonequilibrium steady states

We will derive β from data

- There is a unique effective temperature β^{-1} for finite systems consistent both with Gibbs relation in equilibrium and physical scaling requirements
  - Immediately yields an effective energy function
  - Form suggests application to nonequilibrium steady states
- β and derived quantities useful for data analysis

We will derive β from data

- There is a unique effective temperature β^{-1} for finite systems consistent both with Gibbs relation in equilibrium and physical scaling requirements
  - Immediately yields an effective energy function
  - Form suggests application to nonequilibrium steady states
- β and derived quantities useful for data analysis
- We will exhibit an application to Anosov systems
  - Gallavotti-Cohen chaotic hypothesis: generic systems are morally Anosov

The Gibbs distribution can be derived from symmetry (1)

Ansatz The probability of a state depends only on its energy

- Akin to Faddeev characterization of entropy

Energy only defined up to additive constant ε, so $\exists f$ s.t.

$$\mathbb{P}(E_k) = \frac{f(E_k)}{\sum_j f(E_j)} = \frac{f(E_k + \varepsilon)}{\sum_j f(E_j + \varepsilon)}$$

Define

$$g_E(\varepsilon) := \frac{\sum_j f(E_j + \varepsilon)}{\sum_j f(E_j)}$$

Now $g_E(0) = 1$ and

$$\mathbb{P}(E_k) = \frac{f(E_k)}{\sum_j f(E_j + \varepsilon)} g_E(\varepsilon) = \frac{f(E_k + \varepsilon)}{\sum_j f(E_j + \varepsilon)}$$

The Gibbs distribution can be derived from symmetry (2)

- From preceding slide: $g_E(0) = 1$ and

$$\mathbb{P}(E_k) = \frac{f(E_k)}{\sum_j f(E_j + \varepsilon)} g_E(\varepsilon) = \frac{f(E_k + \varepsilon)}{\sum_j f(E_j + \varepsilon)}$$

$\Rightarrow$ $f(E_k) \cdot g_E(\varepsilon) = f(E_k + \varepsilon)$

$\Rightarrow$ $f(E_k + \varepsilon) - f(E_k) = (g_E(\varepsilon) - 1) \cdot f(E_k)$

$\Rightarrow$ $f'(E_k) = g'_E(0) \cdot f(E_k)$ since $g_E(0) = 1$

$\Rightarrow$ $f(E_k) = C \exp(g'_E(0) E_k)$

The Gibbs distribution can be derived from symmetry (2)

- From preceding slide: $g_E(0) = 1$ and

$$\mathbb{P}(E_k) = \frac{f(E_k)}{\sum_j f(E_j + \varepsilon)} g_E(\varepsilon) = \frac{f(E_k + \varepsilon)}{\sum_j f(E_j + \varepsilon)}$$

$\Rightarrow$ $f(E_k) \cdot g_E(\varepsilon) = f(E_k + \varepsilon)$

$\Rightarrow$ $f(E_k + \varepsilon) - f(E_k) = (g_E(\varepsilon) - 1) \cdot f(E_k)$

$\Rightarrow$ $f'(E_k) = g_E'(0) \cdot f(E_k)$ since $g_E(0) = 1$

$\Rightarrow$ $f(E_k) = C \exp(g_E'(0) E_k)$

- Set (w.l.o.g.) $\beta := -g_E'(0)$ and $C \equiv 1$ for Gibbs distribution
  - Self-consistent argument since $g_E(\varepsilon) = \exp(-\beta\varepsilon) \Rightarrow g_E \equiv g$
  - Derivation is for the canonical ensemble (fixed β)

The Gibbs relation defines coordinates (1)

- We consider a stationary system with
  - $n < \infty$ states
  - Probability distribution $p = (p_1, \ldots, p_n) > 0$
  - Characteristic timescale t_∞ (think mixing time or similar)

The Gibbs relation defines coordinates (1)

- We consider a stationary system with
  - $n < \infty$ states
  - Probability distribution $p = (p_1, \ldots, p_n) > 0$
  - Characteristic timescale t_∞ (think mixing time or similar)
- $t \equiv (t_1, \ldots, t_n) := t_\infty p$
  - $\Rightarrow$ $t_j / t_\infty = p_j$
  - $\Rightarrow$ $t_\infty = \sum_k t_j$
- $H := (E_1, \ldots, E_n, \beta^{-1})$

The Gibbs relation defines coordinates (1)

- We consider a stationary system with
  - $n < \infty$ states
  - Probability distribution $p = (p_1, \ldots, p_n) > 0$
  - Characteristic timescale t_∞ (think mixing time or similar)
- $t \equiv (t_1, \ldots, t_n) := t_\infty p$
  - $\Rightarrow\ t_j / t_\infty = p_j$
  - $\Rightarrow\ t_\infty = \sum_k t_j$
- $H := (E_1, \ldots, E_n, \beta^{-1})$
- Want coordinate map $t \mapsto H$ vs. more common map $H \mapsto p$

The Gibbs relation defines coordinates (1)

- We consider a stationary system with
  - $n < \infty$ states
  - Probability distribution $p = (p_1, \ldots, p_n) > 0$
  - Characteristic timescale t_∞ (think mixing time or similar)
- $t \equiv (t_1, \ldots, t_n) := t_\infty p$
  $\Rightarrow$ $t_j / t_\infty = p_j$
  $\Rightarrow$ $t_\infty = \sum_k t_j$
- $H := (E_1, \ldots, E_n, \beta^{-1})$
- Want coordinate map $t \mapsto H$ vs. more common map $H \mapsto p$
- $e^{-\beta E_j}/Z = p_j^{(H)} = p_j^{(t)} = t_j / t_\infty$
- W.l.o.g., set $\sum_j E_j = 0$
  - Not fixing U or anything physical
  - Can later redefine zero point if desired, e.g. $\sum_j E_j = n\beta^{-1}$

The Gibbs relation defines coordinates (2)

- A line of algebra yields

$$\gamma_j := \beta E_j = \frac{1}{n} \sum_{k=1}^{n} \log p_k - \log p_j$$

The Gibbs relation defines coordinates (2)

- A line of algebra yields

$$\gamma_j := \beta E_j = \frac{1}{n} \sum_{k=1}^{n} \log p_k - \log p_j$$

- $\beta = \|\beta H\| / \|H\| = \sqrt{\|\gamma\|^2 + 1} / \|H\|$
- We will get $\|H\|$ from symmetry and scaling considerations
- This will immediately yield β

β scales as t_∞

- A physically reasonable $t \mapsto H$ must depend on some constant governing parameter x, i.e. $\beta \equiv f(x,t) \equiv f(x, t_\infty, p)$

β scales as t_∞

- A physically reasonable $t \mapsto H$ must depend on some constant governing parameter x, i.e. $\beta \equiv f(x,t) \equiv f(x, t_\infty, p)$
- Π-theorem: $\beta = x^\xi t_\infty^\omega \Psi(p)$ for non-dimensional Ψ

β scales as t_∞

- A physically reasonable $t \mapsto H$ must depend on some constant governing parameter x, i.e. $\beta \equiv f(x,t) \equiv f(x,t_\infty,p)$
- Π-theorem: $\beta = x^\xi t_\infty^\omega \Psi(p)$ for non-dimensional Ψ
- Dilating time by C in a system with Hamiltonian $\mathcal{H}$ induces $t_\infty \mapsto t'_\infty = t_\infty/C$ and the extended canonical transformation

$$X \mapsto X' = X, \quad P \mapsto P' = CP, \quad \mathcal{H} \mapsto \mathcal{H}' = C\mathcal{H}$$

- Since this is a change of units, it leaves $e^{-\beta\mathcal{H}}$ invariant
  - I.e., $\beta' = \beta/C$, so $\omega = 1$ and β scales as t_∞
  - Other arguments (classical gas, KMS, etc.) give same result

β scales as t_∞

- A physically reasonable $t \mapsto H$ must depend on some constant governing parameter x, i.e. $\beta \equiv f(x,t) \equiv f(x,t_\infty,p)$
- Π-theorem: $\beta = x^\xi t_\infty^\omega \Psi(p)$ for non-dimensional Ψ
- Dilating time by C in a system with Hamiltonian $\mathcal{H}$ induces $t_\infty \mapsto t'_\infty = t_\infty/C$ and the extended canonical transformation

$$X \mapsto X' = X, \quad P \mapsto P' = CP, \quad \mathcal{H} \mapsto \mathcal{H}' = C\mathcal{H}$$

- Since this is a change of units, it leaves $e^{-\beta\mathcal{H}}$ invariant
  - I.e., $\beta' = \beta/C$, so $\omega = 1$ and β scales as t_∞
  - Other arguments (classical gas, KMS, etc.) give same result
- Take $x = \hbar$ so $\xi = -1$ and $\beta = \hbar^{-1} t_\infty \Psi(p)$
  - Work in natural units and suppress $\hbar$

