~~Stochastic~~

Langevin dynamics

Suppose I want to sample from

$$\mu\ \cancel{p}(x) \propto \exp(H(x))$$

where $x \in \mathbb{R}^n$.

One option: Glauber dynamics.

Maybe good approach if $\mathcal{L}(x_i \mid x_{\sim i})$ is closed form.

Another option: Langevin dynamics.

Given by SDE: (Stochastic Differential Equation)

$$dX_t = \nabla H(X_t)\,dt + \sqrt{2}\,\underbrace{dB_t}_{\text{differential of Brownian motion}}$$

B_t in $\mathbb{R}^n$, $B_t - B_s \sim \mathcal{N}(0, t-s)$

What does this mean?

~~On~~ limit as $\varepsilon \to 0$ of discretized chain

$$X_{t+\varepsilon} = X_t + \varepsilon \nabla H(X_t) + \mathcal{N}(0, 2\varepsilon)$$

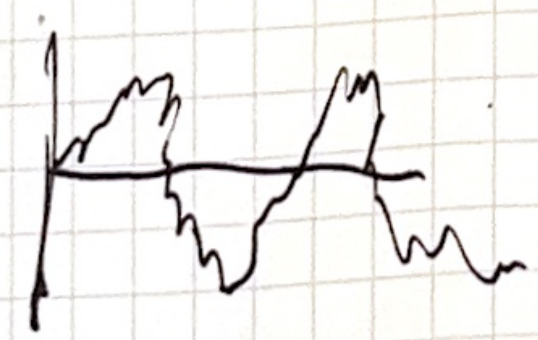
"gradient ascent plus noise"

which is what we actually run on computer.
Taking $\varepsilon \to 0$ makes mathematical analysis nicer.

Example:

$$H(x) = -\|x\|^2/2$$

$$dX_t = -X_t + \sqrt{2}\,dB_t$$



"Ornstein–Uhlenbeck process"

TODO: ~~how to prove~~ • why is e^H the stationary measure?

• What is the Poincaré/spectral gap inequality?

• When does spectral gap hold?

Key to understanding these questions is deriving the generator.

(Analogue of transition matrix for continuous-time m.c.)

Warning: we ignore technical issues about smoothness, etc. See [BGL] book

Def Langevin semigroup

$$P_t f = \mathbb{E}[f(X_t) \mid X_0].$$

Note: P_t is a linear operator, i.e. a Markov kernel.
"Semigroup" b/c $P_{t+s} = P_t P_s$.

Infinitesimal generator of semigroup

$$\mathcal{L}f = \frac{d}{dt}P_t f\Big|_{t=0} = \lim_{t \to 0} \frac{P_t f - f}{t}.$$

We have $P_t = e^{t\mathcal{L}}$. $\frac{d}{dt}P_t f = \mathcal{L} P_t f$, P_t and $\mathcal{L}$ commute.

Thm

$$\mathcal{L}f = \langle \nabla H, \nabla f \rangle + \Delta f$$

Pf For small t, (KEY! Keep second term of Taylor expansion!!!)

$$f(X_t) = f(X_0) + \langle \nabla f(X_0), X_t - X_0 \rangle + \frac{1}{2}\langle \nabla^2 f(X_0), (X_t - X_0)(X_t - X_0)^T \rangle + o(t).$$

and

$$X_t = X_0 + t\nabla H + \underbrace{\mathcal{N}(0, 2t)}_{\sqrt{2}B_t} + o(t).$$

Hence

$$\mathbb{E}(X_t - X_0) = t\,\nabla H + o(t)$$

$$\mathbb{E}(X_t - X_0)(X_t - X_0)^T = \mathbb{E}(t\nabla H + \sqrt{2}B_t)(t\nabla H + \sqrt{2}B_t)^T$$

$$= 2\mathbb{E}B_t^2 + o(t)$$

$$= 2t + o(t)$$

So $\mathbb{E}[f(X_t) - f(X_0) \mid X_0] = \langle \nabla f(X_0), \nabla H \rangle t + \overbrace{\mathrm{Tr}(\nabla^2 f)}^{= \Delta f} t$ $\square$

5

$\mathcal{L}$

$$\frac{\partial}{\partial t} P_t f = \mathcal{L} P_t f = \langle \nabla H, \nabla P_t f \rangle + \Delta (P_t f)$$

"Kolmogorov backward equation."

Let μ_0 be density/law of X_0

$$\mu_t = \mu_0 P_t. \quad \text{law of } X_t$$

Want ~~Fokker-Planck~~ PDE for density μ_t.

