Stochastic Gradient Descent

Gradient Descent

- Repeat until convergence:
  - $\theta := \theta - \epsilon \frac{dL(\theta)}{d\theta}$

Gradient Descent

- Repeat until convergence:
  - $\theta_0 := \theta$
  - for $\mathbf{x}, \mathbf{y} \sim D$:
    - $\theta := \theta - \epsilon \frac{d\ell(f(\mathbf{x}, \theta_0), \mathbf{y})}{d\theta_0}$

Stochastic Gradient Descent

- Repeat until convergence:
  - for $\mathbf{x}, \mathbf{y} \sim D$:
    - $\theta := \theta - \epsilon \dfrac{d\ell(f(\mathbf{x}, \theta), \mathbf{y})}{d\theta}$

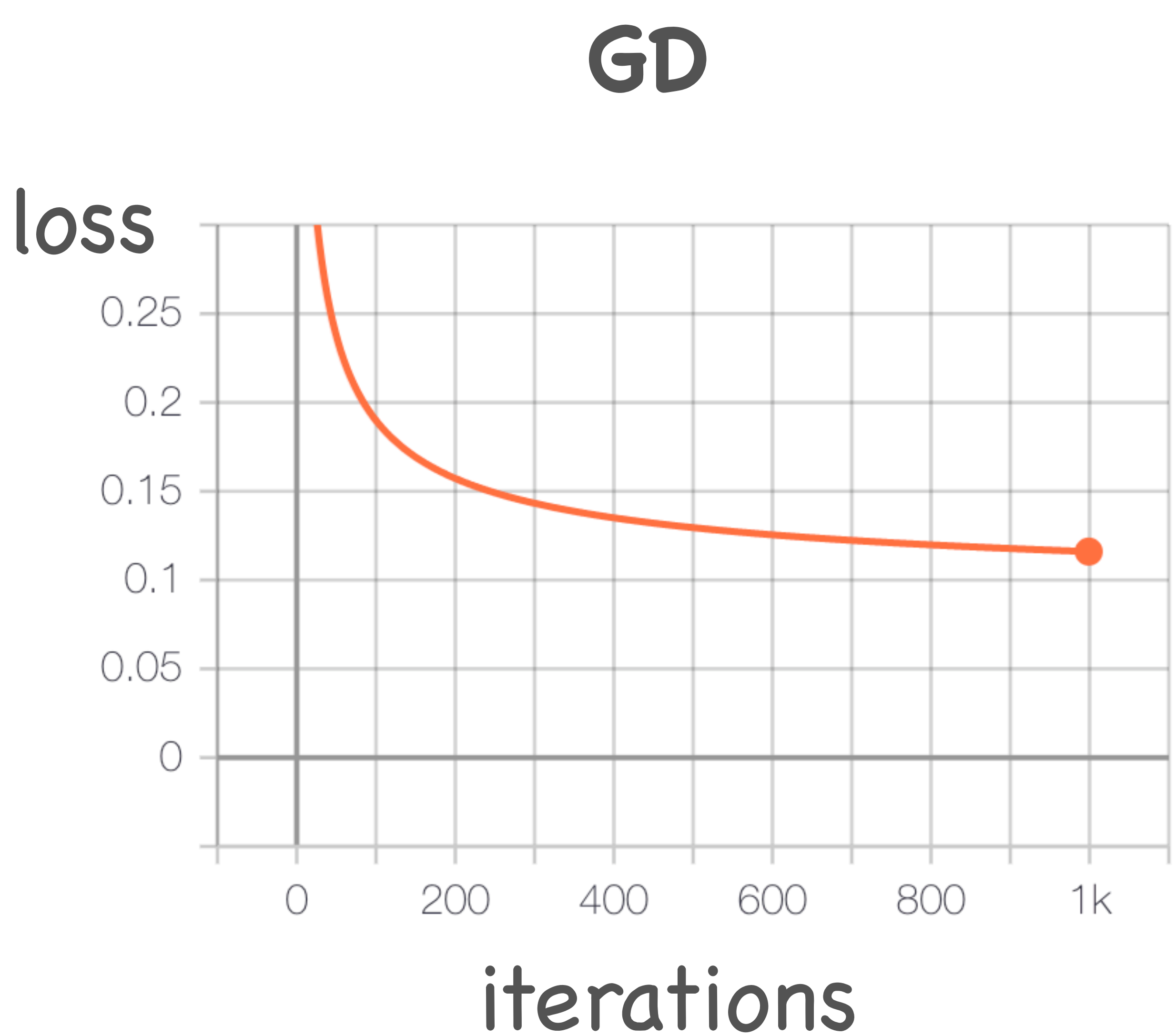
Terminology

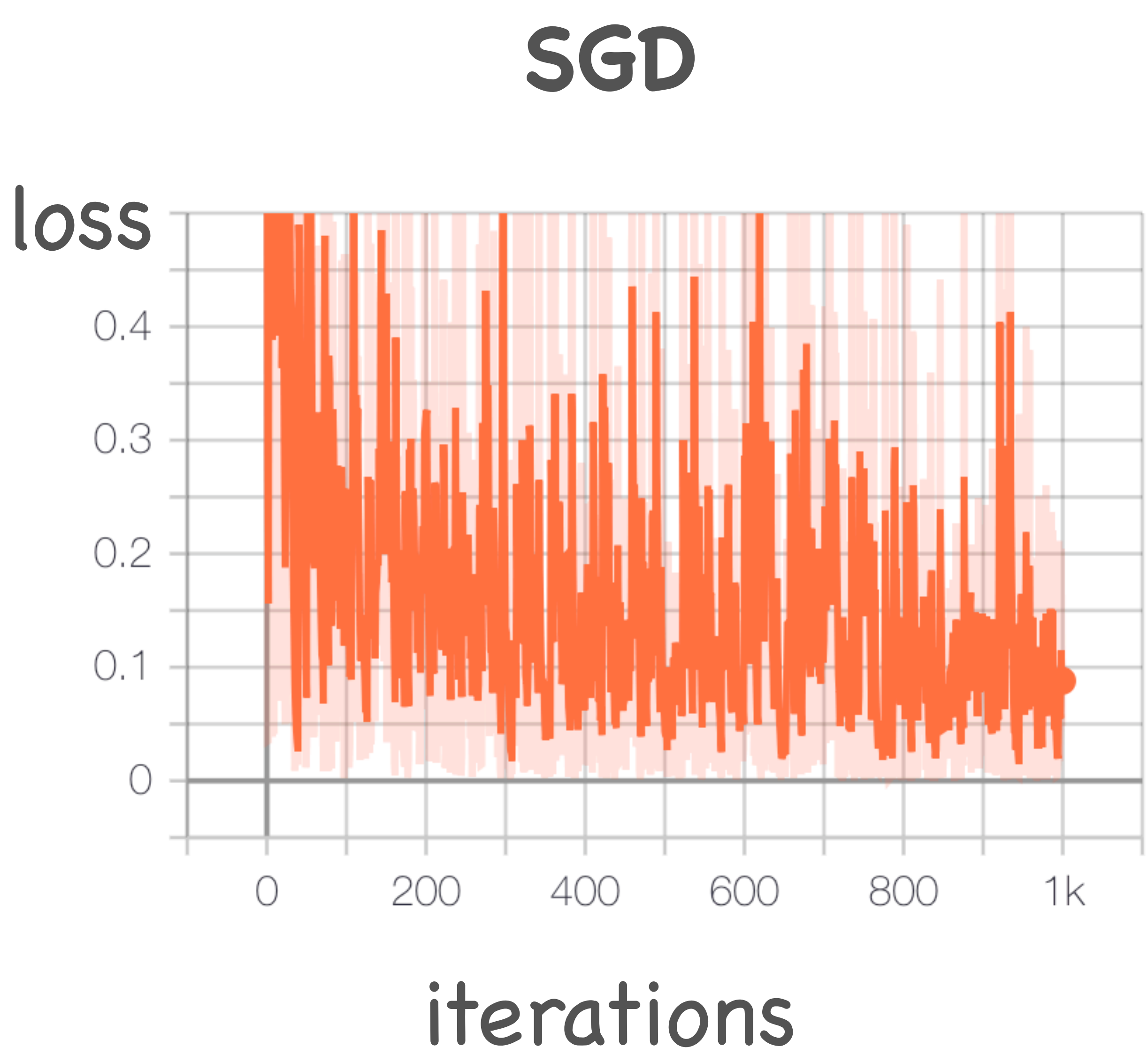
- Repeat until convergence:
  - for $\mathbf{x}, \mathbf{y} \sim D$:
    - $\theta := \theta - \epsilon \dfrac{d\ell(f(\mathbf{x}, \theta), \mathbf{y})}{d\theta}$

Practical SGD

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

Learning curves





The Variance of SGD

$$\frac{d\ell(f(\mathbf{x},\theta),\mathbf{y})}{d\theta} \neq \frac{dL(\theta)}{d\theta}$$

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

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