CPSC 661: Sampling Algorithms in ML

Andre Wibisono

April 5, 2021

Yale University

Last time

- Wasserstein W_2 metric
- Otto calculus
- Optimization of potential energy

Today: Entropy and Brownian Motion

References

- Villani, *Topics in Optimal Transportation*, Springer, 2003
- Villani, *Optimal Transport: Old and New*, Springer, 2008
- Ambrosio, Gigli & Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Springer, 2005
- Wibisono, *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem*, COLT 2018

Entropy

Entropy

Boltzmann (1877): Entropy of ideal gas

$$S = k \log W$$

- $k = 1.380649 \times 10^{-23} J/K$ is Boltzmann constant
- W = number of microstates
- Second law of thermodynamics: Entropy is increasing

Entropy

Entropy of discrete random variable $X \sim p = (p_1, \ldots, p_n)$

$$h(p) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- Shannon (1948), *A Mathematical Theory of Communication*, Bell System Technical Journal

- A measure of randomness, information, surprise

  $\Rightarrow$ Information theory

- *Source coding theorem:* Entropy is minimum description complexity

  To encode $X \sim p$ needs $h(p)$ bits on average

Entropy

If $X_1, \ldots, X_m \sim p$ i.i.d. then

$$\log_2 p(X_1, \ldots, X_m) = \sum_{i=1}^{m} \log_2 p(X_i) \approx m\, \mathbb{E}_p[\log_2 p(X)] = -m\, h(p)$$

so a *typical* sequence $(X_1, \ldots, X_m) \sim p^{\otimes m}$ has almost equal probability

$$p(X_1, \ldots, X_m) \approx 2^{-m\, h(p)}$$

- Asymptotic Equipartition Property (AEP)
- Entropy controls exponential growth rate of typical set
- Large deviations theory

[Cover & Thomas, *Elements of Information Theory*, Wiley, 2006]

Discrete vs continuous entropy

Entropy is defined for distribution $\rho \in \mathcal{P}(\mathcal{X})$ over any space $\mathcal{X}$

- $\mathcal{X}$ can be discrete ($\{1, \ldots, n\}$)
- $\mathcal{X}$ can be continuous ($\mathbb{R}^n$)

Discrete entropy and continuous entropy have similar properties, different values

Continuous entropy inherits geometric structure from $\mathcal{X}$

Discrete entropy

Entropy of discrete distribution $p = (p_1, \ldots, p_n) \in \Delta_{n-1}$

$$h(p) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- Minimum entropy at point mass δ_i: $\delta_i = (0, 0, \ldots, 0, \underset{\substack{\downarrow \\ i}}{1}, 0, \ldots, 0)$

$$h(\delta_i) = 0$$

- Maximum entropy at uniform $u = (\frac{1}{n}, \ldots, \frac{1}{n})$:

$$h(u) = \log_2 n$$

Continuous entropy

Let ρ be a probability distribution on $\mathbb{R}^n$ with density $\rho\colon \mathbb{R}^n \to \mathbb{R}$

Continuous / differential **Entropy**:

$$H(\rho) = -\mathbb{E}_\rho[\log \rho] = -\int_{\mathbb{R}^n} \rho(x) \log \rho(x)\, dx$$

- If $\rho = \text{Uniform}(S)$ for some $S \subset \mathbb{R}^n$

  $\rho(x) = \frac{1}{\text{Vol}(S)}, \; x \in S$

$$H(\rho) = \log \text{Vol}(S)$$

- If $\rho = \mathcal{N}(\mu, \Sigma)$

$$H(\rho) = \frac{1}{2} \log \det(2\pi e \Sigma) = \frac{n}{2} \log(2\pi e) + \frac{1}{2} \log \det \Sigma$$

  If $\Sigma = \lambda I$

$$H(\rho) = \frac{n}{2} \log(2\pi e \lambda)$$

Continuous entropy

$$H(\rho) = -\mathbb{E}_\rho[\log \rho] = -\int_{\mathbb{R}^n} \rho(x) \log \rho(x)\, dx$$

