Some notes on gradient descent

Gradient descent is a method to minimize convex functions. In its basic form, it finds an approximate solution to the unconstrained problem

$$\begin{aligned} &\min f(\mathbf{x}) \\ &s.t. \\ &\quad \mathbf{x} \in \mathbb{R}^n \end{aligned}$$

where $f : \mathbb{R}^n \to R$ is a convex function.

There are a number of variations of gradient descent, such as *accelerated* gradient descent, *conjugate* gradient descent, *stochastic* gradient descent, and *coordinate* descent. There are also methods to apply gradient descent algorithms to *constrained* optimization problems.

Additionally, gradient descent methods often work well on non-convex optimization problems, for reasons that are typically not well understood.

Two algorithms underlie the success of deep learning: careful implementations of stochastic gradient descent on parallel architectures, and the backpropagation algorithm to compute gradients. Training deep neural networks is a non-convex problem, and it is one of the cases in which the good performance of gradient descent methods is poorly understood.

In this notes, we will restrict ourselves to the rigorous analysis of the basic gradient descent algorithm for the problem of unconstrained minimization of strongly convex functions. Before proceeding, we will review the definitions of gradient and convexity, and see how to bound the error in a truncated Taylor series.

1 Calculus Review

1.1 Univariate functions

Let us start with the case of univariate functions $f : \mathbb{R} \to \mathbb{R}$. A function is *convex* if, for every two points x and z, the line segment between $(x, f(x))$ and $(z, f(z))$ is

above the graph of the function f, that is, if for every $c \in [0,1]$ we have

$$f(cx + (1-c)z) \leq cf(x) + (1-c)f(z)$$

The *derivative* of f at x is

$$f'(x) := \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

provided that the limit exists. If $f'(x)$ is well defined for all x, and continuous, then f is called *continuously differentiable*. If f' is differentiable then its derivative is denoted f'', and, if f'' is continuous, f is called *twice–continuously-differentiable*. (We will omit the "continuously" from now on.)

The connection between convexity and derivatives is that if f is differentiable, then f is convex if and only if

$$f(z) \geq f(x) + (z-x) \cdot f'(x)$$

for all x and z. The function $z \to f(x) + (z-x) \cdot f'(x)$ is the tangent of f at x, so the above is saying that a differentiable function is convex if and only if its graph is always above any of its tangents.

If f is twice-differentiable, then it is convex if and only if

$$f''(x) \geq 0$$

for all x.

A function f is *strictly convex* if for every $x, y \in \mathbb{R}$ and every $c \in (0,1)$ we have

$$f(cx + (1-c)z) < cf(x) + (1-c)f(z)$$

The above characterizations of convexity for differentiable and twice-differentiable functions apply also to strong convexity, by replacing $\leq$ with $<$.

The importance of the notion of strong convexity in the context of minimization is that if f is strongly convex, then either it has no minimum (meaning $\inf_x f(x) = -\infty$) or it has exactly one minimum. If f is differentiable and strongly convex, then the minimum, if it exists, is the unique point x^* such that $f'(x^*) = 0$.

To complete this review of univariate calculus, recall Taylor's theorem: if a function is infinitely differentiable, and if we call $f^{(k)}$ the k-th derivative of f, then we have that, for every x and z

$$f(z) = f(x) + \sum_{k=1}^{\infty} \frac{(z-x)^k}{k!} f^{(k)}(x)$$

If we consider only the first two terms of the Taylor's series, that is the expression $f(x)+(z-x)\cdot f'(x)$, then we have what is called the *first order* approximation of $f(z)$. This is just the line tangent to f at x. If f is differentiable, then we can rephrase the condition of being strictly convex as the condition that $f(z)$ is always bigger than its first-order approximation, for every $z \neq x$.

If f is twice differentiable, then the difference between f and its first-order approximation has a very nice expression:

$$f(z) = f(x) + (z-x)\cdot f'(x) + \frac{1}{2}(z-x)^2 \cdot f''(w) \tag{1}$$

For some point w between x and z.

