TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2023

Stochastic Differential Equations (SDEs)

The Diffusion SDE

The Langevin SDE

General SDEs

The SGD SDE

Diffufion

Consider a discrete-time process $z(0), z(1), z(2), z(3), \ldots$ with $z(n) \in \mathbb{R}^d$ defined by

$$\begin{aligned} z(0) &= y, \quad y \sim \mathrm{pop}(y) \\ z(n) &= z(n) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \end{aligned}$$

We can sample from $z(n)$ using

$$\begin{aligned} z(0) &= y, \quad y \sim \mathrm{pop}(y) \\ z(n) &= z(0) + \epsilon\sqrt{n}, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \end{aligned}$$

The Diffusion SDE

Fix a numerical time step Δt and consider a discrete-time process $z(0)$, $z(\Delta t)$, $z(2\Delta t)$, $\ldots$

$$\begin{aligned} z(0) &= y, \quad y \sim \mathrm{pop}(y) \\ z(t+\Delta t) &= z(t) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \Delta t \sigma^2 I) \\ &= z(t) + \epsilon\sqrt{\Delta t}, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \end{aligned}$$

We now take the limit of this numerical simulation as $\Delta t \to 0$.

This limit defines a probability measure on the space of functions $z(t)$.

The Diffusion SDE

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

In the limit of arbitrary small time step numerical simulation, this equation holds for any continuous $t \geq 0$ and any $\Delta t \geq 0$.

The Langevin SDE

Consider gradient flow.

$$\begin{aligned} \frac{d\Phi(t)}{dt} &= g(\Phi) \\ g(\Phi) &= \nabla_\Phi \, \mathcal{L}(\Phi) \\ \mathcal{L}(\Phi) &= E_{(x,y)\sim \mathrm{Pop}} \, \mathcal{L}(\Phi, x, y) \end{aligned}$$

The Langevin SDE

In the Langevin SDE we add Gaussian noise to gradient flow.

$$\Phi(t+\Delta t) \;=\; \Phi(t) + g(\Phi)\Delta t + \epsilon\sqrt{\Delta t}, \quad \epsilon \sim \mathcal{N}(0, \tau I)$$

We will show that the stationary distribution of Langevin Dynamics models a Bayesian posterior probability distrbution on the model parameters where β acts as a temperture parameter.

The Langevin SDE

$$\Phi(t+\Delta t) \;=\; \Phi(t) + g(\Phi)\Delta t + \epsilon\sqrt{\Delta t}, \quad \epsilon \sim \mathcal{N}(0, \tau I)$$

Let $p(\Phi)$ be an probability density on the parameter space Φ.

The density $p(\Phi)$ defines a gradient flow and a diffusion flow.

$$\text{gradient flow} \;=\; -p(\Phi)g(\Phi)$$

$$\text{diffusion flow} \;=\; -\frac{1}{2}\,\tau\;\nabla_\Phi(p(\Phi))$$

A derivation of the diffusion flow is given in the appendix.

The Langevin SDE

$$\text{gradient flow} \; = \; -p(\Phi)g(\Phi)$$

$$\text{diffusion flow} \; = \; -\frac{1}{2}\,\tau\;\nabla_\Phi(p(\Phi))$$

For the stationary distribution these two flows cancel each other out. In one dimention we have

$$\frac{1}{2}\tau\frac{dp}{dx} = -p\frac{d\mathcal{L}}{dx}$$

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 $\hat{g}_i$).

Continuous Time Diffusion (Brownian Motion)

$$\begin{aligned} z(0) &= y, \quad y \sim \mathrm{pop}(y) \\ z(n\Delta t) &= z(0) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, n\Delta t \sigma^2 I) \end{aligned}$$

$$z(t) = z(0) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, t\sigma^2 I)$$

$$z(t + \Delta t) = z(t) +$$

Holding η Fixed

Consider SGD with batch size 1.

$$\Phi_{i+1} = \Phi_i - \eta \hat{g}_i$$

Unlike gradient flow, we now hold $\eta > 0$ fixed.