This is called the

Thm (Fokker-Planck equation),

$$\partial_t \mu_t = \Delta \mu_t - \operatorname{div}(\mu_t \nabla H)$$

Derivation:

Suffices to check $t=0$

$$\frac{d}{dt}\left(\int P_t f\right)\Big|_{t=0} = \int \mu_0 \mathcal{L} f$$

$$= \int \mu_0(x) (\mathcal{L} f)(x)\, dx$$

$$= \int \mu_0(x) \left(\langle \nabla H, \nabla f \rangle + \Delta f \right) dx$$

divergence theorem

$$\operatorname{div}(g \nabla f) = \nabla g \cdot \nabla f + g \Delta f$$

$$0 = \int \operatorname{div}(g \nabla f) = \int \nabla g \cdot \nabla f + \int g \Delta f$$

so $\int \mu_0 \Delta f = -\int \langle \nabla \mu_0, \nabla f \rangle = \int (\Delta \mu_0) f$

(Δ is self-adjoint!)

and

$$\cancel{\int \mu_0 \langle \nabla \log \mu_0, \nabla f \rangle} = \int \langle \nabla \mu_0, \nabla f \rangle = -\int f$$

$$\int \mu_0 \langle \nabla H, \nabla f \rangle = \int \langle \nabla (\mu_0 H) - H \nabla \mu_0, \nabla f \rangle = -\int f \Delta (\mu_0 H) - \int \langle H \nabla \mu_0, \nabla f \rangle$$

$$\int f \operatorname{div}(\mu_0 \nabla H)$$

$$= \int f \left[\langle \nabla \mu_0, \nabla H \rangle + \mu_0 \Delta H \right]$$

$$= \int f \left[\langle \nabla \mu_0, \nabla H \rangle \right] + \int H \Delta (f \mu_0)$$

$$\sum_i \int \mu_0(x) \partial_i H \partial_i f = -\sum_i \int \left[\partial_i (\mu_0 \partial_i H) \right] f$$

$$= -\int \operatorname{div}(\mu_0 \nabla H) f$$

So

$$= \int f \left(\Delta \mu_0 - \operatorname{div}(\mu_0 \nabla H) \right) dx$$

ie

$$\mathcal{L}^* \mu_0 = \Delta \mu_0 - \operatorname{div}(\mu_0 \nabla H) \qquad \square$$

Stationary measure

$$0 = \overset{\operatorname{div}(\nabla \mu)}{\overset{\parallel}{\Delta \mu}} - \operatorname{div}(\mu \nabla H) = \operatorname{div}(\nabla \mu - \mu \nabla H) = \operatorname{div}(\mu (\nabla \log \mu - \nabla H))$$

$\nabla \log \mu = \nabla H + c$

Soln ie. $\mu(x) \propto \exp(H(x))$

Thm

Reversibility Want to check $\mathcal{L}$ is "self-adjoint" w/r μ.

(Warning: technical issues skipped.)

$$\int g \,(\ldots) - \int f (\mathcal{L} g)\, d\mu = -\int \mu(x) f(x) \left(\langle \nabla \log \mu, \nabla g \rangle + \Delta g \right)$$

$$= \int \nabla \mu(x) \cdot f(x) \nabla g(x) + \int f(x) \langle \nabla \mu(x), \nabla g(x) \rangle$$

$$= \underbrace{\int \mu(x) \langle \nabla f(x), \nabla g(x) \rangle}_{\text{Symmetric in } f, g} \quad \square$$

So μ is the stationary measure.

Dirichlet form: $\mathcal{E}(f,g) = -\langle f, \mathcal{L} g \rangle_\mu = \int \mu(x) \langle \nabla f(x), \nabla g(x) \rangle$

Poincaré holds with const C if

$$\mathrm{Var}(f) \le C\, \mathcal{E}(f,f)$$

"Spectral gap $\frac{1}{C}$". ($= 1 - \lambda_2$ if spatial discrete)

[Brascamp-Lieb ~~Bakry-Émery~~]

Thm if $p(x) \propto \exp(H(x))$ and $\nabla^2 H(x) \preceq -\alpha I$

Then Poincaré holds with $C \le \frac{1}{\alpha}$.

"Strongly log-concave"

~~Thm~~

The case $\nabla^2 H(x) \preceq 0$ is also of interest.

Recent progress: "KLS up to polylog"

Thm [C...] if $\nabla^2 H(x) \preceq 0$ and $\mathrm{Cov}(p) = I$,

Then $C \le \mathrm{polylog}(n)$.

Prove Poincaré for Gaussian? and write general ... Thm?

