TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2022

The Thermodynamic Interpretation of Diffusion Models

Why are they called "diffusion" models?

Generative Modeling by Estimating Gradients ...
Song and Erman, July 2019

Langevin Dynamics

Consider a model density defined by a continuous softmax on a model score.

$$\begin{aligned} p_{\mathrm{score}}(y) &= \operatorname*{softmax}_y \ \mathrm{score}(y) \\ &= \frac{1}{Z}\, e^{\mathrm{score}(y)} \\ Z &= \int e^{\mathrm{score}(y)}\, dy \end{aligned}$$

Here $\mathrm{score}(y)$ is a parameterized model computing a score and defining a probability density on R^d.

Langevin Dynamics

If y is discrete, but from an exponentially large space (such as sentences or a semantic image segmentation) we can use MCMC sampling (the Metropolis algorithm or Gibbs sampling).

In the continuous case we can use Langevin dynamics.

Langevin Dynamics

Noisy gradient ascent on score.

$$y(t + \Delta t) \;=\; y(t) + \eta g \Delta t + \sigma \epsilon \sqrt{\Delta t}$$

$$g \;=\; \nabla_y \,\mathrm{score}(y)$$

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

This give a well-defined distribution on functions of time in the limit as $\Delta t \to 0$.

$$dy = \eta g dt + \sigma \epsilon \sqrt{dt} \qquad \epsilon \sim \mathcal{N}(0, I)$$

Langevin Dynamics

$$dy = \eta g dt + \sigma \epsilon \sqrt{dt} \qquad \epsilon \sim \mathcal{N}(0, I)$$

This has stationary (equilibrium) density.

The derivation is mathematically identical to the derivation of the stationary distribution of SGD at a learning rate η and noise covariance Σ.

However, here we have isotropic noise rather than arbitrary gradient noise.

Isotropic noise always yields a Gibbs distribution.

Imposing isotropic noise is called Langevin dynamics.

The Stationary Density

To derive the stationary density we consider a gradient flow and a **diffusion flow** as a function of density $p(y)$.

The gradient flow is $\eta p(y) \nabla_y \mathrm{score}(y)$ and the diffusion flow is $\frac{1}{2}\eta\sigma^2 \nabla_y p(y)$

Setting them to be opposite and solving the resulting differential equation gives

$$p(y) = \frac{1}{Z}\, e^{\frac{2\mathrm{score}(y)}{\eta\sigma^2}}$$

The Stationary Density

$$p(y) = \frac{1}{Z}\, e^{\frac{2\mathrm{score}(y)}{\eta\sigma^2}}$$

Setting $\eta = 1$ and $\sigma^2 = 2$ gives

$$p(y) = \frac{1}{Z}\, e^{\mathrm{score}(y)} \quad = \quad \underset{y}{\mathrm{softmax}}\ \mathrm{score}(y)$$

Running Langevin dynamics long enough will yield a sample from the softmax distribution.

Score Matching

In score matching we train $g(y)$ rather than $\mathrm{score}(y)$ so as to make $g(y) \approx \nabla_y \,\mathrm{score}(y)$

The training objective for the decoder of a diffusion model can be viewed as training an update direction g to approximate $\nabla_y \ln p(y)$.

The score matching interpretation identifies the diffusion model decoding vector $\epsilon(z)$ with $-\nabla_z \ln p(z)$

Warning: The term "score" in score matching technically refers to the gradient vector $\nabla_y \,\mathrm{score}(y)$ rather than to the scalar "score" used in the softmax.

Simulated Annealing

In simulated annealing one tries to avoid local optima by first running at a high temperature and then then gradually reducing the temperature.

In the diffusion model σ_ℓ increases with increasing ℓ which is claimed to be an analogy with simulated annealing.

However, simulated annealing corresponds to adding noise **in sampling** rather than adding noise to a population sample.

Score Matching vs. VAE

The VAE interpretation of diffusion models does not rely on Langevin dynamics, score matching or simulated annealing.

However, the score matching interpretation, which identifies $\epsilon(z_\ell, \ell)$ with $-\nabla_z \, p(z)$, plays a role in "classifier conditioned guidance" used in DALLE-2.

The DDPM Stochastic Differential Equation (SDE)

Consider a DDPM (denoising diffision probabilistic model) for modeling $P(y)$ with $y \in R^d$ where the noise model is defined by

$$z_0 = y$$

$$z_\ell = \alpha z_{\ell-1} + \sqrt{1-\alpha^2}\,\epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

For technical simplicity we take α to be constant for all ℓ and allow $\ell \geq 1$ to be arbitrarily large.

The DDPM SDE

For sampling z_ℓ given z_0 the unit variance constraint gives

$$z_\ell = \alpha^\ell z_0 + \sqrt{1 - \alpha^{2\ell}}\,\epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

For sampling $z_{(\ell+k)}$ given z_ℓ we have

$$z_{(\ell+k)} = \alpha^k z_\ell + \sqrt{1 - \alpha^{2k}}\,\epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

The DDPM SDE

Setting $\alpha = e^{\frac{-1}{N}}$ we have

$$z_\ell = e^{\frac{-\ell}{N}} z_0 + \sqrt{1 - e^{\frac{-2\ell}{N}}}\,\epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

$$z_{(\ell+k)|\ell} = e^{\frac{-k}{N}} z_\ell + \sqrt{1 - e^{\frac{-2k}{N}}}\,\epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

The DDPM SDE

Taking $t = \frac{\ell}{N}$. We have $\ell = Nt$ and the previous slide can be written as

$$z(t) \;=\; e^{-t}z(0) + \sqrt{1-e^{-2t}}\;\epsilon \quad \epsilon \sim \mathcal{N}(0,I)$$

$$z(t+\Delta t) \;=\; e^{-\Delta t}z(t) + \sqrt{1-e^{-2\Delta t}}\;\epsilon \quad \epsilon \sim \mathcal{N}(0,I)$$

The DDPM SDE

For small ϵ we have $e^{-\epsilon} \approx 1 - \epsilon$ and for small Δt the previous slide can be written as

$$z((t + \Delta t)|t) \;\approx\; z(t) - z(t)\Delta t + \sqrt{2\Delta t}\;\epsilon$$

$$\Delta z \;\approx\; -z\Delta t + \sqrt{\Delta t}\;\delta \quad \delta \sim \mathcal{N}(0, 2I)$$

This can be interpreted as the stochastic differential equation for the forward process (the encoder) for diffusion models.

END