CS 263: Counting and Sampling

Nima Anari

Stanford University

slides for

Stochastic Localization

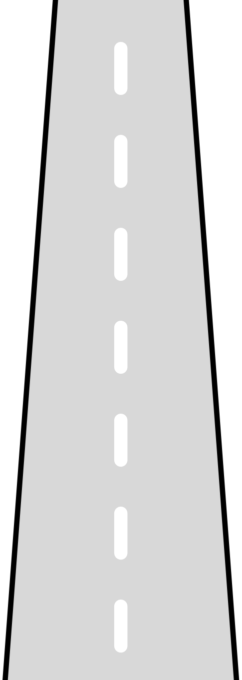
Skipped . . .

Stochastic Calculus

- Localization schemes
- Itô calculus
- Stochastic localization

Conservation

- Sherrington-Kirkpatrick model
- ϕ-entropies in localization scheme
- Approximate conservation

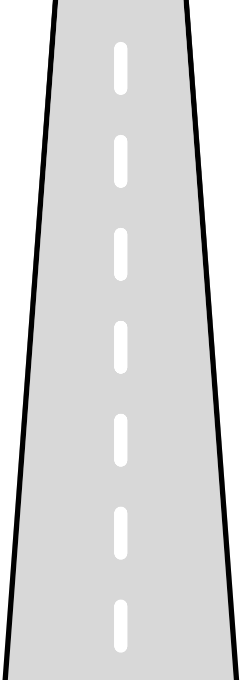


Stochastic Calculus

- Localization schemes
- Itô calculus
- Stochastic localization

Conservation

- Sherrington-Kirkpatrick model
- ϕ-entropies in localization scheme
- Approximate conservation



Simplicial localization

- Imagine μ is on $\binom{[n]}{k} \hookrightarrow \{0,1\}^n$.

Simplicial localization

- Imagine μ is on $\binom{[n]}{k} \hookrightarrow \{0,1\}^n$.
- Denote $p_i = \mathbb{P}_{S\sim\mu}[i \in S]$. Let us choose $i \sim \mu D_{k\to 1} = p/k$.

Simplicial localization

- Imagine μ is on $\binom{[n]}{k} \hookrightarrow \{0,1\}^n$.
- Denote $p_i = \mathbb{P}_{S\sim\mu}[i \in S]$. Let us choose $i \sim \mu D_{k\to 1} = p/k$.
- Let ν be the conditional on $\{i\}$. For $w = \mathbb{1}_i/p_i - \mathbb{1}/k$:

a random measure

$$\nu(x) = \underbrace{(1 + \langle w, x - \mathrm{mean}(\mu)\rangle)}_{\text{linear tilt}}\,\mu(x)$$

Simplicial localization

- Imagine μ is on $\binom{[n]}{k} \hookrightarrow \{0,1\}^n$.
- Denote $p_i = \mathbb{P}_{S\sim\mu}[i \in S]$. Let us choose $i \sim \mu D_{k\to 1} = p/k$.
- Let ν be the conditional on $\{i\}$. For $w = \mathbb{1}_i/p_i - \mathbb{1}/k$:

a random measure

$$\nu(x) = \underbrace{\left(1 + \langle w, x - \text{mean}(\mu)\rangle\right)}_{\text{linear tilt}} \mu(x)$$

- Note that $\mu = \mathbb{E}_i[\nu]$. This is a decomposition of measure.

Simplicial localization

- Imagine μ is on $\binom{[n]}{k} \hookrightarrow \{0,1\}^n$.
- Denote $p_i = \mathbb{P}_{S\sim\mu}[i \in S]$. Let us choose $i \sim \mu D_{k\to 1} = p/k$.
- Let ν be the conditional on $\{i\}$. For $w = \mathbb{1}_i/p_i - \mathbb{1}/k$:

a random measure

$$\nu(x) = \underbrace{(1 + \langle w, x - \text{mean}(\mu)\rangle)}_{\text{linear tilt}}\,\mu(x)$$

- Note that $\mu = \mathbb{E}_i[\nu]$. This is a decomposition of measure.
- Continuing this we get a measure-valued random process: martingale

Simplicial localization

- Imagine μ is on $\binom{[n]}{k} \hookrightarrow \{0,1\}^n$.
- Denote $p_i = \mathbb{P}_{S\sim\mu}[i \in S]$. Let us choose $i \sim \mu D_{k\to 1} = p/k$.
- Let ν be the conditional on $\{i\}$. For $w = \mathbb{1}_i/p_i - \mathbb{1}/k$:

a random measure

$$\nu(x) = \underbrace{(1 + \langle w, x - \text{mean}(\mu)\rangle)}_{\text{linear tilt}}\,\mu(x)$$

- Note that $\mu = \mathbb{E}_i[\nu]$. This is a decomposition of measure.
- Continuing this we get a measure-valued random process: ← martingale

Simplicial localization

Let $S \sim \mu$, and let $e_1, \dots, e_k$ be a u.r. permutation of S. Define μ_i as conditional of μ on $\{e_1, \dots, e_i\}$. Then

$$\mu = \mu_0 \to \mu_1 \to \mu_2 \to \cdots \to \mu_k$$

is called simplicial localization. ← used for local-to-global and trickle down

Stochastic localization

- Same idea applied in continuous time. For some measure μ on $\mathbb{R}^n$, we get measure-valued process $\{\mu_t \mid t \in \mathbb{R}_{\geqslant 0}\}$.

Stochastic localization

- Same idea applied in continuous time. For some measure μ on $\mathbb{R}^n$, we get measure-valued process $\{\mu_t \mid t \in \mathbb{R}_{\geqslant 0}\}$.
- Controlled by (stochastic) differential equation

$$d\mu_t(x) = \underbrace{\langle w_t, x - \text{mean}(\mu)\rangle}_{\text{linear tilt}} \mu_t(x)$$

where now w_t is a mean zero random infinitesimal vector.

think of infinitesimal Gaussian

Stochastic localization

- Same idea applied in continuous time. For some measure μ on $\mathbb{R}^n$, we get measure-valued process $\{\mu_t \mid t \in \mathbb{R}_{\geqslant 0}\}$.
- Controlled by (stochastic) differential equation

$$d\mu_t(x) = \underbrace{\langle w_t, x - \mathrm{mean}(\mu)\rangle}_{\text{linear tilt}} \mu_t(x)$$

where now w_t is a mean zero random infinitesimal vector.

$\uparrow$

think of infinitesimal Gaussian

- Our goal will be to find analogs of local-to-global, etc. for more general, e.g., continuous, distributions.