Thm Let $\mu = \mathcal{N}(0, I)$. Then

$$\mathrm{Var}(f) \le \mathbb{E}\|\nabla f\|^2$$

Pf:

From Thm, $H(x) = -|x|^2/2$

$$\mathcal{L} f = \langle \nabla H, \nabla f \rangle + \Delta f = \langle -x, \nabla f \rangle + \Delta f$$

$$P_t f(x) = \mathbb{E}\left[f\left(e^{-t} x + \sqrt{1 - e^{-2t}}\, \xi\right)\right] \quad \xi \sim \mathcal{N}(0, I)$$

So $\nabla_x P_t f(x) = e^{-t} P_t \nabla f(x)$

b/c $\frac{\partial}{\partial t} P_t f = \langle P_t \nabla f, e^{-t} x \rangle + \mathbb{E}\left[f'(\ldots) \frac{e^{-2t}}{\sqrt{1-e^{-2t}}} \xi \right]$

$\langle \nabla P_t f, -x \rangle$ $\quad = e^{-2t} \mathbb{E}[\nabla f(\ldots)] = e^{-2t} P_t \Delta f \overset{\text{IBP}}{=} \Delta P_t f$

$$\mathrm{Var}(f) = \int_0^\infty 2\, \mathcal{E}(P_t f, P_t f)$$

$$= 2 \int_0^\infty \mathbb{E}\|\nabla P_t f\|^2 = 2 \int_0^\infty e^{-2t}\, \mathbb{E}\|P_t \nabla f\|^2$$

$$\le 2 \int_0^\infty e^{-2t}\, \mathbb{E}\|\nabla f\|^2 = \mathbb{E}\|\nabla f\|^2$$

Using that

$$\frac{d}{dt} \mathrm{Var}(P_t f) = \frac{d}{dt}\left(\mathbb{E}(P_t f)^2 - (\mathbb{E} f)^2 \right)$$

$$= 2\, \mathbb{E}(P_t f)(\mathcal{L} P_t f) = -2\, \mathcal{E}(f,f)$$

Thm ((Caffarelli) contraction)

α-strongly log-concave $\rightsquigarrow$ $\exists \sqrt{\frac{1}{\alpha}}$ Lipschitz map (optimal transport)

from std gaussian

$$\mathrm{Var}(f(\phi(z))) \le \mathbb{E}\,\|\nabla f(\phi(z))\|^2$$
$$= \mathbb{E}\,\| f'(\phi(z))\,\phi'(z)\|^2$$
$$\le \frac{1}{\alpha}\,\mathbb{E}\,\| f'(\phi(z))\|^2$$

Rmk Some proof carries for single transport linearization

MLSI: ~~$\mathrm{Ent}[f] = \mathbb{E}[f\log f] - \mathbb{E}[f]\log\mathbb{E}[f]$~~

$\mathrm{Ent}[f] \le C\,\mathcal{E}(\sqrt{f}, \sqrt{f}) = \frac{C}{4}\,\mathcal{E}(f, \log f)$

Gamma MLSI $\Rightarrow$ strongly log-concave MLSI

$$\frac{d}{dt}\mathrm{Ent}[P_t f] = -\mathcal{E}(\log P_t f, P_t f)$$

(AJKPV'22) Contraction in KL

Thm: C-SI for all $\mathrm{Subdiv} \Rightarrow$ Down-up has MLSI $\frac{1}{k^C}$

Infact $\mathrm{Ent}[\ldots] \le$

[CGM'21] log-concave cont

$\mathrm{Ent}[U_{1\to k-\ell}\, f] \le \frac{C}{k}\,\mathrm{Ent}[f]$

Clarifying remark: $k-2+k-1-c+c-1$

$$c' = \frac{c(k-2)}{k-1-c} = c + \frac{c-1}{k-1-c}$$ "always c"

$$= c + \frac{\frac{c-1}{k-1}}{1-\frac{c}{k-1}}$$

$$= c + \frac{c-1}{k-1}\left(1 + \frac{c}{k-1} + \frac{c^2}{(k-1)^2} + \dots\right)$$

$t = k-1$

So $\frac{dc}{dt} \approx \underbrace{\frac{c-1}{t}} + \frac{c^2-c}{t^2}$

$\frac{dc}{c-1} = \frac{dt}{t}$

$\log(c-1) = \log(t) + k$

$c(t) = kt+1$

Solves to $c(t) = (c(1)-1)\,t+1$

if $c(1) = 1+\varepsilon \rightsquigarrow c(t) = \varepsilon t$

ok Very bad SI!

$\varepsilon = \frac{1}{k}$ is all we can accept.

