# Gradient and Stochastic Gradient Descent

David S. Rosenberg

January 29, 2015

# *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: $\ell(\hat{y}, y) = \frac{1}{2}(y - \hat{y})^2$
- **Hypothesis space:** $\mathcal{F} = \left\{ f : \mathcal{X} \to \mathcal{Y} \mid f(x) = w^T x \right\}$

- 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 $\mathcal{F}$.

# 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

## Definition

The **gradient** $\nabla_x f(x_0)$ of a differentiable function $f(x)$ at the point $x_0$ is the direction to move in for the **fastest increase** in $f(x)$, when starting from $x_0$.

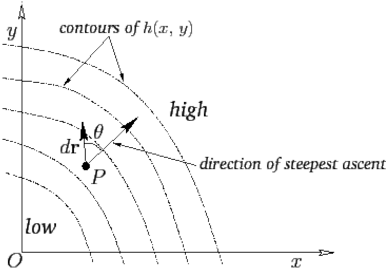



Figure: Figure A.111 from *Newtonian Dynamics*, by Richard Fitzpatrick.

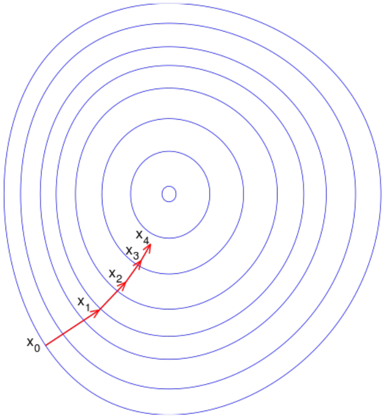
# Gradient Descent

### Gradient Descent

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

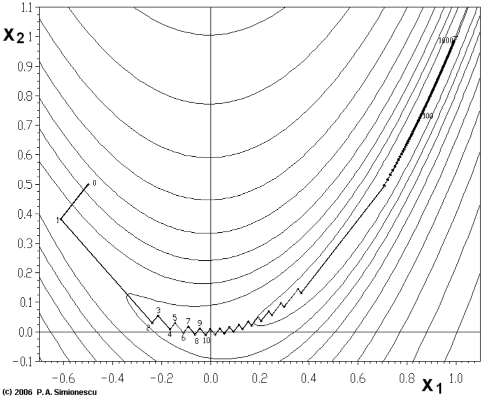
# Gradient Descent Path

Gradient Descent for a nice (convex) function



# Gradient Descent Path

## Gradient Descent Path for the Rosenbrock Function (not convex)




(Figure by P.A. Simionescu from Wikipedia page on gradient descent )

# Gradient Descent - Details

## Step Size

- Empirically $\eta = 0.1$ often works well
- **Better**: Optimize at every step (e.g. backtracking line search)

## Stopping Rule

- Could use a maximum number of steps (e.g. 100)
- Wait until $\|\nabla f(x)\| \leqslant \varepsilon$.
- Test performance on holdout data (in learning setting)

# Gradient Descent for Linear Regression

### Gradient of Objective Function:

The gradient of the objective is

$$\begin{aligned}\nabla_w \hat{R}_n(w) =& \nabla_w \left[ \frac{1}{n} \sum_{i=1}^{n} \left( w^T x_i - y_i \right)^2 \right] \\ =& \frac{2}{n} \sum_{i=1}^{n} \underbrace{\left( w^T x_i - y_i \right)}_{i\text{th residual}} x_i\end{aligned}$$

# Gradient Descent: Does it scale?

- At every iteration, we compute the gradient at current $w$:

$$\nabla_w \hat{R}_n(w) = \frac{2}{n} \sum_{i=1}^{n} \underbrace{\left(w^T x_i - y_i\right)}_{i\text{th residual}} x_i$$

- We have to touch all $n$ training points to take a single step. $[O(n)]$
  - Called a **batch optimization** method
- Can we make progress without looking at all the data?

# Gradient Descent on the Risk

- Real goal is to minimize the risk (expected loss):

$$\arg\min_{f \in \mathcal{F}} \mathbb{E}\left[\ell(f(X), Y)\right]$$

- For linear regression, that's

$$\arg\min_{w} \mathbb{E}\left(w^T X - Y\right)^2$$

- Gradient descent on this?

$$\nabla_w \mathbb{E}\left(w^T X - Y\right)^2 = \mathbb{E}\left[2\left(w^T X - Y\right) X\right]$$

# Gradient Descent on the Risk [approximately]

- Want to find gradient of the risk:

$$\nabla R(w) = \mathbb{E}\left[2\left(w^T X - Y\right) X\right]$$

- Can estimate expectation with a sample:

$$\widehat{\nabla R(w)} = \frac{1}{n}\sum_{i=1}^{n}\left[2\left(\underbrace{w^T x_i - y_i}_{\text{i'th residual}}\right) x_i\right]$$

- Let's return to the general case...

# Gradient Descent on the Risk: General Case

Gradient of Risk:

- Say hypothesis space $\mathcal{F}$ is parameterized by $w \in \mathbf{R}^d$.
- Switching $\nabla_w$ and $\mathbb{E}$ we can write the gradient of risk as

$$\text{Gradient(Risk)} = \nabla_w \mathbb{E}\left[\ell(f(X),Y)\right] = \mathbb{E}\left[\nabla_w \ell(f(X),Y)\right]$$

**Unbiased** estimator for Gradient(Risk):

$$\frac{1}{n}\sum_{i=1}^{n}\left[\nabla_w \ell(f_w(x_i),y_i)\right] \approx \underbrace{\mathbb{E}\left[\nabla_w \ell(f(X),Y)\right]}_{\text{Gradient(Risk)}}$$

# Gradient Descent on the Risk: General Case

- We want Gradient(Risk)
- Estimate it using sample of size $n$
- Bigger $n \implies$ Better estimate
- Bigger $n \implies$ Touching more data (slower!)
- But how big an $n$ do we need?

# Gradient Descent on the Risk [approximately]

- Gradient descent takes a bunch of steps whether we use
  - the perfect step direction $\nabla R(w)$,
  - an empirical estimate using all training data $\nabla \hat{R}_n(w)$, or
  - an empirical estimate using a random subset of data $\nabla \hat{R}_N(w)$ ($N \ll n$)
- What about $N = 1$?
- Even with a sample of size 1, the estimate

$$\nabla_w \ell(f_w(x_i), y_i)$$

is still **unbiased for gradient(Risk).**

# Stochastic Gradient Descent (SGD)

### Stochastic Gradient Descent

- initialize $w = 0$
- repeat
  - randomly choose training point $(x_i, y_i) \in \mathcal{D}_n$
  - $w \leftarrow w - \eta \underbrace{\nabla_w \ell(f_w(x_i), y_i)}_{\text{Grad(Loss on i'th example)}}$
- until stopping criteria met

# SGD: Step Size

- Let $\eta_t$ be the step size at the $t$'th step.
- How should $\eta_t$'s decrease with each step?

### Robbins-Monroe Conditions

Many classical convergence results depend on the following two conditions:

$$\sum_{t=1}^{\infty} \eta_t^2 < \infty \qquad \sum_{t=1}^{\infty} \eta_t = \infty$$

- As fast as $\eta_t = O\left(\frac{1}{t}\right)$ would satisfy this... but should be faster than $O\left(\frac{1}{\sqrt{t}}\right)$.
- A useful reference for practical techniques: Leon Bottou's "Tricks": http://research.microsoft.com/pubs/192769/tricks-2012.pdf