Stochastic localization

- Same idea applied in continuous time. For some measure μ on $\mathbb{R}^n$, we get measure-valued process $\{\mu_t \mid t \in \mathbb{R}_{\geqslant 0}\}$.
- Controlled by (stochastic) differential equation

$$d\mu_t(x) = \underbrace{\langle w_t, x - \text{mean}(\mu)\rangle}_{\text{linear tilt}} \mu_t(x)$$

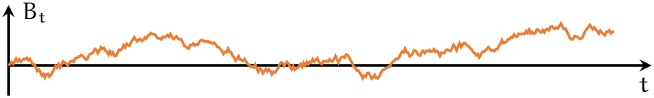
  where now w_t is a mean zero random infinitesimal vector.

  think of infinitesimal Gaussian

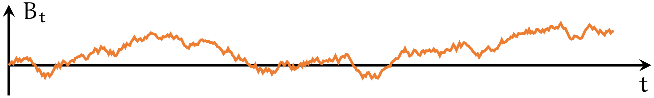
- Our goal will be to find analogs of local-to-global, etc. for more general, e.g., continuous, distributions.
- To make sense of this equation, we need some basics of Itô calculus.

Intro to Itô calculus

- Brownian motion: in nD, the process $\{B_t \mid t \in \mathbb{R}^n_{\geqslant 0}\}$ such that
$$B_t - B_s \sim \mathcal{N}(0, (t-s)I)$$
and for disjoint $[s_1, t_1], \ldots, [s_k, t_k]$ we have $B_{t_i} - B_{s_i}$ are independent.



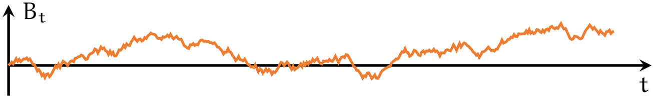
- Brownian motion: in nD, the process $\{B_t \mid t \in \mathbb{R}^n_{\geq 0}\}$ such that
$$B_t - B_s \sim \mathcal{N}(0, (t-s)I)$$
and for disjoint $[s_1, t_1], \ldots, [s_k, t_k]$ we have $B_{t_i} - B_{s_i}$ are independent.



- We think of dB_t intuitively as $B_{t+dt} - B_t$: $dB_t \sim \mathcal{N}(0, dt \cdot I)$

Intro to Itô calculus

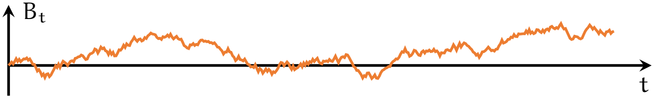
- Brownian motion: in nD, the process $\{B_t \mid t \in \mathbb{R}^n_{\geqslant 0}\}$ such that
$$B_t - B_s \sim \mathcal{N}(0, (t-s)I)$$
and for disjoint $[s_1, t_1], \ldots, [s_k, t_k]$ we have $B_{t_i} - B_{s_i}$ are independent.



- We think of dB_t intuitively as $B_{t+dt} - B_t$: $dB_t \sim \mathcal{N}(0, dt \cdot I)$
- Fact: dB_t is not on the order of dt, but rather on the order of $\sqrt{dt}$!

Intro to Itô calculus

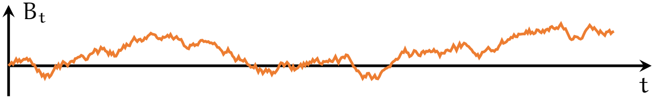
- Brownian motion: in nD, the process $\{B_t \mid t \in \mathbb{R}^n_{\geqslant 0}\}$ such that
$$B_t - B_s \sim \mathcal{N}(0, (t-s)I)$$
and for disjoint $[s_1, t_1], \ldots, [s_k, t_k]$ we have $B_{t_i} - B_{s_i}$ are independent.



- We think of dB_t intuitively as $B_{t+dt} - B_t$: $dB_t \sim \mathcal{N}(0, dt \cdot I)$
- Fact: dB_t is not on the order of dt, but rather on the order of $\sqrt{dt}$!
- Itô process: $\{X_t \mid t \in \mathbb{R}_{\geqslant 0}\}$ derived via stochastic differential equation (SDE):
$$dX_t = u_t dt + C_t dB_t$$
for some "nice" vector and matrix valued processes $\{u_t\}, \{C_t\}$.

Intro to Itô calculus

- Brownian motion: in nD, the process $\{B_t \mid t \in \mathbb{R}^n_{\geqslant 0}\}$ such that
$$B_t - B_s \sim \mathcal{N}(0, (t-s)I)$$
and for disjoint $[s_1, t_1], \ldots, [s_k, t_k]$ we have $B_{t_i} - B_{s_i}$ are independent.



- We think of dB_t intuitively as $B_{t+dt} - B_t$: $dB_t \sim \mathcal{N}(0, dt \cdot I)$
- Fact: dB_t is not on the order of dt, but rather on the order of $\sqrt{dt}$!
- Itô process: $\{X_t \mid t \in \mathbb{R}_{\geqslant 0}\}$ derived via stochastic differential equation (SDE):
$$dX_t = u_t dt + C_t dB_t$$
for some "nice" vector and matrix valued processes $\{u_t\}, \{C_t\}$.
- u_t, C_t can only depend on the past; technical term: adapted.

Itô formula

- Basic question: if we have 1D Itô process X_t defined by

$$dX_t = u_t dt + c_t dB_t$$

and define $Y_t = f(X_t)$, what is the equation defining Y_t?

Itô formula

- Basic question: if we have 1D Itô process X_t defined by
$$dX_t = u_t dt + c_t dB_t$$
and define $Y_t = f(X_t)$, what is the equation defining Y_t?
- Incorrect: if we apply chain rule of calculus, we get
$$dY_t = f'(X_t)dX_t = f'(X_t)u_t dt + f'(X_t)c_t dB_t$$

Itô formula

- Basic question: if we have 1D Itô process X_t defined by
$$dX_t = u_t dt + c_t dB_t$$
and define $Y_t = f(X_t)$, what is the equation defining Y_t?
- Incorrect: if we apply chain rule of calculus, we get
$$dY_t = f'(X_t)dX_t = f'(X_t)u_t dt + f'(X_t)c_t dB_t$$
- This is incorrect because $dY_t = f'(X_t)dX_t$ is only first-order approximation of f, and dX_t has terms of order $\sqrt{dt}$!

Itô formula

- Basic question: if we have 1D Itô process X_t defined by
$$dX_t = u_t dt + c_t dB_t$$
and define $Y_t = f(X_t)$, what is the equation defining Y_t?
- Incorrect: if we apply chain rule of calculus, we get
$$dY_t = f'(X_t)dX_t = f'(X_t)u_t dt + f'(X_t)c_t dB_t$$
- This is incorrect because $dY_t = f'(X_t)dX_t$ is only first-order approximation of f, and dX_t has terms of order $\sqrt{dt}$!
- Correction: expand up to second-order Taylor series, and use $dB_t^2 = dt$, also drop anything of lower order than dt.

