Lecture 13 : Diffusion in Unsupervised Learning

Last week: → Implicit bias of gradient descent: explains why learning is possible in over-parametrized regime, even without reg!
↳ sometimes this leads to "interesting"/efficient algorithms (non-linear max-margin classif), sometimes not really. (double-descent).

Today: Introduction to probabilistic modeling (no labels)
Key mathematical tool: diffusion.

Part I: Basic Notions of Probabilistic Modeling.

- Central object: a probability measure $\mu \in P(\Omega)$, Ω is a high-dimensional data domain.

→ Several situations of practical interest:

(a) μ is unknown; we wish to estimate μ from iid samples $x_i \sim \mu$ [Density Estimation; Generative Modeling].

(b) μ is known, and we wish to sample from it [Sampling].

(c) μ is known up to normalisation; $\frac{d\mu}{dx}(x) \propto G(x)$, and we wish to estimate marginals. [Marginalisation]

→ Several approaches to express probability distributions μ.

(a) "Explicit" models that directly parametrize the density.

↳ Gibbs / Boltzmann Distributions. (aka Energy-Based Models)

Fix a base measure τ in Ω (eg Lebesgue) and consider F of $f: \Omega \to \mathbb{R}$. The Gibbs measure with energy $f \in F$ and inverse temperature $\beta \geq 0$ has density

$$\frac{d\nu}{d\tau}(x) = \frac{1}{Z_{\beta f}} e^{-\beta f}, \text{ with}$$

$$Z_{\beta f} = \int_\Omega e^{-\beta f(x)} \tau(dx).$$

Examples: Gaussian $\hookrightarrow$ $f(x) = \frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)$

Ising Model $\hookrightarrow$ $f(x) = \sum_{i,j} J_{ij} x_i x_j$, $x_i \in \{\pm 1\}$

etc...

↳ Mixture Models: consider a measure with density

$$p(x) = \int_\Theta p(x|h) \cdot dq(h)$$

h: "latent" variable

q: prior over latent.

$p(x|h)$: conditional likelihood.

Ex: Gaussian Mixture Models

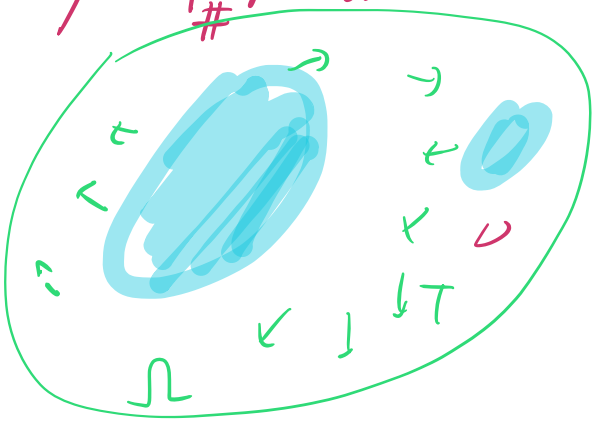
Variational Autoencoders

⋮

↳ Normalizing Flows. Assume $T_\theta : \Omega \to \Omega$ is a diffeomorphism. The pushforward of a measure ν by T is defined as the measure $\nu \circ T^{-1}$ such that

$\forall A$ measurable,

$$\mu(A) = \nu(T^{-1}(A))$$



When ν has density, $p(x)$, then for any test function χ, we have

$$\mathbb{E}_{x\sim\mu}[\chi(x)] = \mathbb{E}_{x\sim\nu}[\chi(T(x))]$$

$$\int \chi(x)\cdot q(x)\,dx \qquad \int \chi(T(x))\cdot p(x)\,dx$$

q: density of μ

$\tilde{x} = T(x)$

$$\int \chi(\tilde{x})\cdot p(T^{-1}(x))\cdot |DT^{-1}(x)|\,dx$$

$$\Rightarrow \quad q(x) = p(T^{-1}(x))\cdot |DT^{-1}(x)|$$

[Normalizing Flows]
[Rezende & Mohamed]
[Tabak, Vanden-Eijn.]

(b) "Implicit" models: define a probability measure indirectly, $\mu = T_{\#}\nu$, where T is now a generic mapping $T: \Theta \to \Omega$ (not 1:1, not mapping from Ω).

Ex: . Generative Adv. Nets (T is given by a NN)

. Score-based Diffusion [Song, Ermon, Sohl-Dickstein, ...]
(T is given by a Reversed-Diffusion process).

→ Several Criteria for Density Estimation: $X_i \sim \nu$

drawn iid $\rightsquigarrow$ $\nu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ empirical data distribution.

μ_θ : one of the probability models from above.

