Convex Optimization Lecture Notes

Darshan Thaker

January 27, 2019

Contents

Lecture 1

Least Squares and Gradient Descent

1.1 Least Squares

Let's take a look at a simple, yet fundamental example: **least squares**. The goal of least squares is to minimize the sum of squared errors, or more concisely

$$\min_{\beta} \|X\beta - y\|_2^2 \tag{1.1}$$

where $\|\cdot\|_2$ signifies the L2/Euclidean norm ($\|x\|_2 = (\sum_i x_i^2)^{1/2}$). Above, $X \in \mathbb{R}^{n \times p}$, where n is the number of examples given as input to least-squares and p is the number of dimensions for each example. For our current example, we assume $n >> p$, so there are much more examples than dimensions. We will get into why we assume this later. Above, $y \in \mathbb{R}^{n \times 1}$ and $\beta \in \mathbb{R}^{p \times 1}$. We are interested in finding the β that minimizes Equation **??**.

Recall that least squares might be used to solve a regression problem, where each point is in a p-dimensional space, and we are given n examples for which we want a least-squares fit line. The sum of squared errors signifies the regression error. For least-squares regression, y is the response variable influenced by each X_i. Our matrix X and y give us equations in the form of $y_i = \beta x_i + b$. We want to find the optimal value $\hat{\beta}$ that minimizes the sum of squared errors. β intuitively holds the slope values for all variables in that dimension.

1.2 Normal Equation

It turns out that there is a closed-form solution to find the optimal $\hat{\beta}$ that minimizes the above formula, known as the **Normal Equation**:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{1.2}$$

From a computational perspective, if we use the Normal Equation, computing $\hat{\beta}$ takes $O(n^3)$ time - this is because matrix inversion is the bottleneck in the equation and matrix inversion takes $O(n^3)$ time.

1.2.1 Derivation

Let's derive the Normal Equation. There are two main ways to derive the Normal Equation:

1. *Linear Algebra/Calculus*

Proof. We suppose that X^TX is invertible. Recall the optimization problem we are trying to solve (Equation **??**). Note that $\|x\|_2 = x \cdot x = \langle x, x \rangle = x^Tx$. Note the following rules for matrix A, B, vectors a, b, and x:

$$(AB)^T = B^TA^T \tag{1.3}$$

$$(A+B)^T = A^T + B^T \tag{1.4}$$

$$\langle a, b \rangle = \langle b, a \rangle = a^Tb = b^Ta \tag{1.5}$$

For symmetric matrix A,

$$\nabla_x(x^TAx) = 2Ax \tag{1.6}$$

$$\nabla_x(\langle x, y \rangle) = y \tag{1.7}$$

Using these rules, we can write Equation **??** as:

$$\begin{aligned}
\langle X\beta - y, X\beta - y \rangle &= (X\beta - y)^T(X\beta - y) \\
&= ((X\beta)^T - y^T)(X\beta - y) && \text{(Equation ??)} \\
&= (\beta^TX^T - y^T)(X\beta - y) && \text{(Equation ??)} \\
&= \beta^TX^TX\beta - \beta^TX^Ty - y^TX\beta + y^Ty \\
&= \beta^TX^TX\beta - 2\beta^TX^Ty + y^Ty && \text{(Equation ??)}
\end{aligned}$$

Since we want to minimize this, we take the gradient w.r.t. β and set to 0.

$$\begin{aligned}
\nabla_\beta(\beta^TX^TX\beta - 2\beta^TX^Ty + y^Ty) &= 0 \\
2X^TX\beta + \nabla_\beta(-2\beta^TX^Ty + y^Ty) &= 0 && \text{(Equation ??)} \\
2X^TX\beta - 2X^Ty + \nabla_\beta(y^Ty) &= 0 && \text{(Equation ??)} \\
2X^TX\beta - 2X^Ty &= 0 \\
2X^TX\beta &= 2X^Ty \\
\beta &= (X^TX)^{-1}X^Ty
\end{aligned}$$

□

Let's look at the 1-D version to gain some intuition. In this case, we want to minimize $(x\beta - y)^2 = x^2\beta^2 - 2x\beta y + y^2$. Taking the derivative and setting to 0 (left as exercise to reader) gives that $\beta = \frac{xy}{x^2}$. In this case, we see that $\frac{x}{x^2} = \frac{1}{x}$ is the 1-D version of $X^TX)^{-1}X^T$ in the Normal Equation.

