Mini-batches

© 2019 Philipp Krähenbühl and Chao-Yuan Wu

Stochastic Gradient Descent

- For n epochs:
  - for $\mathbf{x}, \mathbf{y} \sim D$:
    - $\theta := \theta - \epsilon \frac{d\ell(f(\mathbf{x}, \theta), \mathbf{y})}{d\theta}$

Stochastic Gradient Descent

- For n epochs:
    - for i in $0, \ldots, |D| - 1$
        - $\mathbf{x}, \mathbf{y} := D_i$
        - $\theta := \theta - \epsilon \dfrac{d\ell(f(\mathbf{x}, \theta), \mathbf{y})}{d\theta}$

Mini-batches

- For n epochs:
  - Split dataset D into m mini-batches $B_0, \ldots B_{m-1}$ of size BS
  - for each batch B_i
    - $\theta := \theta - \epsilon \mathbb{E}_{\mathbf{x},\mathbf{y} \sim B_i} \left[\frac{d\ell(f(\mathbf{x},\theta),\mathbf{y})}{d\theta} \right]$

Variance of mini-batches

- Variance of SGD
  - $\mathbb{E}_{\mathbf{x},\mathbf{y}\sim D}\left[\left(\frac{d\ell(f(\mathbf{x},\theta),\mathbf{y})}{d\theta}\right)^2\right] - \left(\frac{dL(\theta)}{d\theta}\right)^2$

- Variance of SGD with mini-batches
  - $\mathbb{E}_{B_i}\left[\left(\mathbb{E}_{\mathbf{x},\mathbf{y}\sim B_i}\left[\frac{d\ell(f(\mathbf{x},\theta),\mathbf{y})}{d\theta}\right]\right)^2\right] - \left(\frac{dL(\theta)}{d\theta}\right)^2$

Always use mini-batches

Variance of mini-batches

Jensen's inequality

$$\left(\mathbb{E}_{\mathbf{x},\mathbf{y}\sim B_i}\left[\frac{d\ell\left(f(\mathbf{x},\theta),\mathbf{y}\right)}{d\theta}\right]\right)^2 \leq \mathbb{E}_{\mathbf{x},\mathbf{y}\sim B_i}\left[\left(\frac{d\ell\left(f(\mathbf{x},\theta),\mathbf{y}\right)}{d\theta}\right)^2\right]$$