CPSC 661: Sampling Algorithms in ML

Andre Wibisono

April 7, 2021

Yale University

Last time

- Wasserstein W_2 metric
- Otto calculus
- Potential energy
- Brownian motion and Entropy

Today: Langevin Dynamics and Relative Entropy

References

- Jordan, Kinderlehrer, & Otto, *The variational formulation of the Fokker-Planck equation*, SIAM Journal on Mathematical Analysis, 1998
- Evans, *An introduction to stochastic differential equation*, AMS, 2013
- Villani, *Topics in Optimal Transportation*, Springer, 2003
- Villani, *Optimal Transport: Old and New*, Springer, 2008
- Wibisono, *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem*, COLT 2018

Langevin Dynamics

Target distribution

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $\int_{\mathbb{R}^n} e^{-f(x)}\, dx < \infty$

Let ν be a probability distribution on $\mathbb{R}^n$ with density $\nu \propto e^{-f}$

$$\nu(x) = \frac{e^{-f(x)}}{\int_{\mathbb{R}^n} e^{-f(y)}\, dy}$$

- f quadratic $\quad\Leftrightarrow\quad$ ν Gaussian
- f convex $\quad\Leftrightarrow\quad$ ν log-concave
- f α-strongly convex $\quad\Leftrightarrow\quad$ ν α-strongly log-concave

Langevin Dynamics

The **Langevin Dynamics** for $\nu \propto e^{-f}$ is the stochastic process $(X_t)_{t\geq 0}$ in $\mathbb{R}^n$ following the stochastic differential equation:

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$$

where $(W_t)_{t\geq 0}$ is the standard Brownian motion in $\mathbb{R}^n$

- Depends on ν via gradient of log-density, doesn't need normalization constant

$$dX_t = \nabla \log \nu(X_t)\,dt + \sqrt{2}\,dW_t$$

Langevin Dynamics

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$$

A mixture of:

1. Gradient flow:

$$\frac{dX_t}{dt} = \dot{X}_t = -\nabla f(X_t)$$

   Converges to a point: $X_t \to x^* = \arg\min_{x \in \mathbb{R}^n} f(x)$

2. Brownian motion:

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

   Diverges via Gaussian noise: $X_t \overset{d}{=} X_0 + \sqrt{2t}\,Z, \quad Z \sim \mathcal{N}(0, I)$

Langevin Dynamics

Fact: The stationary distribution of the Langevin dynamics

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$$

is

$$\nu(x) \propto e^{-f(x)}$$

- If $X_0 \sim \nu$, then along Langevin dynamics, $X_t \sim \nu$ for all $t > 0$. In this case $(X_t)_{t\geq 0}$ is a *stationary process*.
- ν is *attracting*: For any $X_0 \sim \rho_0$, along Langevin dynamics,

$$X_t \sim \rho_t \to \nu \quad \text{as } t \to \infty$$

Fokker-Planck Equation

Lemma: If X_t follows the Langevin Dynamics:

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$$

then the density $X_t \sim \rho_t$ follows the **Fokker-Planck equation**:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t$$

A mixture of:

1. The continuity equation $\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f)$ of gradient flow $\dot{X}_t = -\nabla f(X_t)$
2. The heat equation $\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$ of Brownian motion $dX_t = \sqrt{2}\,dW_t$

Fokker-Planck equation

Itô integral

Let $(W_t)_{t\geq 0}$ be the standard Brownian motion in $\mathbb{R}^n$

An **Itô integral** is an expression of the form

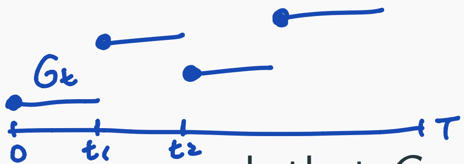
$$\int_0^T G_t \, dW_t \quad \in \mathbb{R}^m$$

for some stochastic process $G_t \in \mathbb{R}^{m\times n}$ which is progressively measurable (G_t depends on the past $(0, t)$) and $\int_0^T \|G_t\|^2 \, dt < \infty$

