CS 263: Counting and Sampling

Nima Anari

slides for

Stochastic Localization
Skipped ...
Stochastic Calculus
- Localization schemes
- Itô calculus
- Stochastic localization

Conservation
- Sherrington-Kirkpatrick model
- $\phi$-entropies in localization scheme
- Approximate conservation
Stochastic Calculus

- Localization schemes
- Itô calculus
- Stochastic localization

Conservation

- Sherrington-Kirkpatrick model
- $\phi$-entropies in localization scheme
- Approximate conservation
Imagine $\mu$ is on $\binom{[n]}{k} \hookrightarrow \{0, 1\}^n$. Let us choose $i \sim \mu_{D_k \to 1} = p/k$. Let $\nu$ a random measure be the conditional on $\{i\}$. For $w = 1/i - 1/k$: $\nu(x) = (1 + \langle w, x - \text{mean}(\mu) \rangle)$. Note that $\mu = E_i[\nu]$.

Continuing this we get a measure-valued random process martingale:

Simplicial localization

Let $S \sim \mu$, and let $e_1, \ldots, e_k$ be a u.r. permutation of $S$. Define $\mu_i$ as conditional of $\mu$ on $\{e_1, \ldots, e_i\}$. 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.
Imagine $\mu$ 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

Let $S \sim \mu$, and let $e_1, \ldots, e_k$ be a u.r. permutation of $S$. Define $\mu_i$ as conditional of $\mu$ on $\{e_1, \ldots, e_i\}$. 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.
Imagine $\mu$ 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 $\nu$ be the conditional on $\{i\}$. For $w = 1_i/p_i - 1/k$:

\[ \nu(x) = \left(1 + \langle w, x - \text{mean}(\mu)\rangle\right) \mu(x) \]

a random measure

linear tilt

Simplicial localization

Let $S \sim \mu$, and let $e_1, \ldots, e_k$ be a u.r. permutation of $S$. Define $\mu_i$ as conditional of $\mu$ on $\{e_1, \ldots, e_i\}$. Then $\mu_0 \rightarrow \mu_1 \rightarrow \mu_2 \rightarrow \cdots \rightarrow \mu_k$ is called simplicial localization used for local-to-global and trickle down.
Imagine $\mu$ is on $\binom{[n]}{k} \hookrightarrow \{0,1\}^n$.

Denote $p_i = \mathbb{P}_{S \sim \mu}[i \in S]$. Let us choose $i \sim \mu D_{k \rightarrow 1} = p/k$.

Let $\nu$ be the conditional on $\{i\}$. For $w = 1_i/p_i - 1/k$:

\[ \nu(x) = (1 + \langle w, x - \text{mean}(\mu) \rangle) \mu(x) \]

Note that $\mu = \mathbb{E}_i[\nu]$. This is a decomposition of measure.
Imagine $\mu$ is on $\binom{[n]}{k} \hookrightarrow \{0, 1\}^n$.

Denote $p_i = P_{S \sim \mu}[i \in S]$. Let us choose $i \sim \mu D_{k \rightarrow 1} = p/k$.

Let $\nu$ be the conditional on $\{i\}$. For $w = 1_i/p_i - 1/k$:

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

Note that $\mu = E_i[\nu]$. This is a decomposition of measure.

Continuing this we get a measure-valued random process: martingale
Imagine $\mu$ is on $\binom{[n]}{k} \hookrightarrow \{0, 1\}^n$.

Denote $p_i = P_{S \sim \mu}[i \in S]$. Let us choose $i \sim \mu D_{k \to 1} = p/k$.

Let $\nu$ be the conditional on $\{i\}$. For $w = 1_i / p_i - 1/k$:

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

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

Continuing this we get a measure-valued random martingale:

---

**Simplicial localization**

Let $S \sim \mu$, and let $e_1, \ldots, e_k$ be a u.r. permutation of $S$. Define $\mu_i$ as conditional of $\mu$ on $\{e_1, \ldots, e_i\}$. 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.
Same idea applied in continuous time. For some measure \( \mu \) on \( \mathbb{R}^n \), we get measure-valued process \( \{ \mu_t \mid t \in \mathbb{R}_{\geq 0} \} \).
Stochastic localization

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

- Controlled by (stochastic) differential equation

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

linear tilt

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 $\mu$ on $\mathbb{R}^n$, we get measure-valued process $\{\mu_t \mid t \in \mathbb{R}_{\geq 0}\}$.