Rays and radii are preserved by $t \mapsto H$

- p is invariant under $t \mapsto t/C$, so γ is also invariant
- *Ansatz* $\beta \equiv \beta(t)$: $t \mapsto t/C \Rightarrow H = \frac{1}{\beta(t)}(\gamma, 1) \mapsto \frac{1}{\beta(t/C)}(\gamma, 1)$
- So p is constant on rays in both t and H coordinates

Rays and radii are preserved by $t \mapsto H$

- p is invariant under $t \mapsto t/C$, so γ is also invariant
- *Ansatz* $\beta \equiv \beta(t)$: $t \mapsto t/C \Rightarrow H = \frac{1}{\beta(t)}(\gamma, 1) \mapsto \frac{1}{\beta(t/C)}(\gamma, 1)$
- So p is constant on rays in both t and H coordinates

Lemma A smooth map between $t \mapsto H$ respecting the Gibbs relation and $\sum_j E_j = 0$ sends rays and sphere orthants in t coordinates to rays and hemispheres in H coordinates, respectively

The effective temperature has a simple closed form

- $u := 1 \cdot \|t\|/\sqrt{n} \Rightarrow \|u\| = \|t\|$; lemma $\Rightarrow \|H(t)\| = \|H(u)\|$

The effective temperature has a simple closed form

- $u := 1 \cdot \|t\|/\sqrt{n} \Rightarrow \|u\| = \|t\|$; lemma $\Rightarrow \|H(t)\| = \|H(u)\|$
- $H(u) = (0, \ldots, 0, 1/\beta(u)) \Rightarrow \|H(t)\| = 1/\beta(u)$

The effective temperature has a simple closed form

- $u := 1 \cdot \|t\|/\sqrt{n} \Rightarrow \|u\| = \|t\|$; lemma $\Rightarrow \|H(t)\| = \|H(u)\|$
- $H(u) = (0, \ldots, 0, 1/\beta(u)) \Rightarrow \|H(t)\| = 1/\beta(u)$
- Follows that $\beta(t) = \beta(u) \cdot \sqrt{\|\gamma\|^2 + 1}$

The effective temperature has a simple closed form

- $u := 1 \cdot \|t\|/\sqrt{n} \Rightarrow \|u\| = \|t\|$; lemma $\Rightarrow \|H(t)\| = \|H(u)\|$
- $H(u) = (0, \ldots, 0, 1/\beta(u)) \Rightarrow \|H(t)\| = 1/\beta(u)$
- Follows that $\beta(t) = \beta(u) \cdot \sqrt{\|\gamma\|^2 + 1}$
- $\beta(u) = K\|t\| = Kt_\infty\|p\|$ ($K =$ constant) since β scales as t_∞

The effective temperature has a simple closed form

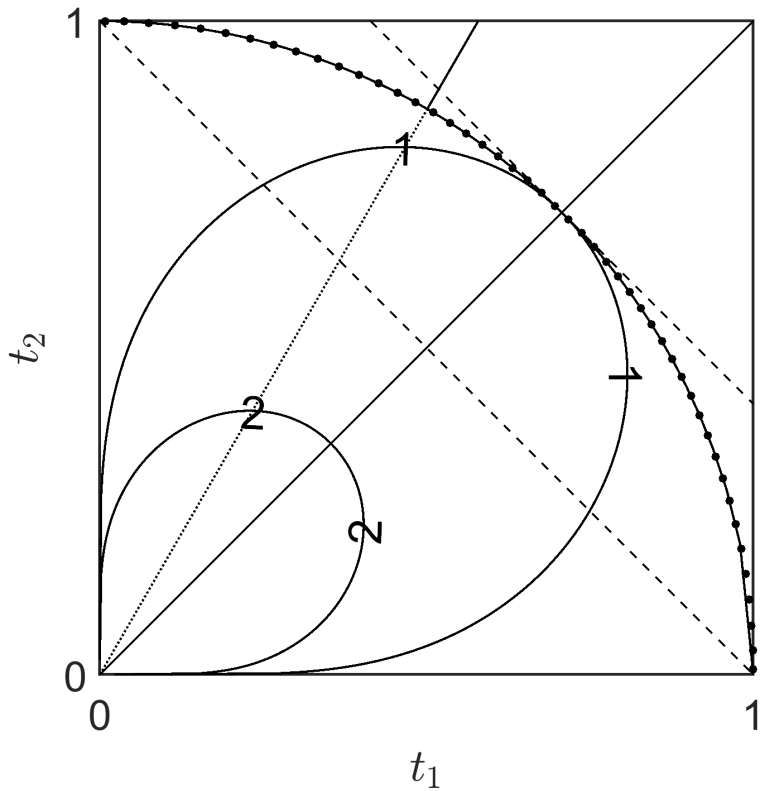
- $u := 1 \cdot \|t\|/\sqrt{n} \Rightarrow \|u\| = \|t\|$; lemma $\Rightarrow \|H(t)\| = \|H(u)\|$
- $H(u) = (0, \ldots, 0, 1/\beta(u)) \Rightarrow \|H(t)\| = 1/\beta(u)$
- Follows that $\beta(t) = \beta(u) \cdot \sqrt{\|\gamma\|^2 + 1}$
- $\beta(u) = K\|t\| = Kt_\infty\|p\|$ ($K =$ constant) since β scales as t_∞
- Taking $K \equiv \hbar^{-1} = 1$ yields

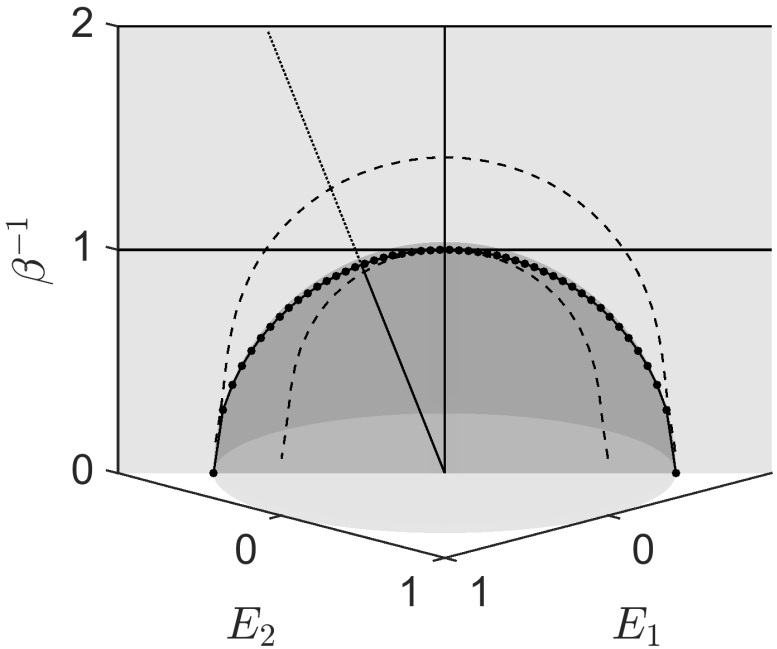
$$\beta(t) = t_\infty\|p\| \cdot \sqrt{\|\gamma\|^2 + 1}$$

- Using $\gamma_j := \beta E_j = \frac{1}{n}\sum_{k=1}^{n} \log p_k - \log p_j$ gives β explicitly in terms of p and t_∞
- Ideas behind derivation of β mostly due to David Ford

Summarizing the bijection $t \leftrightarrow H$

Level curves of $\beta^{-1} = 1, 2$ (solid contours) and of $t_\infty = 1, \sqrt{2}$ (dashed contours) are shown in both coordinate systems.
The bijection is also shown explicitly for circular arcs and rays.





To review: we got here with just a few symmetries

Axiom	Zero point of energy is physically irrelevant
Axiom	The probability of a state depends only on its energy
Derived	Changing unit of time leaves $\beta\mathcal{H}$ invariant
Derived	Any physically nice bijection $t \leftrightarrow H$ preserves rays and radii

What is t_∞? How can we use β?

- Intensivity implies that t_∞ must behave roughly–but not exactly–like a mixing time
  - Precise details still unclear but looking at free energy of discrete memoryless channels offers a possible solution
  - Rest of the talk: L^2 mixing time is a generic surrogate for t_∞

What is t_∞? How can we use β?

- Intensivity implies that t_∞ must behave roughly–but not exactly–like a mixing time
  - Precise details still unclear but looking at free energy of discrete memoryless channels offers a possible solution
  - Rest of the talk: L^2 mixing time is a generic surrogate for t_∞
- Obvious applications to time-varying Markov processes
  - Original motivation of research (started by David Ford in 1998; joint *circa* 2000-2008): analyze Markov processes obtained from computer network traffic

What is t_∞? How can we use β?