Itô formula

- Basic question: if we have 1D Itô process X_t defined by

$$dX_t = u_t dt + c_t dB_t$$

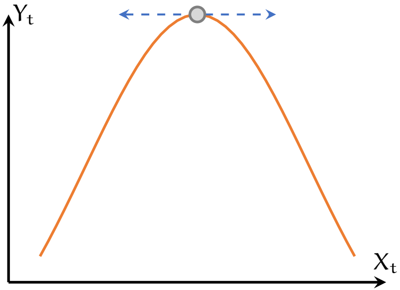
  and define $Y_t = f(X_t)$, what is the equation defining Y_t?
- Incorrect: if we apply chain rule of calculus, we get

$$dY_t = f'(X_t)dX_t = f'(X_t)u_t dt + f'(X_t)c_t dB_t$$

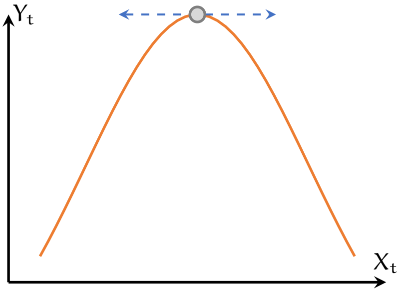
- This is incorrect because $dY_t = f'(X_t)dX_t$ is only first-order approximation of f, and dX_t has terms of order $\sqrt{dt}$!
- Correction: expand up to second-order Taylor series, and use $dB_t^2 = dt$, also drop anything of lower order than dt.
- This gives us the Itô formula:

$$dY_t = \left(f'(X_t)u_t + \underbrace{\frac{1}{2}f''(X_t)c_t^2}_{\text{Itô term}}\right)dt + f'(X_t)c_t dB_t$$

Intuition: curvature creates drift!



Intuition: curvature creates drift!



Itô's lemma (nD to 1D)

For $dX_t = u_t dt + C_t dB_t$ if we have $Y_t = f(X_t)$, then

$$dY_t = \left(\langle \nabla f(X_t), u_t \rangle + \frac{1}{2} \operatorname{tr}(C_t^\intercal \nabla^2 f(X_t) C_t) \right) dt + \langle \nabla f(X_t), C_t dB_t \rangle.$$

Stochastic localization

For μ on subset of $\mathbb{R}^n$, and adapted matrix process C_t, we define $\forall x$

$$d\mu_t(x) = \langle x - \mathrm{mean}(\mu_t), C_t dB_t \rangle \mu_t(x)$$

Stochastic localization

For μ on subset of $\mathbb{R}^n$, and adapted matrix process C_t, we define $\forall x$

$$d\mu_t(x) = \langle x - \mathsf{mean}(\mu_t), C_t dB_t \rangle \mu_t(x)$$

- For continuous μ, we should think of it as density. You can for simplicity assume support is finite.

Stochastic localization

For μ on subset of $\mathbb{R}^n$, and adapted matrix process C_t, we define $\forall x$

$$d\mu_t(x) = \langle x - \mathrm{mean}(\mu_t), C_t dB_t \rangle \mu_t(x)$$

- For continuous μ, we should think of it as density. You can for simplicity assume support is finite.
- It is a martingale, with filtration $\mathcal{F}_t$:

$$\mathbb{E}[\mu_t(x) \mid \mathcal{F}_s] = \mu_s(x) \quad \forall s \leqslant t$$

Stochastic localization

For μ on subset of $\mathbb{R}^n$, and adapted matrix process C_t, we define $\forall x$

$$d\mu_t(x) = \langle x - \mathrm{mean}(\mu_t), C_t dB_t \rangle \mu_t(x)$$

- For continuous μ, we should think of it as density. You can for simplicity assume support is finite.
- It is a martingale, with filtration $\mathcal{F}_t$:
$$\mathbb{E}[\mu_t(x) \mid \mathcal{F}_s] = \mu_s(x) \quad \forall s \leqslant t$$
- If μ is normalized, μ_t remains so:
$$d(\textstyle\sum_x \mu_t(x)) = \langle \sum_x \mu_t(x)(x - \mathrm{mean}(\mu_t)), C_t dB_t \rangle = 0$$

Stochastic localization

For μ on subset of $\mathbb{R}^n$, and adapted matrix process C_t, we define $\forall x$

$$d\mu_t(x) = \langle x - \mathrm{mean}(\mu_t), C_t dB_t \rangle \mu_t(x)$$

- For continuous μ, we should think of it as density. You can for simplicity assume support is finite.
- It is a martingale, with filtration $\mathcal{F}_t$:

$$\mathbb{E}[\mu_t(x) \mid \mathcal{F}_s] = \mu_s(x) \quad \forall s \leqslant t$$

- If μ is normalized, μ_t remains so:

$$d(\textstyle\sum_x \mu_t(x)) = \langle \textstyle\sum_x \mu_t(x)(x - \mathrm{mean}(\mu_t)), C_t dB_t \rangle = 0$$

- Changes in μ_t are proportional to itself. Log-scale? Let's use Itô's lemma for $f = \log$.

Stochastic localization

For μ on subset of $\mathbb{R}^n$, and adapted matrix process C_t, we define $\forall x$

$$d\mu_t(x) = \langle x - \text{mean}(\mu_t), C_t dB_t \rangle \mu_t(x)$$

- For continuous μ, we should think of it as density. You can for simplicity assume support is finite.
- It is a martingale, with filtration $\mathcal{F}_t$:

$$\mathbb{E}[\mu_t(x) \mid \mathcal{F}_s] = \mu_s(x) \quad \forall s \leqslant t$$

- If μ is normalized, μ_t remains so:

$$d(\textstyle\sum_x \mu_t(x)) = \langle \textstyle\sum_x \mu_t(x)(x - \text{mean}(\mu_t)), C_t dB_t \rangle = 0$$

- Changes in μ_t are proportional to itself. Log-scale? Let's use Itô's lemma for $f = \log$.
- If $X_t = \mu_t(x)$, and $Y_t = \log(X_t)$, then $dY_t =$

$$\langle x - \text{mean}(\mu_t), C_t dB_t \rangle + (\text{Itô term})dt$$

where Itô term is

$$\frac{-(x - \text{mean}(\mu_t))^\intercal C_t C_t^\intercal (x - \text{mean}(\mu_t)) \cdot X_t^2}{2X_t^2}$$

Stochastic localization

For μ on subset of $\mathbb{R}^n$, and adapted matrix process C_t, we define $\forall x$

$$d\mu_t(x) = \langle x - \mathrm{mean}(\mu_t), C_t dB_t \rangle \mu_t(x)$$

- For continuous μ, we should think of it as density. You can for simplicity assume support is finite.
- It is a martingale, with filtration $\mathcal{F}_t$:

$$\mathbb{E}[\mu_t(x) \mid \mathcal{F}_s] = \mu_s(x) \quad \forall s \leqslant t$$