2. *Geometry*

We wish to minimize $\|X\beta - y\|_2$ where $n >> p$. To understand the geometric interpretation, let's look at a 1-D example for $n = 2$, $p = 1$.

$$X = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

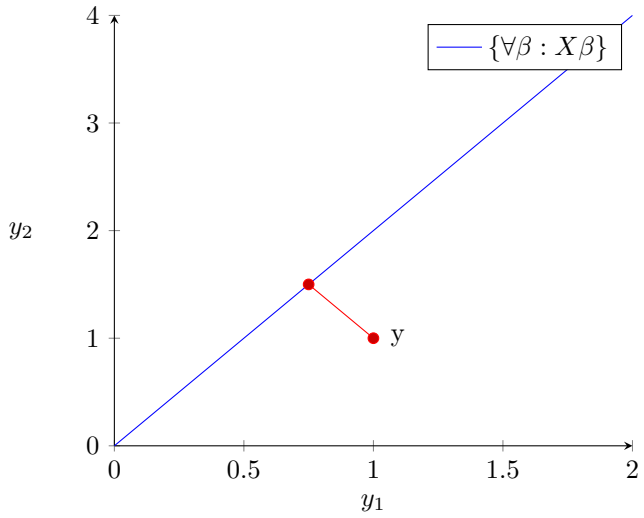


Figure 1.1: 1-D Example in y-space

In Figure **??**, we are visualizing y-space (ranging over all values of y). We are minimizing $\|X\beta - y\|_2$, which is the distance between $X\beta$ and y. In the figure, the visualized line ranges over all values of β given X. Geometrically, we know that the distance between $X\beta$ and y is minimized when $X\beta - y$ (red line in image) is perpendicular to line $X\beta$. Mathematically, this is written as

$$\forall\beta : (y - X\beta) \perp X\beta$$

Only one point $\hat{\beta}$ satisfies this constraint. We give two proofs that $\hat{\beta} = (X^T X)^{-1} X^T y$.

Proof. The set $\{\forall\beta : X\beta\}$ is the column space of X by definition. Observe that the orthogonal vectors of the column space of X falls in the nullspace of X^T. This is because of the linear algebra fact: All vectors in the null space of A are orthogonal to all vectors in the column space of A^T. Thus, if $(y - X\beta) \perp X\beta$, then $y - X\beta$ must be in the null space of X^T, so $X^T(y - x\beta) = 0$ by definition of null space.

$$\begin{aligned} X^T(y - x\hat{\beta}) &= 0 \\ X^T y - X^T X\hat{\beta} &= 0 \\ -X^T X\hat{\beta} &= -X^T y \\ \hat{\beta} &= (X^T X)^{-1} X^T y \end{aligned}$$

□

Proof. If $(y - x\hat{\beta}) \perp \{\forall\beta : X\beta\}$, then their dot products are 0, so $\langle X, y - X\beta\rangle = 0$, which gives that $\hat{\beta} = (X^T X)^{-1} X^T$. □

1.3 Gradient Descent

The above solutions are *closed-form* solutions to Equation **??**. We now turn to another algorithm, called **gradient descent**, to solve the optimization problem. This algorithm is an iterative algorithm that starts off with some β_0 and converges/diverges to some optimum β^*. The update rule from timestep t to $t + 1$ is as follows:

$$\beta_{t+1} = \beta_t - \gamma \nabla f(\beta) \qquad \gamma > 0 \tag{1.8}$$

Above, γ is a step size that dictates how far to move in the direction of the minimum. To move in the direction towards the minimum, we move in the direction of the negative gradient of $f(\beta)$. To illustrate why this is in the direction of the minimum, let us take a example. Suppose we are minimizing the quadratic 2-D function: $y = 2\beta^2 + 1$. Its gradient is 4β. If this is positive, then we will decrease our guess for β according to Equation **??**. This makes sense because the gradient is positive for $\beta > 0$ (since this function is a upward-facing parabola centered at $(0, 1)$ and thus its minimum is $(0, 1)$).

1.3.1 Choice of Step Size

Notice that the choice of γ can drastically affect the convergence of gradient descent. If γ is too high, then we will take very large steps, potentially overshooting the minimum, and leading to divergence (as shown in Figure **??**).