- Controlled by (stochastic) differential equation

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

linear tilt

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.
Stochastic localization

- Same idea applied in continuous time. For some measure $\mu$ on $\mathbb{R}^n$, we get measure-valued process $\{\mu_t \mid t \in \mathbb{R}_{\geq 0}\}$.
- Controlled by (stochastic) differential equation
  \[
d\mu_t(x) = \left\langle w_t, x - \text{mean}(\mu) \right\rangle \mu_t(x)
  \]
  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 $\{B_t | t \in \mathbb{R}_{\geq 0}\}$ 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.
Brownian motion: in $\mathbb{R}^n$, the process $\{B_t \mid t \in \mathbb{R}_{\geq 0}\}$ such that

$$B_t - B_s \sim \mathcal{N}(0, (t - s)\mathbf{I})$$

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 $B_{t+dt} - B_t$: $dB_t \sim \mathcal{N}(0, dt \cdot \mathbf{I})$
Brownian motion: in \( \mathbb{R}^n \), the process \( \{ B_t \mid t \in \mathbb{R}_{\geq 0} \} \) 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 \( B_{t+dt} - B_t: dB_t \sim \mathcal{N}(0, dt \cdot I) \)

Fact: \( dB_t \) is not on the order of \( dt \), but rather on the order of \( \sqrt{dt} \)!
Brownian motion: in $\mathbb{R}^n$, the process $\{B_t \mid t \in \mathbb{R}_{\geq 0}\}$ 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 $B_{t+dt} - B_t$: $dB_t \sim \mathcal{N}(0, dt \cdot I)$

Fact: $dB_t$ is not on the order of $dt$, but rather on the order of $\sqrt{dt}$!

Itô process: $\{X_t \mid t \in \mathbb{R}_{\geq 0}\}$ derived via stochastic differential equation (SDE):

$$dX_t = u_t \, dt + C_t \, dB_t$$

for some “nice” vector and matrix valued processes $\{u_t\}, \{C_t\}$.
Brownian motion: in \( n \)D, the process \( \{B_t \mid t \in \mathbb{R}_{\geq 0}\} \) 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 \( B_{t+dt} - B_t \): \( dB_t \sim \mathcal{N}(0, dt \cdot I) \)

Fact: \( dB_t \) is not on the order of \( dt \), but rather on the order of \( \sqrt{dt} \)! 

Ito process: \( \{X_t \mid t \in \mathbb{R}_{\geq 0}\} \) 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.
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:

$$dY_t = \left( f'(X_t) \, u_t + \frac{1}{2} f''(X_t) \, c_t^2 \right) \, dt + f'(X_t) \, c_t \, dB_t$$

Itô term
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
\[ dY_t = f'(X_t) \, dX_t = f'(X_t) \, u_t \, dt + f'(X_t) \, c_t \, dB_t \]

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:
\[ dY_t = \left( f'(X_t) \, u_t + \frac{1}{2} f''(X_t) \, c_t^2 \right) \, dt + f'(X_t) \, c_t \, dB_t \]
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 $\sqrt{dt}$!
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 $dt$. 
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:

$$dY_t = \left( f'(X_t)u_t + \frac{1}{2}f''(X_t)c_t^2 \right) dt + f'(X_t)c_t \, dB_t$$
Intuition: curvature creates drift!

\[ dX_t = u_t \, dt + C_t \, dB_t \]

If we have \( Y_t = f(X_t) \), then
\[ dY_t = \left( \langle \nabla f(X_t), u_t \rangle + \frac{1}{2} \text{tr} \left( C_t \, \nabla^2 f(X_t) \right) \right) \, dt + \langle \nabla f(X_t), C_t \, dB_t \rangle. \]
Intuition: curvature creates drift!