- If μ is normalized, μ_t remains so:

$$d(\textstyle\sum_x \mu_t(x)) = \langle \textstyle\sum_x \mu_t(x)(x - \mathrm{mean}(\mu_t)), C_t dB_t \rangle = 0$$

- Changes in μ_t are proportional to itself. Log-scale? Let's use Itô's lemma for $f = \log$.
- If $X_t = \mu_t(x)$, and $Y_t = \log(X_t)$, then $dY_t =$

$$\langle x - \mathrm{mean}(\mu_t), C_t dB_t \rangle + (\text{Itô term}) dt$$

where Itô term is

$$\frac{-(x - \mathrm{mean}(\mu_t))^\intercal C_t C_t^\intercal (x - \mathrm{mean}(\mu_t)) \cdot X_t^2}{2X_t^2}$$

- So if we name $\Sigma_t = C_t C_t^\intercal$, then

$$d \log \mu_t(x) = -\tfrac{1}{2} x^\intercal \Sigma_t x dt + \mathrm{affine}(x)$$

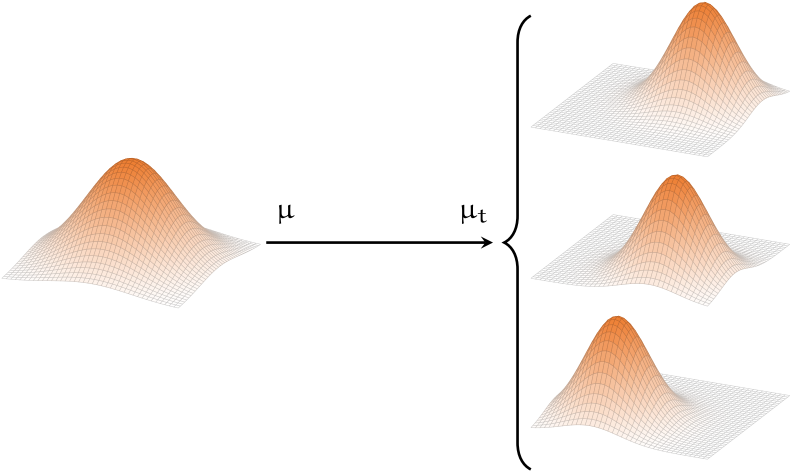
Stochastic localization

For μ on subset of $\mathbb{R}^n$, and adapted matrix process C_t, we define $\forall x$

$$d\mu_t(x) = \langle x - \mathrm{mean}(\mu_t), C_t dB_t \rangle \mu_t(x)$$

- For continuous μ, we should think of it as density. You can for simplicity assume support is finite.
- It is a martingale, with filtration $\mathcal{F}_t$:
$$\mathbb{E}[\mu_t(x) \mid \mathcal{F}_s] = \mu_s(x) \quad \forall s \leqslant t$$
- If μ is normalized, μ_t remains so:
$$d(\textstyle\sum_x \mu_t(x)) = \langle \textstyle\sum_x \mu_t(x)(x - \mathrm{mean}(\mu_t)), C_t dB_t \rangle = 0$$
- Changes in μ_t are proportional to itself. Log-scale? Let's use Itô's lemma for $f = \log$.
- If $X_t = \mu_t(x)$, and $Y_t = \log(X_t)$, then $dY_t =$
$$\langle x - \mathrm{mean}(\mu_t), C_t dB_t \rangle + (\text{Itô term})dt$$
where Itô term is
$$\frac{-(x - \mathrm{mean}(\mu_t))^\intercal C_t C_t^\intercal (x - \mathrm{mean}(\mu_t)) \cdot X_t^2}{2X_t^2}$$
- So if we name $\Sigma_t = C_t C_t^\intercal$, then
$$d \log \mu_t(x) = -\tfrac{1}{2} x^\intercal \Sigma_t x dt + \mathrm{affine}(x)$$
- At any time t, we have $\mu_t(x) \propto$
$$\mu(x) \cdot \exp\left(-\tfrac{1}{2} x^\intercal A_t x + \langle h_t, x \rangle\right)$$
where $A_t = \int_0^t \Sigma_s ds$.

Multiplying by Gaussian density:



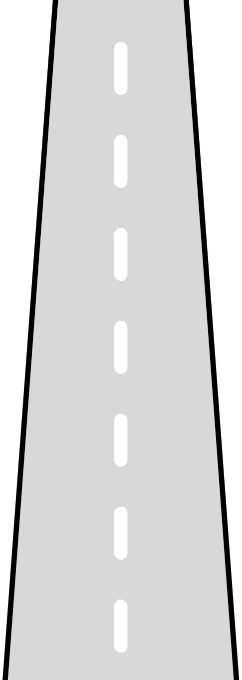
Remark: this process, up to scale/time, same as how diffusion models sample.

Stochastic Calculus

- Localization schemes
- Itô calculus
- Stochastic localization

Conservation

- Sherrington-Kirkpatrick model
- ϕ-entropies in localization scheme
- Approximate conservation

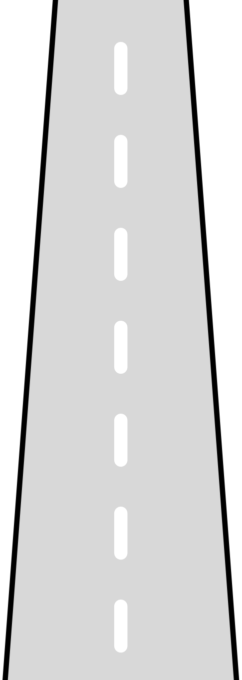


Stochastic Calculus

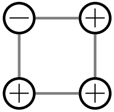
- Localization schemes
- Itô calculus
- Stochastic localization

Conservation

- Sherrington-Kirkpatrick model
- ϕ-entropies in localization scheme
- Approximate conservation



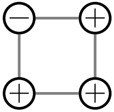
Ising models



$$\mu(x) \propto \exp\left(\frac{1}{2} \sum_{u,v} J_{uv} x_u x_v + \sum_v h_v x_v\right)$$

symmetric matrix

Ising models

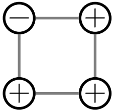


$$\mu(x) \propto \exp\left(\frac{1}{2}\sum_{u,v} J_{uv} x_u x_v + \sum_v h_v x_v\right)$$

symmetric matrix

- Dobrushin: when J has row/col ℓ_1 norms < 1, we get fast mixing.

Ising models

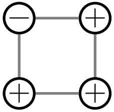


$$\mu(x) \propto \exp\left(\tfrac{1}{2}\textstyle\sum_{u,v} J_{uv} x_u x_v + \sum_v h_v x_v\right)$$

symmetric matrix

- Dobrushin: when J has row/col ℓ_1 norms < 1, we get fast mixing.
- Sherrington-Kirkpatrick model: random Gaussian matrix J with $J_{uv} \sim \mathcal{N}(0, \beta/n)$.

