TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Stochastic Gradient Descent (SGD)

Continuous Time Noise

Stationary Parameter Densities

Modeling the Noise

Can we analytically solve for stationary distributions?

Is the stationary distributin some kind of Gibbs Distribution?

It is possible to model both the stationary distribution and non-stationary stochastic dynamics with a continuous time stochastic differential equation.

Continuous Time Noise

Consider SGD with batch size 1.

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

To model noise we consider holding $\eta > 0$ fixed.

As in gradient flow, taking "time" to be the sum of the learning rates over the updates. For N steps of SGD we define $\Delta t = N\eta$ or equivalently

$$\Phi(t) = \Phi_{i(t)}$$

$$i(t) = \lfloor t/\eta \rfloor$$

Continuous Time Noise

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

We consider Δt small enough so that the gradient distribution does not change during the many-update interval Δt.

Continuous Time Noise

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 law of large numbers a sum (not the average) of N random vectors will approximate a Gaussian distribution where the standard deviation grows like $\sqrt{N}$.

Continuous Time Noise

Let Σ be the covariance matrix of the random variable $\hat{g}$ and assume this is approximately constant over the interval Δt. Let ϵ be a zero mean Gaussian random variable with the same covariance matrix Σ.

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

Continuous Time Noise

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

We can take this last equation to hold for all (arbitrarily small) Δt in which case we get a continuous time stochastic process. This process can be written as

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

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

Continuous Time Noise

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

Note that for $\eta \to 0$ the noise term vanishes. If we then take $\Delta t \to 0$ (at a slower rate) we are back to gradient flow.

To model noise we hold $\eta > 0$ fixed.

Stationary Distributions

SGD (at batch size 1) defines a Markov process

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

We will model the stationary distribution as a continuous density in parameter space.

If the noise covariance matrix is isotropic (all eigenvalues are the same) we get a Gibbs distribution.

The 1-D Stationary Distribution

Consider SGD on a single parameter.

Let p be a probability density on x.

Assume that the gradient $\hat{g}$ has variance σ everywhere.

There is a diffusion flow proportional to $\eta^2\sigma^2\ dp/dx$.

There is a gradient flow equal to $\eta p\ d\mathcal{L}/dx$.

For a stationary distribution the two flows cancel giving.

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

The 1-D Stationary Distribution

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

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

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

$$p(x) = \frac{1}{Z}\exp\left(\frac{-\mathcal{L}(x)}{\alpha\eta\sigma^2}\right) \quad \alpha \approx 1/10$$

We get a Gibbs distribution!

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

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

$$= \frac{1}{Z} \exp\left(-\beta_x \mathcal{L}(x) - \beta_y \mathcal{L}(y) \right)$$

This is not a Gibbs distribution!

It has two different temperature parameters!

Noise Models and RMSProp

Suppose we use parameter-specific learning rates η_x and η_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

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

Noise Models and RMSProp

Suppose we use parameter-specific learning rates η_x and η_y
Setting $\eta_x = \eta'/\sigma_x^2$ and $\eta_y = \eta'/\sigma_y^2$ gives

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

RMSProp sets $\eta_x = \eta'/\sigma_x$ rather than $\eta_x = \eta'/\sigma_x^2$. Empirically RMSProp seems better than the more theoretically motivated algorithm.

END