- Intensivity implies that t_∞ must behave roughly–but not exactly–like a mixing time
  - Precise details still unclear but looking at free energy of discrete memoryless channels offers a possible solution
  - Rest of the talk: L^2 mixing time is a generic surrogate for t_∞
- Obvious applications to time-varying Markov processes
  - Original motivation of research (started by David Ford in 1998; joint *circa* 2000-2008): analyze Markov processes obtained from computer network traffic
- What about applications to physics?

It's hard to find physically relevant examples

- Obvious (but not good) candidate: equilibrium spin systems
  - Single Glauber-Ising spin: β^{-1} = actual temperature
    $\Rightarrow t_\infty \propto 1/$(largest energy scale)
  - Unfortunately, the only point of looking at equilibrium spin systems would be to help understand t_∞ (analytically hard)
  - Spin glasses are very nonstationary ("aging")

It's hard to find physically relevant examples

- Obvious (but not good) candidate: equilibrium spin systems
  - Single Glauber-Ising spin: β^{-1} = actual temperature
    $\Rightarrow t_\infty \propto 1/(\text{largest energy scale})$
  - Unfortunately, the only point of looking at equilibrium spin systems would be to help understand t_∞ (analytically hard)
  - Spin glasses are very nonstationary ("aging")
- For a continuous example, need well-behaved phase space discretization where p and $|\log p|$ are both in $L^1 \cap L^2$
  - No obvious nontrivial examples with physical measure absolutely continuous w.r.t. phase space volume

It's hard to find physically relevant examples

- Obvious (but not good) candidate: equilibrium spin systems
  - Single Glauber-Ising spin: β^{-1} = actual temperature
    $\Rightarrow t_\infty \propto 1/$(largest energy scale)
  - Unfortunately, the only point of looking at equilibrium spin systems would be to help understand t_∞ (analytically hard)
  - Spin glasses are very nonstationary ("aging")
- For a continuous example, need well-behaved phase space discretization where p and $|\log p|$ are both in $L^1 \cap L^2$
  - No obvious nontrivial examples with physical measure absolutely continuous w.r.t. phase space volume
- What about scaling limits of discrete systems?
  - Naive discretizations of ideal gas with obvious boundary conditions, UV cutoff, etc. have no reasonable scaling limit

It turns out that *Anosov systems* are very good examples

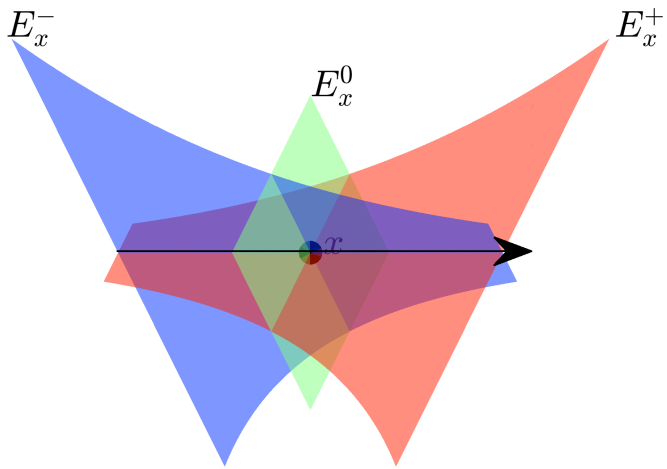
- Physical relevance from Gallavotti-Cohen *chaotic hypothesis*:
  - "For the purpose of studying macroscopic properties, the time evolution map $[T]$ of a many-particle system can be regarded as a mixing Anosov map"
  - *Markov partitions* are natural discretizations that help with the fact that the physical (SRB) probability measure is typically singular w.r.t. phase space volume
  - L^2 mixing time is a computable proxy for t_∞

It turns out that *Anosov systems* are very good examples

- Physical relevance from Gallavotti-Cohen *chaotic hypothesis*:
  - "For the purpose of studying macroscopic properties, the time evolution map $[T]$ of a many-particle system can be regarded as a mixing Anosov map"
  - *Markov partitions* are natural discretizations that help with the fact that the physical (SRB) probability measure is typically singular w.r.t. phase space volume
  - L^2 mixing time is a computable proxy for t_∞
- We have analyzed archetypal examples
  - "Cat map" on the torus
  - Free particle on surfaces of constant negative curvature
- Many general implications, not least by thermostatting

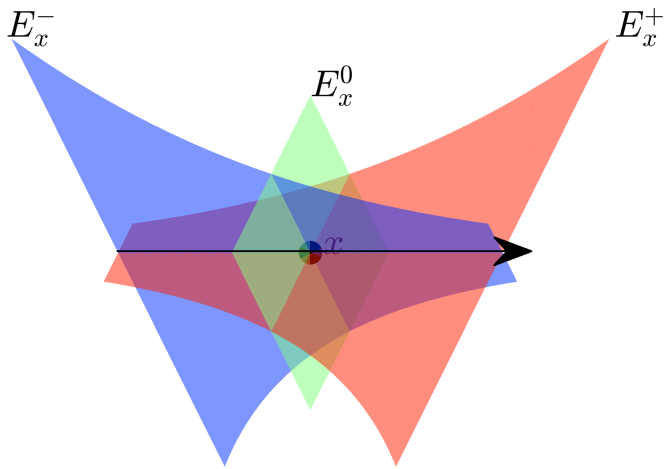
What's an Anosov system?

- A smooth endomorphism T is an *Anosov map* if it is both
  - *Uniformly hyperbolic*, i.e. at every point x there are transverse local stable and unstable surfaces on which points respectively converge and diverge exponentially at a rate independent of x
  - *Invariant*, i.e. the tangent spaces to these surfaces are mapped by the derivative of T into the tangent spaces to the corresponding surfaces at $Tx \equiv T(x)$



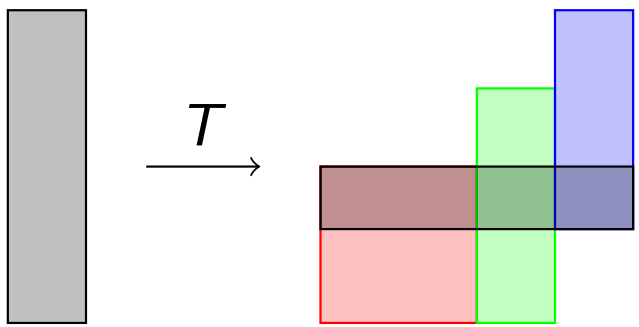
What's an Anosov system?

- A smooth endomorphism T is an *Anosov map* if it is both
  - *Uniformly hyperbolic*, i.e. at every point x there are transverse local stable and unstable surfaces on which points respectively converge and diverge exponentially at a rate independent of x
  - *Invariant*, i.e. the tangent spaces to these surfaces are mapped by the derivative of T into the tangent spaces to the corresponding surfaces at $Tx \equiv T(x)$
- T is *mixing* if global stable and unstable surfaces are dense
- Continuous-time notion of *Anosov flow* is defined similarly



Anosov systems have Markov partitions

- A *rectangle* R is a subset of phase space such that the intersection of a local stable and a local unstable surface consists of a single point also in R: i.e., there is a local product structure compatible with T
  - Generally not a rectangle in the usual geometrical sense
- A partition $\mathcal{R} = \{R_j\}_{j=1}^n$ of phase space into rectangles is *Markov* if (whenever these sets intersect) the images TR_j stretch completely across R_k in the unstable direction and R_k stretches completely across TR_j in the stable direction



The Arnol'd-Avez cat map is Anosov

- Anosov map defined by $T_A x = Ax \mod 1$, where $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$
  - More generally, matrices in $GL(n, \mathbb{Z})$ with no eigenvalues in S^1 correspond to *hyperbolic toral automorphisms* (HTAs)
  - Rectangles for HTAs are geometrically unions of parallelograms

The Arnol'd-Avez cat map is Anosov

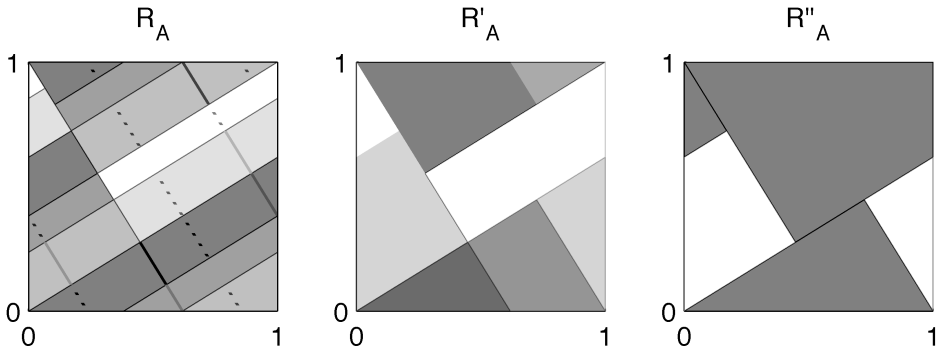
- Anosov map defined by $T_A x = Ax \mod 1$, where $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$
  - More generally, matrices in $GL(n, \mathbb{Z})$ with no eigenvalues in S^1 correspond to *hyperbolic toral automorphisms* (HTAs)
  - Rectangles for HTAs are geometrically unions of parallelograms
- Corresponds to unit-frequency projections for Hamiltonian $\mathcal{H}_A(X, P) = K(P^2 - X^2 + XP)$ with $K = \sinh^{-1}(\sqrt{5}/2)/\sqrt{5}$