Itô’s lemma ($n$D to 1D)

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

$$dY_t = \left( \langle \nabla f(X_t), u_t \rangle + \frac{1}{2} \text{tr}(C_t^T \nabla^2 f(X_t) C_t) \right) \, dt + \langle \nabla f(X_t), C_t \, dB_t \rangle.$$
For $\mu$ on subset of $\mathbb{R}^n$, and adapted matrix process $C_t$, we define $\forall x$

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

For $\mu$ 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 $\mu$, we should think of it as density. You can for simplicity assume support is finite.
Stochastic localization

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

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

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

Changes in $\mu_t$ are proportional to itself. Log-scale? Let’s use Itô’s lemma for $f = \log$. 
Stochastic localization

For $\mu$ 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 $\mu$, we should think of it as density. You can for simplicity assume support is finite.

It is a martingale, with filtration $F_t$:

$$\mathbb{E}[\mu_t(x) \mid F_s] = \mu_s(x) \quad \forall s \leq t$$

If $\mu$ is normalized, $\mu_t$ remains so:

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

Changes in $\mu_t$ 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

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

At any time $t$, we have $\mu_t(x) \propto \mu(x) \cdot \exp(-\frac{1}{2}x^TA_t x + \langle h_t, x \rangle)$

where $A_t = \int_0^t \Sigma_s^2 ds$. 
### Stochastic localization

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

- Changes in $\mu_t$ 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
  $$-(x - \text{mean}(\mu_t))^T C_t C_t^T (x - \text{mean}(\mu_t)) \cdot X_t^2 \quad \frac{1}{2X_t^2}$$

- So if we name $\Sigma_t = C_t C_t^T$, then
  $$d \log \mu_t(x) = -\frac{1}{2} x^T \Sigma_t x \, dt + \text{affine}(x)$$
For $\mu$ 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 $\mu$, we should think of it as density. You can for simplicity assume support is finite.

It is a martingale, with filtration $\mathcal{F}_t$:

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

If $\mu$ is normalized, $\mu_t$ remains so:

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

Changes in $\mu_t$ 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

$$-\frac{1}{2}(x - \text{mean}(\mu_t))^T C_t C_t^T (x - \text{mean}(\mu_t)) \cdot X_t^2$$

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

$$d \log \mu_t(x) = -\frac{1}{2} x^T \Sigma_t x dt + \text{affine}(x)$$

At any time $t$, we have $\mu_t(x) \propto \mu(x) \cdot \exp(-\frac{1}{2} x^T A_t x + \langle h_t, x \rangle)$

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
- $\phi$-entropies in localization scheme
- Approximate conservation
Stochastic Calculus

- Localization schemes
- Itô calculus
- Stochastic localization

Conservation

- Sherrington-Kirkpatrick model
- $\phi$-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 \( \ell_1 \) norms < 1, we get fast mixing.
Ising models

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

- **Dobrushin**: when $J$ has row/col $\ell_1$ norms $< 1$, we get fast mixing.
- **Sherrington-Kirkpatrick model**: random Gaussian matrix $J$ with $J_{uv} \sim \mathcal{N}(0, \beta/n)$. 

**Theorem** [Eldan-Koehler-Zeitouni] If $\lambda_{\text{max}}(J) - \lambda_{\text{min}}(J) < 1$, then Glauber mixes fast.

We now know $O(n \log n)$ mixing [A-Jain-Koehler-Pham-Vuong].
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 \( \ell_1 \) 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 \( \beta \) where Glauber mixes fast w.h.p.
Ising models

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

\textit{symmetric matrix}

- Dobrushin: when J has row/col \( \ell_1 \) 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 \( \beta \) where Glauber mixes fast w.h.p.
Ising models

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

Dobrushin: when \( J \) has row/col \( \ell_1 \) 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 \( \beta \) where Glauber mixes fast w.h.p.

- Dobrushin gives weak bound:
  \[ \beta \leq \Theta(1/n) \implies \text{fast mixing} \]

- [Eldan-Koehler-Zeitouni] got
  \[ \beta \leq \Theta(1) \implies \text{fast mixing} \]
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 \( \ell_1 \) 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 \( \beta \) where Glauber mixes fast w.h.p.