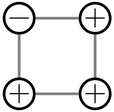
Ising models



$$\mu(x) \propto \exp\Big(\frac{1}{2}\sum_{u,v} \underset{\substack{\uparrow \\ \text{symmetric matrix}}}{J_{uv}} x_u x_v + \sum_v h_v x_v\Big)$$

- Dobrushin: when J has row/col ℓ_1 norms < 1, we get fast mixing.
- Sherrington-Kirkpatrick model: random Gaussian matrix J with $J_{uv} \sim \mathcal{N}(0, \beta/n)$.
- Open: find the exact threshold β where Glauber mixes fast w.h.p.

Ising models



$$\mu(x) \propto \exp\left(\frac{1}{2}\sum_{u,v} J_{uv} x_u x_v + \sum_v h_v x_v\right)$$

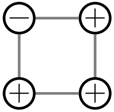
symmetric matrix

- Dobrushin: when J has row/col ℓ_1 norms < 1, we get fast mixing.
- Sherrington-Kirkpatrick model: random Gaussian matrix J with $J_{uv} \sim \mathcal{N}(0, \beta/n)$.
- Open: find the exact threshold β where Glauber mixes fast w.h.p.

- Dobrushin gives weak bound:

  $\beta \leqslant \Theta(1/n) \implies$ fast mixing

Ising models



$$\mu(x) \propto \exp\left(\tfrac{1}{2}\sum_{u,v} J_{uv} x_u x_v + \sum_v h_v x_v\right)$$

symmetric matrix

- Dobrushin: when J has row/col ℓ_1 norms < 1, we get fast mixing.
- Sherrington-Kirkpatrick model: random Gaussian matrix J with $J_{uv} \sim \mathcal{N}(0, \beta/n)$.
- Open: find the exact threshold β where Glauber mixes fast w.h.p.

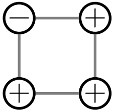
- Dobrushin gives weak bound:

$$\beta \leqslant \Theta(1/n) \implies \text{fast mixing}$$

- [Eldan-Koehler-Zeitouni] got

$$\beta \leqslant \Theta(1) \implies \text{fast mixing}$$

Ising models



$$\mu(x) \propto \exp\left(\tfrac{1}{2}\sum_{u,v} J_{uv} x_u x_v + \sum_v h_v x_v\right)$$

symmetric matrix

- Dobrushin: when J has row/col ℓ_1 norms < 1, we get fast mixing.
- Sherrington-Kirkpatrick model: random Gaussian matrix J with $J_{uv} \sim \mathcal{N}(0, \beta/n)$.
- Open: find the exact threshold β where Glauber mixes fast w.h.p.

- Dobrushin gives weak bound:

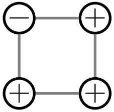
$$\beta \leqslant \Theta(1/n) \implies \text{fast mixing}$$

- [Eldan-Koehler-Zeitouni] got

$$\beta \leqslant \Theta(1) \implies \text{fast mixing}$$

- Within $O(1)$ of optimal.

Ising models



$$\mu(x) \propto \exp\left(\tfrac{1}{2}\sum_{u,v} J_{uv} x_u x_v + \sum_v h_v x_v\right)$$

symmetric matrix

- Dobrushin: when J has row/col ℓ_1 norms < 1, we get fast mixing.
- Sherrington-Kirkpatrick model: random Gaussian matrix J with $J_{uv} \sim \mathcal{N}(0, \beta/n)$.
- Open: find the exact threshold β where Glauber mixes fast w.h.p.
- Dobrushin gives weak bound:

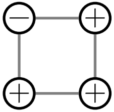
  $$\beta \leqslant \Theta(1/n) \implies \text{fast mixing}$$

- [Eldan-Koehler-Zeitouni] got

  $$\beta \leqslant \Theta(1) \implies \text{fast mixing}$$

- Within $O(1)$ of optimal. ☺
- They only used bounds on spectrum of random matrices:

Ising models



$$\mu(x) \propto \exp\left(\tfrac{1}{2}\sum_{u,v} J_{uv} x_u x_v + \sum_v h_v x_v\right)$$

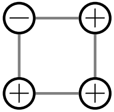
symmetric matrix

- Dobrushin: when J has row/col ℓ_1 norms < 1, we get fast mixing.
- Sherrington-Kirkpatrick model: random Gaussian matrix J with $J_{uv} \sim \mathcal{N}(0, \beta/n)$.
- Open: find the exact threshold β where Glauber mixes fast w.h.p.
- Dobrushin gives weak bound:
  $$\beta \leqslant \Theta(1/n) \implies \text{fast mixing}$$
- [Eldan-Koehler-Zeitouni] got
  $$\beta \leqslant \Theta(1) \implies \text{fast mixing}$$
- Within $O(1)$ of optimal.
- They only used bounds on spectrum of random matrices:

Theorem [Eldan-Koehler-Zeitouni]

If $\lambda_{\max}(J) - \lambda_{\min}(J) < 1$, then Glauber mixes fast.

Ising models



$$\mu(x) \propto \exp\left(\tfrac{1}{2}\sum_{u,v} J_{uv} x_u x_v + \sum_v h_v x_v\right)$$

symmetric matrix

- Dobrushin: when J has row/col ℓ_1 norms < 1, we get fast mixing.
- Sherrington-Kirkpatrick model: random Gaussian matrix J with $J_{uv} \sim \mathcal{N}(0, \beta/n)$.
- Open: find the exact threshold β where Glauber mixes fast w.h.p.

- Dobrushin gives weak bound:

$$\beta \leqslant \Theta(1/n) \implies \text{fast mixing}$$

- [Eldan-Koehler-Zeitouni] got

$$\beta \leqslant \Theta(1) \implies \text{fast mixing}$$

- Within $O(1)$ of optimal.
- They only used bounds on spectrum of random matrices:

Theorem [Eldan-Koehler-Zeitouni]

If $\lambda_{\max}(J) - \lambda_{\min}(J) < 1$, then Glauber mixes fast.

- We now know $O(n \log n)$ mixing [A-Jain-Koehler-Pham-Vuong].

Strategy

Strategy

- We may assume $0 \preceq J \preceq (1-\delta)I$, since diagonals of J do not matter.

Strategy

- We may assume $0 \preceq J \preceq (1-\delta)I$, since diagonals of J do not matter.
- Via stochastic localization we can kill parts of J:

$$\mu_t(x) \propto \exp\left(\tfrac{1}{2}x^\intercal J_t x + \langle h_t, x\rangle\right)$$

where $J_t = J - \int_0^t \Sigma_s \, ds$.

