TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Stochastic Gradient Descent (SGD)

The Classical Convergence Theorem

Vanilla SGD

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

$$\hat{g} = E_{(x,y)\sim\text{Batch}} \nabla_\Phi \text{loss}(\Phi, x, y)$$

$$g = E_{(x,y)\sim\text{Pop}} \nabla_\Phi \text{loss}(\Phi, x, y)$$

η is the "learning rate" hyper-parameter (a parameter not in Φ).

Issues

- **Gradient Estimation.** The accuracy of $\hat{g}$ as an estimate of g.
- **Gradient Drift (second order structure).** The fact that g changes as the parameters change.
- **Convergence.** To converge to a local optimum the learning rate must be gradually reduced toward zero.
- **Exploration.** Since deep models are non-convex we need to search over the parameter space. SGD can behave like MCMC.

A One Dimensional Example

Suppose that y is a scalar, and consider

$$\begin{aligned}
\text{loss}(\beta, y) &= \frac{1}{2}(\beta - y)^2 \\
g &= \nabla_\beta \, E_{y \sim \text{Pop}} \, \frac{1}{2}(\beta - y)^2 \\
&= \beta - E_{y \sim \text{Pop}} \, y \\
\hat{g} &= \beta - E_{y \sim \text{Batch}} \, y
\end{aligned}$$

Even if β is optimal, for a finite batch we will have $\hat{g} \neq 0$.

The Classical Convergence Theorem

$$\Phi \texttt{ -= } \eta_t \nabla_\Phi \operatorname{loss}(\Phi, x_t, y_t)$$

For "sufficiently smooth" non-negative loss with

$$\eta_t \geq 0 \qquad \lim_{t \to \infty} \eta_t = 0 \qquad \sum_t \eta_t = \infty \qquad \sum_t \eta_t^2 < \infty$$

we have that the training loss $E_{(x,y)\sim\text{Train}} \operatorname{loss}(\Phi, x, t)$ converges to a limit and any limit point of the sequence Φ_t is a stationary point in the sense that $\nabla_\Phi E_{(x,y)\sim\text{Train}} \operatorname{loss}(\Phi, x, t) = 0$.

Rigor Police: One can construct cases where Φ diverges to infinity, converges to a saddle point, or even converges to a limit cycle.

Physicist's Proof of the Convergence Theorem

Since $\lim_{t \to 0} \eta_t = 0$ we will eventually get to arbitrarilly small learning rates.

For sufficiently small learning rates any meaningful update of the parameters will be based on an arbitrarily large sample of gradients at essentially the same parameter value.

An arbitrarily large sample will become arbitrarily accurate as an estimate of the full gradient.

But since $\sum_t \eta_t = \infty$, no matter how small the learning rate gets, we still can make arbitrarily large motions in parameter space.

For a rigorous proof see Neuro-Dynamic Programming, Bertsekas and Tsitsiklis, 1996.

SGD as a form of MCMC

Learning Rate as a Temperature Parameter

Gao Huang et. al., ICLR 2017

END