- Dobrushin gives weak bound:
  \[ \beta \leq \Theta(1/n) \implies \text{fast mixing} \]

- [Eldan-Koehler-Zeitouni] got
  \[ \beta \leq \Theta(1) \implies \text{fast mixing} \]

- Within \( O(1) \) of optimal. 😊
Ising models

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

\[ \text{symmetric matrix} \]

- Dobrushin: when \( J \) has row/col \( \ell_1 \) 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 \( \beta \) where Glauber mixes fast w.h.p.

- Dobrushin gives weak bound:
  \[ \beta \leq \Theta(1/n) \implies \text{fast mixing} \]

- [Eldan-Koehler-Zeitouni] got
  \[ \beta \leq \Theta(1) \implies \text{fast mixing} \]

- Within \( O(1) \) of optimal. 😊

- They only used bounds on spectrum of random matrices:

\[ \text{Theorem [Eldan-Koehler-Zeitouni]} \]

\[ \lambda_{\text{max}}(J) - \lambda_{\text{min}}(J) < 1, \text{ then Glauber mixes fast.} \]

\[ \text{We now know } O(n \log n) \text{ mixing [A-Jain-Koehler-Pham-Vuong].} \]
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 \( \ell_1 \) 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 \( \beta \) where Glauber mixes fast w.h.p.

- Dobrushin gives weak bound:
  \[ \beta \leq \Theta(1/n) \implies \text{fast mixing} \]

- [Eldan-Koehler-Zeitouni] got
  \[ \beta \leq \Theta(1) \implies \text{fast mixing} \]

- 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.

[A-Jain-Koehler-Pham-Vuong]
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 \( \ell_1 \) 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 \( \beta \) where Glauber mixes fast w.h.p.

- Dobrushin gives weak bound:
  \[ \beta \leq \Theta(1/n) \implies \text{fast mixing} \]

- [Eldan-Koehler-Zeitouni] got
  \[ \beta \leq \Theta(1) \implies \text{fast mixing} \]

- Within \( O(1) \) of optimal.

- They only used bounds on spectrum of random matrices:

  **Theorem [Eldan-Koehler-Zeitouni]**

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

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

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

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.
Strategy

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

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

- Via stochastic localization we can kill parts of $J$:
  $$\mu_t(x) \propto \exp\left(\frac{1}{2}x^T J_t x + \langle h_t, x \rangle\right)$$
  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 $J$:
  \[
  \mu_t(x) \propto \exp\left(\frac{1}{2}x^T J_t x + \langle h_t, x \rangle\right)
  \]
  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 $J$:
  $$\mu_t(x) \propto \exp\left(\frac{1}{2}x^\top J_t x + \langle h_t, x \rangle\right)$$
  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
Suppose we have a Markov kernel $N$, and would like to show $\forall \nu$:

$$D_\phi(\nu N \parallel \mu N) \leq (1 - \rho) D_\phi(\nu \parallel \mu)$$
Suppose we have a Markov kernel $N$, and would like to show $\forall \nu$:

$$D_\phi(\nu N \parallel \mu N) \leq (1 - \rho) \cdot D_\phi(\nu \parallel \mu)$$

Same as proving $\forall f$:

$$\text{Ent}_{\mu N}^\phi [N \circ f] \leq (1 - \rho) \cdot \text{Ent}_{\mu}^\phi [f]$$
Suppose we have a Markov kernel $N$, and would like to show $\forall \nu$:

$$D_\phi(\nu N \parallel \mu N) \leq (1 - \rho) D_\phi(\nu \parallel \mu)$$

Same as proving $\forall f$:

$$\text{Ent}_\mu^\phi[f] \leq (1 - \rho) \cdot \text{Ent}_\mu^\phi[f]$$

This is equivalent to

$$\text{Ent}_\mu^\phi[f] - \text{Ent}_\mu^\phi[N \circ f] \geq \rho \text{Ent}_\mu^\phi[f]$$
φ-entropies

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

$$D_\phi(\nu N \parallel \mu N) \leq (1 - \rho) D_\phi(\nu \parallel \mu)$$

Same as proving $\forall f$:

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

This is equivalent to

$$\text{Ent}_\mu^\phi[f] - \text{Ent}_\mu^\phi[N \circ f] \geq \rho \text{Ent}_\mu^\phi[f]$$

Lhs is deficit in data processing:

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

- Suppose we have a Markov kernel $N$, and would like to show $\forall \nu$:
  $D_\phi(\nu N \parallel \mu N) \leq (1 - \rho) D_\phi(\nu \parallel \mu)$

- Same as proving $\forall f$:
  $\text{Ent}_{\mu N}^\phi[N^\circ f] \leq (1 - \rho) \cdot \text{Ent}_\mu^\phi[f]$

- This is equivalent to
  $\text{Ent}_\mu^\phi[f] - \text{Ent}_{\mu N}^\phi[N^\circ f] \geq \rho \text{Ent}_\mu^\phi[f]$

- Lhs is **deficit** in data processing:
  $\mathbb{E}_{y \sim \mu N} \left[ \text{Ent}_{N^\circ(y, \cdot)}^\phi[f] \right]$

- Exercise: **concave** in $\mu$. 

- If we know each $\mu'$ contracts $\phi$-divergence by $1 - \rho'$, we get
  $\text{Ent}_\mu^\phi[f] - \text{Ent}_{\mu N}^\phi[N^\circ f] \geq \rho' \text{Ent}_\mu^\phi[f]$

- If we prove $\mathbb{E}_{y \sim \mu'} \left[ \text{Ent}_{\mu}^\phi[f] \right] \geq \gamma \cdot \text{Ent}_\mu^\phi[f]$
  we get to conclude $\rho \geq \gamma \cdot \rho'$.
Suppose we have a Markov kernel $N$, and would like to show $\forall \nu$:

$$D_\phi(\nu N \parallel \mu N) \leq (1 - \rho) D_\phi(\nu \parallel \mu)$$

Same as proving $\forall f$:

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

This is equivalent to

$$\text{Ent}_\mu^\phi[f] - \text{Ent}_{\mu N}^\phi[N^\circ f] \geq \rho \text{Ent}_\mu^\phi[f]$$

Lhs is **deficit** in data processing:

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

Exercise: **concave** in $\mu$.

Now suppose $\mu'$ is a random measure with $\mathbb{E}[\mu'] = \mu$. 

**φ-entropies**
**φ-entropies**

- Suppose we have a Markov kernel $N$, and would like to show $\forall \nu$:
  
  $$
  D_\phi(\nu N \parallel \mu N) \leq (1 - \rho) D_\phi(\nu \parallel \mu)
  $$

- Same as proving $\forall f$:
  
  $$
  \Ent_{\mu N}^\phi [N \circ f] \leq (1 - \rho) \cdot \Ent_\mu^\phi [f]
  $$

- This is equivalent to
  
  $$
  \Ent_\mu^\phi [f] - \Ent_{\mu N}^\phi [N \circ f] \geq \rho \Ent_\mu^\phi [f]
  $$

- Lhs is **deficit** in data processing:
  
  $$
  \mathbb{E}_{y \sim \mu N} \left[ \Ent_{N \circ (y, \cdot)}^\phi [f] \right]
  $$

- Exercise: **concave** in $\mu$.

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

---

**Diagram:**

- **μ** to **μ'** with **decomposed**
- **μN** to **μ'N** with **decomposed**
- **ρ' contraction** from **μ'N**
- **ρ contraction** from **μN**
- **Exercise:** concave in $\mu$. 

---

15/18
Suppose we have a Markov kernel $N$, and would like to show $\forall \nu$:

$$D_\phi(\nu N \parallel \mu N) \leq (1 - \rho) D_\phi(\nu \parallel \mu)$$

Same as proving $\forall f$:

$$\text{Ent}_{\mu N}[N \circ f] \leq (1 - \rho) \cdot \text{Ent}_\mu[f]$$

This is equivalent to

$$\text{Ent}_\mu[f] - \text{Ent}_{\mu N}[N \circ f] \geq \rho \text{Ent}_\mu[f]$$

Lhs is deficit in data processing:

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

Exercise: concave in $\mu$.

Now suppose $\mu'$ is a random measure with $\mathbb{E}[\mu'] = \mu$.

