FML Homework, Mar 8'21

Rami Ahmed

March 9, 2021

Proof that SGD provides unbiased gradient estimator. Given a realization of data, $\mathfrak{D} \in \mathbb{R}^N$, the *MSE* of a *linear regression* model is,

$$MSE = L(\mathbf{y},\ \mathbf{X};\ \boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)})^2 = L\,,$$

and the gradient of this loss is,

$$\mathbf{g} = \nabla_\theta L = \nabla_\theta L(\mathbf{y},\ \mathbf{X};\ \boldsymbol{\theta}) = \nabla_\theta \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)})^2\,, \quad \mathbf{g} \in \mathbb{R}^d$$

Now coming to **SGD**, we want to compute an estimate for the gradient using a sample from that realization; because computing it over the whole dataset is expensive, hence,

$$\hat{\mathbf{g}} = \nabla_\theta(y - \boldsymbol{\theta}^T\mathbf{x})^2; \quad (y, \mathbf{x})\ \sim\ \mathfrak{D}\,, \quad \hat{\mathbf{g}} \in \mathbb{R}^d\,.$$

However, we need an estimator that is unbiased from the true value of the gradient; because we want to land at the same optima that the true gradient leads to.

The bias of the gradient estimator $\hat{\mathbf{g}}$ is calculated as,

$$\mathbb{B}\text{ias}(\hat{\mathbf{g}}) = \mathbb{E}[\hat{\mathbf{g}}] - \mathbf{g},$$

therefore, in order to have an unbiased estimator we need,

$$\mathbb{B}\text{ias}(\hat{\mathbf{g}}) = \mathbf{0} \implies \mathbb{E}[\hat{\mathbf{g}}] = \mathbf{g}.$$

To calculate the expected value of the gradient estimator, we know that in the case of **SGD** we uniformly sample a random variable $(y, \mathbf{x})\ \sim\ \mathfrak{D}$, hence,

$$\mathbb{E}_{(y,\mathbf{x})\,\sim\,\mathfrak{D}}[\hat{\mathbf{g}}] = \mathbb{E}_{(y,\mathbf{x})\,\sim\,\mathfrak{D}}[\nabla_\theta L(y,\ \mathbf{x};\ \boldsymbol{\theta})] = \mathbb{E}_{(y,\mathbf{x})\,\sim\,\mathfrak{D}}[\nabla_\theta(y - \boldsymbol{\theta}^T\mathbf{x})^2],$$

and since we have a finite dataset $\mathfrak{D}$, our expectation is a summation -over samples- of the estimator -calculated at each sampled random variable- times the probability of each sample,

$$\mathbb{E}_{(y,\mathbf{x})\sim\mathfrak{D}}[\nabla_\theta(y-\boldsymbol{\theta}^T\mathbf{x})^2] = \sum_{i=1}^{N}\nabla_\theta(y^{(i)}-\boldsymbol{\theta}^T\mathbf{x}^{(i)})^2\cdot p(y^{(i)},\mathbf{x}^{(i)}),$$

and since the probability of each sample is identical (uniform distribution) then $p(y,\mathbf{x}) = \frac{1}{N}$, therefore,

$$\mathbb{E}_{(y,\mathbf{x})\sim\mathfrak{D}}[\nabla_\theta(y-\boldsymbol{\theta}^T\mathbf{x})^2] = \sum_{i=1}^{N}\nabla_\theta(y^{(i)}-\boldsymbol{\theta}^T\mathbf{x}^{(i)})^2\cdot\frac{1}{N},$$

also, since the sum of the gradients equals the gradient of the sum,

$$\mathbb{E}_{(y,\mathbf{x})\sim\mathfrak{D}}[\nabla_\theta(y-\boldsymbol{\theta}^T\mathbf{x})^2] = \nabla_\theta\frac{1}{N}\sum_{i=1}^{N}(y^{(i)}-\boldsymbol{\theta}^T\mathbf{x}^{(i)})^2,$$

therefore,

$$\mathbb{E}[\hat{\mathbf{g}}] = \mathbf{g} \implies \mathbb{B}\mathrm{i}\mathbb{a}\mathbb{s}(\hat{\mathbf{g}}) = \mathbf{0},$$

then the **SGD** gradient estimates is an unbiased estimator. ■