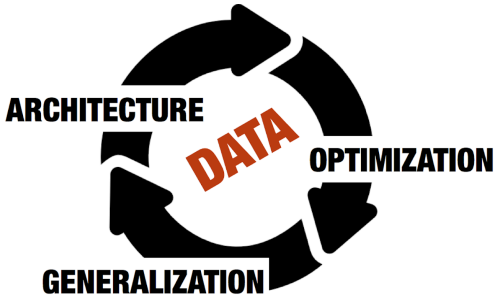
Chapter 9

Background on Optimization, Gradient Descent

We have covered the cliff-notes of the practice of deep learning in the previous eight chapters. It is by no means a complete overview. The practice of deep learning is an enticing, mysterious, and sometimes frustrating. The more time you spend playing with code the more you will learn about deep learning. New ideas are routinely discovered using very simple experiments that each of you is capable of running in your Colab now.

As we discussed, there three main concepts in machine learning. First, the class of functions $f(x; w)$ that you use to make predictions, this is called the hypothesis class or the architecture. Second, the algorithm you use to find the best model in this class of functions that fits your data; this uses tools from optimization theory. Third is the generalization performance of your classifier. Machine Learning is about picking a good hypothesis class, finding the best model within this class and making sure that the model generalizes.



The above process is relatively well-understood for simpler models such as SVMs but the story is quite murky for deep networks. Often in practice, it is never clear which architecture you should pick for your problem (many of you have asked this question in the office hours for instance), training a deep network involves a number of bells and whistles (some of which like Batch-Normalization and Dropout that we have seen) and if at the end of this exercise we get a high validation error, it is unclear how one should change

the parts of the process to improve performance. Disentangling this viscious cycle is what "understanding deep learning" is all about.

Goal Module 2 will develop an understanding of optimization and generalization for more generic machine learning models first. It will end with an insight into understanding their interplay for deep networks. Module 2 has a different flavor, it is more theoretical. Our goal is to grasp the general concepts behind these theoretical results and understand the training process of deep networks better. This will also help us train deep networks much better in practice.

9.1 Convexity

Consider a function $\ell : \mathbb{R}^p \to \mathbb{R}$ that is convex, i.e., for any w, w' that lie in the domain (which is assumed to be a convex set) of f and any $\lambda \in [0, 1]$ we have

$$\ell(\lambda w + (1-\lambda)w') \leq \lambda\ell(w) + (1-\lambda)\ell(w'). \tag{9.1}$$

A function $\ell(w)$ is concave if $-\ell(w)$ is convex. Some examples of convex functions are

- affine functions $Aw + b$, norms $\|w\|_p = (\sum_{i=1}^p |w_i|^p)^{1/p}$, or $\|w\|_\infty = \max_k |w_k|$.
- exponential e^w for $w \in \mathbb{R}$
- powers w^α for $w > 0$ and $\alpha \geq 1$ or $\alpha \leq 0$
- powers of absolute values $|w|^p$ for $w \in \mathbb{R}$ and $p \geq 1$

Strictly convex functions Strictly convex functions have the property that for all $w \neq w'$ in the domain (which is assumed to be convex) and $\lambda \in (0, 1)$

$$\ell(\lambda w + (1-\lambda)w') < \lambda\ell(w) + (1-\lambda)\ell(w').$$

First-order condition for convexity If ℓ is differentiable, the definition of convexity in (9.1) is equivalent to the following first-order condition. A differentiable function ℓ with convex domain is convex iff

$$\ell(w') \geq \ell(w) + \langle \nabla\, \ell(w), w' - w \rangle\,. \tag{9.2}$$

for all w, w' in the domain. Note that the first-order condition is equivalent to the definition of convexity in (9.1) for differentiable functions. The proof is long but easy; you can see https://www.princeton.edu/ aaa/Public/Teaching/ORF523/S16/ORF523_S16_Lec7_gh.pdf for the proof. For strictly convex functions the inequality is strict

$$\ell(w') > \ell(w) + \langle \nabla\, \ell(w), w' - w \rangle\,.$$

Monotonicity of the gradient for convex functions The first-order condition for convexity gives a useful, and equivalent, characterization of the gradient. Write (9.2) for w, w' in two opposite directions

$$\ell(w) \geq \ell(w') + \langle \nabla \ell(w'), w' - w \rangle$$
$$\ell(w') \geq \ell(w) + \langle \nabla \ell(w), w - w' \rangle$$

and add them to get

$$\langle \nabla \ell(w) - \nabla \ell(w'), w - w' \rangle \geq 0. \tag{9.3}$$

It is also true that monotonicity of the gradient implies convexity (try to prove it).

