# Optimization Basics

Find $\bar{w}$ to minimize loss; search over space of params

loss $\mathcal{L}\left(\underbrace{(\bar{x}^{(i)}, y^{(i)})_{i=1}^{D}}_{\text{training data}}, \bar{w}\right) = \sum_{i=1}^{D} \mathcal{L}(\bar{x}^{(i)}, y^{(i)}, \bar{w})$

function of $\bar{w}$

Stochastic gradient descent: repeatedly pick example $i$

$$\bar{w} \leftarrow \bar{w} - \alpha \frac{\partial}{\partial \bar{w}} \mathcal{L}(i, \bar{w})$$

$\alpha$: step size; $\mathcal{L}(i, \bar{w})$: loss on $i$th example

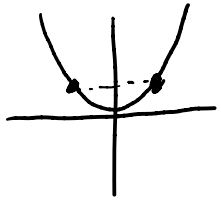
**Step size** Suppose $\mathcal{L}(i, \bar{w}) = w^2$ one feature $\bar{w} = [w]$

$w = -1$

$\frac{\partial}{\partial w} \mathcal{L} = 2w$

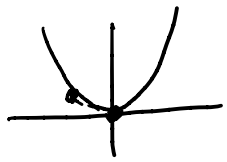
if $\alpha = 1$: $w = -1 - 1 \cdot 2(-1)$

$\Rightarrow w = 1$



Keep $\alpha = 1$: oscillates!

if $\alpha = 1/2$: $w \to 0$



# Choosing Step Size

How to choose step size?

- Try them out: $1e^{0}$ $1e^{-1}$ $1e^{-2}$ ...
- Large → small, e.g. $1/t$ for epoch $t$ ($1/\sqrt{t}$ ...)
  (fixed schedule)

  or decrease step size when performance stagnates on held-out data

Newton's method: $\bar{w} \leftarrow \bar{w} - \underbrace{\left(\frac{\partial^2}{\partial \bar{w}^2}\mathcal{L}\right)^{-1}}_{\text{inverse Hessian}} \frac{\partial}{\partial \bar{w}}\mathcal{L}$ very expensive to compute

DL

Adagrad, Adadelta, **Adam**: "adaptive" methods

approximations to the inverse Hessian (linear in # feats)

Regularization: don't really use