TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2023

Continuous Time Models of SGD

Gradient Flow

The Diffusion SDE

The Langevin SDE

General SDEs

The SGD SDE

Gradient Flow

Gradient flow is a non-stochastic (deterministic) model of stochastic gradient descent (SGD).

Gradient flow is defined by the total gradient differential equation

$$\frac{d\Phi}{dt} = -g(\Phi) \qquad g(\Phi) = \nabla_\Phi \, E_{(x,y)\sim \mathrm{Train}} \, \mathcal{L}(\Phi, x, y)$$

We let $\Phi(t)$ be the solution to this differential equation satisfying $\Phi(0) = \Phi_{\mathrm{init}}$.

Gradient Flow

$$\frac{d\Phi}{dt} = -g(\Phi)$$

For small values of Δt this differential equation can be approximated by

$$\Delta\Phi = -g(\Phi)\Delta t$$

Time as the Sum of the Learning Rates

Consider the update.

$$\Delta\Phi = -g\Delta t$$

Here Δt has both a natural interpretation as time in a numerical simulation of the flow differential equation.

But it also has a natural interpretation as a learning rate.

This leads to interpreting the sum of the learning rates as "time" in SGD.

Gradient Flow and SGD

Consider a sequence of model parameters $\Phi_1, \ldots, \Phi_N$ produced by SGD with

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

and where $\hat{g}_i$ is the gradient of the ith randomly selected training point.

Take $\eta \to 0$ and $N \to \infty$ using $N = t/\eta$. We will show that in this limit for SGD we have that Φ_N converges to $\Phi(t)$ as defined by gradient flow.

Gradient Flow and SGD

For $\Phi_{i+1} = \Phi_i - \eta \hat{g}_i$ we divide Φ_1, …, Φ_N into $\sqrt{N}$ blocks.

$(\Phi_1, \ldots, \Phi_{\sqrt{N}})$ $(\Phi_{\sqrt{N}+1}, \ldots, \Phi_{2\sqrt{N}})$ $\cdots$ $(\Phi_{T-\sqrt{N}+1}, \ldots, \Phi_N)$

For $\eta \to 0$ and $N = t/\eta$ we have $\eta\sqrt{N} \to 0$ which implies

$$\Phi_{\sqrt{N}} \sim \Phi_0 - \eta\sqrt{N} g$$

where g is the average (non-stochastic) gradient.

Since the gradients within each block become non-stochastic, we are back to gradient flow.

Diffusion

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

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

We can sample from $z(n)$ using

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

Diffusion

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

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

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

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

For simple diffusion (Brownian motion) this equation holds for any continuous $t \geq 0$ and $\Delta t \geq 0$.

The Langevin SDE

Consider gradient flow.

$$\frac{d\Phi(t)}{dt} = g(\Phi)$$

$$g(\Phi) = \nabla_\Phi \, \mathcal{L}(\Phi)$$

$$\mathcal{L}(\Phi) = E_{(x,y)\sim \mathrm{Pop}} \, \mathcal{L}(\Phi, x, y)$$

The Langevin SDE

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

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

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

The Langevin SDE

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

Let $p(\Phi)$ be a probability density on the parameter space Φ. The density $p(\Phi)$ defines a gradient flow and a diffusion flow.

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

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

The expression for the diffusion flow follows from the Fokker-Plank equation. A derivation of the diffusion flow expression from first princple is given in the appendix.

The Langevin SDE

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

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

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

$$\frac{1}{2}\sigma^2\nabla_\Phi\, p = -p\nabla_\Phi\mathcal{L}$$

The Langevin Stationary Distribution

$$
\begin{aligned}
\frac{1}{2}\sigma^2 \nabla_\Phi \, p &= -p \nabla_\Phi \mathcal{L} \\
\nabla_\Phi \left(\frac{1}{2}\sigma^2 \ln p \right) &= \nabla_\Phi (-\mathcal{L}) \\
\frac{1}{2}\sigma^2 \ln p &= -\mathcal{L} + C \\
p(\Phi) &= \frac{1}{Z} e^{\frac{-2\mathcal{L}(\Phi)}{\sigma^2}}
\end{aligned}
$$

A Bayesian Interpretation of Langevin Dynamics

If we incorporate a regulaization into the loss function.

$$\mathcal{L}(\Phi) = E_{(x,y)\sim\mathrm{Train}}\left[-\ln P_\Phi(y|x)\right] + \lambda||\Phi||^2$$

Then the Gibbs distribution

$$p(\Phi) = \frac{1}{Z}e^{\frac{-2\mathcal{L}(\Phi)}{\sigma^2}}$$

can be interprete as a Bayesian posterior density $p(\Phi|\mathrm{Train})$.