Claude

Stochastic localization / "Diffusion model"

- Popular approach to sampling w/ a time-inhomogeneous M.C.
- Interesting history. Anderson, Föllmer, Eldan, ... ML people (...) ... tricks ... these

Given $X \sim \mu$, W_t independent Brownian motion

Define $y_t = tX + W_t$. (Equiv. $\frac{y_t}{t} = X + \frac{1}{t}W_t$).

$d\mu_t(x) = \mu_t(x)\langle x - a_t, dW_t\rangle$ $\quad a_t = \mathbb{E}_{\mu_t}[X]$ $\quad dW_t dW_t^T = dt$ (intuitively)

$d\mu_t d\mu_t^T = \mu_t^2 (x-a_t)(x-a_t)^T$

$$d\log\mu_t = \frac{d\mu_t}{\mu_t} + \frac{1}{2}\left(-\frac{1}{\mu_t^2}\right)\langle d\mu_t d\mu_t^T, I\rangle$$

$$= \langle x - a_t, dW_t\rangle - \frac{1}{2}\langle (x-a_t)(x-a_t)^T dt, I\rangle$$

so

$$\mu_t(x) = \exp\left(\underbrace{\int_0^t \langle x - a_s, dW_s\rangle - \frac{1}{2}\int_0^t \|x-a_s\|^2 ds}\right)\mu(x)$$

$$= \langle y_t, x\rangle - \frac{t}{2}\|x\|^2 \; \cdots \;\leftarrow \text{indep't on } x$$

where

$$\langle x, y_t\rangle = \int_0^t \langle x, dW_s\rangle + \int_0^t \langle x, a_s\rangle ds$$

i.e. $dy_t = dW_t + a_t dt$

$y_t = W_t + \int_0^t a_s ds$

~~Lemmas~~

Lemma

1. μ_t is a probability measure (a.s.)
2. $\forall S$, $\mu_t(S)$ is a martingale
3. $\mu_t \to \delta_{a_\infty}$ where $a_\infty \sim \mu$ as $t \to \infty$.

Pf:

1. By Itô calculus, $\mu_t(x) \geq 0$ for all time.

And $\sum_x d\mu_t(x) = \sum_x \mu_t(x)\langle x - a_t, dW_t\rangle = \langle a_t - (\sum_x \mu_t(x)) a_t, dW_t\rangle$

initial $\sum_x \mu_0(x) = 1 \Rightarrow$ Unique soln is $\sum_x \mu_t(x) = 1$

2.

$$\mu_t(S) = \int_S \mu_t(x)\, dx$$

$$d\mu_t(S) = \int_S d\mu_t(x)\, dx = \underbrace{\int_S \mu_t(x)\langle x - a_t, dW_t\rangle}_{\text{mean-zero } \forall t}$$

so $\mu_t(S)$ = martingale.

3. as $t \to \infty$,

$(\mu_t - \delta_{a_t})(f) \to 0$ for any smooth f.

$\mathbb{E}_{\mu_t}[(x-a_t)(x-a_t)^T] \to 0$ like $\frac{1}{t}$

$da_t = \mathbb{E}_{\mu_t}[(x-a_t)(x-a_t)^T]\, dW_t$

$$a_t = \int \underbrace{\mathbb{E}_{\mu_t}[(x-a_t)(x-a_t)^T]}_{O(1/t)}\, dW_t$$

$\|a_s - a_t\|^2$

$$\mathbb{E}\left[\left(\int_T^\infty \mathbb{E}_{\mu_t}[(x-a_t)(x-a_t)^T]\, dW_t\right)^2\right]$$

"Itô isometry"

$$= \mathbb{E}\left[\int_T^\infty \underbrace{\mathbb{E}_{\mu_t}[(x-a_t)(x-a_t)^T]^2}_{O(1/t^2)}\, dt\right]$$

So $\mathbb{E}\|\cdot\|_{op} \lesssim \int_T^\infty \frac{1}{t^2}\, dt = \frac{1}{T}$. $\quad L^2$ is complete Hilbert space $\square$

$X \sim \mu$ follows by martingale property.

$$y_t = W_t + \int_0^t a_s\,ds$$
$$= W_t + t a_\infty + \int_0^t (a_s - a_\infty)\,ds$$

where $\mathbb{E}[a_\infty \mid \mathcal{F}_s] = a_s$

so this is zero by ... variance of WLLN

$$\mathbb{E}\left[\left(\int_0^t (a_s - a_\infty)\,ds\right)^2\right]$$
$$= \mathbb{E}\left[\int_0^t\int_0^t (a_s - a_\infty)(a_{s'} - a_\infty)\,ds\,ds'\right]$$

$\mathbb{E}[(a_s - a_\infty)(a_{s'} - a_\infty)] = \mathbb{E}[(a_s - a_\infty)(a_{s'} - a_{s'})] = 0$ if $s \le s'$ ... orthogonality.