In particular, if there are bounds ℓ and L such that, for every w,

$$\ell \leq f''(w) \leq L$$

then we can get upper and lower bounds to the integral in (1) as follows:

$$f(x) + (z-x)\cdot f'(x) + \frac{\ell}{2}\cdot(z-x)^2 \leq f(z) \leq f(x) + (z-x)\cdot f'(x) + \frac{L}{2}\cdot(z-x)^2$$

1.2 * Eigenvalues of Symmetric Matrices

The content of this section is not necessary to understand what comes next (hence the star in the title), but it motivates some of the definitions that will come later.

For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, their inner product is

$$\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^{n} x_i y_i$$

The inner product can also be written in matrix notation as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$$

The norm of a vector is

$$||\mathbf{x}|| = \sqrt{\sum_i x_i^2} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

If M is an $n \times n$ real matrix, we say that a real number λ is an *eigenvalue* of M if there is a non-zero vector $\mathbf{x} \in \mathbb{R}^n$ such that $M\mathbf{x} = \lambda \mathbf{x}$. If M is symmetric, then (it is a non-trivial theorem that) the smallest eigenvalue of M is

$$\min_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^T M \mathbf{x}}{||\mathbf{x}||^2}$$

and the largest eigenvalue is

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T M\mathbf{x}}{||\mathbf{x}||^2}$$

We say that a symmetric matrix M is *positive semidefinite* if all the eigenvalues are non-negative, and we say that is *positive definite* if all the eigenvalues are strictly positive. By the above characterization of the smallest eigenvalue, it also follows that a symmetric matrix M is positive semidefinite if and only if

$$\forall \mathbf{x} \in \mathbb{R}^n.\ \mathbf{x}^T M\mathbf{x} \geq 0$$

and that it is positive definite if and only if

$$\forall \mathbf{x} \in \mathbb{R}^n - \{\mathbf{0}\}.\ \mathbf{x}^T M\mathbf{x} > 0$$

If M is a positive definite matrix, we call the ratio between largest and smallest eigenvalue the *condition number* of M.

1.3 Multivariate functions

Now we develop the multivariate analogs of all the concepts from Section 1.1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for every two points $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$, and every $c \in [0, 1]$ we have

$$f(c \cdot \mathbf{x} + (1 - c) \cdot \mathbf{z}) \leq c \cdot f(\mathbf{x}) + (1 - c) \cdot f(\mathbf{z})$$

and we say that f is strictly convex if the above inequality is strict for all $c \in (0, 1)$.

The *partial derivative* of f with respect to x_i at $\mathbf{x}$, written $\frac{\partial f}{\partial x_i}(\mathbf{x})$, is the derivative of f if we think of it as a univariate function of the i-th variable, and we think of the other variables as constants.

For example, if $f(x_1, x_2) = x_1^2 - x_1x_2 + 2x_2^2$, then

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 2x_1 - x_2, \quad \frac{\partial f}{\partial x_2}(x_1, x_2) = -x_1 + 4x_2$$

The *directional derivative* of f at $\mathbf{x}$ in the direction $\mathbf{y}$, where $\mathbf{x}, \mathbf{y}$ are n-dimensional vectors, is defined as

$$\frac{df}{d\mathbf{y}}(\mathbf{x}) := \lim_{\epsilon \to 0} \frac{f(\mathbf{x} + \epsilon\mathbf{y}) - f(\mathbf{x})}{\epsilon}$$

(There are half a dozen accepted notations for directional derivatives, including $D_{\mathbf{y}} f(\mathbf{x})$, $\nabla_{\mathbf{y}} f(\mathbf{x})$ and $f_{\mathbf{y}}(\mathbf{x})$. Also, some authors require $\mathbf{y}$ to be a unit vector. Luckily, we do not need directional derivatives in our analysis of gradient descent; we introduce them here just to motivate the notion of gradient.)

