# TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

## Stationary Distributions of SDEs and Temperature

## The Stationary Distribution

$$\Phi(t + \Delta t) \approx \Phi(t) - g(\Phi)\Delta t + \epsilon\sqrt{\Delta t} \qquad \epsilon \sim \mathcal{N}(0, \eta\Sigma)$$

For an SDE we have a stationary continuous density in parameter space.

We have a probability mass flow due to the loss gradient and a diffusion probability mass flow proportional to the density gradient.

## The Stationary Distribution

We consider the one dimensional case — a single parameter $x$ — and a probability density $p(x)$.

The gradient flow is equal to $-p(x)g$.

The diffusion flow is proportional to $-\eta\sigma^2\ dp(x)/dx$ (see the appendix).

For a stationary distribution the sum of the two flows is zero giving.

$$\alpha\eta\sigma^2\frac{dp}{dx} = -p\frac{d\mathcal{L}}{dx}$$

## The 1-D Stationary Distribution

$$\alpha\eta^2\sigma^2\frac{dp}{dx} = -\eta p\frac{d\mathcal{L}}{dx}$$

$$\frac{dp}{p} = \frac{-d\mathcal{L}}{\alpha\eta\sigma^2}$$

$$\ln p = \frac{-\mathcal{L}}{\alpha\eta\sigma^2} + C$$

$$p(x) = \frac{1}{Z}\exp\left(\frac{-\mathcal{L}(x)}{\alpha\eta\sigma^2}\right)$$

We get a Gibbs distribution with $\eta$ as temperature!

## A 2-D Stationary Distribution

Let $p$ be a probability density on two parameters $(x, y)$.

We consider the case where $x$ and $y$ are completely independent with

$$\mathcal{L}(x, y) = \mathcal{L}(x) + \mathcal{L}(y)$$

For completely independent variables we have

$$\begin{aligned} p(x, y) &= p(x)p(y) \\ &= \frac{1}{Z} \exp\left(\frac{-\mathcal{L}(x)}{\alpha\eta\sigma_x^2} + \frac{-\mathcal{L}(y)}{\alpha\eta\sigma_y^2}\right) \end{aligned}$$

# A 2-D Stationary Distribution

$$p(x,y) = \frac{1}{Z}\exp\left(\frac{-\mathcal{L}(x)}{\alpha\eta\sigma_x^2} + \frac{-\mathcal{L}(y)}{\alpha\eta\sigma_y^2}\right)$$

This is not a Gibbs distribution!

It has two different temperature parameters!

## Forcing a Gibbs Distribution

Suppose we use parameter-specific learning rates $\eta_x$ and $\eta_y$

$$p(x,y) = \frac{1}{Z}\exp\left(\frac{-\mathcal{L}(x)}{\alpha\eta_x\sigma_x^2} + \frac{-\mathcal{L}(y)}{\alpha\eta_y\sigma_y^2}\right)$$

Setting $\eta_x = \eta'/\sigma_x^2$ and $\eta_y = \eta'/\sigma_y^2$ gives

$$\begin{aligned} p(x,y) &= \frac{1}{Z}\exp\left(\frac{-\mathcal{L}(x)}{\alpha\eta'} + \frac{-\mathcal{L}(y)}{\alpha\eta'}\right) \\ &= \frac{1}{Z}\exp\left(\frac{-\mathcal{L}(x,y)}{\alpha\eta'}\right) \quad \text{Gibbs!} \end{aligned}$$

## The Case of Locally Constant Noise and Locally Quadratic Loss

In this case we can impose a change of coordinates under which the Hessian is the identity matrix. So without loss of generality we can take the Hessian to be the identity.

We can consider the covariance matrix of the vectors $\hat{g}$ in the Hessian-normalized coordinate system.

## The Case of Locally Constant Noise and Locally Quadratic Loss

If we assume constant noise covariance in the neighborhood of the stationary distribution then, in the Hessian normalized coordinate system, we get a stationary distribution

$$p(\Phi) \propto \exp\left(-\sum_i \frac{\Phi_i^2}{\alpha\eta\sigma_i^2}\right)$$

where $\Phi_i$ is the projection of $\Phi$ onto to a unit vector in the direction of the $i$th eigenvector of the noise covariance matrix and $\sigma_i^2$ is the corresponding noise eigenvalue (the variance of the $\hat{g}_i$).

# END

# Appendix: Diffusion Flow

In the SDE formalism we move stochastically from $x$ to $x + \epsilon\sqrt{\Delta t}$ with $\epsilon \sim \mathcal{N}(0, \eta\sigma^2)$.

To get the order of dependence on $\eta$ and $\sigma$ replace this with drawing $\epsilon$ from the uniform distribution on $[-\sqrt{\eta}\sigma, \sqrt{\eta}\sigma]$.

# Appendix: Diffusion Flow

We will draw $\Delta x$ the uniform distribution on $[-\Delta, \Delta]$ with $\Delta = \sqrt{\eta}\sigma\sqrt{\Delta t}$.

The quantity of mass transfer in the interval $\Delta t$ from values below $x$ to values above $x$ is then the following where $x - z$ is the source of the mass.

$$\begin{aligned} & \int_{z=0}^{\Delta} p(\Delta x \geq z) p(x - z) dz \\ = & \int_{z=0}^{\Delta} \frac{\Delta - z}{2\Delta} \left( p(x) - \frac{dp}{dx} z \right) dz \\ = & \frac{1}{4} p(x) \Delta - \frac{1}{12} \frac{dp}{dx} \Delta^2 \end{aligned}$$

# Appendix: Diffusion Flow

The mass transfer from below $x$ to above $x$ in interval $\Delta t$ is now

$$\frac{1}{4}p(x)\Delta - \frac{1}{12}\frac{dp}{dx}\Delta^2 \quad \text{for} \quad \Delta = \sqrt{\eta}\sigma\sqrt{\Delta t}.$$

Subtracting a similar calculation for the downward mass transfer cancels the first term and doubles the second term.

We then get that the net mass transfer is

$$-\frac{1}{6}\frac{dp}{dx}\eta\sigma^2\Delta t$$

So the mass transfer per unit time — the diffusion flow — is

$$-\frac{1}{6}\eta\sigma^2\frac{dp}{dx}$$

The constant $1/6$ must be adjusted for Gaussian noise rather than uniform noise.

# END