236606 - Deep Learning
Tutorial 2

Yonatan Geifman    Yair Feldman
Ran El-Yaniv

Computer Science Department
Technion - Israel Institute of Technology

Winter 2018-2019
Table of Contents

1 Gradient Based Optimization

2 Linear Regression
   • Analytical Solution
   • Gradient Descent
Table of Contents

1. Gradient Based Optimization

2. Linear Regression
   - Analytical Solution
   - Gradient Descent
Minimize a given objective $f(x)$ (loss)
Minimize a given objective $f(x)$ (loss)

The derivative $f'(x) = \frac{df(x)}{dx}$ tells us how to change $x$ to minimize $f$, by moving right or left depending on $f'$. 
Gradient Based Optimization

- Minimize a given objective $f(x)$ (loss)
- The derivative $f'(x) = \frac{df(x)}{dx}$ tells us how to change $x$ to minimize $f$, by moving right or left depending on $f'$.
- This technique is called **Gradient Descent**.
Critical Points

Points $x$ for which $f'(x) = 0$ are called critical points.
Critical Points

- Points $x$ for which $f'(x) = 0$ are called critical points.
- These points correspond to either (local) minimum, (local) maximum or saddle points.
Multivariate Functions

- $f(x) = f(x_1, ..., x_n)$ is a multivariate function
Multivariate Functions

- \( f(\mathbf{x}) = f(x_1, \ldots, x_n) \) is a multivariate function
- The partial derivative
  \[
  \frac{\partial f(\mathbf{x})}{\partial x_i}
  \]
  measures how \( f \) changes only w.r.t \( x_i \).
Multivariate Functions

- $f(\mathbf{x}) = f(x_1, ..., x_n)$ is a multivariate function
- The partial derivative
  $$\frac{\partial f(\mathbf{x})}{\partial x_i}$$
  measures how $f$ changes only w.r.t $x_i$.
- The gradient of $f$ is the vector of all partial derivatives, that is
  $$\nabla f(\mathbf{x}) = \left[ \frac{\partial f(\mathbf{x})}{\partial x_1}, ..., \frac{\partial f(\mathbf{x})}{\partial x_n} \right]^T$$
Multivariate Functions

- $f(x) = f(x_1, \ldots, x_n)$ is a multivariate function
- The partial derivative
  $$\frac{\partial f(x)}{\partial x_i}$$
  measures how $f$ changes only w.r.t $x_i$.
- The gradient of $f$ is the vector of all partial derivatives, that is
  $$\nabla f(x) = \left[ \frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n} \right]^T$$
- Critical points of a multivariate $f$ are those: $\nabla f = 0$. 
The directional derivative, $\nabla_u f(x_0)$, is the rate at which the function $f(x)$ changes at point $x_0$ in the direction $u$ (a unit vector vector in $\mathbb{R}^n$).
The directional derivative, $\nabla_u f(x_0)$, is the rate at which the function $f(x)$ changes at point $x_0$ in the direction $u$ (a unit vector vector in $\mathbb{R}^n$).

It is a vector form of the standard derivative, and can be defined as $\nabla_u f = \lim_{h \to 0} \frac{f(x+h u) - f(x)}{h}$.
The directional derivative, $\nabla_{\text{u}} f(x_0)$, is the rate at which the function $f(x)$ changes at point $x_0$ in the direction $\text{u}$ (a unit vector vector in $\mathbb{R}^n$).

It is a vector form of the standard derivative, and can be defined as $\nabla_{\text{u}} f = \lim_{h \to 0} \frac{f(x + hu) - f(x)}{h}$.