That is, the directional derivative is just the derivative of the univariate function $t \to f(\mathbf{x}+t\mathbf{y})$ at 0, that is the derivative of f restricted to the line that passes through $\mathbf{x}$ and goes in the direction $\mathbf{y}$. Note also that the partial derivate with respect to x_i is the same as the directional derivative in the direction $(0, 0, \ldots, 1, \ldots, 0)$, the vector that has a 1 in the i-th coordinate and zeroes in the other coordinate.

The analog of $f'(x)$ for multivariate function is the information about all directional derivative at $\mathbf{x}$, along all possible directions. This is an infinite amount of data, but, luckily, it is entirely captured by the n partial derivatives because of the following (non-trivial) theorem: for every $\mathbf{x}$ and every $\mathbf{y}$,

$$\frac{df}{d\mathbf{y}}(\mathbf{x}) = y_1 \cdot \frac{\partial f}{\partial x_1}(\mathbf{x}) + \cdots + y_n \cdot \frac{\partial f}{\partial x_n}(\mathbf{x})$$

assuming that all partial derivates exist at $\mathbf{x}$. In other words, if we define the *gradient* of f at $\mathbf{x}$ as the vector

$$\nabla f(x) := \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \cdots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right)$$

then every directional derivative is given by

$$\frac{df}{d\mathbf{y}}(\mathbf{x}) = \langle \mathbf{y}, \nabla f(\mathbf{x}) \rangle$$

We will not prove this characterization of the gradient, but let us see one example. If, as before, $f(x_1, x_2) = x_1^2 - x_1 x_2 + 2x_2^2$, then its directional derivative in the direction $\mathbf{y} = (1, 1)$ is

$$\begin{aligned}
\frac{df}{d(1,1)}(\mathbf{x}) &= \lim_{\epsilon \to 0} \frac{f(x_1+\epsilon, x_2+\epsilon) - f(x_1, x_2)}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{(x_1+\epsilon)^2 - (x_1+\epsilon)(x_2+\epsilon) + 2(x_2+\epsilon)^2 - x_1^2 + x_1 x_2 - 2x_2^2}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{2\epsilon x_1 + \epsilon^2 - \epsilon x_1 - \epsilon x_2 - \epsilon^2 + 4\epsilon x_2 + 2\epsilon^2}{\epsilon} \\
&= \lim_{\epsilon \to 0} x_1 + 3x_2 + 2\epsilon^2 \\
&= x_1 + 3x_2
\end{aligned}$$

and our formula in terms of the gradient correctly gives

$$\langle (1,1), \nabla f(\mathbf{x}) \rangle = \langle (1,1), (2x_1 - x_2, -x_1 + 4x_2) \rangle = x_1 + 3x_2$$

We say that f is (once) differentiable if $\nabla f(\mathbf{x})$ is well defined for every $\mathbf{x}$ and the function $\mathbf{x} \to \nabla f(\mathbf{x})$ is continuous.

If f is differentiable then it is convex if and only if, for every two points $\mathbf{x}$ and $\mathbf{z}$, we have

$$f(\mathbf{z}) \geq f(\mathbf{x}) + (\mathbf{z} - \mathbf{x})^T \nabla f(\mathbf{x})$$

and it is strictly convex if the above inequality is strict for all $\mathbf{x} \neq \mathbf{z}$. (Note the similarity with the $f(z) \geq f(x) + (z - x) \cdot f'(x)$ condition in the univariate case.)

Let us now move on to the multivariate analog of second derivatives. For two variables x_i and x_j (we allow $i = j$), the second partial derivative with respect to x_i and x_j at $\mathbf{x}$, denoted $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})$ is the partial derivative with respect to x_i of the function $\mathbf{x} \to \frac{\partial f}{\partial x_j}(\mathbf{x})$. A very nice fact is that the order of derivation is not important, that is, if these second partial derivates exist, then

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x})$$

Just like the gradient is the n-dimensional vector that contains all partial derivatives, the *Hessian* of f at $\mathbf{x}$ is the $n \times n$ matrix $Hf(x)$ such that

$$Hf(x)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x})$$

The function f is twice differentiable if the Hessian is always well defined and the function $\mathbf{x} \to Hf(x)$ is continuous.