Strategy

- We may assume $0 \preceq J \preceq (1-\delta)I$, since diagonals of J do not matter.
- Via stochastic localization we can kill parts of J:
$$\mu_t(x) \propto \exp\left(\frac{1}{2}x^\intercal J_t x + \langle h_t, x\rangle\right)$$
where $J_t = J - \int_0^t \Sigma_s \, ds$.
- We will keep $J_t \succeq 0$, and try to get it as close to 0 as possible.

Strategy

- We may assume $0 \preceq J \preceq (1-\delta)I$, since diagonals of J do not matter.
- Via stochastic localization we can kill parts of J:
$$\mu_t(x) \propto \exp\left(\tfrac{1}{2}x^\intercal J_t x + \langle h_t, x\rangle\right)$$
where $J_t = J - \int_0^t \Sigma_s \, ds$.
- We will keep $J_t \succeq 0$, and try to get it as close to 0 as possible.
- 0 would be a product distribution.

ideal

ϕ-entropies

- Suppose we have a Markov kernel N, and would like to show $\forall \nu$:

$$\mathcal{D}_\phi(\nu N \parallel \mu N) \leqslant (1-\rho)\, \mathcal{D}_\phi(\nu \parallel \mu)$$

ϕ-entropies

- Suppose we have a Markov kernel N, and would like to show $\forall \nu$:

  $$\mathcal{D}_\phi(\nu N \parallel \mu N) \leqslant (1-\rho)\, \mathcal{D}_\phi(\nu \parallel \mu)$$

- Same as proving $\forall f$:

  $$\mathrm{Ent}^{\phi}_{\mu N}[N^\circ f] \leqslant (1-\rho) \cdot \mathrm{Ent}^{\phi}_{\mu}[f]$$

ϕ-entropies

- Suppose we have a Markov kernel N, and would like to show $\forall \nu$:
$$\mathcal{D}_{\phi}(\nu N \parallel \mu N) \leqslant (1-\rho)\, \mathcal{D}_{\phi}(\nu \parallel \mu)$$
- Same as proving $\forall f$:
$$\mathrm{Ent}^{\phi}_{\mu N}[N^{\circ} f] \leqslant (1-\rho) \cdot \mathrm{Ent}^{\phi}_{\mu}[f]$$
- This is equivalent to
$$\mathrm{Ent}^{\phi}_{\mu}[f] - \mathrm{Ent}^{\phi}_{\mu N}[N^{\circ} f] \geqslant \rho\, \mathrm{Ent}^{\phi}_{\mu}[f]$$

ϕ-entropies

- Suppose we have a Markov kernel N, and would like to show $\forall \nu$:
  $$\mathcal{D}_\phi(\nu N \parallel \mu N) \leqslant (1-\rho)\, \mathcal{D}_\phi(\nu \parallel \mu)$$
- Same as proving $\forall f$:
  $$\mathsf{Ent}^{\phi}_{\mu N}[N^\circ f] \leqslant (1-\rho) \cdot \mathsf{Ent}^{\phi}_{\mu}[f]$$
- This is equivalent to
  $$\mathsf{Ent}^{\phi}_{\mu}[f] - \mathsf{Ent}^{\phi}_{\mu N}[N^\circ f] \geqslant \rho\, \mathsf{Ent}^{\phi}_{\mu}[f]$$
- Lhs is deficit in data processing:
  $$\mathbb{E}_{y \sim \mu N}\left[\mathsf{Ent}^{\phi}_{N^\circ(y,\cdot)}[f]\right]$$

ϕ-entropies

- Suppose we have a Markov kernel N, and would like to show $\forall \nu$:

$$\mathcal{D}_\phi(\nu N \parallel \mu N) \leqslant (1-\rho)\,\mathcal{D}_\phi(\nu \parallel \mu)$$

- Same as proving $\forall f$:

$$\mathrm{Ent}^\phi_{\mu N}[N^\circ f] \leqslant (1-\rho) \cdot \mathrm{Ent}^\phi_\mu[f]$$

- This is equivalent to

$$\mathrm{Ent}^\phi_\mu[f] - \mathrm{Ent}^\phi_{\mu N}[N^\circ f] \geqslant \rho\, \mathrm{Ent}^\phi_\mu[f]$$

- Lhs is deficit in data processing:

$$\mathbb{E}_{y\sim\mu N}\left[\mathrm{Ent}^\phi_{N^\circ(y,\cdot)}[f]\right]$$

- Exercise: concave in μ.

ϕ-entropies

- Suppose we have a Markov kernel N, and would like to show $\forall \nu$:

$$\mathcal{D}_\phi(\nu N \parallel \mu N) \leqslant (1-\rho)\, \mathcal{D}_\phi(\nu \parallel \mu)$$

- Same as proving $\forall f$:

$$\mathsf{Ent}^{\phi}_{\mu N}[N^\circ f] \leqslant (1-\rho) \cdot \mathsf{Ent}^{\phi}_{\mu}[f]$$

- This is equivalent to

$$\mathsf{Ent}^{\phi}_{\mu}[f] - \mathsf{Ent}^{\phi}_{\mu N}[N^\circ f] \geqslant \rho\, \mathsf{Ent}^{\phi}_{\mu}[f]$$

- Lhs is deficit in data processing:

$$\mathbb{E}_{y \sim \mu N}\left[\mathsf{Ent}^{\phi}_{N^\circ(y,\cdot)}[f]\right]$$

- Exercise: concave in μ.
- Now suppose μ' is a random measure with $\mathbb{E}[\mu'] = \mu$.

ϕ-entropies

- Suppose we have a Markov kernel N, and would like to show $\forall \nu$:

$$\mathcal{D}_\phi(\nu N \parallel \mu N) \leqslant (1-\rho)\, \mathcal{D}_\phi(\nu \parallel \mu)$$

- Same as proving $\forall f$:

$$\mathrm{Ent}^{\phi}_{\mu N}[N^\circ f] \leqslant (1-\rho) \cdot \mathrm{Ent}^{\phi}_{\mu}[f]$$

- This is equivalent to

$$\mathrm{Ent}^{\phi}_{\mu}[f] - \mathrm{Ent}^{\phi}_{\mu N}[N^\circ f] \geqslant \rho\, \mathrm{Ent}^{\phi}_{\mu}[f]$$

- Lhs is deficit in data processing:

$$\mathbb{E}_{y \sim \mu N}\left[\mathrm{Ent}^{\phi}_{N^\circ(y,\cdot)}[f]\right]$$

- Exercise: concave in μ.
- Now suppose μ' is a random measure with $\mathbb{E}[\mu'] = \mu$.

$$\begin{tikzcd}
\mu \arrow[r, "\text{decomposed}"] \arrow[d, "\rho\ \text{contraction}"'] & \mu' \arrow[d, "\rho'\ \text{contraction}"'] \\
\mu N \arrow[r, "\text{decomposed}"] & \mu' N
\end{tikzcd}$$