- If $\rho = \delta_x$ (or has point mass)

$$H(\delta_x) = -\infty$$

- If $\rho = dx$ (Lebesgue measure)

$$H(dx) = +\infty$$

Gaussian as maximum entropy distribution

Gaussian is maximum entropy distribution given second moments

Lemma: If $X \sim \rho$ has $\mathrm{Cov}_\rho(X) = \Sigma$, then

$$H(\rho) \leq H(\mathcal{N}(0,\Sigma)) = \frac{1}{2} \log\det(2\pi e \Sigma)$$

Exponential family

Exponential family distribution:

$$\rho_\theta(x) = \exp\left(\langle T(x), \theta \rangle - A(\theta)\right)$$

where $T\colon \mathbb{R}^n \to \mathbb{R}^d$ is sufficient statistics, $\theta \in \mathbb{R}^d$ is parameter, and $A(\theta) = \log \int_{\mathbb{R}^n} e^{\langle T(x), \theta \rangle} dx$ is log-partition function

- Gaussian, exponential, Poisson, geometric, beta, Dirichlet, ...
- Maximum entropy distribution given sufficient statistics $\mathbb{E}[T(X)]$
- Log-partition function is convex dual negative entropy

[Wainwright & Jordan, *Graphical Models, Exponential Families, and Variational Inference*, Foundations and Trends in Machine Learning, 2008]

Concavity of entropy

$$H(\rho) = -\mathbb{E}_\rho[\log \rho] = -\int_{\mathbb{R}^n} \rho(x) \log \rho(x) \, dx$$

- Entropy is concave in usual sense (along linear combination):

$$H\left((1-t)\rho_0 + t\rho_1\right) \geq (1-t)H(\rho_0) + tH(\rho_1)$$

because $r \longmapsto -r \log r$ is concave

- Entropy is also concave in Wasserstein sense

$$H\left(\rho_t\right) \geq (1-t)H(\rho_0) + tH(\rho_1)$$

for $\rho_t = (T_t)_{\#}\rho_0$ displacement interpolation from ρ_0 to ρ_1

Variants

Boltzmann / Shannon entropy

$$H(\rho) = -\int_{\mathbb{R}^n} \rho(x) \log \rho(x)\, dx$$

is the case $\alpha \to 1$ of:

1. Rényi entropy of order $\alpha > 0$

$$H_\alpha(\rho) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^n} \rho(x)^\alpha \, dx$$

2. Tsallis entropy of order $\alpha > 0$

$$\tilde{H}_\alpha(\rho) = \frac{1}{1-\alpha} \left(\int_{\mathbb{R}^n} \rho(x)^\alpha \, dx - 1 \right)$$

Wasserstein geometry of Entropy

Internal energy

Potential:

$$F(\rho) = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx$$

Internal energy $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$

$$F(\rho) = \int_{\mathbb{R}^n} U(\rho(x))\, dx$$

for some $U : \mathbb{R} \to \mathbb{R}$

- L^2-variation is

$$\frac{\delta F}{\delta \rho}(x) = U'(\rho(x))$$

- Wasserstein gradient is

$$\operatorname{grad} F(\rho) = -\nabla \cdot \left(\rho \nabla \frac{\delta F}{\delta \rho} \right) = -\nabla \cdot (\rho \nabla U'(\rho))$$

Entropy as internal energy

Negative entropy:

$$F(\rho) = -H(\rho) = \int_{\mathbb{R}^n} \rho(x) \log \rho(x)\, dx$$

- Internal energy with $U(r) = r \log r$, $U'(r) = \log r + 1$
- L^2-variation is

$$\frac{\delta F}{\delta \rho}(x) = U'(\rho(x)) = \log \rho(x) + 1$$

Gradient of entropy

$$F(\rho) = -H(\rho) = \int_{\mathbb{R}^n} \rho(x) \log \rho(x) \, dx$$

Lemma: Gradient of entropy is Laplacian

$$\operatorname{grad} F(\rho) = -\Delta \rho$$

Proof:

$$= -\nabla \cdot (\rho \nabla U'(\rho))$$

$$\begin{aligned} \operatorname{grad} F(\rho) &= -\nabla \cdot (\rho \nabla (\log \rho + 1)) \\ &= -\nabla \cdot \left(\rho \frac{\nabla \rho}{\rho} \right) \\ &= -\nabla \cdot (\nabla \rho) \\ &= -\Delta \rho \end{aligned}$$