In general, objective is of the form

$$\min_{\theta} \text{"Dist"}(\mu_\theta, \nu_n)$$

Questions: Which choice of "Dist"?
How about "test" error "Dist"$(\mu_{\hat{\theta}}, \nu)$

→ Kullback-Leibler divergence: $KL(\nu \| \mu_\theta) = \int \log\left(\frac{d\nu}{d\mu_\theta}\right) d\nu$

$KL(\nu \| \mu_\theta) = H(\nu;\nu) - H(\mu_\theta;\nu)$ with

$H(\mu;\tilde{\mu}) = \int \log(d\mu)\, d\tilde{\mu}$ is the cross-entropy.

Thus minimising KL divergence is equivalent to maximizing cross-entropy.

→ $H(\mu_\theta;\nu) = \mathbb{E}_\nu\left[\log \frac{d\mu_\theta(x)}{dx}\right]$, so we can estimate it from our samples x_i as:

$$\max_{\theta} \frac{1}{n}\sum_{i=1}^{n} \log \frac{d\mu_\theta}{dx}(x_i) \quad \longleftrightarrow \text{ Maximum-Likelihood Estimator.}$$

→ Alternatives to KL?

Integral Probability Metrics (IPMs)

$$D_F(\nu,\mu) = \sup_{f} \left\{ \mathbb{E}_\nu f(x) - \mathbb{E}_\mu f(x) \right\}$$

$$\sup_{f \in \mathcal{F}} \left(\right)$$

$\mathcal{F}$: family of test functions

Ex (·) $\mathcal{F} = \{ f : \Omega \to [-1, +1] \} \rightsquigarrow$ TV-distance

(·) $\mathcal{F} = \{ f : \Omega \to \mathbb{R} \ \ 1\text{-Lipschitz} \} \rightsquigarrow$ 1-Wasserstein distance.

In all these cases, we have the same tradeoffs as in SL between approximation, estimation and optimisation errors.

Extra computational challenge: How to efficiently perform training and inference?

Focus on Gibbs/Energy Based Model trained with MLE.

Recall the density of an EBM model is

$$p_\theta(x) = \frac{e^{-\beta f_\theta(x)}}{Z_{\beta,\theta}} = e^{-\beta f_\theta(x) - A(\theta,\beta)}, \text{ with}$$

$$A(\theta,\beta) = \log \int e^{-\beta f_\theta(x)} dx.$$

· The MLE becomes

$$\min_\theta \ \frac{1}{n} \sum_{i=1}^n -\log p_\theta(x_i) = \underbrace{\frac{\beta}{n} \sum_{i=1}^n f_\theta(x_i)}_{\text{empirical energy}} + \underbrace{A(\theta,\beta)}_{\text{free energy}} := L(\theta)$$

"local" term "global" term.

· The gradient $\nabla_\theta L(\theta)$ becomes

$$\nabla_\theta L(\theta) = \frac{\beta}{n} \sum_{i=1}^n \nabla f_\theta(x_i) - \int \beta \nabla f(\theta) \underbrace{\frac{e^{-\beta f_\theta(x)}}{Z_{\beta,\theta}}}_{= p_\theta(x)} dx$$

$$= \beta \left(\underbrace{\mathbb{E}_{\hat{\nu}}[\nabla_\theta f_\theta]}_{\text{easy to compute.}} - \underbrace{\mathbb{E}_{\mu_\theta}[\nabla_\theta f]}_{??} \right)$$

→ Natural tool to estimate $\mathbb{E}_{\mu_\theta}[\nabla f]$ is Monte-Carlo: sample from μ_θ and average gradients. But how? ↓ MCMC

Part II: Langevin Diffusion

→ Let's consider the scheme $x^{(k+1)} = x^{(k)} - \eta \left(\nabla_x \overbrace{f_\theta}^{f}(x^{(k)}) + \sqrt{\frac{2}{\eta}} \cdot \tau \cdot W \right)$

$W \sim N(0, I)$.

→ This stochastic discrete process is the discretization of an underlying SDE:

$$dX_t = -\nabla f(X_t)\, dt + \sqrt{2} \cdot \tau \, dB_t$$ ← Brownian Motion.

↳ This is a Langevin diffusion.

Convergence of this process?

Markov Semigroups

→ We can understand the Markov process $(X_t)_t$ by tracking how measurements $\mathbb{E}[g(X_t) \mid X_0 = x]$ evolve over time.

→ Def: For a Markov process $(X_t)_t$, its semigroup is a

family $(P_t)_t$ of operators acting on test functions

$$(P_t\, g)(x) = \mathbb{E}\left[\, g(X_t) \mid X_0 = x \right]$$

Semigroup properties: (i) $P_0 = id$ $\quad P_t \to id$ as $t \to 0$.