The Arnol'd-Avez cat map is Anosov

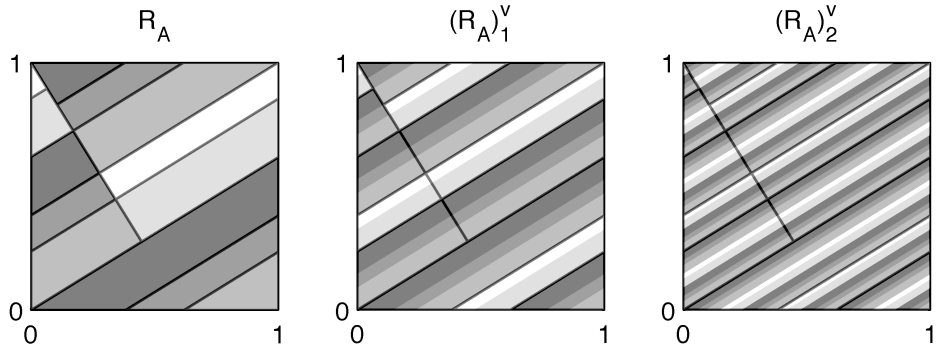
- Anosov map defined by $T_A x = Ax \mod 1$, where $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$
  - More generally, matrices in $GL(n, \mathbb{Z})$ with no eigenvalues in S^1 correspond to *hyperbolic toral automorphisms* (HTAs)
  - Rectangles for HTAs are geometrically unions of parallelograms
- Corresponds to unit-frequency projections for Hamiltonian $\mathcal{H}_A(X, P) = K(P^2 - X^2 + XP)$ with $K = \sinh^{-1}(\sqrt{5}/2)/\sqrt{5}$
- Eigenvalues $\lambda_\pm = \phi^{\pm 2}$, where $\phi = \frac{1+\sqrt{5}}{2}$
- Eigenvectors $e_- = (s, -c)^*$, $e_+ = (c, s)^*$
  - $c = 1/\sqrt{3 - \phi}$ and $s = \sqrt{1 - c^2}$
  - Irrational slopes of eigenvectors imply dense stable and unstable curves on the torus, so cat map is mixing

There are many Markov partitions for the cat map

- E.g., $\mathcal{R}_A$, $\mathcal{R}'_A$, $\mathcal{R}''_A$
  - 3 rounds of "greedy refinements" [defined later] shown for $\mathcal{R}_A$



- Refinements $(\mathcal{R}_A)^{\vee}_m$ formed by intersecting images of $\mathcal{R}_A$



Markov partitions induce probability distributions

- Physical/SRB measure μ for any HTA is just area (or volume)
- Given Markov partition $\mathcal{R} = \{R_j\}_{j=1}^n$, form $p_j := \mu(R_j)$
  - Note that $\beta/t_\infty = \|p\| \cdot \sqrt{\|\gamma\|^2 + 1}$ only depends on p
  - Insofar as β is independent of $\mathcal{R}$, so is *pointwise* E

Markov partitions induce probability distributions

- Physical/SRB measure μ for any HTA is just area (or volume)
- Given Markov partition $\mathcal{R} = \{R_j\}_{j=1}^n$, form $p_j := \mu(R_j)$
  - Note that $\beta/t_\infty = \|p\| \cdot \sqrt{\|\gamma\|^2 + 1}$ only depends on p
  - Insofar as β is independent of $\mathcal{R}$, so is *pointwise* E
- For p corresponding to $\mathcal{R}_m^\vee$, β/t_∞ converges to finite nonzero value for generic 2D HTAs
  - Key step is to count the number of rectangles in $\mathcal{R}_m^\vee$ contained in R_j and with given extents in stable direction

Markov partitions induce probability distributions

- Physical/SRB measure μ for any HTA is just area (or volume)
- Given Markov partition $\mathcal{R} = \{R_j\}_{j=1}^n$, form $p_j := \mu(R_j)$
  - Note that $\beta/t_\infty = \|p\| \cdot \sqrt{\|\gamma\|^2 + 1}$ only depends on p
  - Insofar as β is independent of $\mathcal{R}$, so is *pointwise E*
- For p corresponding to $\mathcal{R}_m^\vee$, β/t_∞ converges to finite nonzero value for generic 2D HTAs
  - Key step is to count the number of rectangles in $\mathcal{R}_m^\vee$ contained in R_j and with given extents in stable direction
- However, detailed calculations show $\lim \beta/t_\infty$ depends on $\mathcal{R}$
  - $\beta/t_\infty \approx 0.3463$ for $\mathcal{R}_A$ and $\mathcal{R}_A''$; $\beta/t_\infty \approx 0.4245$ for $\mathcal{R}_A'$

Any finite nonzero limit for β/t_∞ is already nontrivial

- $\lim \beta/t_\infty = \infty$ if we start with $\mathcal{Y}^{(0)} = [0,1]$ and form $\mathcal{Y}^{(m+1)}$ by subdividing each interval in $\mathcal{Y}^{(m)}$ into two subintervals of relative length q and $1-q$
  - $q = 1/2 \Rightarrow \lim \beta/t_\infty = 0$

Any finite nonzero limit for β/t_∞ is already nontrivial

- $\lim \beta/t_\infty = \infty$ if we start with $\mathcal{Y}^{(0)} = [0,1]$ and form $\mathcal{Y}^{(m+1)}$ by subdividing each interval in $\mathcal{Y}^{(m)}$ into two subintervals of relative length q and $1-q$
  - $q = 1/2 \Rightarrow \lim \beta/t_\infty = 0$
- As mentioned earlier, naive discretization of free particle/ideal gas has no obvious reasonable scaling limit
- 2D HTA limits indicate that while $\beta = K_n \cdot t_\infty \|p\| \sqrt{\|\gamma\|^2 + 1}$ initially appears OK, we should actually enforce $K_n \equiv const$
  - This is not at all obvious: taking $K_n = \sqrt{n}$ (so that $\beta(u)$ does not depend on n) naively appears to be more appropriate

Any finite nonzero limit for β/t_∞ is already nontrivial

- $\lim \beta/t_\infty = \infty$ if we start with $\mathcal{Y}^{(0)} = [0,1]$ and form $\mathcal{Y}^{(m+1)}$ by subdividing each interval in $\mathcal{Y}^{(m)}$ into two subintervals of relative length q and $1-q$
  - $q = 1/2 \Rightarrow \lim \beta/t_\infty = 0$
- As mentioned earlier, naive discretization of free particle/ideal gas has no obvious reasonable scaling limit
- 2D HTA limits indicate that while $\beta = K_n \cdot t_\infty \|p\| \sqrt{\|\gamma\|^2 + 1}$ initially appears OK, we should actually enforce $K_n \equiv const$
  - This is not at all obvious: taking $K_n = \sqrt{n}$ (so that $\beta(u)$ does not depend on n) naively appears to be more appropriate
- Two related issues with Markov partitions of the form $\mathcal{R}_m^\vee$
  - $\lim \beta/t_\infty$ depends on $\mathcal{R}$
  - Phase space volumes (to say nothing of physical measures) of rectangles vary increasingly more as m increases

Greedy refinements are physically natural

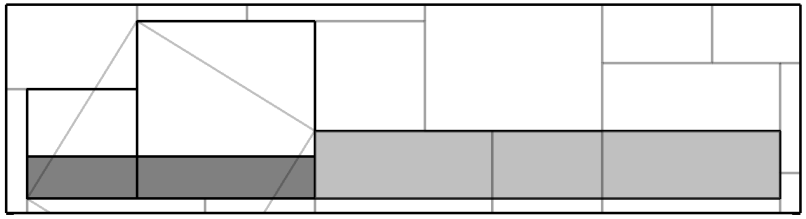
- Physical intuition suggests dealing with Markov partitions that have the most uniform possible phase space volumes
- Even for $\mu \neq$ phase space volume ν, this will tend to minimize β and maximize entropy/minimize effective free energy
  - First indication of a generalized variational principle
  - Can get finite limit for β even as entropy diverges