It would be tempting to guess that, for a twice differentiable function, f is convex if and only if all the entries of the Hessian are ≥ 0. Unfortunately the Hessian containing only non-negative entries is neither a sufficient nor a necessary condition for convexity.

For example, at every point, the Hessian of the function $f(x_1, x_2) = x_1 \cdot x_2$ is

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

but the function is not convex, while the Hessian of the function $f(x_1, x_2) = x_1^2 - x_1 x_2 + 2x_2^2$ is

$$\begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix}$$

and f is convex.

Some insight into the right definition comes from considering second directional derivatives. For every two vectors $\mathbf{y}, \mathbf{z}$, if we take the directional derivative of f in the direction $\mathbf{y}$ and then in the direction $\mathbf{z}$, we get

$$\frac{d^2 f}{d\mathbf{z} d\mathbf{y}}(\mathbf{x}) = \mathbf{z}^T (Hf(x))\mathbf{y}$$

(this is a non-trivial theorem that we will not prove.) In particular, the second derivative of f at $\mathbf{x}$ in the direction $\mathbf{y}$ is $\mathbf{y}^T (Hf(\mathbf{x}))\mathbf{y}$.

Now the correct characterization of convexity in term of Hessian is that if f is twice differentiable, than it is convex if and only for every vectors $\mathbf{x}$ and $\mathbf{y}$,

$$\mathbf{y}^T (Hf(\mathbf{x}))\mathbf{y} \geq 0$$

and it is strictly convex if and only if the above inequality is strict for all $\mathbf{y} \neq \mathbf{0}$. (If you have read the previous subsection, we are saying that f is convex if and only if the Hessian is always positive semidefinite, and it is strictly convex if and only if the Hessian is always positive definite.)

If f is strictly convex, then either it has no minimum or it has exactly one minimum. If it is strictly convex and differentiable, then the minimum, if it exists, is the unique point at which the gradient is zero.

Now we could define higher derivatives, and the analog of the third-derivative will be a 3-tensor (a 3-dimensional array) and so on, and we could state a Taylor theorem for multivariate functions, but we will not need that for our analysis of the gradient descent algorithm. We will need, however, a bound on the error of the first-order Taylor approximation.

We have the following theorem: if f is twice-differentiable, then for every $\mathbf{x}$ and $\mathbf{z}$

$$f(\mathbf{z}) = f(\mathbf{x}) + (\mathbf{z} - \mathbf{x})^T \nabla f(\mathbf{x}) + \frac{1}{2}(\mathbf{z} - \mathbf{x})^T (Hf(\mathbf{w}))(\mathbf{z} - \mathbf{x})$$

For some $\mathbf{w}$ in the line segment between $\mathbf{x}$ and $\mathbf{z}$.

In particular, if there are positive numbers ℓ and L such that, for all $\mathbf{x}$ and $\mathbf{w}$, we have

$$\ell \cdot ||\mathbf{x}||^2 \leq \mathbf{x}^T (Hf(\mathbf{w}))\mathbf{x} \leq L \cdot ||\mathbf{x}||^2$$

(in other word, if the eigenvalues of the Hessian are always between ℓ and L) then we also have

$$f(\mathbf{x}) + (\mathbf{z} - \mathbf{x})^T \nabla f(\mathbf{x}) + \frac{\ell}{2}||\mathbf{z} - \mathbf{x}||^2 \leq f(\mathbf{z}) \leq f(\mathbf{x}) + (\mathbf{z} - \mathbf{x})^T \nabla f(\mathbf{x}) + \frac{L}{2}||\mathbf{z} - \mathbf{x}||^2$$

2 Unconstrained Quadratic Convex Minimization: Two Applications

2.1 Solving Linear Systems

Suppose that we want to solve the unconstrained minimization problem $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$ and that f is a quadratic polynomial. Then it is possible to write f as

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x} + c$$

where A is symmetric. (The reason why we did not write it in the more natural form $\mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ will be clear in a moment.)

