TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2019

Stochastic Gradient Descent (SGD)

Gradient Flow and Langevin Dynamics

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

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

We let $\Phi(t)$ be the solution satisfying $\Phi(0) = \Phi_{\text{init}}$.

Gradient Flow

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

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

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

Here Δt can be interpreted as a learning rate.

Gradient Flow also Models SGD

Consider the SGD update with $\hat{g}$ from a random data point

$$\Delta\Phi = -\hat{g}\Delta t$$

This also converges to deterministic gradient flow as Δt goes to zero.

To see this note that we can divide the updates into $\sqrt{N}$ blocks each of size $\sqrt{N}$. The "time" spent within each block is $t/\sqrt{N}$ which converges to 0. But the number of updates within each block grows as $\sqrt{N}$ and hence the average update within a block converges to g.

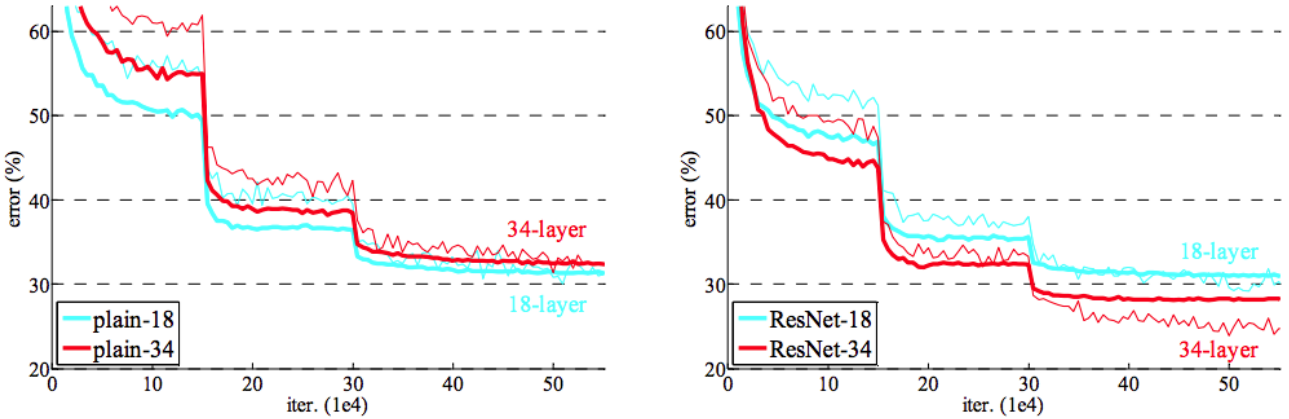
Gradient Flow also Models SGD

Consider the SGD update with $\hat{g}$ from a random data point

$$\Delta\Phi = -\hat{g}\Delta t$$

The sum of the learning rates can be interpreted as "time".

MCMC models of SGD

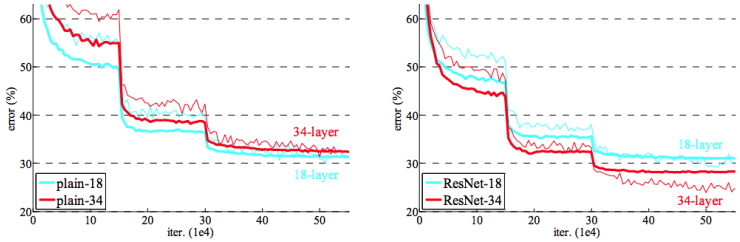


These Plots are from the original ResNet paper. Left plot is for CNNs without residual skip connections, the right plot is ResNet.

Thin lines are training error, thick lines are validataion error.

In all cases η is reduced twice, each time by a factor of 2.

Converged Loss as a Function of η



For each value of η we converge at a loss $\mathcal{L}(\eta)$.

$$\mathcal{L}(0) \doteq \lim_{\eta \to 0} \mathcal{L}(\eta)$$

$$= \mathcal{L}(\Phi^*) \quad \Phi^* \text{ a local optimum}$$

Can we do a Taylor expansion of $\mathcal{L}(\eta)$?

$$\mathcal{L}(\eta) = \mathcal{L}(\Phi^*) + \left(\frac{d\mathcal{L}}{d\eta}\bigg|_{\eta=0} \right) \eta + \ldots$$

A Nice Theorem (TZ)

$$\mathcal{L}(\eta) = \mathcal{L}(\Phi^*) + \left(\frac{d\mathcal{L}}{d\eta}\bigg|_{\eta=0} \right) \eta + \ldots$$

Let i index a training example and let g_i denote $\nabla_\Phi \mathcal{L}_i(\Phi)$ at $\Phi = \Phi^*$.