- This is a random variable in $\mathbb{R}^m$

Itô integral: Definition

- Suppose G_t is a *step process*: There exists a partition



$$0 = t_0 < t_1 < \cdots < t_K = T$$

such that $G_t \equiv G_{t_k}$ for $t_k \leq t \leq t_{k+1}$. Then by definition,

$$\int_0^T G_t \, dW_t := \sum_{k=0}^{K-1} G_{t_k} (\underbrace{W_{t_{k+1}} - W_{t_k}}_{\int_{t_k}^{t_{k+1}} dW_t})$$

- For general G_t, approximate by step processes $G_t^{(\ell)}$ and define

$$\int_0^T G_t \, dW_t := \lim_{\ell \to \infty} \int_0^T G_t^{(\ell)} \, dW_t$$

[Evans, *An introduction to stochastic differential equation*, AMS, 2013]

Itô integral: Properties

1. Linear:

$$\int_0^T (aG_t + bH_t)\, dW_t = a \int_0^T G_t\, dW_t + b \int_0^T H_t\, dW_t$$

2. Zero mean:

$$\mathbb{E}\left[\int_0^T G_t\, dW_t\right] = 0$$

3. Variance:

$$\mathbb{E}\left[\left\| \int_0^T G_t\, dW_t \right\|_2^2\right] = \int_0^T \mathbb{E}[\|G_t\|_{\mathrm{HS}}^2]\, dt$$

where $\|G\|_{\mathrm{HS}}^2 = \mathrm{Tr}(GG^\top) = \sum_{i=1}^n \sum_{j=1}^m G_{ij}^2$.

Stochastic differential equation

Definition: A stochastic process $(X_t)_{t\geq 0}$ in $\mathbb{R}^n$ follows the **stochastic differential equation** (SDE)

$$dX_t = v(X_t)\,dt + G(X_t)\,dW_t$$

for some drift vector field $v\colon \mathbb{R}^n \to \mathbb{R}^n$ and (square-root) covariance matrix $G\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$, if

$$X_T = X_0 + \int_0^T v(X_t)\,dt + \int_0^T G(X_t)\,dW_t$$

for all $T > 0$, where the last term is Itô integral

Fokker-Planck equation

Lemma: Suppose X_t $\in \mathbb{R}^n$ follows the SDE:

G = square root of covariance

$$dX_t = v(X_t)\, dt + G(X_t)\, dW_t \qquad GG^\top = \text{covariance}$$

for some $v\colon \mathbb{R}^n \to \mathbb{R}^n$ and $G\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$ differentiable.

$Z \sim \mathcal{N}(0, I)$

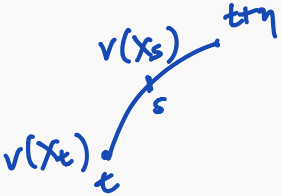
$GZ \sim \mathcal{N}(0, GG^\top)$

Then the density $X_t \sim \rho_t$ follows the **Fokker-Planck equation**:

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v) + \frac{1}{2}\langle \nabla^2, \rho_t GG^\top \rangle_{\mathrm{HS}}$$

- Also known as the *forward Kolmogorov equation*
- This is the continuity equation (evolution of density) of SDE
- $\langle \nabla^2, A\rangle_{\mathrm{HS}}(x) = \sum_{i,j=1}^n \frac{\partial^2}{\partial x_i \partial x_j} A_{ij}(x)$

Proof: By definition, for all $t > 0$ and small $\eta > 0$



$$
\begin{aligned}
X_{t+\eta} &= X_t + \int_t^{t+\eta} v(X_s)\, ds + \int_t^{t+\eta} G(X_s)\, dW_s \\
&\approx X_t + \int_t^{t+\eta} v(X_t)\, ds + \int_t^{t+\eta} G(X_t)\, dW_s \\
&= X_t + \eta v(X_t) + G(X_t)\underbrace{(W_{t+\eta} - W_t)}_{\sim \mathcal{N}(0,\, \eta I)}
\end{aligned}
$$

Write $W_{t+\eta} - W_t = \sqrt{\eta} Z$ where $Z \sim \mathcal{N}(0, I)$ independent of X_t. Then

$$X_{t+\eta} \overset{d}{=} X_t + \eta v(X_t) + \sqrt{\eta}\, G(X_t)\, Z + o(\eta)$$

- Note randomness scales as square root of time: $dW \approx \sqrt{dt}$

Given $h : \mathbb{R}^n \to \mathbb{R}$, compute $\frac{d}{dt} \mathbb{E}[h(X_t)]$

