

# **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**

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

$$p_{\text{score}}(y) = \underset{y}{\text{softmax}} \text{ score}(y)$$

$$= \frac{1}{Z} e^{\text{score}(y)}$$

$$Z = \int e^{\text{score}(y)} dy$$

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

# Sampling from a Continuous Softmax

## 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 for Sampling From a Model

Noisy gradient ascent on score.

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

$$g = \nabla_y \text{score}(y)$$

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

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

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

## Langevin Dynamics for Sampling From a Model

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

This has stationary (equilibrium) density.

The derivation is mathematically identical to the derivation of the stationary distribution of SGD at a learning rate  $\eta$  and noise covariance  $\Sigma$ .

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  $\eta p(y) \nabla_y \text{score}(y)$  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

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

## The Stationary Density

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

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

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

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

## Score Matching

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

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  $\nabla_y \text{score}(y)$  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  $\sigma_\ell$  increases with increasing  $\ell$  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 DALLÉ-2.

**END**