Second-order condition for convexity If ℓ is twice-differentiable with a convex domain, it is convex iff

$$\nabla^2 \ell(w) \succeq 0 \tag{9.4}$$

for all w in the domain. The symbol $\succeq$ denotes positive semi-definiteness of the Hessian matrix $\nabla^2 \ell(w)$

$$\left(\nabla^2 \ell(w)\right)_{ij} = \frac{\partial^2 \ell(w)}{\partial w_i \partial w_j}.$$

For strictly convex functions, the inequality in (9.4) is strict, i.e., the Hessian is positive definite. As an example, the least squares objective $\ell(w) = \|y - Xw\|_2^2$ is convex because

$$\nabla^2 \ell(w) = 2X^\top X$$

which is positive definite for any X (why is it positive definite and not just positive semi-definite?).

Strongly convex functions A function is strongly convex if there exists an $m > 0$ such that

$$\ell(w) - \frac{m}{2}\|w\|_2^2 \text{ is convex.} \tag{9.5}$$

It is easy to see that strict convexity implies convexity. Since the function $\ell(w) - m/2\|w\|^2$ is convex, it satisfies the definition of convexity:

$$\begin{aligned} &\ell(\lambda w + (1-\lambda)w') - \frac{m}{2}\|\lambda w + (1-\lambda)w'\|^2 \\ &\leq \lambda \left(\ell(w) - \frac{m}{2}\|w\|^2\right) + (1-\lambda)\left(\ell(w') - \frac{m}{2}\|w'\|^2\right). \end{aligned} \tag{9.6}$$

But

$$\frac{\lambda m}{2}\|w\|^2 + \frac{(1-\lambda)m}{2}\|w'\|^2 - \frac{m}{2}\|\lambda w + (1-\lambda)w'\|^2 > 0$$

for $\lambda \in (0,1)$ for all $w \neq w'$ because $\|w\|^2$ is strictly convex which shows that if we have a strongly convex function ℓ it also satisfies

$$\ell(\lambda w + (1-\lambda)w') \leq \lambda \ell(w) + (1-\lambda)\ell(w').$$

In other words, we have

$$\text{strong convexity} \;\Rightarrow\; \text{strict convexity} \;\Rightarrow \text{convexity}.$$

Observe that strongly convexity in (9.6) is a stronger version of Jensen's inequality. Strongly convex functions are easier to optimize for algorithms. It will also always be much easier to prove a result in optimization on strongly convex functions. It is easy to see using the second-order condition for convexity that an m-strongly convex function has

$$\nabla^2 \ell(w) \succeq m I_{p\times p}.$$

We will use the following first-order condition for strongly convex functions often. A function is m-strongly convex if and only if

$$\ell(w') \geq \ell(w) + \langle \nabla \ell(w), w' - w\rangle + \frac{m}{2}\|w' - w\|^2 \tag{9.7}$$

for any w, w' in the domain.

9.2 Introduction to Gradient Descent

In this chapter, we will write $\ell(w)$ to denote the training objective, i.e., if we have a classifier $f(x; w)$ and a dataset $D = \left\{(x^i, y^i)\right\}_{i=1,\ldots,n}$ of n samples we will denote

$$\ell(w) := \frac{1}{n}\sum_{i=1}^{n} \ell(w; x^i, y^i).$$

The objective ℓ will always be a function of the entire dataset but we will keep the dependence implicit. Note that the number of samples n is usually quite large in deep learning, so the summation above has a large number of terms on the right-hand side.

Gradient descent is a simple algorithm to minimize $\ell(w)$. Before we study its properties, it will help to refresh the following few facts.

9.2.1 Conditions for optimality

Local and global minima A point w is a local minimum of the function $\ell(w)$ for all all w' in a neighborhood of w we have $\ell(w) \leq \ell(w')$. The point is a global minimum of the function ℓ if this condition is true for all w' in the domain, not just the ones in the neighborhood.