$$\nabla \log \rho = \frac{\nabla \rho}{\rho}$$

$\square$

Gradient flow of entropy is heat equation

$$F(\rho) = -H(\rho) = \int_{\mathbb{R}^n} \rho(x) \log \rho(x) \, dx$$

Theorem: Gradient flow for minimizing negative entropy $F(\rho)$ ($\Leftrightarrow$ for maximizing entropy $H(\rho)$) is the **heat equation**

$$\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$$

<u>Proof:</u> Gradient flow for minimizing F is

$$\frac{\partial \rho_t}{\partial t} = \dot{\rho}_t = -\mathrm{grad}\, F(\rho_t) = \Delta \rho_t$$

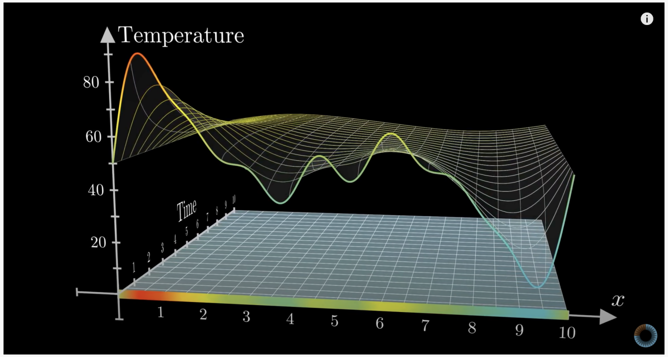
$\square$

Heat equation

Heat equation $\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$ is a PDE for $\rho(t,x) = \rho_t(x)$

$$\frac{\partial \rho}{\partial t}(t,x) = \sum_{i=1}^{n} \frac{\partial^2 \rho}{\partial x_i^2}(t,x)$$

- Modeling diffusion of heat



3Blue1Brown, *But what is a partial differential equation?*, Youtube, 2019,
https://www.youtube.com/watch?v=ly4S0oi3Yz8

Solution to heat equation

Theorem: The solution to the heat equation

$$\frac{\partial \rho_t}{\partial t}(x) = \Delta \rho_t(x)$$

is given by convolution with Gaussian density (heat kernel):

$$\rho_t = \rho_0 * \mathcal{N}(0, 2tI)$$

$$\rho_t(x) = \frac{1}{(4\pi t)^{n/2}} \int_{\mathbb{R}^n} \rho_0(y) \exp\left(-\frac{\|x-y\|^2}{4t}\right) dy$$

<u>Proof:</u> Compute $\frac{\partial \rho_t}{\partial t}$ and $\Delta \rho_t$, check both are equal $\square$

Probabilistic interpretation

Theorem: The solution to the heat equation

$$\frac{\partial \rho_t}{\partial t}(x) = \Delta \rho_t(x)$$

is given by convolution with Gaussian density (heat kernel):

