

Stochastic gradient descent

GD minimizes:

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \underbrace{e(h(\mathbf{x}_n), y_n)}_{\ln(1+e^{-y_n \mathbf{w}^T \mathbf{x}_n})} \leftarrow \text{in logistic regression}$$

by iterative steps along $-\nabla E_{\text{in}}$:

$$\Delta \mathbf{w} = -\eta \nabla E_{\text{in}}(\mathbf{w})$$

∇E_{in} is based on all examples (\mathbf{x}_n, y_n)

“batch” GD

The stochastic aspect

Pick one (\mathbf{x}_n, y_n) at a time. Apply GD to $\mathbf{e}(h(\mathbf{x}_n), y_n)$

“Average” direction:

$$\begin{aligned}\mathbb{E}_n \left[-\nabla \mathbf{e}(h(\mathbf{x}_n), y_n) \right] &= \frac{1}{N} \sum_{n=1}^N -\nabla \mathbf{e}(h(\mathbf{x}_n), y_n) \\ &= -\nabla E_{\text{in}}\end{aligned}$$

randomized version of GD

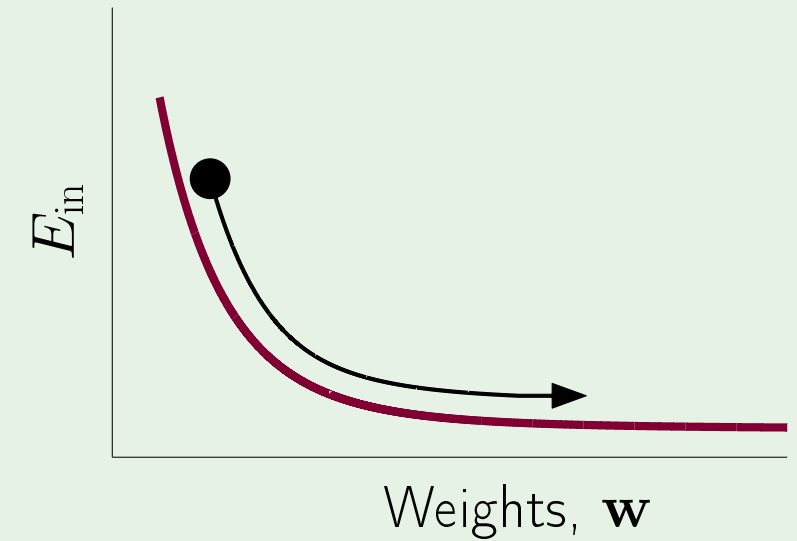
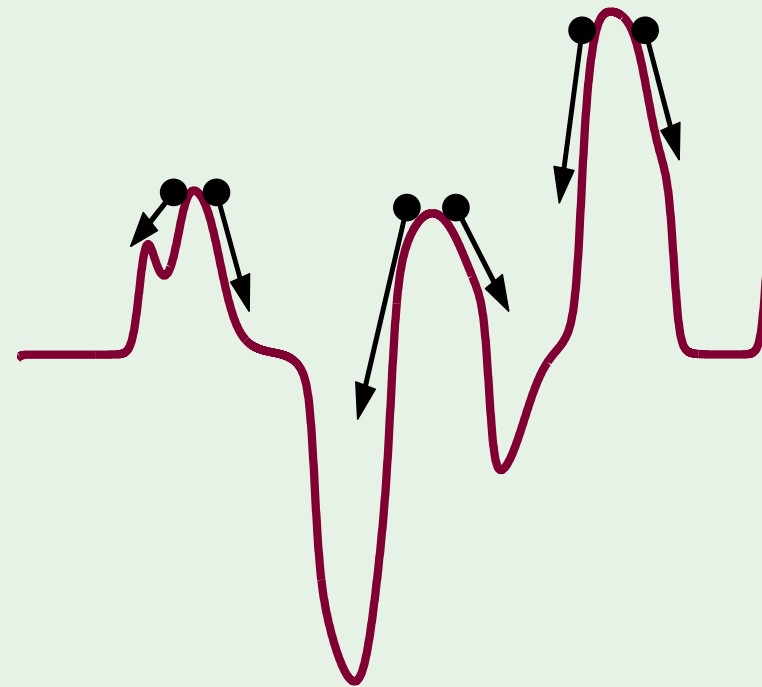
stochastic gradient descent (SGD)

Benefits of SGD

1. cheaper computation
2. randomization
3. simple

Rule of thumb:

$$\eta = 0.1 \text{ works}$$



randomization helps

SGD in action

Remember movie ratings?

$$e_{ij} = \left(r_{ij} - \sum_{k=1}^K u_{ik} v_{jk} \right)^2$$