Local minima are global minima for convex functions This is easy to see using an argument by contradiction. If w is a local minimum that is not the global minimum, there exists a point w' in the domain such that $\ell(w') < \ell(w)$. The domain of the function is convex, so pick a point $v = \lambda w' + (1-\lambda)w$ and see that

$$\ell(v) - \ell(w) \leq \lambda\left(\ell(w') - \ell(w)\right)$$

using the definition of convexity. Since w is only a local minimum, we can pick λ to be small enough that the left hand side is non-negative. This shows that $\ell(w') \geq \ell(w)$ but this means that w is a global minimum and we have a contradiction.

Global minimum is unique for strictly convex functions If a function is strictly convex on a convex domain the optimal solution (if it exists) must be unique. Indeed, if there were two solutions w, w' that were both minimizers we would have

$$\ell(w) = \ell(w') \leq \ell(w'') \quad \forall w''. \tag{9.8}$$

We can now apply the definition of convexity to the point $v = (w + w')/2$ to get

$$\ell(v) < \frac{1}{2}\ell(w) + \frac{1}{2}\ell(w') = \ell(w).$$

which contradicts (9.8). The least-squares objective is strictly convex, so the solution is unique global minimizer of the objective.

First-order optimality condition If w is a local minimum of a continuously differentiable function ℓ, then it satisfies

$$\nabla \ell(w) = 0. \tag{9.9}$$

If further ℓ is convex, then $\nabla \ell(w) = 0$ is a sufficient condition for global optimality from the above discussion.

9.2.2 Different types of convergence

Let us assume that we have a continuously differentiable convex function ℓ and let

$$w^* = \operatorname*{argmin}_w \ell(w)$$

be the global minimizer of this function.

We would like to develop an iterative scheme that takes in the initialization of the weights w^0 and updates them to obtain a sequence

$$w^0, w^1, \ldots, w^t, \ldots$$

Along this sequence we are interested in understanding the

1. convergence of the function value $\ell(w^t)$ to the minimal value $\ell(w^*)$, and
2. convergence of the iterates $\|w^t - w^*\|$.

Descent direction We are going to perform a sequence of updates given by

$$w^{t+1} = w^t + \eta\, d^t \tag{9.10}$$

where d^t is called the descent direction and the scaler parameter $\eta > 0$ is called the step-size and determines how far we travel using this descent direction. Any direction such that

$$\left\langle \nabla \ell(w^t), d^t \right\rangle < 0$$

is a good descent direction because this leads to a reduction in the value of the function $\ell(w^{t+1})$ after the weight update. There are numerous ways to pick a good descent direction. Among the simplest ones is gradient descent which

descents along the direction of the negative gradient and thereby performs the following set of updates

$$w^{t+1} = w^t - \eta \nabla \ell(w^t) \tag{9.11}$$

given an initial value w^0. The step-size (also called the learning rate) is chosen by the user. The step-size need not always be fixed, for instance you chose it to be a function of the number of weight updates t in the homework. A good step-size is one that does not overshoot the minimum w^*. For instance, after having chosen a particular descent direction d^t we can compute the best step-size to use at time t by solving

$$\eta^t = \operatorname*{argmin}_{\eta \geq 0} \ell(w^t + \eta\, d^t).$$

This is known as line-search in the optimization literature. You may have seen Newton's method

$$w^{t+1} = w^t - \left(\nabla^2 \ell(w^t)\right)^{-1} \nabla \ell(w^t). \tag{9.12}$$

which does not have a user-tuned step-size and further modifies the descent direction to be the product of the inverse Hessian with the gradient.

⚠ Draw a picture of overshooting using a large step-size.

❓ Can you think of an algorithm for minimizing a function that does not use the gradient of the function to compute the descent direction?

9.3 Convergence rate for gradient descent

We will next understand how quickly gradient descent converges to the global minimum. There are two concrete goals of this analysis

1. to be able to pick the step-size to avoid overshooting without doing line-search, and
2. characterize how many iterations of gradient descent to run until we are guaranteed to be within some distance of the global minimum.

9.3.1 Some assumptions

Before we begin, we will make a few simplifying assumptions on the function $\ell(w)$. These are quite typical in optimization and ensure that we are not dealing with pathological functions that make minimizing them arbitrarily hard.

1. **Lipschitz continuity/bounded gradients** We will assume that ℓ is Lipschitz continuous

$$|\ell(w) - \ell(w')| \leq B\|w - w'\|_2. \tag{9.13}$$

for some $B > 0$. You might also see this condition written as

$$\|\nabla \ell(w)\| \leq B$$

for differentiable functions.

2. **Smoothness** We will always consider functions such that their gradients are L-Lipschitz, i.e.,

$$\|\nabla \ell(w) - \nabla \ell(w')\|_2 \leq L\|w - w'\|_2. \tag{9.14}$$

If ℓ is twice-differentiable, this is equivalent to assuming

$$\nabla^2 \ell(w) \preceq L\, I_{p\times p}. \tag{9.15}$$

From the Cauchy-Schwarz inequality which states that

$$\langle u, v\rangle \leq \|u\|\, \|v\|$$

for two vectors u, v, we have the following implication of smoothness:

$$\langle \nabla \ell(w) - \nabla \ell(w^*), w - w^*\rangle \leq L\|w - w^*\|^2. \tag{9.16}$$

A related concept is called **co-coercivity** of the gradient. The gradient being L-Lipschitz is equivalent to co-coercivity of the gradient with parameter $1/L$

$$\frac{1}{L}\|\nabla \ell(w) - \nabla \ell(w)\|^2 \leq \langle \nabla \ell(w) - \nabla \ell(w'), w - w'\rangle\,. \tag{9.17}$$

We can see that co-coercivity implies Lipschitz continuity of the gradients $\nabla \ell(w)$ using (9.16) and (9.17). The reverse is also true, Lipschitz-continuity of the gradient implies the Descent Lemma Lemma 1 which is seen by applying the Descent Lemma twice for the two functions $g(u) = \ell(u) - \langle \nabla \ell(w'), u\rangle$ and $h(u) = \ell(u) - \langle \nabla \ell(w), u\rangle$.

9.3.2 GD for convex functions

We begin with the so-called Descent Lemma.

Lemma 1 (Descent Lemma). For an L-smooth function, we have

$$\ell(w') \leq \ell(w) + \langle \nabla \ell(w), w' - w\rangle + \frac{L}{2}\|w' - w\|^2. \tag{9.18}$$

for any two w, w' in the domain.

Proof. First, you should compare this with the first-order characterization of convexity

$$\ell(w') \geq \ell(w) + \langle \nabla \ell(w), w' - w\rangle\,.$$

The two conditions can be used to sandwich the value of $\ell(w^{t+1})$ given the value of $\ell(w^t)$ in gradient descent with room for a quadratic term $\frac{L}{2}\|w' - w\|^2$. This also gives some intuition as to what L-smooth really means; a large value of L means that the function ℓ decreases quickly. Let $v = w + \lambda(w' - w)$ and use Taylor's theorem to see that

$$\ell(w') = \ell(w) + \int_0^1 \langle \nabla \ell(v), w' - w\rangle\ \mathrm{d}\lambda \tag{9.19}$$

Subtract $\langle \nabla \ell(w), w' - w\rangle$ from both sides to get

$$\ell(w') - \ell(w) - \langle \nabla \ell(w), w' - w\rangle = \int_0^1 \langle \nabla \ell(v) - \nabla \ell(w), w' - w\rangle\ \mathrm{d}\lambda.$$

Observe that

$$\begin{aligned}
|\ell(w') - \ell(w) - \langle \nabla\, \ell(w), w' - w\rangle| &= \left| \int_0^1 \langle \nabla\, \ell(v) - \nabla\, \ell(w), w' - w\rangle \ \mathrm{d}\lambda \right| \\
&\leq \int_0^1 |\langle \nabla\, \ell(v) - \nabla\, \ell(w), w' - w\rangle| \, \mathrm{d}\lambda \\
&\leq \int_0^1 \|\nabla\, \ell(v) - \nabla\, \ell(w)\| \ \|w' - w\| \mathrm{d}\lambda \\
&\leq L \int_0^1 \lambda \ \|w' - w\|^2 \mathrm{d}\lambda \\
&= \frac{L}{2} \|w' - w\|^2.
\end{aligned}$$

This completes the proof after removing the absolute value on the left-hand side. □

We can use the Descent Lemma twice on two points to w, w' to get (9.16). Another direct consequence of the Descent Lemma is the following corollary that relates the value $\ell(w)$ at any point w in the domain to that of the global minimum.

Corollary 2. For L-smooth convex function ℓ, if w^* is the global minimizer, then

$$\frac{1}{2L} \|\nabla\, \ell(w)\|^2 \leq \ell(w) - \ell(w^*) \leq \frac{L}{2} \|w - w^*\|^2. \tag{9.20}$$

Proof. Since $\nabla\, \ell(w^*) = 0$, the right-hand side follows directly from the Descent Lemma. To get the left-hand size, let us optimize the upper bound in the Descent Lemma using $w' = w + \lambda v$ with $\|v\| = 1$ as follows

$$\begin{aligned}
\ell(w^*) = \inf_{w'} \ell(w') &\leq \inf_{w'} \left\{ \ell(w) + \langle \nabla\, \ell(w), w' - w\rangle + \frac{L}{2} \|w' - w\|^2 \right\} \\
&= \inf_{\|v\|=1} \ \inf_{\lambda} \left\{ \ell(w) + \lambda \, \langle \nabla\, \ell(w), v\rangle + \frac{L}{2} \lambda^2 \right\} \\
&= \inf_{\|v\|=1} \left\{ \ell(w) - \frac{1}{2L} \left(\langle \nabla\, \ell(w), v\rangle \right)^2 \right\} \\
&= \ell(w) - \frac{1}{2L} \|\nabla\, \ell(w)\|^2.
\end{aligned}$$

□

In other words, the gap between the function values $\ell(w) - \ell(w^*)$ is upper-bounded by the gap to the minimizer $\frac{L}{2} \|w - w^*\|^2$ and lower-bounded by the norm of the gradient $\frac{1}{2L} \|\nabla\, \ell(w)\|^2$.

Lemma 3 (Monotonic progress for gradient descent). For gradient descent $w^{t+1} = w^t - \eta\, \nabla\, \ell(w^t)$, if we pick the step-size

$$\eta \leq \frac{1}{L} \tag{9.21}$$

we have

$$\ell(w^{t+1}) \leq \ell(w^t) - \frac{\eta}{2} \ \|\nabla\, \ell(w^t)\|^2. \tag{9.22}$$

Further,

$$\ell(w^{t+1}) - \ell(w^*) \leq \frac{1}{2\eta}\left(\|w^t - w^*\|^2 - \|w^{t+1} - w^*\|^2\right) \tag{9.23}$$

which implies

$$\|w^{t+1} - w^*\|^2 \leq \|w^t - w^*\|^2. \tag{9.24}$$

Proof. Substitute $\eta \leq 1/L$ in the Descent Lemma and simplify to get (9.22). The second result is obtained by

$$\begin{aligned}
0 \leq \ell(w^{t+1}) - \ell(w^*) &\leq \ell(w^t) - \ell(w^*) - \frac{\eta}{2}\|\nabla \ell(w^t)\|^2 \\
&\leq \left\langle \nabla \ell(w^t), w^t - w^* \right\rangle - \frac{\eta}{2}\|\nabla \ell(w^t)\|^2 \\
&= \frac{1}{2\eta}\left(\|w^t - w^*\|^2 - \|w^t - w^* - \eta \nabla \ell(w^t)\|^2\right) \\
&= \frac{1}{2\eta}\left(\|w^t - w^*\|^2 - \|w^{t+1} - w^*\|^2\right).
\end{aligned}$$

Observe that since the left-hand side is positive, the claim in (9.24) is true. □

We have therefore shown that if the step-size is not too large (the smoothness parameter of the function determines how large the step-size can be) then gradient descent always improves the value of the function with each iteration (9.22). It also improves the distance of the weights to the global minimum at each iteration (9.24).

Lemma 4 (Convergence rate for gradient descent, convex function). For gradient descent $w^{t+1} = w^t - \eta \nabla \ell(w^t)$ with step-size $\eta < 1/L$, we have

$$\ell(w^{t+1}) - \ell(w^*) \leq \frac{1}{2\,t\,\eta}\,\|w^0 - w^*\|^2. \tag{9.25}$$

Proof. We sum up the expression in (9.23) for all times t to get

$$\begin{aligned}
\sum_{s=1}^{t} \ell(w^s) - \ell(w^*) &\leq \frac{1}{2\eta}\sum_{s=1}^{t}\left(\|w^{s-1} - w^*\|^2 - \|w^s - w^*\|^2\right) \\
&= \frac{1}{2\eta}\left(\|w^0 - w^*\|^2 - \|w^t - w^*\|^2\right) \\
&\leq \frac{1}{2\eta}\|w^0 - w^*\|^2.
\end{aligned}$$

We know from (9.22) that $\ell(w^t)$ is non-increasing, so we can write

$$\ell(w^t) - \ell(w^*) \leq \frac{1}{t}\sum_{s=1}^{t}\left(\ell(w^s) - \ell(w^*)\right) \leq \frac{1}{2\,t\,\eta}\,\|w^0 - w^*\|^2.$$

□

If we want to find a weights with

$$\ell(w^t) - \ell(w^*) \leq \epsilon$$

for a convex function, we need to run gradient descent for at least

$$\mathcal{O}(1/\epsilon)$$

iterations. This is an important result to remember.

9.3.3 Gradient descent for strongly convex functions

Things are much better if the function we are minimizing is strongly convex. First we have the following lemma for strongly-convex functions which involves a rewriting co-coercivity condition for strongly convex functions.

Lemma 5 (Co-coercivity for strongly convex function). If $\ell(w)$ is m-strongly convex, and L-smooth, then the function $g(w) = \ell(w) - \frac{m}{2}\|w\|^2$ is convex and $L - m$-smooth. The co-coercivity condition for $\nabla\, g(w)$ can therefore be re-written as

$$\langle \nabla\, \ell(w) - \nabla\, \ell(w'), w - w' \rangle \quad \geq \frac{mL}{m+L}\|w - w'\|^2 + \frac{1}{m+L}\|\nabla\, \ell(w) - \nabla\, \ell(w')\|^2. \tag{9.26}$$

Proof. The convexity of $g(w)$ is immediate to see from the definition of strong convexity of $\ell(w)$. Use the monotonicity of the gradient of $g(w)$ to get

$$\begin{aligned} 0 &\leq \langle \nabla\, g(w) - \nabla\, g(w'), w - w' \rangle \\ &= \langle \nabla\, \ell(w) - \nabla\, \ell(w'), w - w' \rangle - m\|w - w'\|^2 \\ &\leq (L - m)\|w - w'\|^2. \end{aligned}$$

We can now rewrite the co-coercivity condition for $\nabla\, g(w)$ with the smoothness parameter $L - m$ and simplify to get (9.26). □

Lemma 6 (Convergence rate of gradient descent for strongly convex functions). For strongly convex functions we have pick a step-size

$$0 < \eta < \frac{2}{m+L}$$

to get

$$\|w^{t+1} - w^*\|^2 \leq \left(1 - \eta\frac{2mL}{m+L}\right)\|w^t - w^*\|^2. \tag{9.27}$$

which gives

$$\|w^t - w^*\|^2 \leq c^t\, \|w^0 - w^*\|^2 \tag{9.28}$$

where $c = \left(1 - \eta\frac{2mL}{m+L}\right)$.

Proof. We expand the left hand-side in (9.27) to get

$$\begin{aligned} \|w^{t+1} - w^*\|^2 &= \|w^t - \eta\, \nabla\, \ell(w^t) - w^*\|^2 \\ &= \|w^t - w^*\|^2 - 2\eta\, \langle \nabla\, \ell(w^t), w^t - w^* \rangle + \eta^2\|\nabla\, \ell(w^t)\|^2 \\ &\leq \left(1 - \eta\frac{2mL}{m+L}\right)\|w^t - w^*\|^2 + \eta\left(\eta - \frac{2}{m+L}\right)\|\nabla\, \ell(w^t)\|^2 \\ &\leq \left(1 - \eta\frac{2mL}{m+L}\right)\|w^t - w^*\|^2. \end{aligned}$$

We have substituted the co-coercivity condition from (9.26) for the inner-product with $w' := w^t$ and $w := w^*$ to get the first inequality. This implies that the distance to the global minimum $\|w^t - w^*\|$ decreases multiplicatively; compare this with (9.24) where the progress is additive. The additional assumption of strong convexity therefore means that we are making very quick progress towards the global minimum. We can use this inequality repeatedly for all iterations t to get

$$\|w^t - w^*\|^2 \leq c^t \, \|w^0 - w^*\|^2$$

where $c = \left(1 - \eta \frac{2mL}{m+L}\right)$. □

Strong convexity enables much faster progress towards the global minimum. If we want $\|w^t - w^*\| \leq \epsilon$ we need

$$\mathcal{O}(\log(1/\epsilon))$$

iterations of gradient descent. This is *much* less than that for a convex function. This is called *linear* convergence because we need a constant number of iterations to reduce the gap to the optimal in half. The naming convention is a bit unusual here but you will see that if we plot $\log\|w^t - w^*\|$ (or $\log\left(\ell(w^t) - \ell(w^*)\right)$) on the Y-axis and number of iterations t on the X-axis, we get a straight line for gradient descent on strongly-convex functions; you can see this from (9.28).

We say that the convergence rate of gradient descent for non-strongly convex functions is *sub-linear*. The longer we run GD for convex functions, the slower its progress.

Further, if we pick the largest step-size $\eta = 2/(m + L)$ we get

$$c = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 < 1. \tag{9.29}$$

where $\kappa = L/m$ is the condition number of the Hessian (it is the ratio of the largest eigenvalue and the smallest eigenvalue). Larger the condition number κ, closer to 1 the multiplicative constant c and *slower* the convergence rate of gradient descent.

⚠ Plot the convergence rate of gradient descent for convex and strongly-convex functions.

⚠ In the optimization literature, an algorithm with

$$\lim_{t \to \infty} \frac{\ell(w^{t+1}) - \ell(w^*)}{\ell(w^t) - \ell(w^*)} = \rho$$

is said to be sub-linear if $\rho \in (0, 1)$, linear if $\rho = 1$ and super-linear if $\rho = 0$.

A few more points to note

1. The step-size if limited by $m + L$ but the convergence rate depends on $\kappa = L/m$. Smaller the value of c, faster the convergence.
2. Larger the L, smaller the ideal step-size η
3. You can get the upper bound

$$\ell(w^t) - \ell(w^*) \leq \frac{L}{2}\|w^t - w^*\|^2 \leq \frac{c^t L}{2}\|w^0 - w^*\|^2 \tag{9.30}$$

using (9.20).

You will also see the convergence rate written in many papers as

$$\|w^t - w^*\| \leq e^{-4t/\kappa} \, \|w^0 - w^*\|. \tag{9.31}$$

You can get this inequality by using the fact that $1 + x \leq e^x$ in (9.29). We can use this to pull out the dependence on κ in the convergence rate; for strongly convex functions, gradient descent requires

$$\mathcal{O}(\kappa \log(1/\epsilon))$$

iterations to reach within an ϵ-neighborhood of the global minimum $\ell(w^*)$. This suggests that smaller the condition number κ fewer the iterations required.

We can intuitively understand why convergence of gradient descent is slower for a large condition number. A large condition number means that some directions of the objective ℓ are highly curved while some others are very flat. It is difficult to pick one single scalar step-size in such situations that makes quick progress along the flat directions but also does not overshoot the highly curved directions. You might imagine that clever schemes to change the step-size depending upon the local geometry of the function $\ell(w^t)$ could help solve this issue and indeed it does, but such adaptive schemes are expensive to implement computationally. We will see some algorithms that pick different step-sizes for different weights in Chapter 11.

⚠ Draw a picture of this phenomenon for a quadratic objective $\ell(w) = \langle w, Aw \rangle$ for matrices $A \succ 0$ with different condition numbers κ.

9.4 Limits on convergence rate of first-order methods

It is a powerful and deep result that we cannot do better than a linear convergence rate for optimization methods that only use the gradient of the function $\ell(w)$. More precisely, for any first-order method, i.e., any method where the iterate at step t given by w^t is chosen to be

$$w^t \in w^0 + \text{span}\left\{\nabla \ell(w^0), \ldots, \nabla \ell(w^t)\right\},$$

we have the following theorem by Yurii Nesterov.

Theorem 7 (Nesterov's lower bound). If $w \in \mathbb{R}^p$, for any $t \leq (p-1)/2$ and every initialization of weights w^0 there exist functions $\ell(w)$ that are convex, differentiable, L-smooth with finite optimal value $\ell(w^*)$ such that any first-order method has

$$\ell(w^t) - \ell(w^*) \geq \frac{3}{32} \frac{L\|w^0 - w^*\|^2}{(t+1)^2}.$$

p is the dimension of the parameter space

Let us read the statement of the theorem carefully. It states that *fixed* a time t and initial condition w^0, we can *find* a convex function $\ell(w)$ such that it takes gradient descent at least $\mathcal{O}(1/\epsilon^2)$ to reach an ϵ-neighborhood of the optimal value $\ell(w^*)$. The implication of this theorem is as follows. The convergence rate $\mathcal{O}(1/\epsilon)$ we obtained for convex functions is not the best rate we can get. Nesterov's lower bound suggests that there should be gradient-based algorithms that only require $\mathcal{O}(1/\sqrt{\epsilon})$ iterations. Such methods will be the topic of the next Chapter.