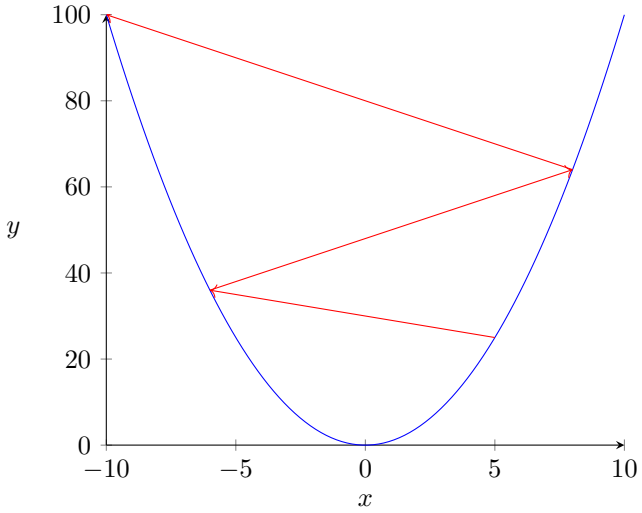


Figure 1.2: Divergence with large γ

So how can we pick the appropriate γ? Let us consider the example of a quadratic function: $f(\beta) = a\beta^2 + c$.

Theorem 1.3.1. *For gradient descent convergence on a quadratic function $f(\beta) = a\beta^2 + c$, the step size γ must be less than $\frac{1}{a}$.*

Proof. The gradient of $f(\beta)$ w.r.t β is $2a\beta$. This gives an expanded update rule:

$$\begin{aligned}
\beta_{t+1} &= \beta_t - \gamma 2a\beta_t \\
&= (1 - 2a\gamma)\beta_t \\
&= (1 - 2a\gamma)[(1 - 2a\gamma)\beta_{t-1}] && \text{(Expand } \beta_t \text{ again)} \\
&\;\;\vdots \\
&= (1 - 2a\gamma)^{t+1}\beta_0
\end{aligned}$$

We don't want β_{t+1} to diverge, so we want a converging series. This is a geometric series, thus $|1-2a\gamma| < 1$. Consider the two cases for the absolute value, which give us two constraints:

Case 1:

$$\begin{aligned} 1 - 2a\gamma &< 1 \\ -2a\gamma &< 0 \\ \gamma &> 0 \end{aligned}$$

Case 2:

$$\begin{aligned} 2a\gamma - 1 &< 1 \\ a\gamma &< 1 \\ \gamma &< \frac{1}{a} \end{aligned}$$

The first case is simply an assumption we make for gradient descent anyways. The second case gives us the important constraint, which is that $\gamma < \frac{1}{a}$ in order for gradient descent to converge for $f(\beta)$. □

Note that this γ is dependent upon the actual function $f(\beta)$ since it depends upon a. In 1-D, it is easy to interpret what the quantity $\frac{1}{a}$ represents. It encodes the curvature of the quadratic function (i.e. how steep the function is). This makes sense because if the function is very steep, our step size should not be very large or we will diverge. Now, we know that γ cannot be too large, but also does not need to be too small for convergence.

Gradient descent seems to have an interesting **self-tuning property**, which is encoded by the gradient. It appears we don't need to dynamically adjust step size because the gradient will automatically be smaller closer to the minimum, so we will take smaller steps anyways. This seems like a rock-solid argument, however it is not always true. Consider the non-smooth function $f(x) = |x|$. The gradient is $+1$ for $x > 0$, -1 for $x < 0$, and not defined at $x = 0$. This translates into the update rule:

$$x_t = x_{t-1} - \gamma \cdot \text{sign}(x)$$

Clearly, this update rule has no self-tuning property, since the sign of x is always $+1$ or -1, which does not change as we get near the minimum.

Finally, consider a quadratic in arbitrary higher dimensions, which can be written as $f(\beta) = \frac{1}{2}\beta^T Q\beta + q^T b$ for some matrix Q and vector q. In this case, the curvature is the largest eigenvalue of Q. We will come back to this key observation.

Lecture 2

Gradient Descent Convergence and Convex Analysis

Typically, in convergence analysis for an algorithm like gradient descent, we consider the error ϵ with which we want the correct answer (the optimal minimum value x^* in our case). We then prove some bound on this error ϵ to prove convergence.

2.1 Convergence Proofs

Clearly, convergence of gradient descent depends upon the function considered. For example, if the function has no minimum value or infinite minimum values, convergence cannot be guaranteed. For now, we examine gradient descent convergence for two functions we have discussed before: the absolute value function and quadratic function. Later on, we will generalize these proofs for a more general class of functions.