We see that the gradient of f is

$$\nabla f(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$$

and that, at every point, the Hessian of f is just the matrix A. This means that f is strictly convex if and only if A is positive definite. In such a case, there is a unique minimum, which is achieved at the unique point $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{b}$.

This means that if we want to minimize a strictly convex quadratic function, we can do so by solving, for example using Gaussian elimination, a linear system where the matrix of constraints is positive definite.

But it also means that if we want to solve the linear system $A\mathbf{x} = \mathbf{b}$ and A is positive definite, we can do so by solving the unconstrained minimization problem for $f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x}$. It turns out that, for well-conditioned matrices (i.e. matrices whose condition number is small), applying gradient descent to minimize $\frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x}$ is faster, more numerically stable, and more memory efficient than applying Gaussian elimination.

2.2 Least Squares Regression

Suppose we are given n points in the plane $(a_1, b_1), \ldots, (a_n, b_n)$ and we want to find a line that passes as closely as possible to those points. That is, we want to find a linear function $t \to tx + y$ such that the b_i are as close as possible to the values $a_i x + y$. If we measure the error by the sum of squared-differences, we have the optimization problem

$$\min \sum_{i=1}^{n} (a_i x + y - b_i)^2$$
$$s.t.$$
$$x, y \in \mathbb{R}$$

Which is strictly convex provided that $n \geq 2$. Since the cost function is quadratic, the problem can be solved either by solving a system of two linear equations in two variables, or applying gradient descent to a function of two variables. In this case, solving the linear system is preferable, but for higher-dimensional analog it is generally better to apply gradient descent.

3 A Good Exercise and a Useful Lemma

In this section we prove the following result.

Lemma 1 *Let* $\mathbf{b}$ *and* $\mathbf{v}$ *be an arbitrary* n*-dimensional vector,* $r > 0$ *be a positive real, and define the function*

$$f(\mathbf{x}) := \langle \mathbf{b}, \mathbf{x} \rangle + r||\mathbf{x}||^2$$

Then f *is strongly convex, its unique minimum is the point* $\mathbf{x}^* = -\frac{1}{2r}\mathbf{b}$*, and, for every* $\mathbf{x}$*,*

$$\langle \mathbf{b}, \mathbf{x} \rangle + r||\mathbf{x}||^2 \geq -\frac{1}{4r}||\mathbf{b}||^2$$

Proving this result will be a good exercise in applying the definitions of convexity, Hessian, and gradient, and it will play a key role in the analysis of gradient descent for general strongly convex functions that we develop in the next section.

Let us start by computing the Hessian of f. The Hessian of $\langle \mathbf{b}, \mathbf{x} \rangle$ is zero, because all second derivatives of a linear function are zero. The Hessian of $||\mathbf{x}||^2 = \sum_i x_i^2$ is $2I$, and so

$$Hf(\mathbf{x}) = 2rI$$

for all $\mathbf{x}$ and it is clearly positive definite. So we have proved that f is strongly convex.

The gradient of f is

$$\nabla f(\mathbf{x}) = \mathbf{b} + 2r\mathbf{x}$$

and so the unique point $\mathbf{x}^*$ such that $\nabla f(\mathbf{x}^*) = 0$ is $\mathbf{x}^* = -\frac{1}{2r}\mathbf{b}$.

Finally, we have that, for every $\mathbf{x}$,

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) = -\frac{1}{2r}||\mathbf{b}||^2 + \frac{r}{4r^2}||\mathbf{b}||^2 = -\frac{1}{4r}||\mathbf{b}||^2$$

4 Analysis of Gradient Descent

The structure of a gradient descent algorithm is as follows. To implement the algorithm one needs to specify the parts that we wrote in square brackets. We will show how to do so assuming that we have upper and lower bounds to the eigenvalues of the Hessian.

