Momentum

- © 2019 Philipp Krähenbühl and Chao-Yuan Wu

Stochastic Gradient Descent

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

Momentum

- $\mathbf{v} := 0$
- For n epochs:
  - for $\mathbf{x}, \mathbf{y} \sim D$
    - $\mathbf{v} := \rho\mathbf{v} + \frac{d\ell(f(\mathbf{x}, \theta), \mathbf{y})}{d\theta}$
    - $\theta := \theta - \epsilon\mathbf{v}$

Variance reduction of Momentum

- Variance of SGD

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

Momentum