$$\rho_t = \rho_0 * \mathcal{N}(0, 2tI)$$

- If $X_0 \sim \rho_0$, can generate $X_t \sim \rho_t$ via

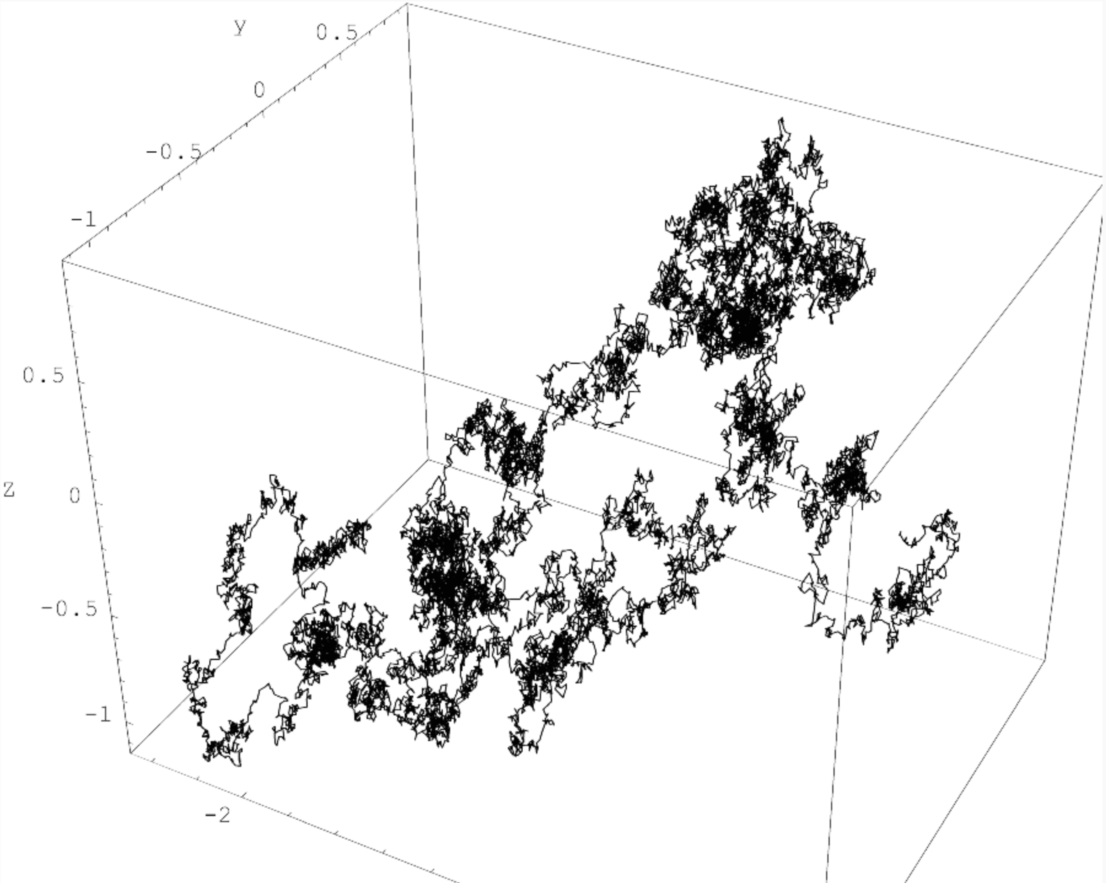
  $$X_t = X_0 + \sqrt{2t}\, Z$$

  where $Z \sim \mathcal{N}(0, I)$ is independent of X_0
- Can also generate $(X_t)_{t \geq 0}$ via *Brownian motion*

Brownian Motion

Brownian motion

- Brown (1828): *"pollen grains suspended in water perform a continual swarming motion"*



Brownian motion

- Bachelier (1900): Model fluctuations in stock prices
- Einstein (1905): Model fluctuations of particles from random collisions of atoms
  - Mathematical basis for the atomic theory of matters
  - Perrin (1908): Experiment to compute Avogadro's number
    $\Rightarrow$ Nobel prize in Physics (1926)
- Black & Scholes (1973): Risk-neutral option pricing using geometric Brownian motion (GBM)

  $\Rightarrow$ Nobel prize in Economics (1997)

Brownian motion

Standard **Brownian motion** (*Wiener process*) $(W_t)_{t \geq 0}$ in $\mathbb{R}^n$:

- $W_0 = 0$
- Independent increments: If $t_0 < t_1 < t_2 < \cdots$ then $W_{t_0}, W_{t_1} - W_{t_0}, W_{t_2} - W_{t_1}, \ldots$ are independent
- Gaussian increments: $W_t - W_s \sim \mathcal{N}(0, (t-s)I)$ for all $s < t$
- Continuous path: $t \mapsto W_t$ is continuous

[Durrett, *Probability: Theory and Examples*, Cambridge University Press, 2019]

[Evans, *An Introduction to Stochastic Differential Equations*, 2003]

Brownian motion

Write Brownian motion as Stochastic Differential Equation (SDE):

$$dX_t = dW_t$$

This means

$$X_t = X_0 + \int_0^t dW_s = X_0 + W_t$$

where $(W_t)_{t \geq 0}$ is standard Brownian motion independent of X_0

$W_t \sim \mathcal{N}(0, tI)$

- If $X_0 \sim \rho_0$, then $X_t \sim \rho_t = \rho_0 * \mathcal{N}(0, tI)$
- ρ_t satisfies the heat equation

$$\frac{\partial \rho_t}{\partial t} = \frac{1}{2} \Delta \rho_t$$

Recap: Gradient flow of entropy

Entropy:

$$H(\rho) = \mathbb{E}_\rho[\log \rho] = \int_{\mathbb{R}^n} \rho(x) \log \rho(x)\, dx$$