Greedy refinements are physically natural

- Physical intuition suggests dealing with Markov partitions that have the most uniform possible phase space volumes
- Even for $\mu \neq$ phase space volume ν, this will tend to minimize β and maximize entropy/minimize effective free energy
  - First indication of a generalized variational principle
  - Can get finite limit for β even as entropy diverges
- For a rectangle $R_j \in \mathcal{R}$ with $\nu(R_j)$ maximal, the intersection of TR_j with rectangles in $\mathcal{R}$ determines subrectangles of $T\mathcal{R}$ that in turn determine various refinements of $\mathcal{R}$ under T^{-1}
- We call such a refinement of maximal entropy w.r.t. ν *greedy*
  - Generally not unique

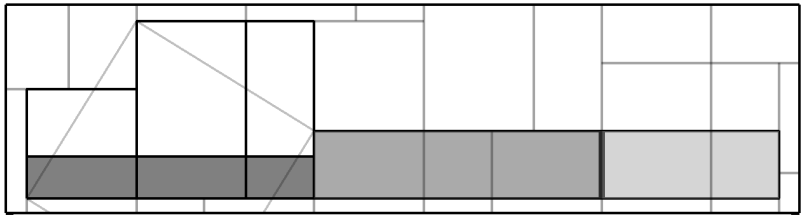
Greedy refinements of $\mathcal{R}''_A$ in eigencoordinates

$\mathcal{R}''_{A,0} \equiv \mathcal{R}''_A$



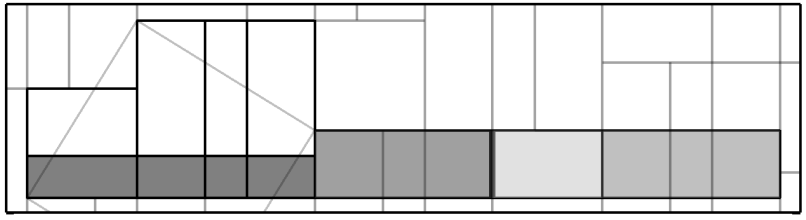
$\{(\nu/\nu_{\text{min}}, \text{multiplicity})\} = \{(\phi^2, 1), (1, 1)\}$

1 GR = 1 round $\mapsto \mathcal{R}''_{A,1}$



$\{(\nu/\nu_{\text{min}}, \text{multiplicity})\} = \{(\phi, 1), (1, 2)\}$

1 GR = 1 round $\mapsto \mathcal{R}''_{A,2}$



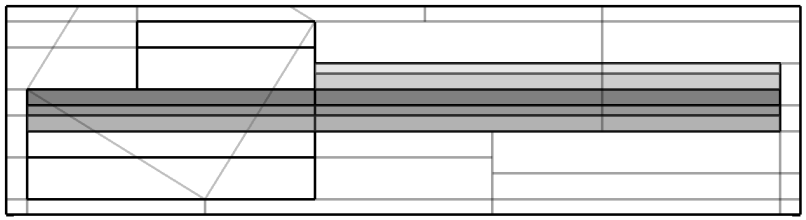
$\{(\phi, 3), (1, 1)\}$

3 GRs = 1 round $\mapsto \mathcal{R}''_{A,3}$

$\{(\phi, 4), (1, 3)\}$

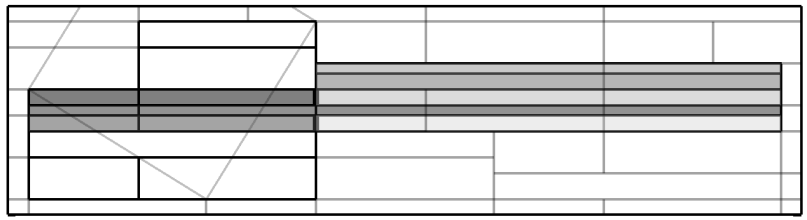
Greedy refinements of $\mathcal{R}_A$ and $\mathcal{R}'_A$ in eigencoordinates

$\mathcal{R}_{A,0} \equiv \mathcal{R}_A$



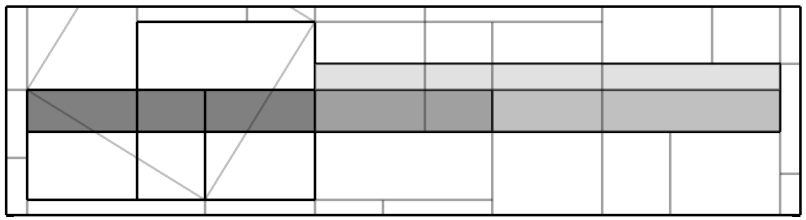
$\{(\nu/\nu_{\text{min}}, \text{multiplicity})\} = \{(\phi^2, 2), (\phi, 2), (1, 1)\}$

2 GRs $=$ 1 round $\mapsto \mathcal{R}_{A,1}$



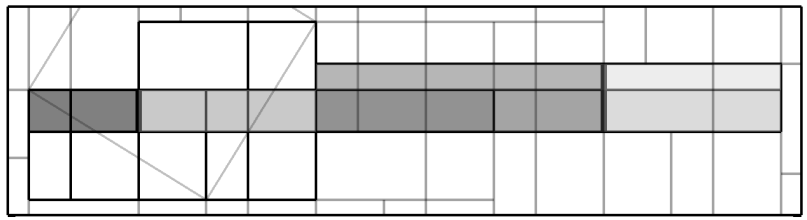
$\{(\nu/\nu_{\text{min}}, \text{multiplicity})\} = \{(\phi, 4), (1, 3)\}$

$\mathcal{R}'_{A,0} \equiv \mathcal{R}'_A$



$\{(\phi, 3), (1, 1)\}$

3 GRs $=$ 1 round $\mapsto \mathcal{R}'_{A,1}$



$\{(\phi, 4), (1, 3)\}$

BAE SYSTEMS

Greedy refinements stabilize rectangle measures

- For $m > 0$, both $\mathcal{R}_{A,m}$ and $\mathcal{R}'_{A,m}$ contain L_{m+1} and L_{m+2} rectangles of relative measure 1 and ϕ, respectively
  - *Lucas numbers* obey $L_{m+2} = L_{m+1} + L_m$ with $L_1 = 1$, $L_2 = 3$
- For $m > 1$, $\mathcal{R}''_{A,m}$ contains L_{m-1} and L_m rectangles of relative measure 1 and ϕ, respectively

Greedy refinements stabilize rectangle measures

- For $m > 0$, both $\mathcal{R}_{A,m}$ and $\mathcal{R}'_{A,m}$ contain L_{m+1} and L_{m+2} rectangles of relative measure 1 and ϕ, respectively
  - *Lucas numbers* obey $L_{m+2} = L_{m+1} + L_m$ with $L_1 = 1$, $L_2 = 3$
- For $m > 1$, $\mathcal{R}''_{A,m}$ contains L_{m-1} and L_m rectangles of relative measure 1 and ϕ, respectively
- Good reason to think that similar results hold more generally
  - E.g., the common limit of $\beta/t_\infty \approx 0.2393$ for all the cases above is apparently minimal/universal for the cat map
  - Even if this turned out not to hold in other cases, we could still take an extremum over Markov partitions with diminishing size

There are two archetypal Anosov flows

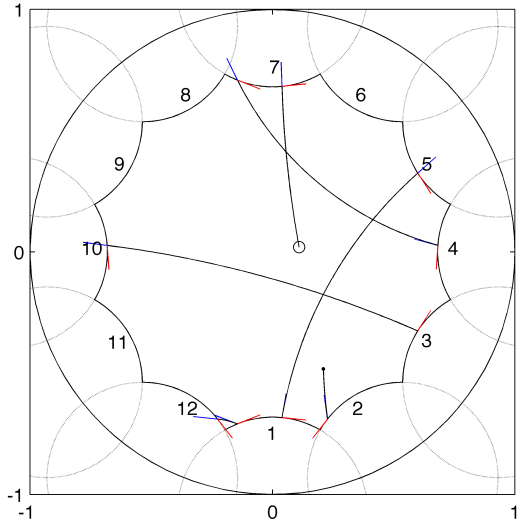
- "Suspension" of cat map generated by vector field e_z under twisted periodic boundary condition $(T_A x, z) \sim (x, z+1)$
  - The *cat flow* can be analyzed in a manner similar to that of the cat map, and we get exactly the same limiting behavior
  - However unlike the cat map, the cat flow is not mixing, so its utility as a model physical system is comparatively limited

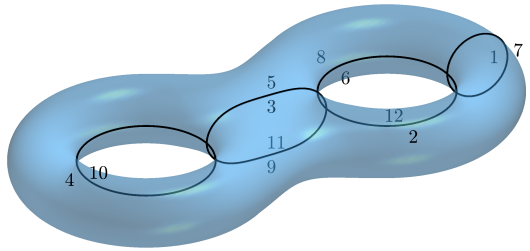
There are two archetypal Anosov flows

- "Suspension" of cat map generated by vector field e_z under twisted periodic boundary condition $(T_A x, z) \sim (x, z+1)$
  - The *cat flow* can be analyzed in a manner similar to that of the cat map, and we get exactly the same limiting behavior
  - However unlike the cat map, the cat flow is not mixing, so its utility as a model physical system is comparatively limited
- Geodesic flow on surface of constant negative curvature
  - Corresponds to free particle Hamiltonian $\mathcal{H} = \frac{1}{2m} \sum_{jk} g^{jk} P_j P_k$
  - Geodesic flow is mixing and will give apparently geometry-independent effective temperature of free particle