$y_t = W_t + t a_\infty$ "reverse diffusion"

Thm Let p be joint law of W_t, a_∞

$\mu_t = p(a_\infty = \cdot \mid \mathcal{F}_t)$ $\quad \mathcal{F}_t$ = info up to time t.

① $\mu_t = p(a_\infty = \cdot \mid y_t)$ ② y_t a Markov chain ... (just Bayes: y_s ... $a_\infty \mid y_t$) etc of Gaussian process

Follows from "forward diffusion" rep...

$$X_r = r\,y_{\frac{1}{r}} = rW_{\frac{1}{r}} + a_\infty$$

Thm (classical) Let $B_r = rW_{\frac{1}{r}}$. Then B_r is a Brownian motion.

so $X_r = a_\infty + B_r$ "forward diffusion"

$\quad = r y_{\frac{1}{r}}$

$y_t = t a_\infty + W_t$ "reverse diffusion"

$\longrightarrow$ $dy_t = a_t\,dt + dW_t$

Can implement reverse diffusion provided an estimate for a_t "score function", "denoiser"

Rmk: Can also write for $t \in (0,T]$

$$y_t = W_t + t a_T + \int_0^t (a_s - a_T)\,ds.$$
$$= W_t + t a_T = W_t - \frac{t}{T} W_T + t(a_T + \frac{W_T}{T})$$

Brownian bridge

$B_t = W_t - \frac{t}{T} W_T$ is a Brownian bridge

y_t is "follmer drift"

$\mu \propto \mathcal{N}(0, \frac{1}{t})$ and "entropy-optimal coupling" of Wiener measure

Pf: ① B_r is a jointly Gaussian process

(optional?) ① Markov property of W_t: $\{W_{t<T}\} \perp \{W_{t>T}\} \mid W_T$

$\Updownarrow$

of B_t

② $\mathrm{Cov}(B_s, B_t) = \mathrm{Cov}(sW_{1/s}, tW_{1/t}) = st\min(\frac{1}{s}, \frac{1}{t}) = \min(s,t)$ □

SL used as ① direct sampler ② tool for analyzing models

Warmup $\frac{dx}{dt} = \sqrt{x}$ $\quad x(0)=0$

Coleman alpha
Mathematica soln: $x(t) = \frac{t^2}{4}$.

But $x(t)=0$ also works... (naive Euler discretization soln)

Example applications of SL

① Sampling SK model (El Alaoui - Montanari - Sellke)

$\mu(x) \propto \exp(\frac{1}{2}\beta\langle x, Jx\rangle)$

$J_{ij} \sim N(0, \frac{1}{n})$ symmetric

AMP "approximately" can compute $\mathbb{E}[X \mid X + N(0,\sigma^2)]$

Importance sampling diffusion

$\beta < 1$

② Analyzing Glauber in SK model + other Ising models.

$\beta < 0.295$

③ KLS conjecture

$p(x) \propto \exp(-H(x))$

H concave but not strongly convex

$\mathbb{E}_p(x) = 0$
$\mathbb{E}_p[xx^T] = I$ "isotropic position"

Thm [..., Chen '22, Klartag-Lehec '23]

$\mathrm{Var}(f) \lesssim \log^4(n)\ \mathbb{E}\|\nabla f\|^2$

Pf idea: Suffices to show Σ_t doesn't blow up

$dx_t = \Sigma_t\, dW_t$

$d\Sigma_t = \mathbb{E}[(x-x_t)^{\otimes 3}]\,dW_t - \Sigma_t^2\, dt$

after $\Omega(\frac{1}{\mathrm{polylog}(n)})$ time strongly log-concave $\Rightarrow$ Σ_t good

at O time Σ_t good by ...

Claim: $\mathbb{E}\|\Sigma_t\|_q \le \left(\frac{t}{s}\right)^{2q} \mathbb{E}\|\Sigma_s\|_q$ $t \geq s$, $q \geq 3$ th

SL

s $\quad$ t (dur asgmt) $\quad$ mode agreement are flatness...

$0 \;—\; T_0 \;—\; T_1$

BK

Slack loc

HDX

Wm

Analogy - GSI: if $\|\Sigma\|_{op}$ small $\approx$ SI

Code

|

G

~~"Local-to-global"~~

"Local-to-global" argument:

μ_0 ③ good param

② $\|\Sigma_t\|$ small

Exp $\quad$ ①+② $\rightsquigarrow$ ③

μ_T good ① Phase

C