- $\mathbf{x}^0 :=$ [initial point]
- $i := 0$
- while not [termination condition]
  - compute $\nabla f(\mathbf{x}^i)$
  - $\epsilon_i :=$ [step length at time i]
  - $\mathbf{x}^{i+1} := \mathbf{x}^i - \epsilon_i \cdot \nabla f(\mathbf{x}^i)$
  - $i := i + 1$
- return $\mathbf{x}^i$

The intuition for the algorithm is that $\nabla f(\mathbf{x})$ points in the direction in which f changes most quickly near $\mathbf{x}$, and f increases in the direction of $\nabla f(\mathbf{x})$ and decreases in the direction of $-\nabla f(\mathbf{x})$. Thus, if, at time t, we move from $\mathbf{x}^t$ by a small multiple of $-\nabla f(\mathbf{x}^t)$ we should land in a new point with a smaller value of f, meaning that we are making progress toward reaching the minimum. To get the algorithm to work correctly, we do not want to move too much, because we may overshoot the minimum, and actually end up in a worse point than the one we started from, but we also do not want to move too little, or else it will take a huge number of iterations to reach the minimum.

In general, finding a good value of ϵ_t is more an art than a science, and there are several good heuristics. In the following analysis, we will assume that we know positive numbers ℓ, L such that the eigenvalues of the Hessian are always between ℓ and L and, in this case, we can make the simple choice of always picking $\epsilon_i = 1/L$ at every step.

4.1 Quadratic Functions

Let us begin with the special case in which we want to minimize a strictly convex quadratic function f. We write

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x} + c$$

where A is a symmetric matrix. Suppose that we are given positive numbers ℓ, L such that, for every non-zero vector $\mathbf{y}$,

$$\ell \leq \frac{\mathbf{y}^T A \mathbf{y}}{||\mathbf{y}||^2} \leq L$$

and suppose that we run the gradient descent algorithm from an initial point $\mathbf{x}^0$ using $\epsilon_i = \frac{1}{2L}$.

Let $\mathbf{x}^*$ be the unique optimum. Then we have

$$\begin{aligned}
||\mathbf{x}^{t+1} - \mathbf{x}^*|| &= ||\mathbf{x}^t - \epsilon_t \cdot (A\mathbf{x}^t - b) - \mathbf{x}^*|| \\
&= ||\mathbf{x}^t - \epsilon_t A\mathbf{x}^t + \epsilon_t A\mathbf{x}^* - \mathbf{x}^*|| \\
&= ||(I - \epsilon_t A) \cdot (\mathbf{x}^t - \mathbf{x}^*)|| \\
&\leq ||I - \epsilon_t A|| \cdot ||\mathbf{x}^t - \mathbf{x}^*|| \\
&\leq \left(1 - \frac{\ell}{2L}\right) \cdot ||\mathbf{x}^t - \mathbf{x}^*||
\end{aligned}$$

The last two lines above require a bit of knowledge of linear algebra. It is ok if you take them on faith and you skip ahead to the next paragraph. The norm of a matrix M is defined as $||M|| := \sup_{\mathbf{x} \neq \mathbf{0}} \frac{||M\mathbf{x}||}{||\mathbf{x}||}$ so that, by definition, we always have $||M\mathbf{x}|| \leq ||M|| \cdot ||\mathbf{x}||$, explaining the second-to-last line. If M is symmetric, as is the case for $I - \epsilon_t A$, then $||M||$ is the largest absolute value of the eigenvalues. The eigenvalues of $I - \epsilon_t A$ are precisely $1 - \epsilon_t \lambda$ for every eigenvalue λ of A, and so they are in the range between $1 - \epsilon_t L$ and $1 - \epsilon_t \ell$. If we choose $\epsilon_t = \frac{1}{2L}$, then all the eigenvalues are between $\frac{1}{2}$ and $1 - \frac{\ell}{2L}$ and so the norm is at most $1 - \frac{\ell}{2L}$.

So

$$||\mathbf{x}^t - \mathbf{x}^*|| \leq \left(1 - \frac{\ell}{2L}\right)^t ||\mathbf{x}^0 - \mathbf{x}^*||$$

meaning that if we want to guarantee

$$||\mathbf{x}^t - \mathbf{x}^*|| \leq \delta ||\mathbf{x}^*||$$

it is enough to start gradient descent at $\mathbf{x}^0 = \mathbf{0}$ and then continue for $t = O\left(\frac{L}{\ell} \log \frac{1}{\delta}\right)$ steps.