(ii) $P_s P_t = P_{s+t} = P_{t+s} = P_t \cdot P_s$

→ This semigroup has an associated infinitesimal generator (also a derivative operator):

$$\mathcal{L} g = \lim_{t \to 0^+} \frac{P_t g - g}{t}$$

→ Generator of Langevin diffusion? ($\sigma = 1$)

$$dX_t = -\nabla f(X_t)\, dt + \sqrt{2}\, dB_t \qquad B_t \sim N(0, t \cdot Id)$$

$$X_t = X_0 - \int_0^t \nabla f(X_s)\, ds + \sqrt{2}\, B_t$$

$$= X_0 - t \cdot \nabla f(X_0) + \sqrt{2}\, B_t + o(t)$$

Taylor Expansion of test function g (Ito lemma)

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

$$= g(X_0) + \langle \nabla g(X_0), -t \cdot \nabla f(X_0) + \sqrt{2}\, B_t \rangle + B_t^T \nabla^2 g(X) B_t + o(t)$$

$o(t^2)$.

since $\mathbb{E}[B_t B_t^T]$ is of order t.

→ Thus $\mathbb{E}\left[g(X_t) \mid X_0 = x \right] =$

$$= g(x) - t \langle \nabla g(x), \nabla f(x) \rangle + t \cdot \mathrm{Tr}(\nabla^2 g(x)) + o(t)$$

so $\mathcal{L} g(x) = \lim_{t \to 0^+} \frac{\mathbb{E}[g(X_t) \mid X_0 = x] - g(x)}{t} =$

$$= - \langle \nabla g(x), \nabla f(x) \rangle + \Delta g(x).$$

→ observe that $\frac{P_{t+h} g - P_t g}{h} = P_t \cdot \frac{P_h g - g}{h} \to P_t \mathcal{L}$

$$\| \quad \frac{P_h g - g}{h} \cdot P_t \to \mathcal{L} \cdot P_t \text{ , so}$$

$$\boxed{\partial_t P_t g = \lim_{h \to 0^+} \frac{P_{t+h} g - P_t g}{h} = \mathcal{L} P_t g}$$

↳ this is called the Backwards Kolmogorov Eq.

→ Let $(\pi_t)_t$ be the law of the process X_t. We can write

$$\mathbb{E}\, g(X_t) = \int P_t g(x) \, d\pi_0(x) = \int g \, d P_t^* \pi_0$$

(P_t^*: formal adjoint of P_t)

$$\| \quad \int g \cdot d\pi_t \qquad \text{so} \qquad \pi_t = P_t^* \pi_0 .$$

. We have

$$\partial_t \int g \, d(P_t^* \pi_0) = \partial_t \int P_t g \, d\pi_0 = \int \mathcal{L} P_t g \, d\pi_0 =$$

$$= \int P_t \mathcal{L} g \, d\pi_0 = \int g \, d(\mathcal{L}^* P_t^* \pi_0) \quad , \text{ so}$$

$$\partial_t P_t^* \pi_0 = \mathcal{L}^* P_t^* \pi_0 = \mathcal{L}^* \pi_t .$$

$$\| \quad \partial_t \pi_t$$

$$\boxed{\partial_t \pi_t = \mathcal{L}^* \pi_t}$$

Forward-Kolmogorov Eq.
(Fokker-Plank eq.)

→ Let $g, h : \mathbb{R}^d \to \mathbb{R}$. smooth.

$$\int \mathcal{L} g \cdot h \, dx = \int (\Delta g - \langle \nabla g, \nabla f \rangle) \cdot h \, dx$$

$$\overset{?}{=} \int g \cdot (\text{"something that depends on } h\text{"}).$$

$$= \int g \left(\Delta h + \operatorname{div}(h \cdot \nabla f) \right) dx.$$

$=$ ↑ integration by parts

So $\mathcal{L}^* h = \Delta h + \operatorname{div}(h \cdot \nabla f)$.

So $\partial_t \pi_t = \mathcal{L}^* \pi_t$.

→ So the equilibrium / stationary solution of $\partial_t \pi = \mathcal{L}^* \pi$ solves $\mathcal{L}^* \pi = 0$

