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 with Constant Gradiant Noise

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

We will assume the stationary distribution is limited to a region where the gradient noise is effectively constant.

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

The diffusion flow is $-\frac{1}{2}\,\eta\sigma^2\,dp(x)/dx$ (see the appendix).

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

$$\frac{1}{2}\eta\sigma^2\frac{dp}{dx} = -p\frac{d\mathcal{L}}{dx}$$

The 1-D Stationary Distribution

$$\begin{aligned}
\frac{1}{2}\eta\sigma^2\frac{dp}{dx} &= -p\frac{d\mathcal{L}}{dx} \\
\frac{dp}{p} &= \frac{-2d\mathcal{L}}{\eta\sigma^2} \\
\ln p &= \frac{-2\mathcal{L}}{\eta\sigma^2} + C \\
p(x) &= \frac{1}{Z}\exp\left(\frac{-2\mathcal{L}(x)}{\eta\sigma^2}\right)
\end{aligned}$$

We get a Gibbs distribution with η 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

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

A 2-D Stationary Distribution

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

This is not a Gibbs distribution!

It has two different temperature parameters!

Forcing a Gibbs Distribution

Suppose we use parameter-specific learning rates η_x and η_y

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

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

$$\begin{aligned} p(x,y) &= \frac{1}{Z}\exp\left(\frac{-2\mathcal{L}(x)}{\eta'} + \frac{-2\mathcal{L}(y)}{\eta'}\right) \\ &= \frac{1}{Z}\exp\left(\frac{-2\mathcal{L}(x,y)}{\eta'}\right) \quad \text{Gibbs!} \end{aligned}$$

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

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

where Φ_i is the projection of Φ onto to a unit vector in the direction of the ith eigenvector of the noise covariance matrix and σ_i^2 is the corresponding noise eigenvalue (the variance of the $\hat{g}_i$).

END

Appendix: Diffusion Flow

We consider the one dimensional case. In the SDE formalism we move stochastically from x to $x + \epsilon\sqrt{\Delta t}$ with $\epsilon \sim \mathcal{N}(0, \eta\sigma^2)$ where η is the learning rate and σ^2 is the variance of the random gradients $\hat{g}_{t,b}$.

To avoid confusion between different probability densities we will us $\rho(x)$ for the probability mass distribution in x and use $p_\epsilon(\Psi)$ for the probability that Ψ holds under a random draw of ϵ.

Appendix: Diffusion Flow

We move stochastically from x to $x + \epsilon\sqrt{\Delta t}$ with $\epsilon \sim \mathcal{N}(0, \eta\sigma^2)$

This is the same as moving stochastically from x to $x + \epsilon\sqrt{\eta}\sigma\sqrt{\Delta t}$ with $\epsilon \sim \mathcal{N}(0, 1)$.

The quantity of mass transfer in the time interval Δt from values above x to values below x is

$$\begin{aligned} & \int_{z=0}^{\infty} \rho(x+z)\; p_\epsilon(\epsilon\sqrt{\eta}\sigma\sqrt{\Delta t} \leq -z) dz \\ = & \int_{z=0}^{\infty} \rho(x+z)\; p_\epsilon\left(\epsilon \leq \frac{-z}{\sqrt{\eta}\sigma\sqrt{\Delta t}}\right) dz \\ = & \int_{z=0}^{\infty} \rho(x+z)\; \Phi\left(\frac{-z}{\sqrt{\eta}\sigma\sqrt{\Delta t}}\right) dz \end{aligned}$$

where Φ is the cummulative function of the Gaussian.

Appendix: Diffusion Flow

The quantity of mass transfer in the time interval Δt from values above x to values below x is

$$\int_{z=0}^{\infty} \rho(x+z)\ \Phi\left(\frac{-z}{\sqrt{\eta}\sigma\sqrt{\Delta t}}\right) dz$$

By a change of variables $u = z/(\sqrt{\eta}\sigma\sqrt{\Delta t})$ we get

$$\int_{u=0}^{\infty} \rho(x+\sqrt{\eta}\sigma\sqrt{\Delta t}\ u)\ \Phi(-u)\sqrt{\eta}\sigma\sqrt{\Delta t}\ du$$

As $\Delta t \to 0$ we can use the first order Taylor expansion of the density.

$$\sqrt{\eta}\sigma\sqrt{\Delta t}\int_{u=0}^{\infty}\left(\rho(x)+\sqrt{\eta}\sigma\sqrt{\Delta t}\ u\frac{d\rho(x)}{dx}\right)\ \Phi(-u)\ du$$

Appendix: Diffusion Flow

$$\sqrt{\eta}\sigma\sqrt{\Delta t}\int_{u=0}^{\infty}\left(\rho(x)+\sqrt{\eta}\sigma\sqrt{\Delta t}\,u\frac{d\rho(x)}{dx}\right)\,\Phi(-u)\,du$$

$$= \sqrt{\eta}\sigma\sqrt{\Delta t}\,\rho(x)\left(\int_{u=0}^{\infty}\Phi(-u)\,du\right)+\eta\sigma^2\Delta t\frac{d\rho(x)}{dx}\left(\int_{u=0}^{\infty}u\Phi(-u)du\right)$$

A similar alanysis shows that the mass transfer from lower values to higher values is

$$= \sqrt{\eta}\sigma\sqrt{\Delta t}\,\rho(x)\left(\int_{u=0}^{\infty}\Phi(-u)\,du\right)-\eta\sigma^2\Delta t\frac{d\rho(x)}{dx}\left(\int_{u=0}^{\infty}u\Phi(-u)du\right)$$

The net mass transfer in the positive x direction is the second minus the first or

$$= -2\eta\sigma^2\Delta t\frac{d\rho(x)}{dx}\left(\int_{u=0}^{\infty}u\Phi(-u)du\right)$$

Appendix: Diffusion Flow

The net mass transfer in the positive x direction is

$$-2\eta\sigma^2\Delta t\frac{d\rho(x)}{dx}\left(\int_{u=0}^{\infty} u\Phi(-u)du\right)$$

Note that the mass transfer is proportional to Δt. Dividing by Δt gives the flow per unit time.

$$\text{Diffusion flow} \;\; = -\alpha\eta\sigma^2\frac{d\rho(x)}{dx} \qquad \alpha = 2\int_{u=0}^{\infty} u\Phi(-u)du$$

α can be calculated using integration by parts.

$$\begin{aligned}
\alpha &= 2\int_0^{\infty} u\Phi(-u)du \\
&= \int_0^{\infty} \Phi(-u)du^2 \\
&= u^2\Phi(-u)|_0^{\infty} + \int_0^{\infty} u^2\phi(-u)du \;\; \text{where } \phi \text{ is the Gaussian density} \\
&= \int_0^{\infty} u^2\phi(-u)du \\
&= \frac{1}{2}
\end{aligned}$$

Appendix: Diffusion Flow

We now have that the diffusion flow is

$$\text{Diffusion flow} \quad = -\frac{1}{2}\,\eta\sigma^2\frac{d\rho(x)}{dx}$$

For dimension larger than 1 we have

$$\text{Diffusion flow} \quad = -\frac{1}{2}\,\eta\Sigma\nabla_x\rho$$

Where $\Sigma = E\,(\hat{g}-g)(\hat{g}-g)^\top$ is the covariance matrix of the gradient noise.

Here we have derived this from first principle but it also follows from the Fokker–Planck equation (see Wikipedia).

END