2.1.1 Absolute Value Function

Recall the function we are trying to find the minimum for: $f(x) = |x|$, or the absolute value function. For this function, we observed last lecture that gradient descent has no self-tuning property. Thus, γ must be proportional to ϵ.

Theorem 2.1.1. *If γ is proportional to ϵ, then the number of steps required for convergence of gradient descent is approximately $\frac{1}{\epsilon}$ to reach x^* with error ϵ.*

Proof. Suppose we start off at some point x_0 and wish to get to x^*, the optimal point, with some error ϵ. First, we prove with a non-formal argument to understand intuition. To get from x_0 to x^* given γ proportional to ϵ, we must take $\frac{x_0 - x^*}{\epsilon}$ steps, since we divide total journey from start to optimal point, which is of size $x_0 - x^*$, into steps of size ϵ. This shows that the number of steps is proportional to $\frac{1}{\epsilon}$.

More formally (but still not 100% rigorously), consider the update rule:

$$
\begin{aligned}
x_t &= x_{t-1} - \gamma \cdot \text{sign}(x_{t-1}) \\
&= x_{t-2} - \gamma \cdot \text{sign}(x_{t-2}) - \gamma \cdot \text{sign}(x_{t-1}) \\
&\cdot \\
&\cdot \\
&\cdot \\
&= x_0 - \sum_{k=0}^{t} \gamma \cdot \text{sign}(x_k) \\
&\approx x_0 \pm t\gamma && \text{(Upper/lower bounds)}
\end{aligned}
$$

Now, assume that x_t is within ϵ error of the optimal point, i.e. $x_t \in (x^* - \epsilon, x^* + \epsilon)$. Also assume that $x^* \to 0$, which is something we assume in most convergence proofs to make the math a little easier, but is not required. Let $x_t = x^* + \epsilon$. Equating the upper bound shown above to this and solving for t gives our desired result:

$$
\begin{aligned}
x_t &= x_0 \pm t\gamma \\
x^* + \epsilon &= x_0 \pm t\epsilon && (\gamma \propto \epsilon) \\
\epsilon &= x_0 \pm t\epsilon && (x^* \to 0) \\
\pm\frac{\epsilon - x_0}{\epsilon} &= t \\
\pm(1 - \frac{x_0}{\epsilon}) &= t
\end{aligned}
$$

Thus, the number of timesteps t is proportional to $\frac{1}{\epsilon}$. □

2.1.2 2-D Quadratic Function

Recall our 2-D quadratic function $f(x) = ax^2 + c$.

Theorem 2.1.2. *Given an appropriate step size for convergence, the number of steps required for convergence of gradient descent to reach x^* with error ϵ is proportional to $\ln\left(\frac{1}{\epsilon}\right)$.*

Proof. Consider the update rule for gradient descent, from which we solve for t - the number of timesteps at which we reach $x_t \in (x^* - \epsilon, x^* + \epsilon)$. As always, we assume that x^* is 0 and $x_t = x^* + \epsilon$ to simplify the proofs.

$$
\begin{aligned}
x_t &= (1 - 2a\gamma)^t x_0 \\
\frac{x_t}{x_0} &= (1 - 2a\gamma)^t \\
\ln(x_t) - \ln(x_0) &= t\ln(1 - 2a\gamma) \\
t &= \frac{\ln(x_t) - \ln(x_0)}{\ln(1 - 2a\gamma)}
\end{aligned}
$$

Recall from Theorem **??** that for gradient descent to converge for a quadratic, the step size γ must be less than $\frac{1}{a}$. Thus, the quantity $1 - 2a\gamma$ must be in the range $(-1, 1)$, so the denominator $\ln(1 - 2a\gamma)$ must be negative. Thus, transferring this negative from the denominator to the numerator, we obtain