Let $u_t = G(X_t)Z$ and $v_t = v(X_t)$, so

$$X_{t+\eta} = X_t + \sqrt{\eta}\, u_t + \eta\, v_t + o(\eta)$$

Note $\mathbb{E}[u_t] = 0$ and $\mathbb{E}[u_t u_t^\top] = \mathbb{E}[G(X_t)G(X_t)^\top]$

For any test function $h\colon \mathbb{R}^n \to \mathbb{R}$, by second-order Taylor expansion,

$$\begin{aligned} h(X_{t+\eta}) &= h(X_t + \sqrt{\eta} u_t + \eta v_t + o(\eta)) \\ &= h(X_t) + \sqrt{\eta}\,\langle \nabla h(X_t), u_t\rangle + \eta \langle \nabla h(X_t), v_t\rangle \\ &\qquad + \frac{1}{2}\eta\langle u_t, \nabla^2 h(X_t)\, u_t\rangle + o(\eta) \end{aligned}$$

$$h(x+a) = h(x) + \langle \nabla h(x), a\rangle + \frac{1}{2}\langle a, \nabla^2 h(x) a\rangle + o(\|a\|^2)$$

$$a = \sqrt{\eta}\, u + \eta v$$

$$\langle a, \nabla^2 h(x)\, a\rangle = (\sqrt{\eta}\, u + \eta v)^\top \nabla^2 h(x)(\sqrt{\eta}\, u + \eta v)$$

$$= \eta\, u^\top \nabla^2 h(x) u + 2\eta^{3/2} u^\top \nabla^2 h(x) v + \eta^2 v^\top \nabla^2 h(x) v$$

$= o(\eta)$ $= o(\eta)$

Taking expectation:

$$\begin{aligned}\mathbb{E}[h(X_{t+\eta})] = \mathbb{E}[h(X_t)] + \eta\,\mathbb{E}[\langle \nabla h(X_t), {\color{blue}v(X_t)}\rangle] \\ + \frac{\eta}{2}\mathbb{E}[\langle \nabla^2 h(X_t), {\color{red}G(X_t)G(X_t)^\top}\rangle_{\mathrm{HS}}] + o(\eta)\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{d}{dt}\mathbb{E}[h(X_t)] &= \lim_{\eta\to 0}\frac{\mathbb{E}[h(X_{t+\eta})] - \mathbb{E}[h(X_t)]}{\eta} \\ &= \mathbb{E}\left[\langle \nabla h(X_t), {\color{blue}v(X_t)}\rangle + \frac{1}{2}\langle \nabla^2 h(X_t), {\color{red}G(X_t)G(X_t)^\top}\rangle_{\mathrm{HS}}\right] \\ &= \int_{\mathbb{R}^n}\left(\langle \nabla h(x), {\color{blue}v(x)}\rangle + \frac{1}{2}\langle \nabla^2 h(x), {\color{red}G(x)G(x)^\top}\rangle_{\mathrm{HS}}\right)\rho_t(x)dx \\ &= \int_{\mathbb{R}^n} h(x)\left(-\nabla\cdot(\rho_t {\color{blue}v})(x) + \frac{1}{2}\langle \nabla^2, \rho_t {\color{red}GG^\top}\rangle_{\mathrm{HS}}(x)\right)dx\end{aligned}$$

$$= \frac{d}{dt} \int_{\mathbb{R}^n} h(x)\, \rho_t(x)\, dx$$

On the other hand,

$$\frac{d}{dt}\mathbb{E}[h(X_t)] = \int_{\mathbb{R}^n} h(x) \frac{\partial \rho_t}{\partial t}(x)\, dx$$

Therefore,

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v) + \frac{1}{2}\langle \nabla^2, \rho_t G G^\top \rangle_{\mathrm{HS}}$$

which is the Fokker-Planck equation. $\square$

Fokker-Planck Equation

SDE: $dX_t = v(X_t)\,dt + G(X_t)\,dW_t$

$\Rightarrow$ Fokker-Planck equation: $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v) + \frac{1}{2}\langle \nabla^2, \rho_t G G^\top \rangle_{\mathsf{HS}}$

Fokker-Planck Equation

SDE: $dX_t = v(X_t)\,dt + G(X_t)\,dW_t$

$\Rightarrow$ Fokker-Planck equation: $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v) + \frac{1}{2}\langle \nabla^2, \rho_t G G^\top \rangle_{\mathrm{HS}}$