We will still take “time” to be the sum of the learning rates over the updates.

For N steps of SGD we define $\Delta t = N\eta$

Holding η Fixed

We consider Δt large enough so that Δt corresponds to many SGD updates.

We consider Δt small enough so that the gradient estimate distribution does not change over the interval Δt.

Applying the Central Limit Theorem

If the mean gradient $g(\Phi)$ is approximately constant over the interval $\Delta t = N\eta$ we have

$$\Phi(t + \Delta t) \approx \Phi(t) - g(\Phi)\Delta t + \eta \sum_{i=1}^{N} (g(\Phi) - \hat{g}_i)$$

The random variables in the last term have zero mean.

By the central limit theorem a sum (not the average) of N random vectors will approximate a Gaussian distribution where the variance grows like N.

Modeling Noise

The mean of $\hat{g}$ is the true gradient $g(\Phi)$.

$$\begin{aligned}
\Delta t &= \eta N \\
\Phi(t + \Delta t) &= \Phi(t) - \sum_{j=1}^{N} \eta \hat{g}_i \\
&= \Phi(t) - g(\Phi)\Delta t + \eta \sum_{j=1}^{N} (g(\Phi) - \hat{g}_i)
\end{aligned}$$

Modeling Noise

$$\Delta t \; = \; \Phi(t) - g(\Phi)\Delta t + \eta \sum_{j=1}^{N} (g(\Phi) - \hat{g}_i)$$

By the central limit theorem $\sum_{j=1}^{N}(g(\Phi) - \hat{g}_i)$ is approximately Gaussian.

In One Dimension

We first consider the case of one parameter (a one dimensional optimization problem) so that $\hat{g}$ is a scalar.

To model the noise as Gaussian we take $\epsilon \sim \mathcal{N}(0, \sigma^2)$ where σ^2 is the variance of $\hat{g}$.

$$\Phi(t + \Delta t) \approx \Phi(t) - g(\Phi)\Delta t + \eta\epsilon\sqrt{N}$$

In One Dimension

$$\begin{aligned}
\Phi(t+\Delta t) &\approx \Phi(t) - g(\Phi)\Delta t + \eta\epsilon\sqrt{N} \\
&= \Phi(t) - g(\Phi)\Delta t + \eta\epsilon\sqrt{\frac{\Delta t}{\eta}} \\
&= \Phi(t) - g(\Phi)\Delta t + \sqrt{\eta}\epsilon\sqrt{\Delta t} \\
&= \Phi(t) - g(\Phi)\Delta t + \epsilon'\sqrt{\Delta t}
\end{aligned}$$

With $\epsilon' \sim \mathcal{N}(0, \eta\sigma^2)$.

The Stochastic Differential Equation (SDE)

$$\begin{aligned}\Phi(t+\Delta t) &\approx \Phi(t) - g(\Phi)\Delta t + \epsilon\sqrt{\Delta t} \\ \epsilon &\sim \mathcal{N}(0, \eta\sigma^2)\end{aligned}$$

We can take this last equation to hold in the limit of arbitrarily small Δt in which case we get a continuous time stochastic process. This process can be written as

$$d\Phi = -g(\Phi)dt + \epsilon\sqrt{dt} \qquad \epsilon \sim \mathcal{N}(0, \eta\sigma^2)$$

Higher Dimension

In the higher dimensional case we get

$$d\Phi = -g(\Phi)dt + \epsilon\sqrt{dt} \qquad \epsilon \sim \mathcal{N}(0, \eta\Sigma)$$

Where Σ is the covariance matrix of $\hat{g}$.

In one dimension Σ is σ^2.

For $g(\Phi) = 0$ and $\Sigma = I$ we get Brownian motion.

But in general we do not have $\Sigma = I$.

END

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.

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

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).