ϕ-entropies

- Suppose we have a Markov kernel N, and would like to show $\forall \nu$:
  $$\mathcal{D}_\phi(\nu N \,\|\, \mu N) \leqslant (1-\rho)\,\mathcal{D}_\phi(\nu \,\|\, \mu)$$
- Same as proving $\forall f$:
  $$\mathrm{Ent}^{\phi}_{\mu N}[N^\circ f] \leqslant (1-\rho)\cdot \mathrm{Ent}^{\phi}_{\mu}[f]$$
- This is equivalent to
  $$\mathrm{Ent}^{\phi}_{\mu}[f] - \mathrm{Ent}^{\phi}_{\mu N}[N^\circ f] \geqslant \rho\,\mathrm{Ent}^{\phi}_{\mu}[f]$$
- Lhs is deficit in data processing:
  $$\mathbb{E}_{y\sim\mu N}\left[\mathrm{Ent}^{\phi}_{N^\circ(y,\cdot)}[f]\right]$$
- Exercise: concave in μ.
- Now suppose μ' is a random measure with $\mathbb{E}[\mu'] = \mu$.

$$\begin{tikzcd}
\mu \arrow[r, "\text{decomposed}"] \arrow[d, "\rho\ \text{contraction}"'] & \mu' \arrow[d, "\rho'\ \text{contraction}"'] \\
\mu N \arrow[r, "\text{decomposed}"] & \mu' N
\end{tikzcd}$$

- If we know each μ' contracts ϕ-divergence by $1-\rho'$, we get
  $$\mathrm{Ent}^{\phi}_{\mu}[f] - \mathrm{Ent}^{\phi}_{\mu N}[N^\circ f] \geqslant \rho'\,\mathbb{E}[\mathrm{Ent}^{\phi}_{\mu'}[f]]$$

ϕ-entropies

- Suppose we have a Markov kernel N, and would like to show $\forall \nu$:
$$\mathcal{D}_\phi(\nu N \parallel \mu N) \leqslant (1-\rho)\,\mathcal{D}_\phi(\nu \parallel \mu)$$
- Same as proving $\forall f$:
$$\mathrm{Ent}^{\phi}_{\mu N}[N^\circ f] \leqslant (1-\rho)\cdot \mathrm{Ent}^{\phi}_{\mu}[f]$$
- This is equivalent to
$$\mathrm{Ent}^{\phi}_{\mu}[f] - \mathrm{Ent}^{\phi}_{\mu N}[N^\circ f] \geqslant \rho\, \mathrm{Ent}^{\phi}_{\mu}[f]$$
- Lhs is deficit in data processing:
$$\mathbb{E}_{y\sim\mu N}\left[\mathrm{Ent}^{\phi}_{N^\circ(y,\cdot)}[f]\right]$$
- Exercise: concave in μ.
- Now suppose μ' is a random measure with $\mathbb{E}[\mu'] = \mu$.

$$\begin{tikzcd}
\mu \arrow[r, "\text{decomposed}"] \arrow[d, "\rho\text{ contraction}"'] & \mu' \arrow[d, "\rho'\text{ contraction}"'] \\
\mu N \arrow[r, "\text{decomposed}"] & \mu' N
\end{tikzcd}$$

- If we know each μ' contracts ϕ-divergence by $1-\rho'$, we get $\mathrm{Ent}^{\phi}_{\mu}[f] - \mathrm{Ent}^{\phi}_{\mu N}[N^\circ f] \geqslant \rho'\, \mathbb{E}[\mathrm{Ent}^{\phi}_{\mu'}[f]]$
- If we prove
$$\mathbb{E}[\mathrm{Ent}^{\phi}_{\mu'}[f]] \geqslant \gamma \cdot \mathrm{Ent}^{\phi}_{\mu}[f]$$
we get to conclude $\rho \geqslant \gamma \cdot \rho'$.

Approximate conservation [Chen-Eldan]

- Suppose we have a discrete/continuous time localization scheme $\{\mu_t\}$.

Approximate conservation [Chen-Eldan]

- Suppose we have a discrete/continuous time localization scheme $\{\mu_t\}$.
- Approximate conservation: at every step $\mathbf{Ent}^{\phi}_{\mu_t}[f]$ does not shrink by much on average.

Approximate conservation [Chen-Eldan]

- Suppose we have a discrete/continuous time localization scheme $\{\mu_t\}$.
- Approximate conservation: at every step $\mathrm{Ent}^{\phi}_{\mu_t}[f]$ does not shrink by much on average.
- In discrete time

$$\mathbb{E}\left[\mathrm{Ent}^{\phi}_{\mu_{t+1}}[f] \;\middle|\; \mathcal{F}_t\right] \geqslant (1-\alpha_t)\,\mathrm{Ent}^{\phi}_{\mu_t}[f]$$

Approximate conservation [Chen-Eldan]

- Suppose we have a discrete/continuous time localization scheme $\{\mu_t\}$.
- Approximate conservation: at every step $\mathrm{Ent}^{\phi}_{\mu_t}[f]$ does not shrink by much on average.
- In discrete time

$$\mathbb{E}\left[\mathrm{Ent}^{\phi}_{\mu_{t+1}}[f] \;\middle|\; \mathcal{F}_t\right] \geqslant (1-\alpha_t)\,\mathrm{Ent}^{\phi}_{\mu_t}[f]$$

- In continuous time

$$\mathbb{E}\left[d\,\mathrm{Ent}^{\phi}_{\mu_t}[f] \mid \mathcal{F}_t\right] \geqslant -\alpha_t\,\mathrm{Ent}^{\phi}_{\mu_t}[f]dt$$

Approximate conservation [Chen-Eldan]

- Suppose we have a discrete/continuous time localization scheme $\{\mu_t\}$.
- Approximate conservation: at every step $\mathrm{Ent}^{\phi}_{\mu_t}[f]$ does not shrink by much on average.
- In discrete time

$$\mathbb{E}\left[\mathrm{Ent}^{\phi}_{\mu_{t+1}}[f] \;\middle|\; \mathcal{F}_t\right] \geqslant (1-\alpha_t)\,\mathrm{Ent}^{\phi}_{\mu_t}[f]$$

- In continuous time

$$\mathbb{E}\left[d\,\mathrm{Ent}^{\phi}_{\mu_t}[f] \mid \mathcal{F}_t\right] \geqslant -\alpha_t\,\mathrm{Ent}^{\phi}_{\mu_t}[f]dt$$

- Then we get to transfer contraction rates on μ_t to contraction rates on μ with loss:

$$\gamma \geqslant (1-\alpha_0)(1-\alpha_1)\cdots(1-\alpha_{t-1}) \quad \text{or} \quad \gamma \geqslant \exp\left(-\int_0^t \alpha_s\,ds\right)$$