1. $G = 0$: Deterministic dynamics $\dot{X}_t = v(X_t)$

   $\Rightarrow$ Continuity equation:

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v)$$

Fokker-Planck Equation

SDE: $dX_t = v(X_t)\,dt + G(X_t)\,dW_t$

$\Rightarrow$ Fokker-Planck equation: $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v) + \frac{1}{2}\langle \nabla^2, \rho_t G G^\top \rangle_{\text{HS}}$

1. $G = 0$: Deterministic dynamics $\dot{X}_t = v(X_t)$

   $\Rightarrow$ Continuity equation:

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v)$$

2. $v = 0$, $G = \sqrt{2} I$: Brownian motion $dX_t = \sqrt{2} dW_t$

   $\Rightarrow$ Heat equation:

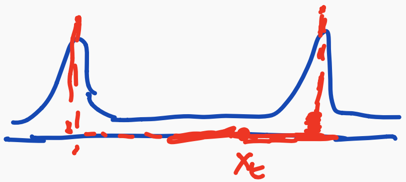
$$\frac{\partial \rho_t}{\partial t} = \underbrace{\frac{1}{2}\langle \nabla^2, \rho_t 2I \rangle_{\text{HS}}}_{= \sum_{i,j=1}^{n} \frac{\partial^2}{\partial x_i x_j} \left(\rho_t(x) \overbrace{I_{ij}}^{= \begin{cases} 1 & \text{if } i=j \\ 0 & \text{else} \end{cases}} \right) = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} \rho_t(x)} = \Delta \rho_t$$

Langevin Dynamics

Langevin dynamics

Lemma: If X_t follows the Langevin Dynamics for $\nu \propto e^{-f}$:

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$$



then the density $X_t \sim \rho_t$ follows the **Fokker-Planck equation**:

$$\begin{aligned}\frac{\partial \rho_t}{\partial t} &= \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t \\ &= \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu} \right)\end{aligned}$$

In particular, $\rho_t = \nu$ is a stationary solution.

Proof: The first line follows from general Fokker-Planck equation with $v(x) = -\nabla f(x)$ and $G(x) = \sqrt{2}I$.

The second line follows since $\nu \propto e^{-f} \Rightarrow \nabla f = -\nabla \log \nu$:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t$$

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

$$= -\nabla \cdot (\rho_t \nabla \log \nu) + \nabla \cdot (\rho_t \nabla \log \rho_t)$$

$$= \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu} \right)$$

In particular, if $\rho_t = \nu$, then $\nabla \log \frac{\rho_t}{\nu} = \nabla \log 1 = 0$, so $\frac{\partial \rho_t}{\partial t} = 0$.

$\square$

Relative entropy

Relative entropy

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

Relative entropy with respect to ν is $H_\nu \colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ given by

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

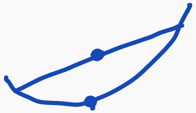
- Also called Kullback-Leibler (KL) divergence, denoted $\mathrm{KL}(\rho \,\|\, \nu)$
- Requires $\rho \ll \nu$, otherwise $H_\nu(\rho) = +\infty$
- Not a distance (not symmetric: $H_\nu(\rho) \neq H_\rho(\nu)$)
- But a good *divergence* to distinguish ρ from ν

Relative entropy

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

Lemma: $H_\nu(\rho) \geq 0$ for all $\rho \in \mathcal{P}(\mathbb{R}^n)$, and $H_\nu(\rho) = 0$ iff $\rho = \nu$

<u>Proof:</u> Let $h = \frac{\rho}{\nu}$, so $\mathbb{E}_\nu[h] = 1$. Then



$$H_\nu(\rho) = \mathbb{E}_\nu[h \log h] \geq (\mathbb{E}_\nu[h]) \log \mathbb{E}_\nu[h] = 1 \log 1 = 0$$

by Jensen's inequality for the convex function $r \mapsto r \log r$.