Gradient flow w.r.t. W_2 is heat equation: in $\mathcal{P}(\mathbb{R}^n)$

$$\frac{\partial \rho_t}{\partial t}(x) = \Delta \rho_t(x)$$

$$\rho_t = \rho_0 * \mathcal{N}(0, 2tI)$$

This is implemented by Brownian motion: in $\mathbb{R}^n$

$$dX_t = \sqrt{2}\, dW_t$$

$$X_t = X_0 + \sqrt{2}\, W_t$$

$$\overset{d}{=} X_0 + \sqrt{2t}\, Z, \quad Z \sim \mathcal{N}(0, I)$$

Gradient descent of entropy

Gradient descent of entropy

$$\rho_{k+1} = \mathrm{Exp}_{\rho_k}(\eta \, \mathrm{grad}\, H(\rho_k))$$

- $\mathrm{grad}\, H(\rho_k) = \nabla \cdot (\rho_k \nabla \log \rho_k)$
- **Lemma:** Assume ρ_k is K-log-semiconcave for some $K \in \mathbb{R}$:

  $$-\nabla^2 \log \rho_k \succeq KI$$

  For $0 < \eta \leq \frac{1}{\max\{0, -K\}}$, gradient descent of entropy is

  $$\rho_{k+1} = (I - \eta \nabla \log \rho_k)_{\#} \rho_k$$

  which is implemented by

  $$x_{k+1} = x_k - \eta \nabla \log \rho_k(x_k)$$

- Requires knowing density ρ_k

Gradient descent of entropy with Gaussian data

- If $\rho_0 = \mathcal{N}(\mu_0, \Sigma_0)$, then $\rho_k = \mathcal{N}(\mu_k, \Sigma_k)$ stays Gaussian

$$-\nabla \log \rho_k(x) = \Sigma_k^{-1}(x - \mu_k)$$

- Gradient descent of entropy becomes

$$\begin{aligned} x_{k+1} &= x_k - \eta \nabla \log \rho_k(x_k) \\ &= (I + \eta \Sigma_k^{-1}) x_k - \eta \Sigma_k^{-1} \mu_k \end{aligned}$$

$\hookrightarrow \mathbb{E} \Rightarrow \mu_{k+1} = (I + \cancel{\eta \Sigma_k^{-1}}) \mu_k - \cancel{\eta \Sigma_k^{-1} \mu_k}$
$= \mu_k$

- Therefore, $\mu_k = \mu_0$ and

$$\begin{aligned} \Sigma_{k+1} &= \Sigma_k (I + \eta \Sigma_k^{-1})^2 \\ &= \Sigma_k + 2\eta I + \eta^2 \Sigma_k^{-1} \end{aligned}$$

$> \Sigma_k + 2\eta I$

- Covariance grows faster than along heat equation

Proximal method of entropy

- Proximal method of entropy:

$$\rho_{k+1} = \arg \min_{\rho \in \mathcal{P}(\mathbb{R}^n)} \left\{ -H(\rho) + \frac{1}{2\eta} W_2(\rho, \rho_k)^2 \right\}$$

- Cannot implement except in special cases, e.g. Gaussian data
- If $\rho_0 = \mathcal{N}(\mu_0, \Sigma_0)$, then $\rho_k = \mathcal{N}(\mu_k, \Sigma_k)$ where $\mu_k = \mu_0$ and

$$\Sigma_{k+1} = \Sigma_k + 2\eta I - \eta^2 \Sigma_k^{-1} + O(\eta^3)$$

$$< \Sigma_k + 2\eta I$$

- Covariance grows slower than along heat equation

Fisher information

Entropy along Heat Equation

Theorem (de Bruijn's identity): Along the heat equation

$$\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$$

entropy is increasing:

$$\frac{d}{dt} H(\rho_t) = J(\rho_t) > 0$$

where $J(\rho)$ is the **Fisher information**

$$J(\rho) = \mathbb{E}_\rho[\|\nabla \log \rho\|^2]$$

$= \nabla \cdot (\nabla \rho)$

Proof: Since $\Delta\rho = \nabla \cdot (\rho \nabla \log \rho)$, by integration by parts,

$$\begin{aligned}
\frac{d}{dt} H(\rho_t) &= -\frac{d}{dt} \int_{\mathbb{R}^n} \rho_t(x) \log \rho_t(x)\, dx \\
&= -\int_{\mathbb{R}^n} \frac{\partial \rho_t}{\partial t}(x) \log \rho_t(x)\, dx \underbrace{- \int_{\mathbb{R}^n} \cancel{\rho_t(x)} \overbrace{\frac{1}{\cancel{\rho_t(x)}} \frac{\partial \rho_t}{\partial t}(x)}^{\frac{\partial}{\partial t} \log \rho_t(x)}\, dx}_{\substack{= \int_{\mathbb{R}^n} \frac{\partial \rho_t}{\partial t}(x)\, dx \\ = \frac{d}{dt} \int_{\mathbb{R}^n} \rho_t(x)\, dx \\ = \frac{d}{dt} 1 = 0}} \\
&= -\int_{\mathbb{R}^n} \Delta \rho_t(x) \log \rho_t(x)\, dx \\
&= \int_{\mathbb{R}^n} \langle \rho_t(x) \nabla \log \rho_t(x), \nabla \log \rho_t(x) \rangle\, dx \\
&= \int_{\mathbb{R}^n} \rho_t(x) \|\nabla \log \rho_t(x)\|^2\, dx \\
&= J(\rho_t)
\end{aligned}$$

$\square$

Fisher information

Fisher information:

$$J(\rho) = \mathbb{E}_\rho[\|\nabla \log \rho\|^2] = \int_{\mathbb{R}^n} \rho(x) \|\nabla \log \rho(x)\|^2 \, dx$$

- Gaussian: $\rho = \mathcal{N}(\mu, \Sigma)$

$$J(\rho) = \mathbb{E}_\rho[\|\Sigma^{-1}(x-\mu)\|^2] = \mathrm{Tr}(\Sigma^{-1})$$

- Uncertainty principle / Cramer-Rao lower bound:

$$J(\rho) \cdot \mathrm{Var}(\rho) \geq n^2$$

- Related to Fisher information matrix for parameterized distribution

[Wibisono, Jog, & Loh, *Information and estimation in Fokker-Planck channels*, ISIT 2017]

Fisher information and entropy

Lemma: Fisher information is squared norm of gradient of entropy:

$$J(\rho) = \|\operatorname{grad} H(\rho)\|_\rho^2$$

<u>Proof:</u> Gradient of entropy is Laplacian:

$$\operatorname{grad} H(\rho) = \Delta\rho = \nabla \cdot (\rho \nabla \log \rho)$$

By definition of Wasserstein metric:

if $\phi = -\nabla\cdot(\rho\nabla u)$

$$\|\operatorname{grad} H(\rho)\|_\rho^2 = \mathbb{E}_\rho[\|\nabla \log \rho\|^2] = J(\rho)$$

then

$\|\phi\|_\rho^2 = \mathbb{E}_\rho[\|\nabla u\|^2]$

$\square$

Optimization interpretation of de Bruijn's identity

de Bruijn's identity along heat equation:

$$\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$$
$$\Rightarrow \quad \tfrac{d}{dt} H(\rho_t) = J(\rho_t)$$

is instance of abstract identity along gradient flow to maximize H:

$$\dot{\rho}_t = \mathrm{grad}\, H(\rho_t)$$
$$\Rightarrow \quad \tfrac{d}{dt} H(\rho_t) = \|\mathrm{grad}\, H(\rho_t)\|^2_{\rho_t}$$

$$= \langle \mathrm{grad}\, H(\rho_t), \dot{\rho}_t \rangle_{\rho_t}$$
$$= \langle \mathrm{grad}\, H(\rho_t), \mathrm{grad}\, H(\rho_t) \rangle_{\rho_t}$$
$$= \| \mathrm{grad}\, H(\rho_t) \|^2_{\rho_t}$$

