Practical Optimization

A Slight Change

$$D=\{(\mathrm{x}_{1,}y_1),(\mathrm{x}_{2,}y_2),\ldots,(\mathrm{x}_\mathrm{N},y_N)\}$$

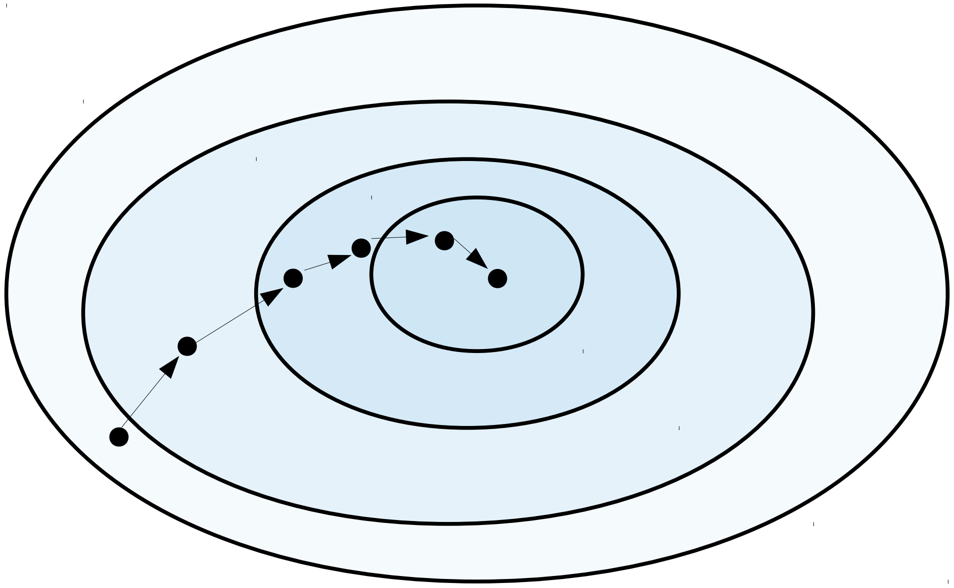
$$L(\theta)=\sum_i \ell(\theta;\mathrm{x}_i,y_i)$$

$$L(\theta)=\frac{1}{|D|}\sum_i \ell(\theta;\mathrm{x}_i,y_i)$$

$$=\mathrm{E}_{\mathrm{x},y\sim D}[\ell(\theta;\mathrm{x},y)]$$

Gradient Descent

How do we find $\operatorname{argmin}_{\theta} L(\theta)$



Slow!

Initialize random θ

N iterations:

Compute $\nabla_{\theta} L(\theta)$

$\theta \leftarrow \theta - \epsilon \nabla_{\theta} L(\theta)$

$\theta_0 \leftarrow \theta$

For $\mathrm{x}, y \in D$

$$\theta \leftarrow \theta - \frac{\epsilon}{|D|} \nabla_{\theta_0} \ell(\theta_0; f(\mathrm{x}), y)$$

Stochastic Gradient Descent

<u>Gradient Descent:</u>

Initialize random θ

N iterations:

$\cancel{\theta_0 \leftarrow \theta}$

For $x, y \in D$

$$\theta \leftarrow \theta - \frac{\epsilon}{|D|} \nabla_{\theta_{\cancel{0}}} \ell(\theta_{\cancel{0}}; f(x), y)$$

<u>Stochastic Gradient Descent:</u>

Initialize random θ

N iterations:

For $x, y \in D$

Iteration

$$\boxed{\theta \leftarrow \theta - \frac{\epsilon}{|D|} \nabla_{\theta} \ell(\theta; f(x), y)}$$

Epoch

Noisy but much faster than standard GD

SGD – Variance

$$\mathrm{E}_{\mathrm{x},y\sim D}\left[\nabla_\theta \ell(\theta;\mathrm{x},y)\right]=\nabla_\theta L(\theta)$$

$$\nabla_\theta \ell(\theta;\mathrm{x}_i,y_i)\neq\nabla_\theta L(\theta)$$

$$\begin{aligned}\mathrm{Var}\left[\nabla_\theta \ell(\theta;\mathrm{x},y)\right]&=\mathrm{E}_{\mathrm{x},y\sim D}\left[\left\|\nabla_\theta \ell(\theta;\mathrm{x},y)-\nabla_\theta L(\theta)\right\|^2\right]\\&=\mathrm{E}_{\mathrm{x},y\sim D}\left[\left\|\nabla_\theta \ell(\theta;\mathrm{x},y)\right\|^2\right]-\left\|\nabla_\theta L(\theta)\right\|^2\end{aligned}$$

Mini-batches

<u>Stochastic Gradient Descent:</u>

Initialize random θ

N iterations:

For $\mathrm{x}, y \in D$

$$\theta \leftarrow \theta - \frac{\epsilon}{|D|} \nabla_\theta \ell(\theta; f(\mathrm{x}), y)$$

<u>Minibatch Gradient Descent:</u>

Initialize random θ

N iterations:

