Gradient Descent

David S. Rosenberg

New York University

January 23, 2018

Gradient Descent

Unconstrained Optimization

Setting

Objective function $f : \mathbf{R}^d \to \mathbf{R}$ is *differentiable.*
Want to find

$$x^* = \arg\min_{x \in \mathbf{R}^d} f(x)$$

The Gradient

- Let $f : \mathbf{R}^d \to \mathbf{R}$ be differentiable at $x_0 \in \mathbf{R}^d$.
- The **gradient** of f at the point x_0, denoted $\nabla_x f(x_0)$, is the direction to move in for the **fastest increase** in $f(x)$, when starting from x_0.

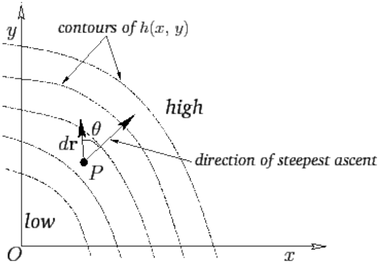


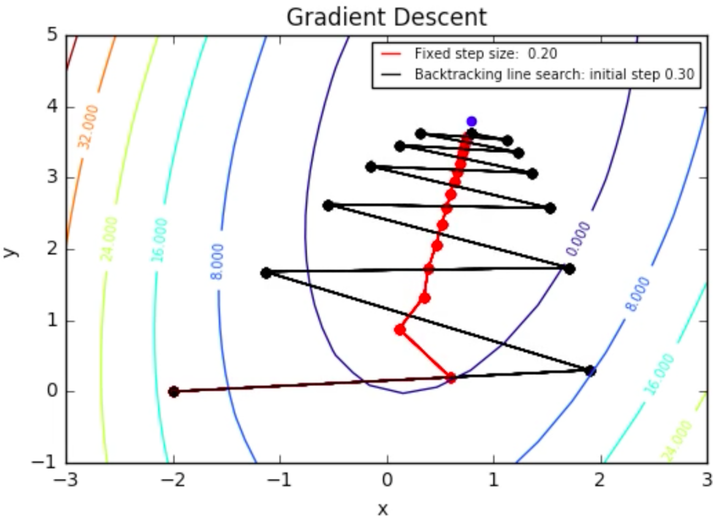
Figure A.111 from Newtonian Dynamics, by Richard Fitzpatrick.

Gradient Descent

Gradient Descent

- Initialize $x = 0$
- repeat
  - $x \leftarrow x - \underbrace{\eta}_{\text{step size}} \nabla f(x)$
- until stopping criterion satisfied

Gradient Descent Path



Gradient Descent: Step Size

- A fixed step size will work, eventually, as long as it's small enough (roughly - details to come)
  - Too fast, may diverge
  - In practice, try several fixed step sizes
- Intuition on when to take big steps and when to take small steps?
  - Demo.

Convergence Theorem for Fixed Step Size

Theorem

*Suppose $f : \mathbf{R}^d \to \mathbf{R}$ is convex and differentiable, and ∇f is **Lipschitz continuous** with constant $L > 0$, i.e.*

$$\|\nabla f(x) - \nabla f(x')\| \leqslant L\|x - x'\|$$

*for for any $x, x' \in \mathbf{R}^d$. Then gradient descent with fixed step size $\eta \leqslant 1/L$ **converges**. In particular,*

$$f(x^{(k)}) - f(x^*) \leqslant \frac{\|x^{(0)} - x^*\|^2}{2\eta k}.$$

Step Size: Practical Note

- Although a $1/L$ step-size guarantees convergence,
  - it may be **much slower** than necessary.
- May be worth trying larger step sizes as well.
- But math tells us, no need for anything smaller.

Gradient Descent: When to Stop?

- Wait until $\|\nabla f(x)\|_2 \leqslant \varepsilon$, for some ε of your choosing.
    - (Recall $\nabla f(x) = 0$ at minimum.)
- For learning setting,
    - evalute performance on validation data as you go
    - stop when not improving, or getting worse

Gradient Descent for Empirical Risk (And Other Averages)

Linear Least Squares Regression

Setup

- Input space $\mathcal{X} = \mathbf{R}^d$
- Output space $\mathcal{Y} = \mathbf{R}$
- Action space $\mathcal{Y} = \mathbf{R}$
- Loss: $\ell(\hat{y}, y) = (y - \hat{y})^2$
- **Hypothesis space:** $\mathcal{F} = \left\{ f : \mathbf{R}^d \to \mathbf{R} \mid f(x) = w^T x,\, w \in \mathbf{R}^d \right\}$

- Given data set $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,
  - Let's find the ERM $\hat{f} \in \mathcal{F}$.

Linear Least Squares Regression

Objective Function: Empirical Risk

The function we want to minimize is the empirical risk:

$$\hat{R}_n(w) = \frac{1}{n}\sum_{i=1}^{n}\left(w^T x_i - y_i\right)^2,$$

where $w \in \mathbf{R}^d$ parameterizes the hypothesis space $\mathcal{F}$.

- Now let's think more generally...