A General SDE

$$x(t+\Delta t) \;=\; f(t) + \mu(x,t)\Delta t + \sigma(x,t)\epsilon\sqrt{\Delta t}, \quad \epsilon \sim \mathcal{N}(0,I) \quad (1)$$

Here $\sigma(x,t)$ is a matrix equal to $\Sigma(x,t)^{1/2}$ where $\Sigma = \sigma^2$ is the covarinace matrix of the random variable $\sigma(x,t)\epsilon$.

This is conventionally written as

$$dx = \mu(x,t)dt + \sigma(x,t)dB \quad (2)$$

where B denotes a Weiner process (simple diffusion, aka Brownian motion)

I find (1) more intuitive than (2) but they are the same thing.

The SGD SDE

We now consider SGD

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

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

Holding $\eta > 0$ fixed introduces gradient noise.

Gradient noise is shaped differently from the Brownian motion noise of Langevin dynamics.

The SGD SDE

As in Gradient flow we take "time" to be the sum of the learning rates over the updates.

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

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.

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.

Modeling Noise

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

Let $\sigma(\Phi)$ be the square root of the covariance matrix of the random variable $g - \hat{g}$ at parameter setting Φ.

In one dimension we can approximately sample from $\sum_{i=1}^{N}(g(\Phi) - \hat{g}_i)$ as

$$\sigma(\Phi)\epsilon\sqrt{N}, \quad \epsilon \sim \mathcal{N}(0,1)$$

In high dimension this becomes

$$\sigma(\Phi)\epsilon\sqrt{N}, \quad \epsilon \sim \mathcal{N}(0,I)$$

The SGD SDE

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

$$\begin{aligned}
\Phi(t + \Delta t) &\approx \Phi(t) - g(\Phi)\Delta t + \eta\sigma(\Phi)\epsilon\sqrt{N} \\
&= \Phi(t) - g(\Phi)\Delta t + \eta\sigma(\Phi)\epsilon\sqrt{\Delta t/\eta} \\
&= \Phi(t) - g(\Phi)\Delta t + \eta\sigma(\Phi)\epsilon\sqrt{\Delta t}
\end{aligned}$$

This is the SGD SDE

The SGD SDE in One Dimension

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

In one dimension, if the the gradient noise $\sigma(\Phi)$ is constant, then the SGD SDE has the same form as Langevin dynamics and we get.

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

This is Gibbs and provides an interpretation of the learning rate as temperature.

The SGD SDE in Higher Dimension

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

This is almost the general case of an SDE.

Here $g(\Phi)$ is the gradient of a scalar function. This is not true for a general SDE.

But the matrix $\sigma(\Phi)$ is arbitrary.

Here the learning rate η controls the level of noise but we do not in general have a Gibbs distribution.

The SGD SDE, A Counter Example

If we have two dimensions x and y where the loss separates as $\mathcal{L}(x,y) = \mathcal{L}(x) + \mathcal{L}(y)$, and the matrix $\sigma(\Phi)$ is constant and diagonal, each dimension behaves as an independent one dimensional SGD and we get.

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

This is not Gibbs.

END

Appendix: Diffusion Flow

We consider the one dimensional case where we have a function $x(t) \in \mathbb{R}$. We consider a very small time step Δt and consider only the diffusion flow.

$$x(t + \Delta t) \;=\; x(t) + \sigma \epsilon \sqrt{\Delta t}, \quad \epsilon \sim \mathcal{N}(0, 1)$$

We assume a density p_x of values of x and let $p_\epsilon(\epsilon)$ be the normal distribution $\mathcal{N}(0, 1)$ on ϵ.

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

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

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

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

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

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

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

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

Appendix: Diffusion Flow

$$\begin{aligned}
&\sigma\sqrt{\Delta t}\int_{u=0}^{\infty}\left(p_x(x)+\sigma\sqrt{\Delta t}\;u\frac{dp_x(x)}{dx}\right)\;\Phi(-u)\;du \\
=&\;\;\sigma\sqrt{\Delta t}\;p_x(x)\left(\int_{u=0}^{\infty}\Phi(-u)\;du\right)+\sigma^2\Delta t\frac{dp_x(x)}{dx}\left(\int_{u=0}^{\infty}u\Phi(-u)du\right)
\end{aligned}$$

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

$$= \;\;\sigma\sqrt{\Delta t}\;p_x(x)\left(\int_{u=0}^{\infty}\Phi(-u)\;du\right)-\sigma^2\Delta t\frac{dp_x(x)}{dx}\left(\int_{u=0}^{\infty}u\Phi(-u)du\right)$$

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

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

Appendix: Diffusion Flow

The net mass transfer in the positive x direction is

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

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

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

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

$$\text{Diffusion flow} \;\; = -\frac{1}{2}\,\sigma^2 \frac{dp_x(x)}{dx}$$

For dimension larger than 1 we have

$$\text{Diffusion flow} \;\; = -\frac{1}{2}\,\Sigma \nabla_x p_x$$

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