Equality holds if and only if $h \equiv 1$, or equivalently $\rho = \nu$. $\square$

Relative entropy inequalities

Recall for probability distributions ρ, ν on $\mathbb{R}^n$

$$0 \leq 2\mathrm{TV}(\rho,\nu)^2 \leq H_\nu(\rho) \leq \chi^2_\nu(\rho)$$

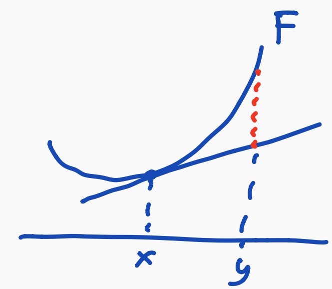
where

- $\mathrm{TV}(\rho,\nu) = \frac{1}{2}\int_{\mathbb{R}^n} \nu(x) \left| \frac{\rho(x)}{\nu(x)} - 1 \right| dx$ is total variation distance
- $H_\nu(\rho) = \int_{\mathbb{R}^n} \rho(x) \log \frac{\rho(x)}{\nu(x)}\, dx$ is relative entropy
- $\chi^2_\nu(\rho) = \int_{\mathbb{R}^n} \nu(x) \left(\frac{\rho(x)}{\nu(x)} - 1 \right)^2 dx$ is χ^2-divergence

Relative entropy as Bregman divergence

Relative entropy is also the **Bregman divergence** of negative entropy:

$$H_\nu(\rho) = -H(\rho) + H(\nu) + \left\langle \frac{\delta H}{\delta \nu}, \rho - \nu \right\rangle$$



$D_F(y,x) =$

$F(y) - F(x) - \langle \nabla F(x), y-x \rangle$

where

- $H_\nu(\rho) = \int_{\mathbb{R}^n} \rho(x) \log \frac{\rho(x)}{\nu(x)}\, dx$ is relative entropy
- $H(\rho) = -\int_{\mathbb{R}^n} \rho(x) \log \rho(x)\, dx$ is Shannon entropy
- $\frac{\delta H}{\delta \nu}(x) = -\log \nu(x) - 1$ is L^2-derivative
- Inner product is in $L^2(\mathbb{R}^n, dx)$

Since $\rho \mapsto -H(\rho)$ is convex in L^2, this also shows $H_\nu(\rho) \geq 0$

Decomposition of relative entropy

Let $\nu = e^{-f}$ be a probability distribution on $\mathbb{R}^n$, so $f = -\log \nu$

Decomposition of relative entropy into potential energy and entropy:

$$H_\nu(\rho) = \mathbb{E}_\rho[f] - H(\rho)$$

since indeed

$$\int_{\mathbb{R}^n} \rho(x) \log \frac{\rho(x)}{\nu(x)}\, dx = \int_{\mathbb{R}^n} \rho(x) f(x)\, dx + \int_{\mathbb{R}^n} \rho(x) \log \rho(x)\, dx$$

- Note: If $\nu \propto e^{-f}$, there is a constant term $\log \int_{\mathbb{R}^n} e^{-f(x)}\, dx$

Relative entropy along Langevin dynamics

Lemma: (de Bruijn's identity)

Along the Langevin dynamics for $\nu \propto e^{-f}$:

$$dX_t = -\nabla f(X_t)\,dt + \sqrt{2}\,dW_t$$

relative entropy $H_\nu(\rho_t)$ is decreasing:

$$\frac{d}{dt} H_\nu(\rho_t) = -J_\nu(\rho_t) \le 0$$

where $J_\nu(\rho)$ is the **relative Fisher information**:

$$J_\nu(\rho) = \mathbb{E}_\rho\left[\left\|\nabla \log \frac{\rho}{\nu}\right\|^2\right]$$

$\square$

Relative Fisher information

<u>Proof:</u> Fokker-Planck equation is

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu} \right)$$

By integration by parts,

$$\begin{aligned}
\frac{d}{dt} H_\nu(\rho_t) &= \frac{d}{dt} \int_{\mathbb{R}^n} \rho_t \log \frac{\rho_t}{\nu} \, dx \\
&= \int_{\mathbb{R}^n} \frac{\partial \rho_t}{\partial t} \log \frac{\rho_t}{\nu} \, dx \quad + \overbrace{\int_{\mathbb{R}^n} \rho_t \left(\frac{\partial}{\partial t} \log \frac{\rho_t}{\nu} \right) dx}^{=0} \\
&= \int_{\mathbb{R}^n} \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu} \right) \log \frac{\rho_t}{\nu} \, dx \\
&= - \int_{\mathbb{R}^n} \rho_t \left\langle \nabla \log \frac{\rho_t}{\nu}, \nabla \log \frac{\rho_t}{\nu} \right\rangle dx \\
&= -J_\nu(\rho_t)
\end{aligned}$$

