Unconstrained optimization II

Topics we'll cover

1. Why does gradient descent work?
2. Setting the step size
3. Gradient descent for logistic regression

Gradient descent

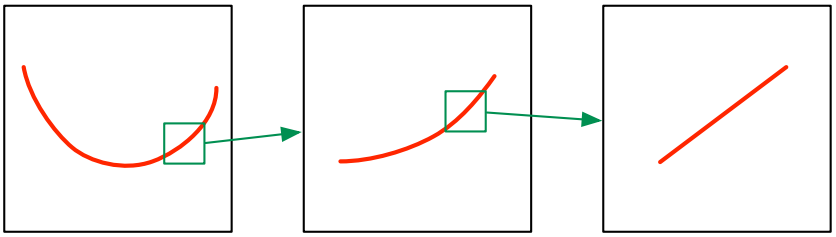
For minimizing a function $L(w)$, over $w \in \mathbb{R}^d$:

- $w_o = 0, t = 0$
- while $\nabla L(w_t) \not\approx 0$:
  - $w_{t+1} = w_t - \eta_t \nabla L(w_t)$
  - $t = t + 1$

Here η_t is the *step size* at time t.

Gradient descent: rationale

"Differentiable" $\implies$ "locally linear".



For *small* displacements $u \in \mathbb{R}^d$,

$$L(w+u) \approx L(w) + u \cdot \nabla L(w) \quad .$$

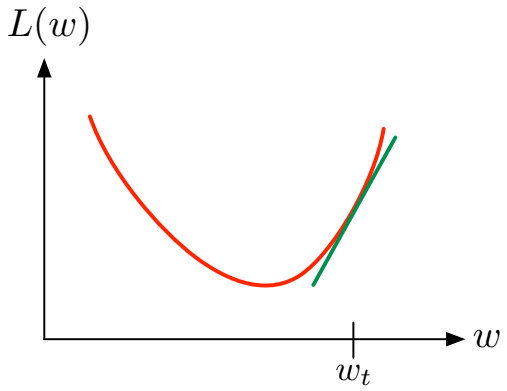
Therefore, if $u = -\eta \nabla L(w)$ is small,

$$L(w+u) \approx L(w) - \eta \|\nabla L(w)\|^2 < L(w)$$

The step size matters

Update rule: $w_{t+1} = w_t - \eta_t \nabla L(w_t)$

- Step size η_t too small: not much progress
- Too large: overshoot the mark



Some choices:

- Set η_t according to a fixed schedule, like $1/t$
- Choose by line search to minimize $L(w_{t+1})$

Example: logistic regression

For $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$, loss function

$$L(w) = \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})})$$

What is the derivative?

Gradient descent for logistic regression

- Set $w_0 = 0$
- For $t = 0, 1, 2, \ldots$, until convergence:

$$w_{t+1} \;=\; w_t \;+\; \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} \Pr_{w_t}(-y^{(i)} | x^{(i)})$$