Gradient Descent Convex Optimization Machine Learning +1

How does one show that the expected value of a mini-batch in SGD is equal to the true empirical gradient?

This question previously had details. They are now in a comment.

1 Answer

Conner Davis, Data Scientist at Microsoft
Answered May 4, 2017

The same way you show that the mean of any simple random sample is an unbiased estimator of the population mean: linearity of expectation.

Linearity of expectation just means that $E[X+Y]=E[X]+E[Y]$

We have that

$J(X)=\frac{1}{n}\sum_{i=1}^{n} Loss(f(x_i),y_i)$

Differentiate both sides and use the linearity of differentiation to move the ∇ inside the summation

$\nabla J(X)=\frac{1}{n}\sum_{i=1}^{n} \nabla Loss(f(x_i),y_i)$

We want to evaluate

$E_A[\frac{1}{m}\sum_{i=1}^{m} \nabla Loss(f(x_i),y_i)]$

Apply linearity of expectation

$=\frac{1}{m}\sum_{I=1}^{m} E_A[\nabla Loss(f(x_i),y_i)]$

What's $E_A[\nabla Loss(f(x_i),y_i)]$?

$E[X]=\sum_x x*P(X=x)$.

Since the examples are chosen uniformly at random, all their probabilities ($P(X=x)$) are equal to $\frac{1}{n}$, so it's just the average value of the gradient over all examples.

Mathematically, that's:

$E_A[\nabla Loss(f(x_i),y_i)]=\sum_{j=1}^{n} P(i=j)*\nabla Loss(f(x_j),y_j)$

Where $P(i=j)=\frac{1}{n}$, so

$E_A[\nabla Loss(f(x_i),y_i)]=\frac{1}{n}\sum_{j=1}^{n} \nabla Loss(f(x_j),y_j)=\nabla J(X)$

Plugging that back in to

$E_A[\frac{1}{m}\sum_{I=1}^{m} \nabla Loss(f(x_i),y_i)]=\frac{1}{m}\sum_{i=1}^{m} E_A[\nabla Loss(f(x_i),y_i)]$

We get

$E_A[\frac{1}{m}\sum_{i=1}^{m} \nabla Loss(f(x_i),y_i)]$

$=\frac{1}{m}\sum_{i=1}^{m} \nabla J(X)$

$=\nabla J(X)$

Just as we wanted.