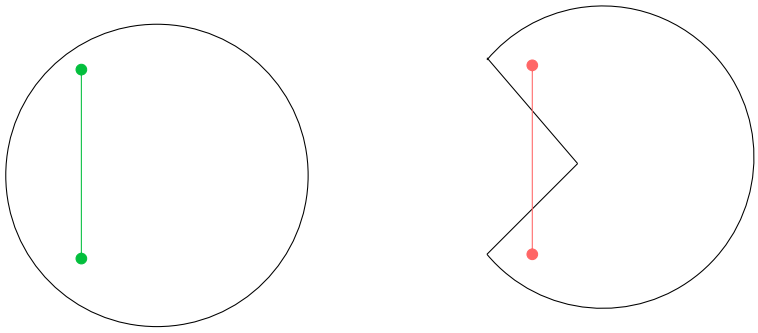
$$
\begin{aligned}
t &= \frac{-\ln(x_t) + \ln(x_0)}{|\ln(1-2a\gamma)|} \\
&= \frac{-\ln(x^* + \epsilon) + \ln(x_0)}{|\ln(1-2a\gamma)|} \\
&= \frac{-\ln(\epsilon) + \ln(x_0)}{|\ln(1-2a\gamma)|} \\
&\approx \ln(\frac{1}{\epsilon})
\end{aligned}
$$

□

2.2 Convex Analysis

We go through some basic definitions. There are a lot of definitions here, but they will all come in handy later, so work through the examples carefully.

Definition 2.2.1. Convex set: A set $X \subset \mathbb{R}^n$ is convex if $\forall x_1, x_2 \in X, \forall \lambda \in [0,1] : \lambda x_1 + (1-\lambda)x_2 \in X$.



(a) Convex Set (b) Non-Convex Set

Figure 2.1

Geometrically, we can interpret this as the following: if $x_1, x_2 \in X$, then the line segment $[x1, x2]$ must also be fully contained in X in order for X to be convex. This is shown in Figure **??**.

Example: The set of symmetric PSD matrices is convex.
Recall that a *positive semi-definite* (PSD) matrix is one whose eigenvalues are greater than or equal to 0. To prove the above claim, we will first require some linear algebra review, which we present below.

Theorem 2.2.1 (**Spectral Theorem**). *If M is a symmetric $n \times n$ matrix, then for $\lambda_i \in \mathbb{R}, v_i \in \mathbb{R}^{n \times 1}, v_i \perp v_j \forall i \neq j, \|v_i\|_2 = 1$:*

$$M = \sum_{i=0}^{n} \lambda_i v_i v_i^T \tag{2.1}$$

We will defer proof of this theorem to a linear algebra textbook. What needs to be taken away from the Spectral Theorem is this very simple corollary.

Corollary 2.2.2. *In the Spectral Theorem decomposition, each λ_i is an eigenvalue of M and v_i are orthonormal eigenvectors of M.*

Proof. Let j be an integer in the range from 1 to n. By definition of an eigenvalue/eigenvector, we wish to show that $Mv_j = \lambda_j v_j$.

$$\begin{aligned} Mv_j &= \left(\sum_{i=0}^{n} \lambda_i v_i v_i^T\right) v_j && \text{(Spectral Theorem)} \\ &= \lambda_j v_j v_j^T v_j && (\forall i \neq j : \langle v_i, v_j \rangle = 0) \\ &= \lambda_j v_j && (\langle v_j, v_j \rangle = 1) \end{aligned}$$

The steps used above also demonstrate these eigenvectors are orthonormal, thus completing the proof. □

Now that we have the background to prove our claim that the set of all symmetric PSD matrices is convex, we start with a simple lemma that will aid us in our proof.

Lemma 2.2.3. *M is a $n \times n$ symmetric positive semi-definite matrix if and only if for all $x \in \mathbb{R}^n$,*

$$x^T M x \geq 0$$

Proof. We show one direction first. Let M be a $n \times n$ symmetric positive semi-definite matrix. Let x be any value in $\mathbb{R}^n$. Now, consider the value $x^T M x$:

$$\begin{aligned} x^T M x &= x^T \left(\sum_{i=0}^{n} \lambda_i v_i v_i^T\right) x && \text{(Spectral Theorem)} \\ &= \sum_{i=0}^{n} x^T \lambda_i v_i v_i^T x \\ &= \sum_{i=0}^{n} \lambda_i x^T v_i v_i^T x \\ &= \sum_{i=0}^{n} \lambda_i (v_i^T x)^T v_i^T x \\ &= \sum_{i=0}^{n} \lambda_i \|v_i^T x\|_2^2 \\ &\geq 0 && (\lambda_i \geq 0, \|\cdot\|_2^2 \geq 0) \end{aligned}$$

These steps are all bidirectional, so the converse also holds. □

Now, we are finally ready to show our claim.

Claim: Suppose M is the set of all symmetric $n \times n$ matrices with eigenvalues greater than or equal to 0. The set M is a convex set.