Geodesic flow in Poincaré disk model is tractable

- Differential arclength $ds = dr/(1 - r^2)$
- Geodesics correspond to circular arcs intersecting S^1 at right angles
- Surface of constant negative curvature obtained by identifying pairs of edges s_j of hyperbolic polygon such as shown in top figure via maps $T_j(s_j) = s_{\sigma(j)}^{-1}$
  - Here s_j^{-1} is orientation reversal of s_j
  - Pairing $\sigma(j)$ indicated in bottom figure
  - Note that the pairing is not "twisted"
  - $8g - 4$ edges $\Rightarrow$ genus $g = \#$ of holes
- Hamiltonian $\mathcal{H} = (1 - r^2)^2 \cdot P^2/2m$

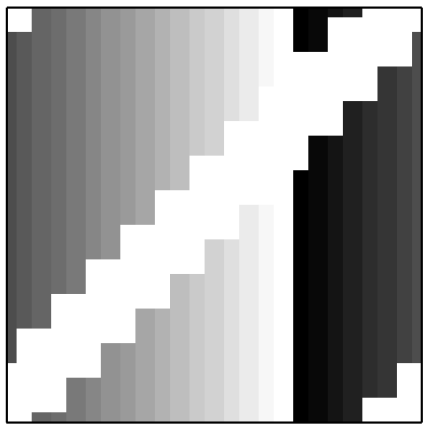
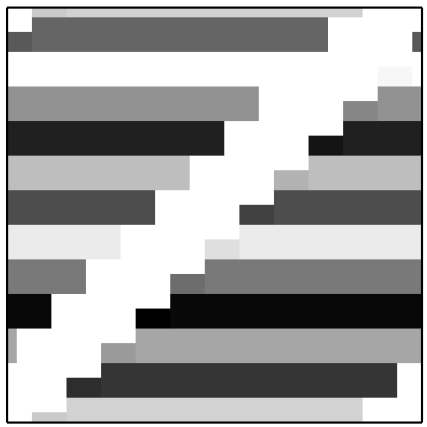
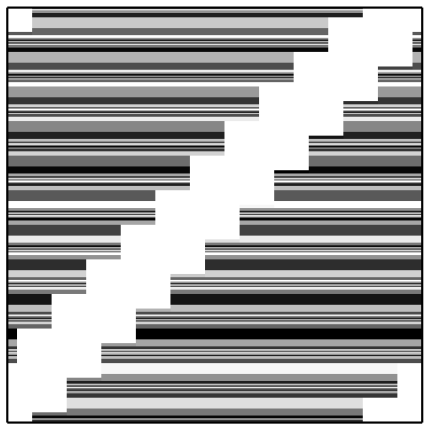




BAE SYSTEMS

Timing map and Markov partition for geodesic flow

- Following Adler and Weiss, we instantiate edge pairing maps T_j *en route* to $T_R = (\text{timing/Poincaré map}) \circ (\text{isometry})$
  - Isometry $\Rightarrow$ T_R is equivalent to timing map for our purposes
- We also instantiate a Markov partition $\mathcal{R}$ for T_R
  - $T_R^m \mathcal{R}$ for $g = 2$, $m = 0, 1, 2$: rectangles consistently shaded
  - Get $\mathcal{R}_m^\vee$ by intersecting rectangles in $T_R^0 \mathcal{R}, \ldots, T_R^m \mathcal{R}$

We numerically compute β for geodesic flow

- Unlike the cat map, T_R is highly nonlinear
- Rationale of T_R vs. timing map: "rectangles are rectangles"
- Although $\mu = \nu$ in this case (as with HTAs), it is nontrivial:

$$\mu([x_1, x_2] \times [y_1, y_2]) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} \frac{|dx\ dy|}{|e^{ix} - e^{iy}|^2}$$

- We exploit a few tricks to numerically compute measures of rectangles in refinements of $\mathcal{R}$

We numerically compute β for geodesic flow

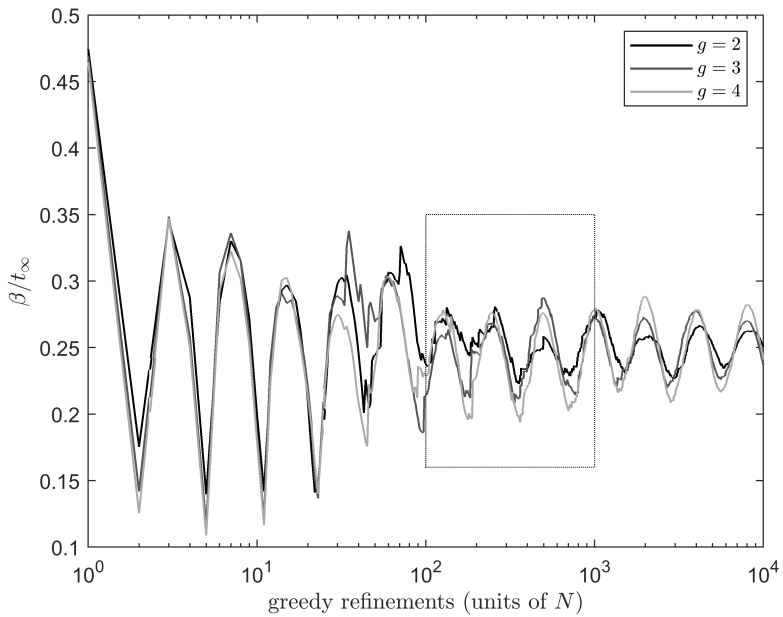
- Unlike the cat map, T_R is highly nonlinear
- Rationale of T_R vs. timing map: "rectangles are rectangles"
- Although $\mu = \nu$ in this case (as with HTAs), it is nontrivial:

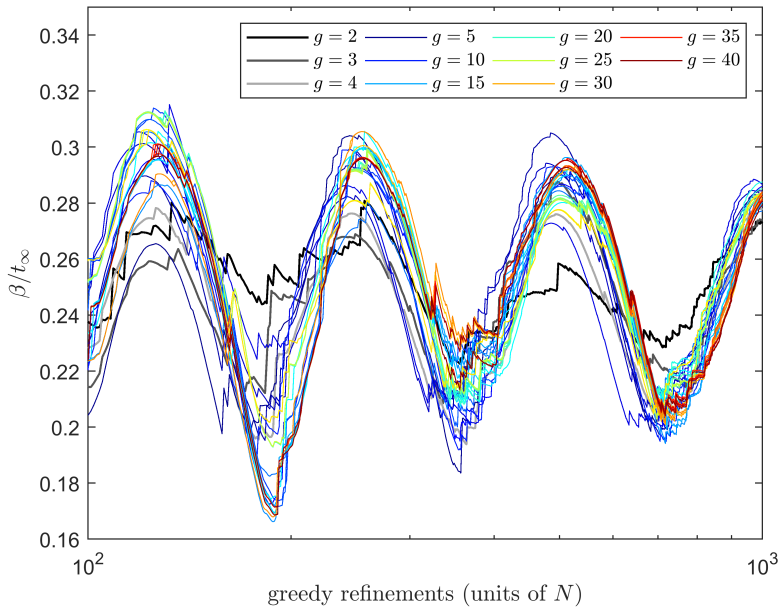
$$\mu([x_1, x_2] \times [y_1, y_2]) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} \frac{|dx\ dy|}{|e^{ix} - e^{iy}|^2}$$

- We exploit a few tricks to numerically compute measures of rectangles in refinements of $\mathcal{R}$
- First result: β/t_∞ diverges nearly exponentially for $\mathcal{R}_m^\vee$
  - Difference in behavior vs. T_A due to lack of linear structure