$\square$

Wasserstein geometry of H_ν

Gradient of relative entropy

Recall the Wasserstein gradient of $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is

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

Lemma: The Wasserstein gradient of $H_\nu(\rho) = \int_{\mathbb{R}^n} \rho \log \frac{\rho}{\nu}\, dx$ is

$$\operatorname{grad} H_\nu(\rho) = -\nabla \cdot \left(\rho \nabla \log \frac{\rho}{\nu}\right)$$

Proof: L^2-derivative is $\underbrace{\frac{\delta H_\nu}{\delta \rho}} = \log \frac{\rho}{\nu} + 1$ □

$$\frac{\delta H_\nu}{\delta \rho}(x) = \frac{\partial H_\nu(\rho)}{\partial \rho(x)} = \frac{\partial}{\partial \rho(x)} \left(\rho(x) \log \frac{\rho(x)}{\nu(x)} \right)$$

Gradient flow of relative entropy

Theorem: The gradient flow dynamics of relative entropy:

$$\dot{\rho}_t = -\mathrm{grad}\, H_\nu(\rho_t)$$

is the Fokker-Planck equation: $\nu \propto e^{-f}$

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu} \right) = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t$$

which is implemented by the Langevin dynamics:

$$dX_t = -\nabla f(X_t)\, dt + \sqrt{2}\, dW_t$$

- Jordan, Kinderlehrer, & Otto, *The variational formulation of the Fokker-Planck equation*, SIAM Journal on Mathematical Analysis, 1998

Relative Fisher information and relative entropy

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

$$J_\nu(\rho) = \|\mathrm{grad}\, H_\nu(\rho)\|_\rho^2$$

Proof: Gradient of relative entropy is

$$\mathrm{grad}\, H_\nu(\rho) = -\nabla \cdot \left(\rho \nabla \log \frac{\rho}{\nu}\right)$$

if $\phi = -\nabla\cdot(\rho \nabla u)$
then
$\|\phi\|_\rho^2 = \mathbb{E}_\rho[\|\nabla u\|^2]$

By definition of Wasserstein metric:

$$\|\mathrm{grad}\, H_\nu(\rho)\|_\rho^2 = \mathbb{E}_\rho\left[\left\|\nabla \log \frac{\rho}{\nu}\right\|^2\right] = J_\nu(\rho)$$

$\square$

Optimization interpretation of de Bruijn's identity

de Bruijn's identity along Langevin dynamics for sampling from ν:

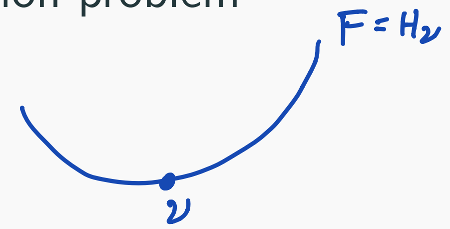
$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu} \right)$$
$$\Rightarrow \quad \tfrac{d}{dt} H_\nu(\rho_t) = -J_\nu(\rho_t)$$

is instance of abstract identity along gradient flow to minimize H_ν:

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

Sampling as optimization

Encode sampling from $\nu \in \mathcal{P}(\mathbb{R}^n)$ as an optimization problem



$$\min_{\rho \in \mathcal{P}(\mathbb{R}^n)} F(\rho)$$

for some $F \colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ which is minimized at ν.

Relative entropy $H_\nu \colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is a good objective function

- Minimized at ν: $H_\nu(\rho) \geq 0$ and $H_\nu(\nu) = 0$
- No local minima: $\|\mathrm{grad}\, H_\nu(\rho)\|_\rho^2 = J_\nu(\rho)$, so $\mathrm{grad}\, H_\nu(\rho) = 0$ if and only if $\rho = \nu$
- Can be optimized efficiently: Gradient flow is Langevin dynamics