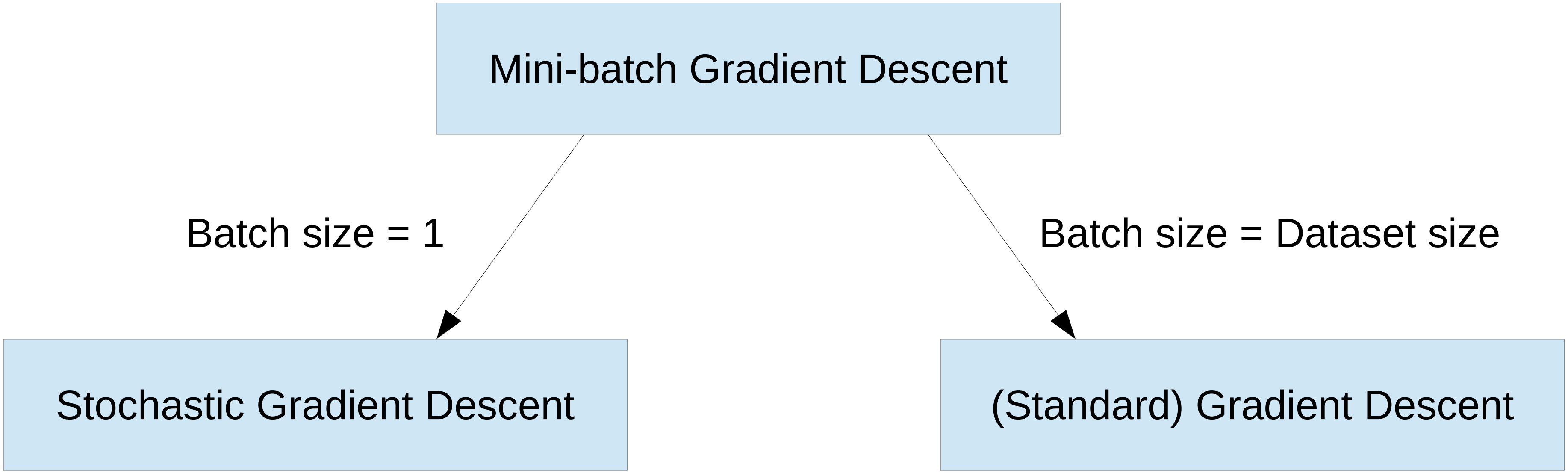
Partition D into $D_1, \ldots, D_k$

For $1 \leq i \leq k$

$$\theta \leftarrow \theta - \epsilon \mathrm{E}_{\mathrm{x}, y \sim D_i} \left[\nabla_\theta \ell(\theta; f(\mathrm{x}), y) \right]$$

Batch size = $|D_1| = |D_2| = \ldots = |D_k|$

Gradient Descent Algorithms



Mini-batch Variance

Stochastic Gradient Descent

$$\mathrm{Var}\left[\nabla_\theta \ell(\theta;\mathrm{x},y)\right]=\mathrm{E}_{\mathrm{x},y\sim D}\left[\left\|\nabla_\theta \ell(\theta;\mathrm{x},y)\right\|^2\right]-\left\|\nabla_\theta L(\theta)\right\|^2$$

Minibatch Gradient Descent

$$\mathrm{Var}\left[\nabla_\theta \ell(\theta;\mathrm{x},y)\right]=\mathrm{E}_{D_i}\left[\mathrm{E}_{\mathrm{x},y\sim D_i}\left[\left\|\nabla_\theta \ell(\theta;\mathrm{x},y)\right\|\right]^2\right]-\left\|\nabla_\theta L(\theta)\right\|^2$$

Smaller

Momentum

Initialize random θ

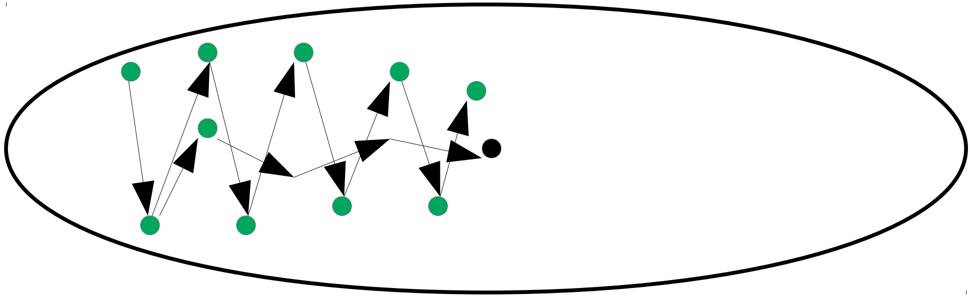
N iterations:

For $\mathrm{x}, y \in D$

$\cancel{\mathrm{v} \leftarrow \nabla_\theta \ell(\theta; f(\mathrm{x}), y)}$

$\mathrm{v} \leftarrow \boxed{\rho}\mathrm{v} + \nabla_\theta \ell(\theta; f(\mathrm{x}), y)$

$\theta \leftarrow \theta - \epsilon \mathrm{v}$



$\rho = 0.9$