Gradient and Stochastic Gradient Descent

David S. Rosenberg

January 29, 2015
**Linear Least Squares Regression**

**Setup**

- Input space $\mathcal{X} = \mathbb{R}^d$
- Output space $\mathcal{Y} = \mathbb{R}$
- Action space $\mathcal{Y} = \mathbb{R}$
- Loss: $\ell(\hat{y}, y) = \frac{1}{2} (y - \hat{y})^2$
- **Hypothesis space:** $\mathcal{F} = \{ f : \mathcal{X} \rightarrow \mathcal{Y} \mid f(x) = w^T x \}$

Given data set $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,

- Let’s find the ERM $\hat{f} \in \mathcal{F}$. 
Objective Function: Empirical Risk

The function we want to minimize is the empirical risk:

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

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

Setting

Objective function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is \textit{differentiable}. Want to find

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

- Initialize $x = 0$
- repeat
  - $x \leftarrow x - \eta \nabla f(x)$
    - step size
- 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 $\|\nabla f(x)\| \leq \epsilon$.
- Test performance on holdout data (in learning setting)
Gradient Descent for Linear Regression

Gradient of Objective Function:
The gradient of the objective is

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

\[ = \frac{2}{n} \sum_{i=1}^{n} (w^T x_i - y_i) x_i \]

\text{i-th residual}
Gradient Descent: Does it scale?

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

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

- 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):
  \[
  \arg\min_{f \in \mathcal{F}} \mathbb{E} [\ell(f(X), Y)]
  \]

- For linear regression, that’s
  \[
  \arg\min_w \mathbb{E} (w^T X - Y)^2
  \]

- Gradient descent on this?
  \[
  \nabla_w \mathbb{E} (w^T X - Y)^2 = \mathbb{E} \left[ 2 (w^T X - Y) X \right]
  \]
Want to find gradient of the risk:

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

Can estimate expectation with a sample:

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

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 $w \in \mathbb{R}^d$.

- Switching $\nabla_w$ and $\mathbb{E}$ we can write the gradient of risk as

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

Unbiased estimator for Gradient(Risk):

\[
\frac{1}{n} \sum_{i=1}^{n} [\nabla_w \ell(f_w(x_i), y_i)] \approx \mathbb{E} [\nabla_w \ell(f(X), Y)]
\]
Gradient Descent on the Risk: General Case

- We want $\text{Gradient}(\text{Risk})$
- 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 $(x_i, y_i) \in \mathcal{D}_n$
  - $w \leftarrow w - \eta \nabla_w \ell(f_w(x_i), y_i)$
    - 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:

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

- 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”: