# Theoretical ML - Convergence and Bounds for Gradient Descent, p1

Souradip Chakraborty, Walmart Labs
Google Developers Expert in Machine Learning

April 2021

**Gradient Descent Convergence for Quadratic Functions**

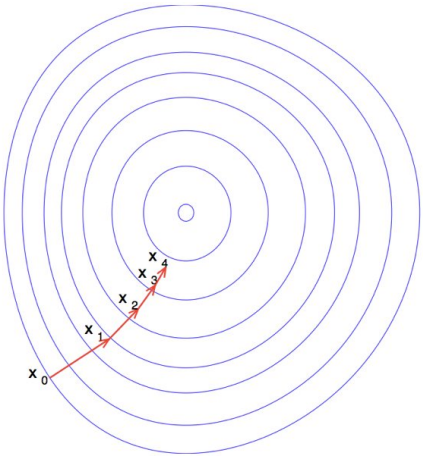


Figure 1: Gradient Descent Convergence

Gradient Descent is a first-order iterative optimization algorithm to find the local minima of differentiable functions. The primary intuition of gradient descent is that for any multivariate differentiable function $f(x)$, if we move in the direction of its negative gradient then the value of the function $f(x)$ decreases fastest (Wikipedia). It has been immensely successfully in finding the optimal weights for deep neural models and has been widely used in a vast majority of the neural architectures. I will explore certain mathematical aspects of gradient descent and its variants and try to understand the convergence and bounds of the same for different types of functions.
In this article, I focus on deriving the order and theoretical bounds on the convergence of the gradient descent algorithm for Quadratic functions.
Let's start with the basic definition of a Quadratic function in it's general

form.

$$f(x) = 1/2 \times x^T Q x - b^T x$$

where $x, b \in R^d, Q \in R^{d*d}$ and $Q$ is positive definite i.e $z^t Q z > 0, \forall z \neq 0$. All the eigen values of $Q$ are strictly positive and the largest eigen value of Q is represented as $\lambda_{max}$ and the smallest eigen value is represented as $\lambda_{min}$. The condition number $\kappa$ is given by $\kappa = \lambda_{max}$ / $\lambda_{min}$. $Q$ is a full rank matrix and invertible.
The objective is to minimize the Quadratic function and find the optimal values of x.

$$\underset{x}{\text{minimize}}\ f(x) = 1/2 \times x^T Q x - b^T x$$

Now, let's consider gradient descent with fixed step size $\alpha$ which is referred to as the learning rate for the gradient descent. As stated above, gradient descent is an iterative optimization algorithm and incrementally converges to the local minima. The objective of our problem is to theoretically guarantee that gradient descent will converge for $f(x)$ and talk about the order and complexity of the same.
A very interesting aspect of the quadratic functions ($f(x)$ in our case) with $Q$ a PD matrix, lie in the fact that they are convex and hence satisfy the wonderful property that any local minima is a global minima. Let's try to prove that :

Lemma 1 : Any local minima of a convex function is a global minima.
Let's assume $\bar{x}$ be a local minima of the convex optimization problem :

$$\underset{x \in C}{\text{minimize}}\ f(x), f : R^n -> R$$

where, $f(x)$ is a convex function and $C$ is a convex set. Since, $\bar{x}$ is a local minima, then we can say

$$f(x) \geq f(\bar{x})\ \forall x \in B_\epsilon(\bar{x}) \cap C$$

where $B_\epsilon(\bar{x})$ is the epsilon neighbourhood of $\bar{x}$. Now, let's take any point $y \in C$, then we can express any $z$ as $z = \delta * y + (1 - \delta) * \bar{x}, z \in C$ and $\lambda \in (0, 1)$ since C is a convex set. So, as we reduce the value of $\delta$ from 1 towards 0, $z$ approaches $x$ from $y$ and let's say when the value of $\delta = \delta'$, $z$ reaches the epsilon neighbourhood of $\bar{x}$ i.e $B_\epsilon(\bar{x})$ which means we can find a $z'$ in $B_\epsilon(\bar{x})$ such that $z' = \delta'' * y + (1 - \delta'') * \bar{x}$ and $\delta'' \in (0, \delta')$. Now since $z'$ lies in $B_\epsilon(\bar{x})$, we can say that

$$f(z') \geq f(\bar{x})$$
$$f(\delta'' y + (1 - \delta'')\bar{x}) \geq f(\bar{x})$$

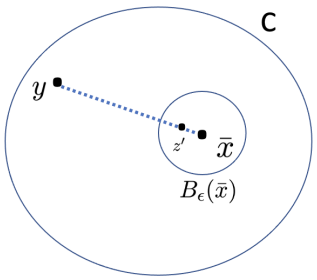


Figure 2: Convex Set Properties

since $\bar{x}$ is a local minima.

$$f(\delta'' y + (1 - \delta'')\bar{x}) \leq \delta'' f(y) + (1 - \delta'') f(\bar{x})$$

since $f(x)$ is convex. Hence, combining the above two equations we can write

$$\begin{gathered} \delta'' f(y) + (1 - \delta'') f(\bar{x}) \geq f(\bar{x}) \\ \delta'' f(y) \geq \delta'' f(\bar{x}) \\ f(y) \geq f(\bar{x}) \end{gathered}$$

which implies that for any $y \in C$, $f(y) \geq f(\bar{x})$ is true which proves that $\bar{x}$ is a global minima. This completes the first part of our proof. This is an extremely important result especially in the context of gradient descent as gradient descent converges to local optima. As for convex functions all local optima are global, hence it is guaranteed to converge with gradient descent.
Now, the above lemma has been proven for any convex function but it has not been shown yet that all quadratic functions with a positive definite matrix $Q$ are convex. So let's do the very simple proof using the property of convex functions. For our problem, the quadratic function is $f(x) = 1/2 \times x^T Q x - b^T x$. Now, lets differentiate the function and determine the gradient and hessian of the same.

$$\begin{gathered} f(x) = 1/2 \times x^T Q x - b^T x \\ \nabla(f(x)) = Qx - b \\ H(f(x)) = Q \end{gathered}$$

Since the hessian Q is positive definite, it implies that $f(x)$ is convex (by definition). Now, it brings to the most important part of the proof i.e to derive the convergence bounds for gradient descent in this setting. The

objective is to minimize the Quadratic function and find the optimal values of x using gradient descent and deriving the order and guarantees for its convergence.

$$\underset{x}{\text{minimize}}\ f(x) = 1/2 \times x^T Q x - b^T x$$

Let's say the iterative value of $x$ at the $t_{th}$ iteration is represented by $x_t$ and then based on the gradient descent optimization procedure, we can write

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

where, $x_{t+1}$ & $x_t$ are the iterative values of $x$ at the $t_{th}$ & $(t+1)_{th}$ iterations respectively and $\alpha$ is the learning rate. This is how gradient descent iteratively updates the values of x, but it has not yet been shown why? So, the question stays that how to show that $f(x_{t+1})$ is always lesser than $f(x_t)$ $\forall t$. To show that, the Taylor's series can be approximated as : (ignoring the higher terms)

$$f(y) \approx f(x) + \nabla(f(x))^T * (y - x)$$

replacing, $x_{t+1} = y$ and $x_t = x$ in the first equation and assuming that $\delta x = x_{t+1} - x_t$ is small.

$$f(x_{t+1}) = f(x_t) + \nabla(f(x_t))^T (x_{t+1} - x_t)$$
$$f(x_{t+1}) - f(x_t) = \nabla(f(x_t))^T (x_{t+1} - x_t)$$

Now, taking the value of $x_{t+1} - x_t$ from the gradient descent equation, we have

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

and putting this value in the equation of the updated Taylor's series we have

$$f(x_{t+1}) - f(x_t) = \nabla(f(x_t))^T (x_{t+1} - x_t)$$
$$f(x_{t+1}) - f(x_t) = \nabla(f(x_t))^T \times (-\alpha \nabla(f(x)))$$
$$f(x_{t+1}) - f(x_t) = -\alpha \nabla(f(x_t))^T \nabla(f(x))$$
$$f(x_{t+1}) - f(x_t) = -\alpha \|\nabla(f(x))\|^2$$

Since $\|\nabla(f(x))\|^2$ is always positive and $\alpha \geq 0$, we can conclude that $f(x_{t+1})$ is always lesser than $f(x_t)$, assuming $\delta x$ is small. Now, finally we will derive the theoretical bounds on gradient descent.

Theorem 1 : If we are running gradient descent on a quadratic objective function $f(x) = 1/2 \times x^T Q x - b^T x$ where the largest and smallest eigen values of Q be $\lambda_{max}$ & $\lambda_{min}$ respectively, $\alpha = 2/(\lambda_{max} + \lambda_{min})$, then gradient descent satisfies

$$\|x_t - x_*\| \leq (\frac{1 - \frac{1}{\kappa}}{1 + \frac{1}{\kappa}})^t \times \|x_0 - x_*\|$$

where $x_*$ is the optimal value of $x$ and $x_0$ is the initial value of $x$ with which the gradient descent optimization algorithm begins. The above theorem states that there exists an upper bound on the norm of $x_t - x_*$ i.e eventually with iterations, $x^t$ reaches the optimum $x_*$. Now, the point lies in how fast it converges or reaches the optimal $x^*$ and that depends on two factors : 1. How good or bad our initialisation $x_0$ is from the optimal $x_*$ which is very clear that if we initialise at a value very far from our optimal point, it will take more number of iterations to converge. 2. Secondly, it depends on the condition number $\kappa = \lambda_{max}$ / $\lambda_{min}$, where $\kappa \geq 1$. But it is clear from the equation, that lower the value of $\kappa$, faster will be the convergence i.e if $\kappa = 1$, the first part of the product becomes 0 which indicates immediate convergence at the first iteration with $\alpha = 2/(\lambda_{max} + \lambda_{min})$ as the learning rate. Similarly, with higher values of $\kappa$ the norm $\|x_t - x_*\|$ becomes independent of the first part of the product and as $\kappa$ tends to infinity, the norm $\|x_t - x_*\|$ only depends on the initialisation.
Proof : Let's try to prove Theorem 1.

$$f(x) = 1/2 * x^T Q x - b^T x$$
$$\nabla f(x) = Qx - b$$

We know that at the global minimal point $x_*$, $\nabla f(x_*) = 0$ and since $Q$ is invertible, the optimal $x^*$ is given by $x_* = Q^{-1} b$.

$$x_{t+1} - x_* = x_t - \alpha \nabla(f(x_t)) - x_*$$
$$x_{t+1} - x_* = x_t - \alpha(Qx_t - Qx_*) - x_*$$
$$x_{t+1} - x_* = x_t - \alpha Q x_t + \alpha Q x_* - x_*$$
$$x_{t+1} - x_* = (I - \alpha Q)x_t - (I - \alpha Q)x_*$$
$$x_{t+1} - x_* = (I - \alpha Q)x_t - (I - \alpha Q)x_*$$
$$x_{t+1} - x_* = (I - \alpha Q)(x_t - x_*)$$

Now taking the (induced) norms on both sides of the equation and using the property that $\|Ax\| \leq \|A\|\|x\|$ where $A$ is a matrix $\in R^{n*n}$ $\forall x \in R^n$ and $x \neq 0$. Using this property, the equation can be written as

$$\|x_{t+1} - x_*\| \leq \|(I - \alpha Q)\| \times \|(x_t - x_*)\|$$

Now, lets focus on the $\|(I - \alpha Q)\|$ part and try to find out the induced matrix norm of the above expression. The induced matrix norm can be defined as following :

$$\|A\| = \underset{x, x \neq 0}{\text{Max}} \frac{\|Ax\|_p}{\|x\|_p}$$

where $A \in R^{m*n}$ and $x \in R^n$. For our case, lets take $A \in R^{n*n}$. Let's try to find a connection between the matrix norm and the eigen values for the above matrix $I - \alpha Q$ and for our case $A = I - \alpha Q$. Now. the induced norm equation can be written in the form of an optimization framework :

$$\|A\| = \underset{x, x \neq 0}{\text{Max}} \|Ax\|_p$$
$$x : \|x\|_p = 1$$

This can be shown for any p but for our case we take $p = 2$ or the $L2$ norm. Then the norm can be written as

$$\|A\| = \underset{x, x \neq 0}{\text{Max}} \|Ax\|_2$$
$$x : \|x\|_2 = 1$$

Hence, the equation can be restructured as

$$\|A\| = \underset{x, \|x\|_2 = 1}{\text{Max}} (x^T A^T A x)^{1/2}$$

Now, since square root is a strictly monotonically increasing function , we can use the property of strict monotonic transformation of functions i.e the solution sets for both the below equations will be the same.

$$\underset{x \in C}{\text{Max}} f(x)$$
$$\underset{x \in C}{\text{Max}} g(f(x))$$

where $f : X \rightarrow R$ and $g : R \rightarrow R$ and $g$ is a strictly monotonically increasing function. Then we can modify our problem as:

$$\underset{x,\|x\|_2=1}{\text{Max}}(x^T A^T A x)$$

Since $A^T A \in R^{n*n}$ is a symmetric matrix and hence can be decomposed into its eigen vectors and eigen values as $A^T A = U \Lambda U^T$ where Q is matrix with eigen vectors of $A^T A$ and $\Lambda$ is the diagonal matrix with corresponding eigen values and $UU^T = I$.

$$\underset{x,\|x\|_2=1}{\text{Max}}(x^T A^T A x)$$
$$\underset{x,\|x\|_2=1}{\text{Max}}(x^T U \Lambda U^T x)$$
$$\underset{x,\|x\|_2=1}{\text{Max}}((U^T x)^T \Lambda (U^T x))$$

Now, for simplicity of calculations let's take $U^T x$ as $z \in R^n$ and the equation can be written as

$$\underset{z,z\neq 0,\|z\|_2=1}{\text{Max}}(z^T \Lambda z)$$

Also, its important to note that

$$z = U^T x$$
$$z^T z = x^T U U^T x$$
$$z^T z = x^T x$$
$$z^T z = 1$$
$$\|z\|_2 = 1$$

Hence our maximization equation can be written as

$$\underset{z,z\neq 0,\|z\|_2=1}{\text{Max}}(z^T \Lambda z)$$
$$\underset{z,z\neq 0,\|z\|_2=1}{\text{Max}} \sum_{n=1}^{n} z_i^2 \times \beta_i$$

where $\beta_i$ are the eigen values of $A^TA$ as mentioned above and let $\beta_{max}$ be the maximal eigen value, then we know that :

$$z_i^2\beta_i \leq z_i^2\beta_{max}, \forall i \in n$$
$$\sum_{n=1}^{n} z_i^2\beta_i \leq \beta_{max}\sum_{n=1}^{n} z_i^2$$
$$\sum_{n=1}^{n} z_i^2\beta_i \leq \beta_{max}$$

From this we can also say that

$$\sum_{n=1}^{n} z_i^2\beta_i \leq \beta_{max}$$
$$z^T\Lambda z \leq \beta_{max}$$
$$x^TA^TAx \leq \beta_{max}$$

where $\beta_{max}$ is the maximal eigen value of the matrix $A^TA$. It is important to note that we have to find the maximum value of $(x^TA^TAx)^{1/2}$. Let $x^TA^TAx = L^2$

$$x^TA^TAx \leq \beta_{max}$$
$$L^2 \leq \beta_{max}$$
$$L \leq (\beta_{max})^{1/2}$$
$$(x^TA^TAx)^{1/2} \leq (\beta_{max})^{1/2}$$

Since $L \geq 0$. Now, using these above properties let's try to derive the norm of $\|(I - \alpha Q)\|$ to prove Theorem 1. So, we know that the induced norm of $\|(I - \alpha Q)\|$ can be written as :

$$\|(I - \alpha Q)\| = \underset{x, x\neq 0, \|x\|_2=1}{\text{Max}} (x^T(I - \alpha Q)^T(I - \alpha Q)x)^{1/2}$$

We have seen from the previous section that it reduces to finding the square root of the maximum eigen value of the matrix $(I - \alpha Q)^T(I - \alpha Q)$.
In the earlier section we had assumed that $Q$ is positive definite i.e $z^tQz > 0, \forall z \neq 0$. All the eigen values of $Q \in R^{d*d}$, are strictly positive and the largest eigen value of Q is represented as $\lambda_{max}$ and the smallest eigen value is represented as $\lambda_{min}$. The condition number $\kappa$ is given by $\kappa = \lambda_{max} \ / \ \lambda_{min}$. Let the eigen values of $Q$ be represented by $\lambda$ and let $t$ be

an eigen vector of $(I - \alpha Q)$ with value $\theta$

$$(I - \alpha Q)t = \theta t$$
$$It - \alpha Qt = \theta t$$
$$(1 - \alpha\lambda)t = \theta t$$

Hence, it's clear that the eigen value of $(I - \alpha Q)$ is $\theta = (1 - \alpha\lambda)$ Also, since $(I - \alpha Q)^T = (I - \alpha Q)$ i.e it's a symmetric matrix and hence we can write

$$(I - \alpha Q)^T(I - \alpha Q) = (I - \alpha Q)^2$$
$$(I - \alpha Q)^T(I - \alpha Q)t = (I - \alpha Q)^2 t$$
$$(I - \alpha Q)^2 t = (I - \alpha Q)((I - \alpha Q)t$$
$$(I - \alpha Q)((I - \alpha Q)t) = (I - \alpha Q)\theta t$$
$$(I - \alpha Q)\theta t = \theta(I - \alpha Q)t$$
$$\theta(I - \alpha Q)t = \theta^2 t$$

Hence, it can be concluded that the eigen value of $(I - \alpha Q)^T(I - \alpha Q)$ is $\theta^2$ and $t$ is the eigen vector, where $\theta = (1 - \alpha\lambda)$, the eigen value of $(I - \alpha Q)$. Hence, the problem of finding the norm of the matrix boils down to finding the square root of the maximum eigen value of the matrix $(I - \alpha Q)^T(I - \alpha Q)$ i.e square root of the maximum value of $(1 - \alpha\lambda)^2$. This can also be thought of as finding the maximum values of square root of $(1 - \alpha\lambda)^2$ due to the monotonic transformation property of the square root function, discussed above.

$$\|(I - \alpha Q)\| = \text{Max } ((1 - \alpha\lambda)^2)^{1/2}$$
$$\|(I - \alpha Q)\| = \text{Max } (|1 - \alpha\lambda|)$$

The possible values from the RHS of the above equation can be $(1 - \alpha\lambda_{min}), (1 - \alpha\lambda_2), (1 - \alpha\lambda_3), .....(1 - \alpha\lambda_{max})$ and $\alpha(\lambda_{min} - 1), (\alpha\lambda_2 - 1), (\alpha\lambda_2 - 1), ....(\alpha\lambda_{max} - 1)$. Since $\lambda_{max}$ and $\lambda_{min}$ are the largest and smallest eigen values of the $Q$, we can write

$$\|(I - \alpha Q)\| = \text{Max } (\alpha\lambda_{max} - 1, 1 - \alpha\lambda_{min})$$

Hence, our norm equation can be written as :

$$\|x_{t+1} - x_*\| \leq \|(I - \alpha Q)\| \times \|(x_t - x_*)\|$$
$$\|x_{t+1} - x_*\| \leq \text{Max } (\alpha\lambda_{max} - 1, 1 - \alpha\lambda_{min}) \times \|(x_t - x_*)\|$$

Now, as per Theorem 1 if we replace the value of $\alpha = \frac{2}{\lambda_{max}+\lambda_{min}}$ in the above equation, we get :

$$\|x_{t+1} - x_*\| \leq \frac{1 - \frac{\lambda_{min}}{\lambda_{max}}}{1 + \frac{\lambda_{min}}{\lambda_{max}}} \times \|(x_t - x_*)\|$$

Now, the matrix condition number $\kappa = \lambda_{max}/\lambda_{min}$ and hence replacing it in the above equation

$$\|x_{t+1} - x_*\| \leq \frac{1 - \frac{1}{\kappa}}{1 + \frac{1}{\kappa}} \times \|(x_t - x_*)\|$$

Now, rewriting the above equation for multiple iterations $(t)$ :

$$\|x_{t+1} - x_*\| \leq \frac{1 - \frac{1}{\kappa}}{1 + \frac{1}{\kappa}} \times \|(x_t - x_*)\|$$
$$\|x_t - x_*\| \leq \frac{1 - \frac{1}{\kappa}}{1 + \frac{1}{\kappa}} \times \|(x_{t-1} - x_*)\|$$
$$\|x_{t-1} - x_*\| \leq \frac{1 - \frac{1}{\kappa}}{1 + \frac{1}{\kappa}} \times \|(x_{t-2} - x_*)\|$$
$$.....$$
$$\|x_1 - x_*\| \leq \frac{1 - \frac{1}{\kappa}}{1 + \frac{1}{\kappa}} \times \|(x_0 - x_*)\|$$

Hence, replacing subsequent values in the right-hand side of the above equation iteratively, we get

$$\|x_{t+1} - x_*\| \leq (\frac{1 - \frac{1}{\kappa}}{1 + \frac{1}{\kappa}})^t \times \|(x_0 - x_*)\|$$

This proves the Theorem 1 i.e error rate decays exponentially i.e first-order convergence, GD will exhibit first order convergence.