Convergence rate of entropy

Upper bound 1

Lemma: Let $\Sigma_0 = \mathrm{Cov}(\rho_0)$. Along the heat equation $\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$,

$$H(\rho_t) \le \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log\det(\Sigma_0 + 2tI) \sim \frac{n}{2}\log t$$

- Covariance increases linearly: $\mathrm{Cov}(\rho_t) = \mathrm{Cov}(\rho_0) + 2tI = \Sigma_0 + 2tI$
- Gaussian is maximum entropy distribution given covariance, so $H(\rho_t) \le H(\mathcal{N}(0, \mathrm{Cov}(\rho_t)))$

Upper bound 2

A better bound with correct dependence at $t = 0$

Lemma: Along the heat equation $\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$,

$$H(\rho_t) \le H(\rho_0) + \frac{n}{2} \log\left(1 + \frac{2t}{n} J(\rho_0)\right) \sim \frac{n}{2} \log t$$

- From relation between first and second derivatives of entropy
- Equality achieved by Gaussian
- Entropy increases *at most* logarithmically

Lower bound

Lemma: Along the heat equation $\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$,

$$H(\rho_t) \geq H(\mathcal{N}(0, 2tI)) = \frac{n}{2} \log(4\pi e t) \sim \frac{n}{2} \log t$$

- From mutual information $I(X_0; X_t) = H(X_t) - H(X_t \mid X_0) \geq 0$
- Entropy increases *at least* logarithmically

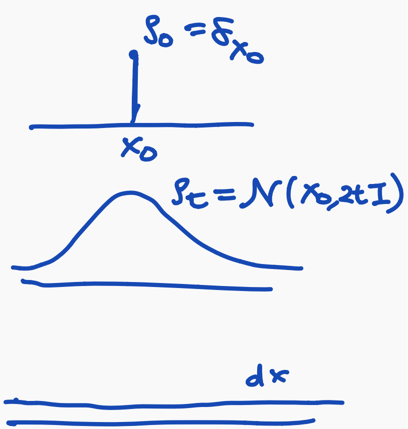
Convergence rate of entropy

Conclusion: Along the heat equation

$$\frac{\partial \rho_t}{\partial t} = \Delta \rho_t \quad \Longleftrightarrow \quad \dot{\rho}_t = \text{grad } H(\rho_t)$$

entropy is increasing at logarithmic rate as $t \to \infty$:

for any ρ_0 :
$$H(\rho_t) = \Theta\left(\frac{n}{2}\log t\right)$$

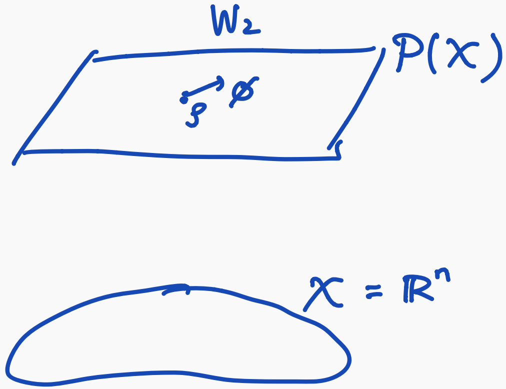


Solution: $\rho_t = \rho_0 * \mathcal{N}(0, 2t\, I)$

$\rho_t \longrightarrow$ Lebesgue measure dx as $t \to \infty$

$$\rho_t = \rho_0 * \mathcal{N}(0, 2tI)$$

ρ_0



Displacement convexity of entropy

Hessian of entropy

$$F(\rho) = -H(\rho) = \mathbb{E}_\rho[\log \rho]$$

Lemma: Hessian of negative entropy is a quadratic form

$$\mathrm{Hess}\, F(\rho) \colon \mathrm{T}_\rho \mathcal{P} \times \mathrm{T}_\rho \mathcal{P} \to \mathbb{R}$$

that sends $\phi = -\nabla \cdot (\rho \nabla u) \in \mathrm{T}_\rho \mathcal{P}$ to

$$\frac{d^2}{dt^2} F(\rho_t)\Big|_{t=0} = \mathrm{Hess}\, F(\rho)(\phi, \phi) = \mathbb{E}_\rho[\|\nabla^2 u\|_{\mathsf{HS}}^2] = \int_{\mathbb{R}^n} \rho(x)\, \|\nabla^2 u(x)\|_{\mathsf{HS}}^2\, dx$$