Approximate conservation of variance

- Let us specialize to $\mathrm{Ent}^{\phi} = \mathrm{Var}$ and stochastic localization: $\leftarrow$ works for discrete too
$$d\mu_t(x) = \langle C_t \, dB_t, x - \mathrm{mean}(\mu_t) \rangle \mu_t(x).$$

Approximate conservation of variance

- Let us specialize to $\mathsf{Ent}^{\phi} = \mathsf{Var}$ and stochastic localization: ← works for discrete too
$$d\mu_t(x) = \langle C_t\, dB_t, x - \mathsf{mean}(\mu_t)\rangle \mu_t(x).$$
- We have $\mathbb{E}_{\mu_t}[f^2]$ and $\mathbb{E}_{\mu_t}[f]$ are both martingales. Evolution:
$$d\,\mathbb{E}_{\mu_t}[f] = \sum\nolimits_x \langle C_t\, dB_t, x - \mathsf{mean}(\mu_t)\rangle \mu_t(x) f(x) = \langle C_t\, dB_t, v_t\rangle$$
for the vector $v_t = \mathbb{E}_{x\sim\mu_t}[f(x)(x - \mathsf{mean}(\mu_t))]$.

Approximate conservation of variance

- Let us specialize to $\mathsf{Ent}^{\Phi} = \mathsf{Var}$ and stochastic localization: ← works for discrete too
$$d\mu_t(x) = \langle C_t \, dB_t, x - \mathsf{mean}(\mu_t) \rangle \mu_t(x).$$
- We have $\mathbb{E}_{\mu_t}[f^2]$ and $\mathbb{E}_{\mu_t}[f]$ are both martingales. Evolution:
$$d\,\mathbb{E}_{\mu_t}[f] = \sum_x \langle C_t \, dB_t, x - \mathsf{mean}(\mu_t) \rangle \mu_t(x) f(x) = \langle C_t \, dB_t, v_t \rangle$$
for the vector $v_t = \mathbb{E}_{x \sim \mu_t}[f(x)(x - \mathsf{mean}(\mu_t))]$.
- This means that
$$d\,\mathsf{Var}_{\mu_t}[f] = (\text{martingale term}) - v_t^{\intercal} \Sigma_t v_t \, dt$$

Approximate conservation of variance

- Let us specialize to $\mathsf{Ent}^{\phi} = \mathsf{Var}$ and stochastic localization: ← works for discrete too
$$d\mu_t(x) = \langle C_t\, dB_t, x - \mathsf{mean}(\mu_t)\rangle \mu_t(x).$$
- We have $\mathbb{E}_{\mu_t}[f^2]$ and $\mathbb{E}_{\mu_t}[f]$ are both martingales. Evolution:
$$d\,\mathbb{E}_{\mu_t}[f] = \sum_x \langle C_t\, dB_t, x - \mathsf{mean}(\mu_t)\rangle \mu_t(x) f(x) = \langle C_t\, dB_t, v_t\rangle$$
for the vector $v_t = \mathbb{E}_{x\sim\mu_t}[f(x)(x - \mathsf{mean}(\mu_t))]$.
- This means that
$$d\,\mathsf{Var}_{\mu_t}[f] = (\text{martingale term}) - v_t^{\intercal}\Sigma_t v_t\, dt$$
- As long as Σ_t and v_t are orthogonal, we get that $\mathsf{Var}_{\mu_t}[f]$ is a martingale! ☺

Application to Sherrington-Kirkpatrick

- Going back to Ising models

$$\mu_t(x) \propto \exp\left(\tfrac{1}{2}x^\intercal J_t x + \langle h_t, x\rangle\right)$$

Application to Sherrington-Kirkpatrick

- Going back to Ising models
$$\mu_t(x) \propto \exp\left(\tfrac{1}{2}x^\intercal J_t x + \langle h_t, x\rangle\right)$$
- As long as $J_t \succeq 0$ and $\mathsf{rank}(J_t) \geqslant 2$, we can choose nonzero $\Sigma_t \succeq 0$ such that
$$\mathsf{span}(\Sigma_t) \subseteq \mathsf{span}(J_t)$$
and $\Sigma_t v_t = 0$.

Application to Sherrington-Kirkpatrick

- Going back to Ising models
$$\mu_t(x) \propto \exp\left(\tfrac{1}{2}x^\intercal J_t x + \langle h_t, x\rangle\right)$$
- As long as $J_t \succeq 0$ and $\mathsf{rank}(J_t) \geqslant 2$, we can choose nonzero $\Sigma_t \succeq 0$ such that
$$\mathsf{span}(\Sigma_t) \subseteq \mathsf{span}(J_t)$$
and $\Sigma_t v_t = 0$.
- The process stops when J_t becomes rank 1, not quite $J_t = 0$ ☹

Application to Sherrington-Kirkpatrick

- Going back to Ising models
$$\mu_t(x) \propto \exp\left(\tfrac{1}{2}x^\top J_t x + \langle h_t, x\rangle\right)$$
- As long as $J_t \succeq 0$ and $\mathsf{rank}(J_t) \geqslant 2$, we can choose nonzero $\Sigma_t \succeq 0$ such that
$$\mathsf{span}(\Sigma_t) \subseteq \mathsf{span}(J_t)$$
and $\Sigma_t v_t = 0$.
- The process stops when J_t becomes rank 1, not quite $J_t = 0$ ☹
- However, note that for rank 1 matrices $J_t = uu^\top$ we have Dobrushin++:
$$\mathcal{I}[i \to j] \leqslant |u_i u_j|$$
and $\lambda_{\mathsf{max}}(\mathcal{I}) \leqslant \sum_i |u_i|^2 = \|u\|^2$.

Application to Sherrington-Kirkpatrick

- Going back to Ising models
$$\mu_t(x) \propto \exp\left(\tfrac{1}{2}x^\intercal J_t x + \langle h_t, x\rangle\right)$$
- As long as $J_t \succeq 0$ and $\mathrm{rank}(J_t) \geqslant 2$, we can choose nonzero $\Sigma_t \succeq 0$ such that
$$\mathrm{span}(\Sigma_t) \subseteq \mathrm{span}(J_t)$$
and $\Sigma_t v_t = 0$.
- The process stops when J_t becomes rank 1, not quite $J_t = 0$ ☹
- However, note that for rank 1 matrices $J_t = uu^\intercal$ we have Dobrushin++:
$$\mathcal{J}[i \to j] \leqslant |u_i u_j|$$
and $\lambda_{\mathsf{max}}(\mathcal{J}) \leqslant \sum_i |u_i|^2 = \|u\|^2$.
- This shows contraction of χ^2 under Glauber. ☺