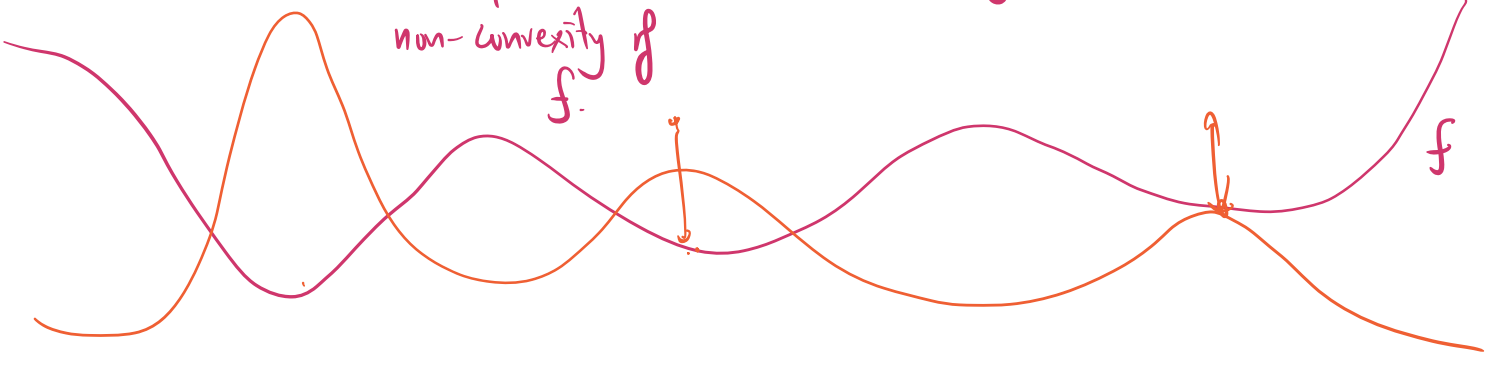
$$0 = \Delta \pi + \operatorname{div}(\pi \nabla f) = \operatorname{div}(\nabla \pi + \pi \nabla f)$$

$$= \operatorname{div}(\pi(\nabla \log \pi + \nabla f))$$

$$\Rightarrow \log \pi = -f + \text{const.} \Rightarrow \pi = \frac{1}{Z} \exp(-f)$$

→ In summary, the unique stationary measure of Langevin diffusion $dX_t = -\nabla f(X_t)\,dt + \sqrt{\frac{2}{\beta}}\,dB_t$ is the Gibbs measure $\pi_{\beta f}(x) = \frac{1}{Z_\beta} \exp(-\beta f(x))$.

→ Although the global convergence to $\pi_{\beta f}$ holds for very general energies, the rate of convergence is generally cursed by dimension. ↪ presence of "metastability"



→ Rate of convergence of $\pi_t \to \pi_\infty$ is controlled by so-called functional inequalities, measuring how "contractive" the dynamics are in a certain metric

↳ spectral Gap (χ^2- divergence)

↳ Log- Sobolev Inequality (KL- divergence)

→ Punch-line / Conclusion: slow rate is unavoidable in high-dim.

→ Sampling hard ⇝ EBM learning hard!

Alternatives?

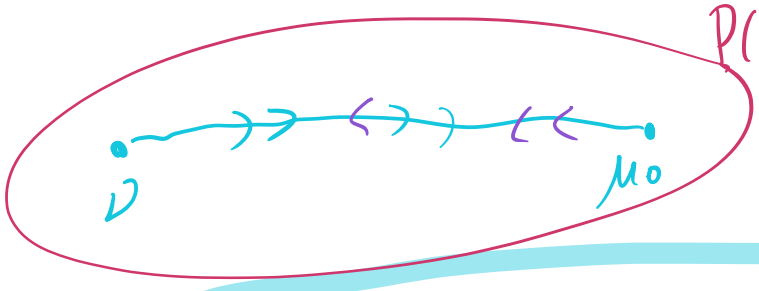
Score- Based Diffusion

↳ we have seen that in order to sample from $\mu = \frac{1}{Z} e^{-f}$, Langevin diffusion only needs access to $\nabla \log \rho = \nabla f$ the "score function".

Idea: Build a diffusion process $\{X(t)\}_{t=0}^{T}$ such that:

(i) $X(0) \sim \nu$ (data distribution) for which we have iid samples.

(ii) $X(T) \approx \mu_0$ base measure, for which we can sample.



$$dX_t = f(X_t, t)\cdot dt + g(t)\, dB_t$$

Running backwards in time: we sample from μ_0, then run backwards, until time $-T$, gives a sample from ν.

→ Reverse Diffusion [Anderson, Follmer]

$$dY_t = \left[f(X,t) - g(t)^2 \nabla_x \log \Pi_t(x) \right] dt + g(t)\, dB_t.$$

↳ if we can estimate the scores of Π_t for $0 \le t \le T$, then we can go back in time.

→ When f is linear, the Forward diffusion is an Orstein-Ullhenbeck process, where $\nabla_x \log \Pi_t(x)$ can be efficiently estimated from data!

[Song, Ermon, Sohl-Dickstein et al].

Questions / Summary:

→ Convergence guarantees for score-based wrt Langevin?

→ Iterative Sampling schemes ↔ Diffusion

Direct sampling schemes ↔ Transport. [Flows, GAN].