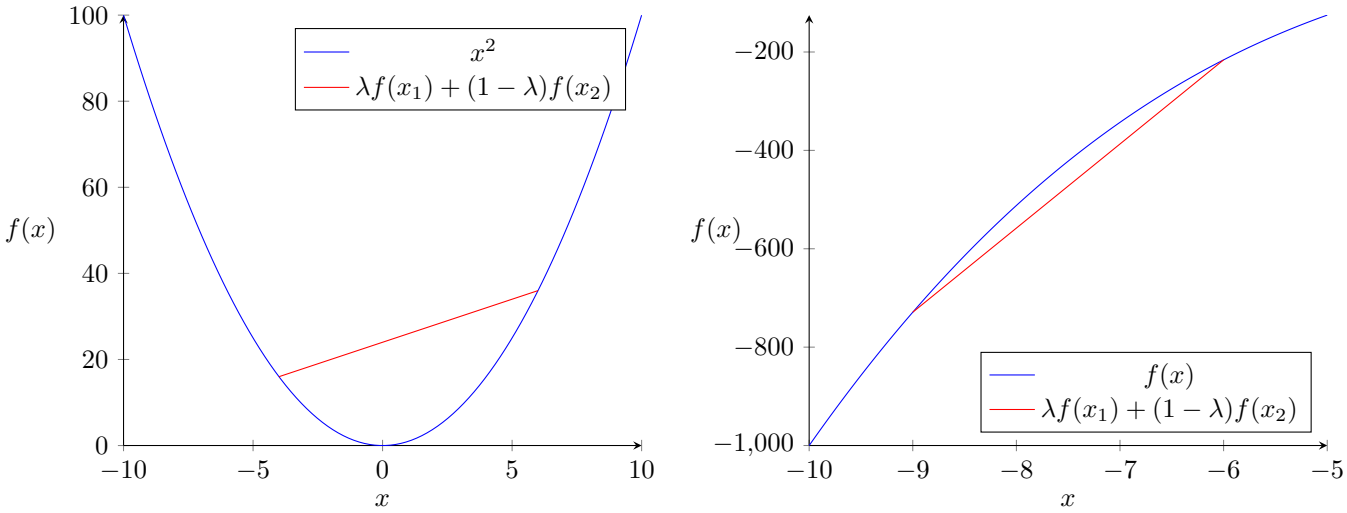
Proof. Let $M_1, M_2 \in M$. We will show, by definition of a convex set, that $\forall \lambda \in [0,1] : \lambda M_1 + (1-\lambda) M_2 \in M$. From Lemma **??**, $M_1 \in M$ if and only if for all $x \in \mathbb{R}^n$, $x^T M x \geq 0$. Let x be some vector in $\mathbb{R}^n$. Then, we know that $x^T M_1 x \geq 0$ and $x^T M_2 x \geq 0$. Let $\lambda \in [0,1]$.

$$\lambda x^T M_1 x + (1-\lambda) x^T M_2 x \geq 0 \qquad (x^T M_1 x \geq 0,\ x^T M_2 x \geq 0)$$
$$\implies x^T(\lambda M_1 + (1-\lambda) M_2) x \geq 0$$
$$\implies \lambda M_1 + (1-\lambda) M_2 \in M \qquad \text{(Lemma ??)}$$

□

Next, we examine three equivalent definitions of a convex function.

Definition 2.2.2 (Convex Function). A function $f : X \to Y$ is convex iff $\forall x_1, x_2 \in X, \forall \lambda \in [0,1]$: $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$.



(a) Convex Function: $f(x) = x^2$

(b) Non-Convex function: $f(x) = x^3$

Figure 2.2

Geometrically, we can interpret this as the following: The actual function value between any two points should always be below the chord between the two points. This is illustrated in Figure **??**. Note that this definition assumes nothing about the regularity of f (i.e. when it is defined, whether it is smooth, etc.). Alternatively, a **concave function** has an identical definition to Definition **??** with the inequality reversed.