ρ_t = geodesic from $\rho_0 = \rho$ along $\dot{\rho}_0 = \phi$

- In particular, $\mathrm{Hess}\, F(\rho)(\phi, \phi) \geq 0$ for all ϕ, denoted $\mathrm{Hess}\, F(\rho) \succeq 0$

$\|A\|_{HS}^2 = \mathrm{Tr}(AA^T)$

Convexity of negative entropy

Lemma: $F(\rho) = -H(\rho) = \mathbb{E}_\rho[\log \rho]$ is displacement convex (geodesically convex in W_2 metric)

Proof: Hessian is non-negative: $\mathrm{Hess}\, F(\rho) \succeq 0$ $\square$

- $F = -H$ is not strictly convex in general
- $F = -H$ is strongly convex along geodesics with constant mean and satisfying Poincaré inequality

[Carlen & Gangbo, *Constrained steepest descent in the 2-Wasserstein metric*, Annals of Mathematics, 2003]

Second-order Fisher information

Second-order Fisher information:

$$K(\rho) = \mathbb{E}_\rho \left[\|\nabla^2 \log \rho\|_{\mathsf{HS}}^2 \right]$$

- Example: $\rho = \mathcal{N}(\mu, \Sigma)$

$$K(\rho) = \|\Sigma^{-1}\|_{\mathsf{HS}}^2 = \mathsf{Tr}(\Sigma^{-2})$$

- Hessian of entropy along gradient $\mathrm{grad}\, H(\rho) = \nabla \cdot (\rho \nabla \log \rho)$

$$K(\rho) = \mathrm{Hess}\, F(\rho_t)(\mathrm{grad}\, H(\rho_t),\, \mathrm{grad}\, H(\rho_t))$$

Acceleration of entropy along heat equation

Lemma: Along heat equation

$$\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$$

acceleration of entropy is

$$\frac{d^2}{dt^2} H(\rho_t) = -2K(\rho_t)$$

- Follows from differentiation and integration by parts
- Instance of abstract gradient flow identity

$$\frac{d^2}{dt^2} H(\rho_t) = 2\,\mathrm{Hess}\, H(\rho_t)(\mathrm{grad}\, H(\rho_t),\, \mathrm{grad}\, H(\rho_t))$$

[Villani, *A short proof of the concavity of entropy power*, IEEE Transactions on Information Theory, 2000]

Entropy and Fisher information

Along heat equation: $\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$

$$\frac{d}{dt} H(\rho_t) = J(\rho_t)$$
$$\frac{d}{dt} J(\rho_t) = -2K(\rho_t)$$

Entropy: $H(\rho) = \mathbb{E}_\rho[\log \rho]$

Fisher information: $J(\rho) = \mathbb{E}_\rho[\|\nabla \log \rho\|^2]$

Second-order Fisher information: $K(\rho) = \mathbb{E}_\rho\left[\|\nabla^2 \log \rho\|_{\mathsf{HS}}^2\right]$

$K(\rho) \gtrsim \frac{J(\rho)^2}{n} \Rightarrow$ Convergence Rate $H(\rho_t) \lesssim \frac{n}{2} \log t$

Recap: Optimization of entropy

$$H(\rho) = -\mathbb{E}_\rho[\log \rho] = -\int_{\mathbb{R}^n} \rho(x) \, \log \rho(x) \, dx$$

- Geodesically concave in Wasserstein metric
- Gradient flow is the heat equation: $\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$

  $\Rightarrow$ Can be implemented by Brownian motion: $dX_t = \sqrt{2}\, dW_t$
- Gradient descent, proximal method cannot be implemented
  - except in special cases, e.g. with Gaussian data
  - Other cases / approximations?

Heat equation as gradient flow of Dirichlet energy

The heat equation

$$\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$$

can also be interpreted as the gradient flow of the *Dirichlet energy*:

$$\mathcal{E}(\rho) = \int_{\mathbb{R}^n} \|\nabla \rho(x)\|^2 \, dx$$

with respect to $L^2(\mathbb{R}^n, dx)$ structure