TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Stochastic Gradient Descent (SGD)

and Stochastic Differential Equations (SDEs)

Modeling the SGD as a Stochastic Process

If we randomly select training points then SGD is a stochastic process.

Can we analytically solve for stationary distributions?

Is the stationary distribution Gibbs — is it $\frac{1}{Z}e^{-\frac{\mathcal{L}}{kT}}$ where the temperature T is determined by the learning rate?

Modeling the SGD as a Stochastic Process

It is possible to model both the stationary distribution and non-stationary stochastic dynamics with a continuous time stochastic differential equation such as Brownian motion or Langevin Dynamics.

Langevin Dynamics is the special case where the stationary distribution is Gibbs.

We will show here that in general the stationary distribution of SGD is not Gibbs and hence does not correspond to Langevin dynamics.

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