Example: Ornstein-Uhlenbeck

Ornstein-Uhlenbeck process

Let $\nu = \mathcal{N}(\mu, \Sigma)$ so

$$-\log \nu(x) = \quad f(x) = \frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) + \frac{1}{2}\log\det(2\pi\Sigma)$$

$$\nabla f(x) = \Sigma^{-1}(x-\mu)$$

The Langevin dynamics for ν is known as the **Ornstein-Uhlenbeck process**:

$$dX_t = -\underbrace{\Sigma^{-1}(X_t-\mu)}_{\nabla f(X_t)}\, dt + \sqrt{2}\, dW_t$$

- Linear drift, can solve exactly
- Provides interpolation between any ρ_0 and Gaussian $\rho_\infty = \nu$

Ornstein-Uhlenbeck solution

Let $\nu = \mathcal{N}(0, \frac{1}{\alpha})$ on $\mathbb{R}^1$.

Lemma: The solution to the Ornstein-Uhlenbeck process

$$dX_t = -\alpha X_t \, dt + \sqrt{2}\, dW_t$$

is

$$X_t = e^{-\alpha t} X_0 + \sqrt{2} \int_0^t e^{-\alpha(t-s)} \, dW_s$$

$$\stackrel{d}{=} e^{-\alpha t} X_0 + \sqrt{\frac{1 - e^{-2\alpha t}}{\alpha}} Z$$

$$\lim_{\alpha \to 0} \frac{1 - e^{-2\alpha t}}{\alpha} = 2t$$

where $Z \sim \mathcal{N}(0,1)$ is independent of X_0.

- $\alpha \to 0$ recovers Brownian motion $dX_t = \sqrt{2}\, dW_t$, $X_t = X_0 + \sqrt{2t} Z$

<u>Proof:</u> Let $Y_t = e^{\alpha t} X_t$. Then

$$\begin{aligned} dY_t &= d(e^{\alpha t})\, X_t + e^{\alpha t}\, dX_t \\ &= e^{\alpha t} \alpha X_t\, dt + e^{\alpha t}(-\alpha X_t\, dt + \sqrt{2}\, dW_t) \\ &= \sqrt{2}\, e^{\alpha t}\, dW_t \end{aligned}$$

Therefore,

$$\begin{aligned} & Y_t = Y_0 + \sqrt{2} \int_0^t e^{\alpha s}\, dW_s \\ \Leftrightarrow \quad & X_t = e^{-\alpha t} X_0 + \sqrt{2} \int_0^t e^{-\alpha(t-s)}\, dW_s \end{aligned}$$

Note $\int_0^t e^{-\alpha(t-s)}\, dW_s$ is a Gaussian random variable with mean

$$\mathbb{E}\left[\int_0^t e^{-\alpha(t-s)}\, dW_s\right] = 0$$

and variance

$$\mathbb{E}\left[\left(\int_0^t e^{-\alpha(t-s)}\, dW_s\right)^2\right] = \int_0^t e^{-2\alpha(t-s)}\, ds = \frac{1 - e^{-2\alpha t}}{2\alpha}$$

Therefore, can write

$$\begin{aligned} X_t &= e^{-\alpha t} X_0 + \sqrt{2} \int_0^t e^{-\alpha(t-s)}\, dW_s \\ &\overset{d}{=} e^{-\alpha t} X_0 + \sqrt{\frac{1 - e^{-2\alpha t}}{\alpha}} Z \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$ is independent of X_0 $\square$

Ornstein-Uhlenbeck

Let $\nu = \mathcal{N}(0, \Sigma)$ on $\mathbb{R}^n$. Ornstein-Uhlenbeck is

$$dX_t = -\Sigma^{-1} X_t \, dt + \sqrt{2} \, dW_t$$

The solution is

$$\begin{aligned} X_t &= e^{-\Sigma^{-1} t} X_0 + \sqrt{2} \int_0^t e^{-\Sigma^{-1}(t-s)} \, dW_s \\ &\overset{d}{=} e^{-\Sigma^{-1} t} X_0 + \sqrt{\Sigma(1 - e^{-2\Sigma^{-1} t})} Z \end{aligned}$$

where $Z \sim \mathcal{N}(0, I)$ is independent of X_0.

- Observe: $X_t \xrightarrow{d} \nu = \mathcal{N}(0, \Sigma)$ exponentially fast
- Rate controlled by $\lambda_{\min}(\Sigma^{-1}) = 1/\lambda_{\max}(\Sigma)$ (also the strong log-concavity constant of ν)