Theorem:

$$\frac{\partial \mathcal{L}(\eta)}{\partial \eta}\bigg|_{\eta=0} = \frac{1}{4} E_i \, ||g_i||^2$$

Proof Step 1 (TZ)

Let i index a training example and let $\mathcal{L}_i(\Phi^* + \Delta\Phi)$ be the loss on training example i with model parameters $\Phi^* + \Delta\Phi$.

We take a second order Taylor expansion.

$$\mathcal{L}(\Phi) = E_i \; \mathcal{L}_i(\Phi)$$

$$\mathcal{L}_i(\Phi^* + \Delta\Phi) = \mathcal{L}_i + g_i \Delta\Phi + \frac{1}{2} \Delta\Phi^\top H_i \Delta\Phi$$

$$E_i \; g_i = 0$$

$$E_i \; H_i \qquad \text{is positive definite}$$

Proof: Step 2 (TZ)

$$
\begin{aligned}
\mathcal{L}(\eta) = & \quad E_{\Delta\Phi\sim P_\eta}\, E_i \quad \mathcal{L}_i + g_i\Delta_\Phi + \frac{1}{2}\,\Delta\Phi^\top H_i\Delta\Phi \\
& = E_i\,\mathcal{L}_i + E_{\Delta\Phi\sim P_\eta} \quad (E_i\; g_i)\;\Delta\Phi + \frac{1}{2}\,\Delta\Phi^\top(E_i\; H_i)\Delta\Phi \\
& = \mathcal{L}(\Phi^*) \;\; + \;\; E_{\Delta\Phi\sim P_\eta} \;\; \frac{1}{2}\,\Delta\Phi^\top(E_i\; H_i)\Delta\Phi
\end{aligned}
$$

Proof: Step 3 (TZ)

Because P_η is a stationary distribution we must have

$$E_{\Delta\Phi\sim P_\eta} E_i \; ||\Delta\Phi - \eta(g_i + H_i\Delta\Phi)||^2 = E_{\Delta\Phi\sim P_\eta} \; ||\Delta\Phi||^2$$

$$E_{\Delta\Phi\sim P_\eta} E_i \; - 2\eta\Delta\Phi^\top(g_i + H_i\Delta\Phi) + \eta^2||(g_i + H_i\Delta\Phi)||^2 = 0$$

$$E_{\Delta\Phi\sim P_\eta} \left(\frac{1}{2}\Delta\Phi^\top(E_i \; H_i)\Delta\Phi\right) = \frac{\eta}{4} \; E_{\Delta\Phi\sim P_\eta} E_i \; ||(g_i + H_i\Delta\Phi)||^2$$

$$\mathcal{L}(\eta) = \mathcal{L}(\Phi^*) + \frac{\eta}{4} \; E_{\Delta\Phi\sim P_\eta} E_i \; ||(g_i + H_i\Delta\Phi)||^2$$

Proof Step 4 (TZ)

$$\mathcal{L}(\eta) = \mathcal{L}(\Phi^*) + \frac{\eta}{4} E_{\Delta\Phi\sim P_\eta} E_i \, ||(g_i + H_i \Delta\Phi)||^2$$

$$\left.\frac{\partial \mathcal{L}(\eta)}{\partial \eta}\right|_{\eta=0} = \frac{1}{4} \lim_{\eta\to 0} E_{\Delta\Phi\sim P_\eta} E_i \, ||(g_i + H_i \Delta\Phi)||^2$$

$$= \frac{1}{4} E_i \, ||g_i||^2$$

Langevin Dynamics

Can we analytically solve for stationary distributions?

Is the stationary distributin some kind of Gibbs Distribution?

Langevin dynamics models both the stationary distribution and non-stationary stochastic dynamics with a continuous time stochastic differential equation.

Langevin Dynaimcs

Consider SGD with $B = 1$.

$$\Phi \texttt{ -= } \eta \hat{g}$$

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

In Langevin dynamics we hold $\eta > 0$ fixed.

We then consider Δt large compared to η (so that it corresponds to many SGD updates) but small enough so that the gradient distribution does not change during the interval Δt.

Langevin Dynamics

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

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

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 $\sqrt{N}$.

Langevin Dynamics

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.

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

Langevin Dynamics

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

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

This can be modeled by a continuous time stochastic process — Langevin dynamics — defined by the notation

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

This is a stochastic differential equation.

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

Langevin Dynamics

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

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.

In Langevin dynamics we hold $\eta > 0$ fixed.

Stationary Distributions

SGD at $B = 1$ defines a Markov process

$$\Phi \texttt{ -= } \eta \hat{g}$$

Under Langevin dynamics the stationary distribution is a continuous density in parameter space.

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

The 1-D Langevin 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 Langevin 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 Langevin 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 Langevin 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!

Langevin 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

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

Langevin 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 that the more theoretically motivated algorithm.

END