From Inf 2:

$$\nabla_{\text{u}} f(x) = \nabla f(x) \cdot \text{u}$$
Directional Derivative

- The directional derivative, $\nabla_u f(x_0)$, is the rate at which the function $f(x)$ changes at point $x_0$ in the direction $u$ (a unit vector vector in $\mathbb{R}^n$).
- It is a vector form of the standard derivative, and can be defined as $\nabla_u f = \lim_{h \to 0} \frac{f(x+hu) - f(x)}{h}$
- From Inf 2:
  $$\nabla_u f(x) = \nabla f(x) \cdot u$$
- To which direction $u$ should we go to minimize $f$?
Steepest Slopes

- Find a \( \mathbf{u} \) for which \( \nabla_{\mathbf{u}} f(\mathbf{x}) \) is minimized.
Steepest Slopes

- Find a $\mathbf{u}$ for which $\nabla_{\mathbf{u}} f(\mathbf{x})$ is minimized.
- $\theta$ is the angle between the vectors $\nabla f(\mathbf{x})$ and $\mathbf{u}$. 

$$\cos \theta = \nabla f(\mathbf{x}) \cdot \mathbf{u} / \| \nabla f(\mathbf{x}) \|$$

The cosine is minimized when the direction of $\mathbf{u}$ is the opposite of the direction of the gradient $\nabla f(\mathbf{x})$ (at which the cosine is -1).

We would like to take small steps in direction opposite of the gradient.
Steepest Slopes

- Find a $\mathbf{u}$ for which $\nabla_{\mathbf{u}} f(\mathbf{x})$ is minimized.
- $\theta$ is the angle between the vectors $\nabla f(\mathbf{x})$ and $\mathbf{u}$.
- $\cos \theta = \frac{\nabla f(\mathbf{x}) \cdot \mathbf{u}}{||\nabla f(\mathbf{x})||}$
Steepest Slopes

- Find a $\mathbf{u}$ for which $\nabla_{\mathbf{u}} f(\mathbf{x})$ is minimized.
- $\theta$ is the angle between the vectors $\nabla f(\mathbf{x})$ and $\mathbf{u}$.
- $\cos \theta = \frac{\nabla f(\mathbf{x}) \cdot \mathbf{u}}{||\nabla f(\mathbf{x})||}$
- The cosine is minimized when the direction of $\mathbf{u}$ is the opposite of the direction of the gradient $\nabla f(\mathbf{x})$ (at which the cosine is -1).
Steepest Slopes

- Find a $\mathbf{u}$ for which $\nabla_{\mathbf{u}} f(\mathbf{x})$ is minimized.
- $\theta$ is the angle between the vectors $\nabla f(\mathbf{x})$ and $\mathbf{u}$.
- $\cos \theta = \frac{\nabla f(\mathbf{x}) \cdot \mathbf{u}}{\|\nabla f(\mathbf{x})\|}$
- The cosine is minimized when the direction of $\mathbf{u}$ is the opposite of the direction of the gradient $\nabla f(\mathbf{x})$ (at which the cosine is -1).
- We would like to take small steps in direction opposite of the gradient.
Gradient Descent

- To minimize $f$ we update $\mathbf{x}$ as follows,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x})$$
Gradient Descent

- To minimize $f$ we update $x$ as follows,

$$x_{t+1} = x_t - \eta \nabla f(x)$$

- $\eta$ is called the learning rate.
Gradient Descent

- To minimize $f$ we update $\mathbf{x}$ as follows,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x})$$

- $\eta$ is called the **learning rate**.
- Convergence: critical point achieved $\nabla f = 0$. 

Gradient Based Optimization
Linear Regression
Gradient Descent

- To minimize $f$ we update $\mathbf{x}$ as follows,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x})$$

- $\eta$ is called the **learning rate**.
- Convergence: critical point achieved $\nabla f = 0$.
- This could be a saddle point.
Gradient descent optimization is called a **first-order method**, because it uses only first derivative information.
Gradient descent optimization is called a **first-order method**, because it uses only first derivative information.

Optimization methods that also use second derivative information (Hessian) are called second-order methods, e.g. newton method.
Gradient Descent

- Gradient descent optimization is called a **first-order method**, because it uses only first derivative information.

- Optimization methods that also use second derivative information (Hessian) are called second-order methods, e.g. newton method.