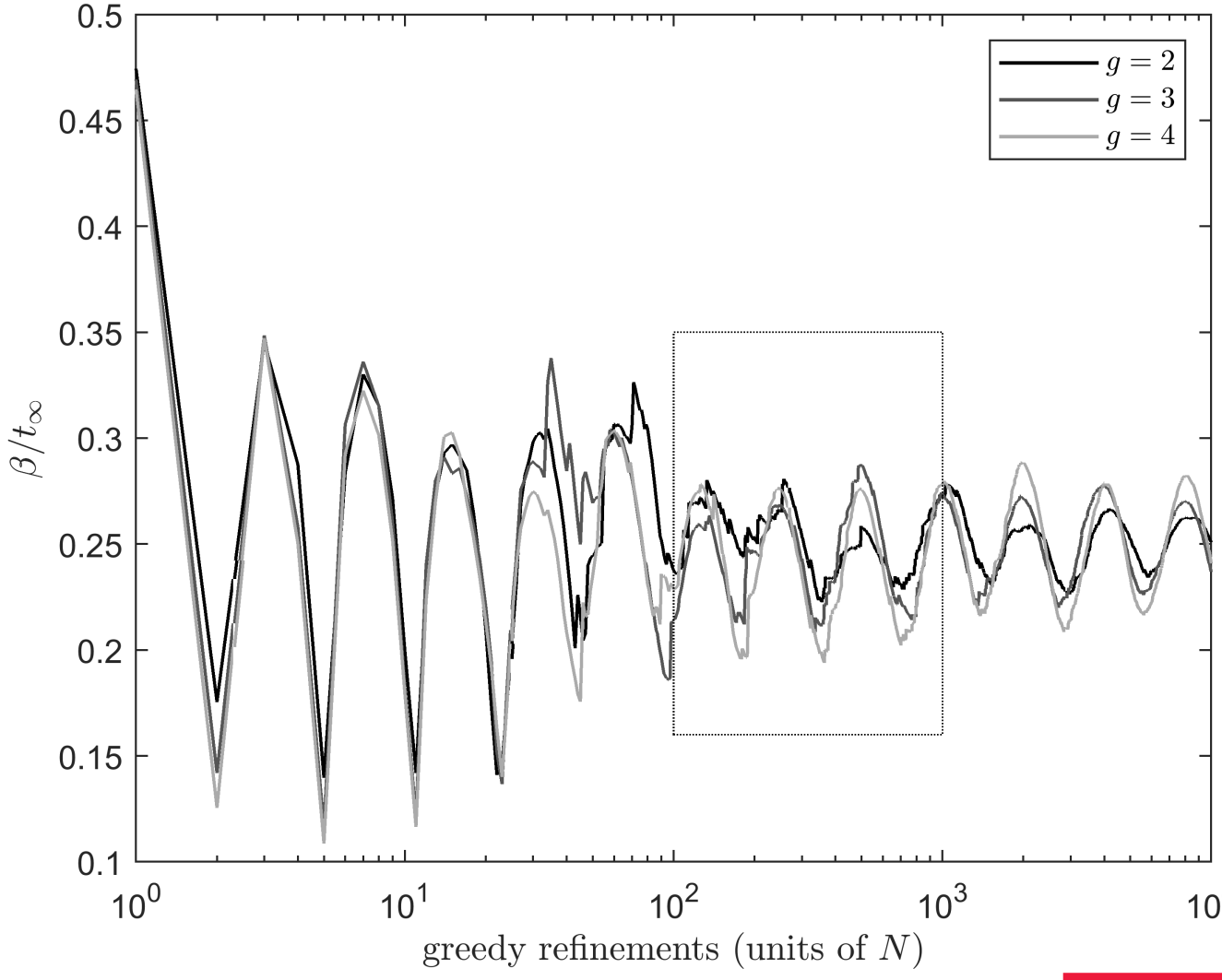
We numerically compute β for geodesic flow

- For greedy refinements, we have strong numerical evidence that $\lim \beta$ is nonzero, finite, and independent of genus g
  - Actually computing $\lim \beta / t_\infty$, but in fact mixing time $\equiv 1/2$
- Copies with different initial conditions give ideal gas
  - Weak coupling $\Rightarrow$ thermometer



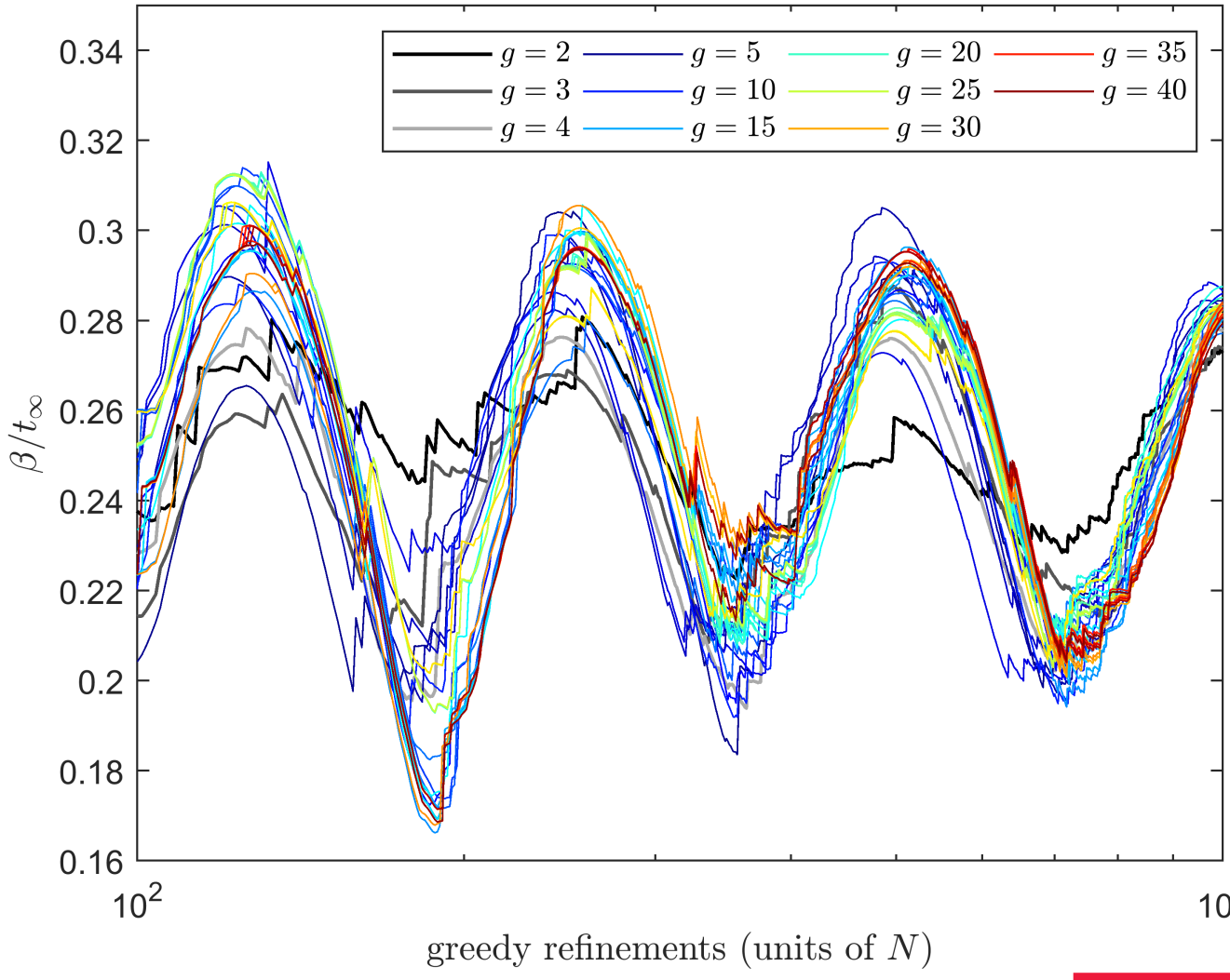


We numerically compute β for geodesic flow

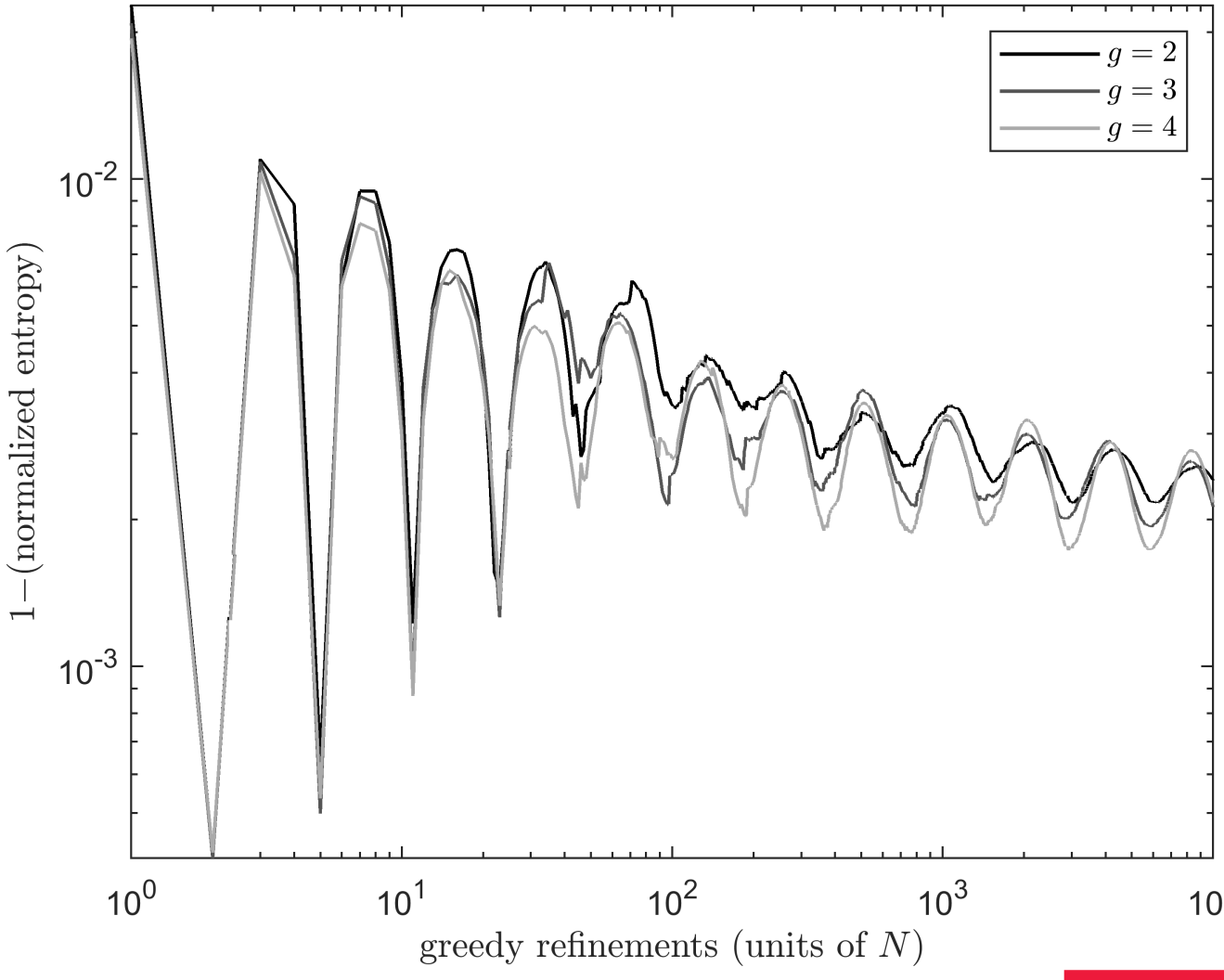


BAE SYSTEMS

We numerically compute β for geodesic flow



We numerically compute entropy for geodesic flow

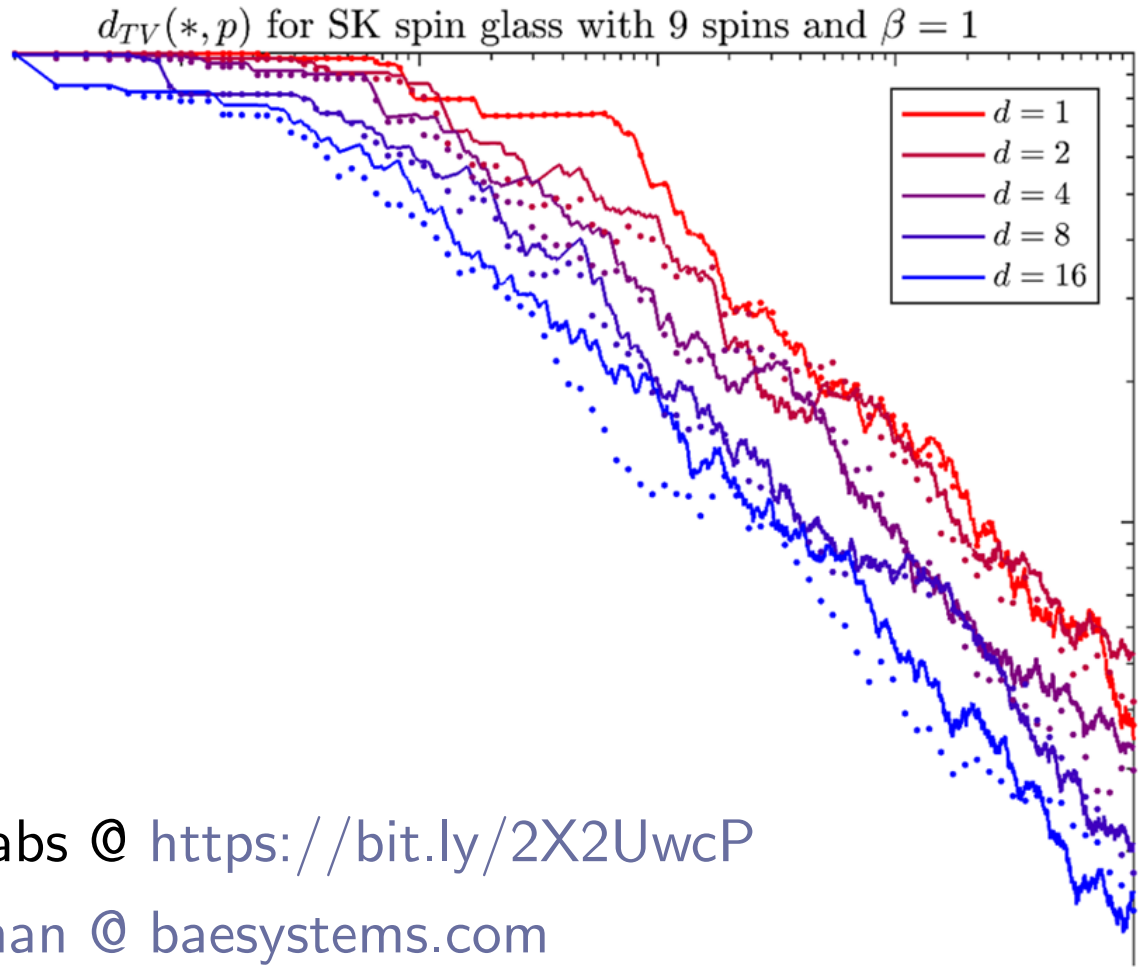


BAE SYSTEMS

A conjecture for nonequilibrium statistical physics

- Intrinsic β/t_∞ via (extremal limit over?) greedy refinements
- t_∞ similar (but not identical) to mixing time
- Conjecture: the effective temperature for classical steady-state systems satisfying the chaotic hypothesis is well-defined and is equivalent to physical temperature
  - This would extend Ruelle's thermodynamical formalism to a more complete theory of statistical physics for nonequilibrium steady states in which not only entropy production rates but also temperature and energy could be meaningfully interpreted
- Some major obstructions to any proof
  - The simplicial complex of Markov partitions seems complicated
  - Hard to develop nonlinear estimates, spectral techniques, etc.

Thanks



BAE Systems FAST Labs @ https://bit.ly/2X2UwcP

steve.huntsman @ baesystems.com

MCMC paper @ https://arxiv.org/abs/1901.08606

Anosov paper @ https://arxiv.org/abs/1009.2127

BAE SYSTEMS

Physical background for Anosov systems

- Both Anosov and Markov systems obey a *fluctuation theorem*
- Generic form of FT: $\mathbb{P}\left(t^{-1}\Sigma_t = z\right) = e^{tz} \cdot \mathbb{P}\left(t^{-1}\Sigma_t = -z\right)$
  - $t^{-1}\Sigma_t$ is trajectory's mean entropy production rate to time t
- Generalizes Onsager and Green-Kubo relations linking fluxes and transport coefficients; gives 2nd law behavior
- Vast majority of work on the chaotic hypothesis concerned with things like entropy production rate, FT, etc.
  - Precise meaning of "Anosov-like" in hypothesis not yet known
- We explore the chaotic hypothesis in an entirely different direction, providing evidence that reasonable notions of temperature and energy are defined for systems that obey it

Comparison with other approaches

- We do everything in terms of time data alone
  - That said, physical measures still provide a dynamical basis
  - Take seriously: "there is no conceptual difference between stationary states in equilibrium and out of equilibrium"
- In another approach, dynamical rates of expansion/contraction "provide an 'energy function' that assigns relative probabilistic weights to the coarse grained cells"
  - Define an effective temperature of a thermostat by $\dot{W}/\dot{\Sigma}$, where $\dot{W}$ is the work rate of external forces on the system and $\dot{\Sigma}$ is the entropy production rate
  - This still requires *a priori* knowledge of an energy function of some sort in order to define a sensible notion of work rate–but we don't need anything like that in our approach

Towards generalization of these results

- Markov partitions, physical measures, etc. exhibit great regularity w.r.t. small perturbations of dynamics
- Although perturbed and especially weakly coupled lattices of cat maps or perturbed geodesic flows are not easily treated explicitly, they still behave nicely
  - Hence while our explicit examples deal with "microcanonical" ensembles, our results generalize to related systems in what amount to both canonical and nonequilibrium ensembles
  - Relevant ideas: thermostats, weakly coupled map lattices, etc.
  - Only case of obvious desired generalization with elusive approach is coupled geodesic flows (read: interacting gas)