If we know each $\mu'$ contracts $\phi$-divergence by $1 - \rho'$, we get

$$\text{Ent}_\mu[f] - \text{Ent}_{\mu N}[N \circ f] \geq \rho' \mathbb{E}[\text{Ent}_\mu'[f]]$$
\(\phi\)-entropies

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

\[
\begin{aligned}
\mu &\quad \xrightarrow{\text{decomposed}} & \quad \mu' \\
\mu N &\quad \xrightarrow{\rho \text{ contraction}} & \quad \mu' N \\
\mu N &\quad \xrightarrow{\rho' \text{ contraction}} & \quad \mu' N
\end{aligned}
\]

- If we know each \(\mu'\) contracts \(\phi\)-divergence by \(1 - \rho'\), we get
  \[\text{Ent}_{\mu}^\phi[f] - \text{Ent}_{\mu N}^\phi[N \circ f] \geq \rho' \mathbb{E}[\text{Ent}_{\mu'}^\phi[f]]\]
- If we prove
  \[\mathbb{E}[\text{Ent}_{\mu'}^\phi[f]] \geq \gamma \cdot \text{Ent}_{\mu}^\phi[f]\]
  we get to conclude \(\rho \geq \gamma \cdot \rho'\).
Approximate conservation [Chen-Eldan]

Suppose we have a discrete/continuous time localization scheme \( \{\mu_t\} \).
Suppose we have a discrete/continuous time localization scheme \(\{\mu_t\}\).

**Approximate conservation:** at every step \(\operatorname{Ent}_{\mu_t}[f]\) 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 \( \text{Ent}_{\mu_t}^\phi [f] \) does not shrink by much on average.
- In discrete time

\[
\mathbb{E} \left[ \text{Ent}_{\mu_{t+1}}^\phi [f] \ \bigg| \ F_t \right] \geq (1 - \alpha_t) \text{Ent}_{\mu_t}^\phi [f]
\]
Suppose we have a discrete/continuous time localization scheme \( \{ \mu_t \} \).

**Approximate conservation:** at every step \( \text{Ent}^\phi_{\mu_t} [f] \) does not shrink by much on average.

**In discrete time**

\[
\mathbb{E} \left[ \text{Ent}^\phi_{\mu_{t+1}} [f] \mid \mathcal{F}_t \right] \geq (1 - \alpha_t) \text{Ent}^\phi_{\mu_t} [f]
\]

**In continuous time**

\[
\mathbb{E} \left[ d \text{Ent}^\phi_{\mu_t} [f] \mid \mathcal{F}_t \right] \geq -\alpha_t \text{Ent}^\phi_{\mu_t} [f] dt
\]
Approximate conservation [Chen-Eldan]

- Suppose we have a discrete/continuous time localization scheme \( \{\mu_t\} \).
- **Approximate conservation:** at every step \( \text{Ent}^{\phi}_{\mu_t}[f] \) does not shrink by much on average.
- In discrete time
  \[
  \mathbb{E}\left[\text{Ent}^{\phi}_{\mu_{t+1}}[f] \mid \mathcal{F}_t\right] \geq (1 - \alpha_t) \text{Ent}^{\phi}_{\mu_t}[f]
  \]
- In continuous time
  \[
  \mathbb{E}\left[ d\text{Ent}^{\phi}_{\mu_t}[f] \mid \mathcal{F}_t\right] \geq -\alpha_t \text{Ent}^{\phi}_{\mu_t}[f]dt
  \]
- Then we get to transfer contraction rates on \( \mu_t \) to contraction rates on \( \mu \) with loss:
  \[
  \gamma \geq (1 - \alpha_0)(1 - \alpha_1) \cdots (1 - \alpha_{t-1}) \quad \text{or} \quad \gamma \geq \exp\left(-\int_0^t \alpha_s ds\right)
  \]
Let us specialize to $\text{Ent}^\Phi = \text{Var}$ and stochastic localization:

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

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

We have $\mathbb{E}_{\mu_t}[f^2]$ and $\mathbb{E}_{\mu_t}[f]$ are both martingales. Evolution:

$$d \mathbb{E}_{\mu_t}[f] = \sum_x \langle C_t dB_t, x - \text{mean} (\mu_t) \rangle \mu_t(x) f(x) = \langle C_t dB_t, \nu_t \rangle$$

for the vector $\nu_t = \mathbb{E}_{x \sim \mu_t}[f(x)(x - \text{mean} (\mu_t))].$
Let us specialize to $\text{Ent}^\Phi = \text{Var}$ and stochastic localization: works for discrete too

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

We have $E_{\mu_t}[f^2]$ and $E_{\mu_t}[f]$ are both martingales. Evolution:

$$dE_{\mu_t}[f] = \sum_x \langle C_t dB_t, x - \text{mean}(\mu_t) \rangle \mu_t(x) f(x) = \langle C_t dB_t, v_t \rangle$$

for the vector $v_t = E_{x \sim \mu_t}[f(x)(x - \text{mean}(\mu_t))]$.

This means that

$$d\text{Var}_{\mu_t}[f] = (\text{martingale term}) - v_t^T \Sigma_t v_t dt$$
Let us specialize to $\text{Ent}^\phi = \text{Var}$ and stochastic localization: works for discrete too

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

We have $E_{\mu_t}[f^2]$ and $E_{\mu_t}[f]$ are both martingales. Evolution:

$$d E_{\mu_t}[f] = \sum_x \langle C_t dB_t, x - \text{mean}(\mu_t) \rangle \mu_t(x) f(x) = \langle C_t dB_t, \nu_t \rangle$$

for the vector $\nu_t = E_{x \sim \mu_t}[f(x)(x - \text{mean}(\mu_t))].$

This means that

$$d \text{Var}_{\mu_t}[f] = (\text{martingale term}) - \nu_t^T \Sigma_t \nu_t dt$$

As long as $\Sigma_t$ and $\nu_t$ are orthogonal, we get that $\text{Var}_{\mu_t}[f]$ is a martingale! 😊
Application to Sherrington-Kirkpatrick

Going back to Ising models

\[ \mu_t(x) \propto \exp\left( \frac{1}{2} x^T J_t x + \langle h_t, x \rangle \right) \]
Going back to Ising models

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

As long as $J_t \succeq 0$ and $\text{rank}(J_t) \geq 2$, we can choose nonzero $\Sigma_t \succeq 0$ such that

$$\text{span}(\Sigma_t) \subseteq \text{span}(J_t)$$

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++:

$$I[i \rightarrow j] \leq |u_i u_j|$$

and

$$\lambda_{\text{max}}(I) \leq \sum_i |u_i|^2 = \|u\|^2.$$
Going back to Ising models

\[ \mu_t(x) \propto \exp\left( \frac{1}{2} x^T J_t x + \langle h_t, x \rangle \right) \]

As long as \( J_t \succeq 0 \) and \( \text{rank}(J_t) \geq 2 \), we can choose nonzero \( \Sigma_t \succeq 0 \) such that

\[ \text{span}(\Sigma_t) \subseteq \text{span}(J_t) \]

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

\[ \mu_t(x) \propto \exp\left( \frac{1}{2} x^T J_t x + \langle h_t, x \rangle \right) \]

- As long as \( J_t \succeq 0 \) and \( \text{rank}(J_t) \geq 2 \), we can choose nonzero \( \Sigma_t \succeq 0 \) such that \( \text{span}(\Sigma_t) \subseteq \text{span}(J_t) \)

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 \( J_t = uu^T \) we have Dobrushin++:

\[ J[i \rightarrow j] \leq |u_i u_j| \]

and \( \lambda_{\text{max}}(J) \leq \sum_i |u_i|^2 = \|u\|^2 \).
Going back to Ising models

\[ \mu_t(x) \propto \exp \left( \frac{1}{2} x^T J_t x + \langle h_t, x \rangle \right) \]

As long as \( J_t \succeq 0 \) and \( \text{rank}(J_t) \geq 2 \), we can choose nonzero \( \Sigma_t \succeq 0 \) such that

\[ \text{span}(\Sigma_t) \subseteq \text{span}(J_t) \]

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 \( J_t = uu^T \) we have Dobrushin++:

\[ J[i \rightarrow j] \leq |u_i u_j| \]

and \( \lambda_{\max}(J) \leq \sum_i |u_i|^2 = \|u\|^2 \).

This shows contraction of \( \chi^2 \) under Glauber.