- In high dimensional problems calculating the Hessian is expensive \((Od^2)\)
Gradient Descent

- Gradient descent optimization is called a **first-order method**, because it uses only first derivative information.
- Optimization methods that also use second derivative information (Hessian) are called second-order methods, e.g. newton method.
- In high dimensional problems calculating the Hessian is expensive \( (Od^2) \)
- It is evident that first order methods are sufficient for deep learning
Gradient Descent

- The learning rate is a crucial parameter for the convergence of GD
Gradient Descent

- The learning rate is a crucial parameter for the convergence of GD
- 1D example:
Table of Contents

1 Gradient Based Optimization

2 Linear Regression
   • Analytical Solution
   • Gradient Descent
Linear Regression

- Linear regression - \( f : \mathcal{X} \rightarrow \mathbb{R} \) with \( f \) linear.
Linear Regression

- Linear regression - $f : \mathcal{X} \rightarrow \mathbb{R}$ with $f$ linear.
- The vanilla linear model is an affine vector function specified by a weight vector $\mathbf{w}$ and a bias term (scalar) $b$,

$$\hat{y} = \sum_{i=1}^{d} (w_i x_i) + b = \mathbf{w} \cdot \mathbf{x} + b$$
Let $y$ be the true value and $\hat{y}$ the predicted value.
Linear Regression

- Let $y$ be the true value and $\hat{y}$ the predicted value.
- The squared loss is defined as:

$$
\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2
$$
Linear Regression

- Let \( y \) be the true value and \( \hat{y} \) the predicted value.
- The squared loss is defined as:
  \[
  \ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2
  \]
- Given a training set \( S_m = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \), we train the model by minimizing the empirical risk with respect to the training set.

  \[
  \hat{L}(w, b) = \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, \hat{y}_i) = \frac{1}{2m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \\
  = \frac{1}{2m} \sum_{i=1}^{m} (y_i - (w \cdot x_i + b))^2
  \]
We would like to solve the optimization problem

$$\min_{w,b}(\hat{L}(w, b))$$
Linear Regression

- We would like to solve the optimization problem

\[
\min_{w,b} (\hat{L}(w, b))
\]

- We will show two methods to solve it.
We would like to solve the optimization problem

$$\min_{w,b}(\hat{L}(w, b))$$

We will show two methods to solve it.

- Analytical solution
We would like to solve the optimization problem

\[
\min_{w,b}(\hat{L}(w, b))
\]

We will show two methods to solve it.
- Analytical solution
- Gradient descent
We arrange the training set $S_m$ as a matrix $X$ where each row $i$ is $x_i$, the dimensions are $m \times d$, we put the $y$ observations in a vector.
Analytical Solution

- We arrange the training set $S_m$ as a matrix $X$ where each row $i$ is $x_i$, the dimensions are $m \times d$, we put the $y$ observations in a vector.
- In this setting our prediction is $\hat{y} = Xw + 1b$. 
We arrange the training set $S_m$ as a matrix $X$ where each row $i$ is $x_i$, the dimensions are $m \times d$, we put the $y$ observations in a vector.

In this setting our prediction is $\hat{y} = Xw + 1b$.

The loss function is $\hat{L}(w, b) = \frac{1}{2m}||y - \hat{y}||^2$
Analytical Solution

- We arrange the training set $S_m$ as a matrix $X$ where each row $i$ is $x_i$, the dimensions are $m \times d$, we put the $y$ observations in a vector.
- In this setting our prediction is $\hat{y} = Xw + 1b$.
- The loss function is $\hat{L}(w, b) = \frac{1}{2m} ||y - \hat{y}||^2$
- As seen in lectures, we can embed the bias in the feature vector by concatenating 1 to each sample and extending $w$ to $w \in \mathbb{R}^{d+1}$. 
Analytical Solution

Optimizing $\hat{L}$ analytically
Gradient Descent for Regression

To optimize $w$ calculate $\frac{\partial \hat{L}}{\partial w}$.

\[
\frac{\partial \hat{L}}{\partial w} = \frac{1}{m} X^T (Xw - y)
\]
Gradient Descent for Regression

- To optimize $w$, calculate $\frac{\partial \hat{L}}{\partial w}$.
- We use $\frac{\partial \hat{L}}{\partial w}$ for GD updates.

$$\frac{\partial \hat{L}}{\partial w} = \frac{1}{m} X^T (Xw - y)$$
Analytical Solution VS. GD

Example in python notebook