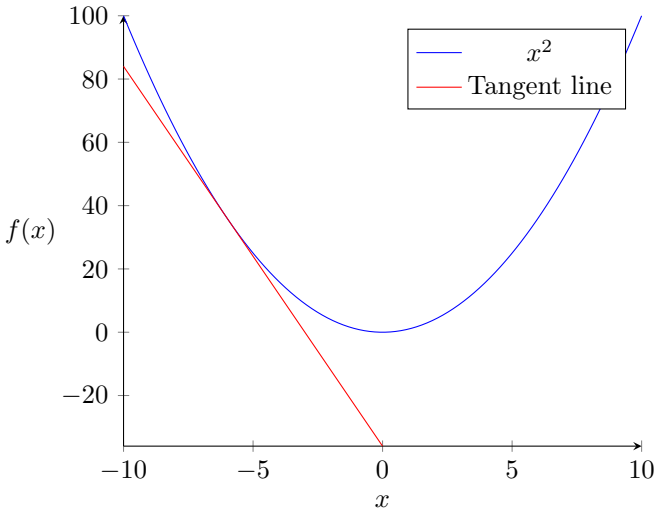
Definition 2.2.3 (Convex Function). A function $f : X \to Y$ is convex iff $\forall x, y \in X : f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

To get some intuition into this definition, recall that the 1st order Taylor approximation of a function at a point is the line tangent to that point. Algebraically, at any point $y \in X$, the 1st order Taylor approximation at that point is the line $\forall x : f(y) = f(x) + \langle \nabla f(x), y - x \rangle$. The definition of a convex function states that the 1st order Taylor approximation at any point on the function (or the tangent line at that point) should always underestimate the function everywhere in order for that function to be convex. This is shown in Figure **??**. Note that f must be differentiable once for this definition to apply.

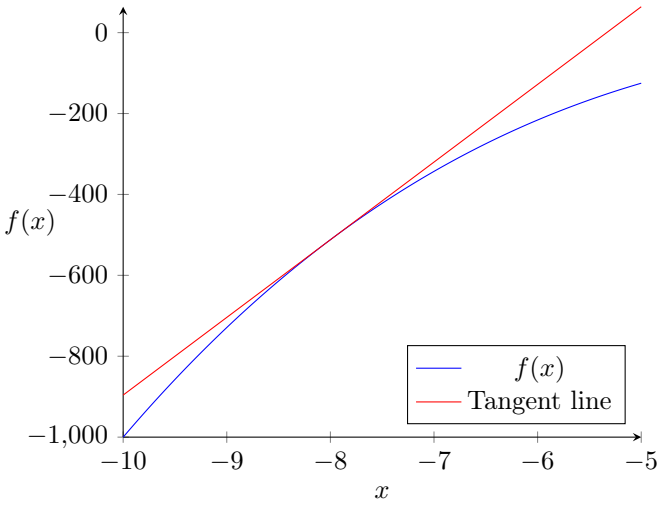
Definition 2.2.4 (Convex Function). A function $f : X \to Y$ is convex iff $\nabla^2 f(x) \succeq 0$ i.e. the Hessian (matrix of second derivatives) has non-negative eigenvalues.

In 2-D, this definition translates to saying that the second derivative of the function needs to be non-negative in order for the function to be convex. Note that f must be differentiable twice for this definition to apply.

This implies that convexity need not hold for all x in the domain of a function, but could hold for a subset (we could have concavity changes in $\nabla^2 f(x)$. For example, consider the quartic function $f(x) = ax^2 + bx + c + d$.



(a) Convex Function: $f(x) = x^2$



(b) Non-Convex function: $f(x) = x^3$

Figure 2.3

The second derivative of this function is: $f''(x) = bax + 2b$, which is only greater than 0 for some x in the domain of the function. This means that for these x in the domain, $f(x)$ is convex.

Theorem 2.2.4. *Suppose $f : X \to Y$ is convex and differentiable. Then, if we are at some point $\hat{x}$ s.t. $f'(\hat{x}) = 0$, then $\hat{x}$ is a global optimum. This theorem implies that any local optimum on a convex function is always the global optimum.*

Proof. Let $x \in X$. By Definition **??**, we know the following:

$$\begin{aligned} f(x) &\geq f(\hat{x}) + \langle \nabla f(\hat{x}), x - \hat{x} \rangle \\ &\geq f(\hat{x}) \qquad (\nabla f(\hat{x}) = 0) \end{aligned}$$

This implies that for all $x \in X$, the function value at x is greater than or equal to the function value at $\hat{x}$, implying that $\hat{x}$ is the global optimum. □

Definition 2.2.5 (Hyperplane). Given $a \in \mathbb{R}^n$ such that $a \neq 0$ and $b \in \mathbb{R}$, a *hyperplane* is a set or subspace of a vector space of the form

$$H = \{x : a^T x = b\} \tag{2.2}$$