4.2 General Functions

Let now f be an arbitrary strongly convex functions, and suppose that we are given two positive numbers ℓ, L such that, for every non-zero vector $\mathbf{y}$ and every input $\mathbf{x}$

we have

$$\ell \leq \frac{\mathbf{y}^T(Hf(\mathbf{x}))\mathbf{y}}{||\mathbf{y}||^2} \leq L$$

Then recall that we also have, for all $\mathbf{x}$ and $\mathbf{z}$,

$$f(\mathbf{x}) + (\mathbf{z}-\mathbf{x})^T\nabla f(\mathbf{x}) + \frac{\ell}{2}||\mathbf{z}-\mathbf{x}||^2 \leq f(\mathbf{z}) \leq f(\mathbf{x}) + (\mathbf{z}-\mathbf{x})^T\nabla f(\mathbf{x}) + \frac{L}{2}||\mathbf{z}-\mathbf{x}||^2 \quad (2)$$

Suppose that we run gradient descent starting from a point $\mathbf{x}^0$ using $\epsilon_i = \frac{1}{L}$ at every step, for $t+1$ steps. Let $\mathbf{x}^*$ the unique minimum. Then by using the upper bound in (2) we have

$$\begin{aligned} f(\mathbf{x}^{t+1}) &= f(\mathbf{x}^t - \epsilon_t \nabla f(\mathbf{x}^t)) \\ &\leq f(\mathbf{x}^t) - \epsilon_t \nabla f(\mathbf{x}^t)^T \nabla f(\mathbf{x}^t) + \frac{L}{2}|| - \epsilon_t \nabla f(\mathbf{x}^t)||^2 \\ &= f(\mathbf{x}^t) - \left(\epsilon_t - \frac{L}{2}\epsilon_t^2\right)||\nabla f(\mathbf{x}^t)||^2 \\ &= f(\mathbf{x}^t) - \frac{1}{2L}||\nabla f(\mathbf{x}^t)||^2 \end{aligned}$$

Note that choice of $\epsilon_t = \frac{1}{L}$ maximizes the gain $\epsilon_t - \epsilon_t^2 \frac{L}{2}$.
Using the lower bound in (2) we have

$$\begin{aligned} f(\mathbf{x}^*) &\geq f(\mathbf{x}^t) + \nabla f(\mathbf{x}^t)^T(\mathbf{x}^* - \mathbf{x}^t) + \frac{\ell}{2}||\mathbf{x}^* - \mathbf{x}^t||^2 \\ &\geq f(\mathbf{x}^t) - \frac{1}{2\ell}||\nabla f(\mathbf{x}^t)||^2 \end{aligned}$$

where the second inequality is due to the fact that

$$\nabla f(\mathbf{x}^t)^T(\mathbf{x}^* - \mathbf{x}^t) + \frac{\ell}{2}||\mathbf{x}^* - \mathbf{x}^t||^2 \geq -\frac{1}{2\ell}||\nabla f(\mathbf{x}^t)||^2$$

which is a special case of the last part of Lemma 1 (set $\mathbf{b} = \nabla f(\mathbf{x}^t)$, $\mathbf{x} = \mathbf{x}^* - \mathbf{x}^t$, and $r = \frac{\ell}{2}$).
We can also write what we just discovered as

$$||\nabla f(\mathbf{x}^t)||^2 \geq 2\ell \cdot (f(\mathbf{x}^t) - f(\mathbf{x}^*))$$

and, putting everything together,

$$f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) - \frac{\ell}{L} \cdot ((f(\mathbf{x}^t) - f(\mathbf{x}^*))$$

which is the same as

$$f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*) \leq \left(1 - \frac{\ell}{L}\right) \cdot ((f(\mathbf{x}^t) - f(\mathbf{x}^*))$$

and so

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \left(1 - \frac{\ell}{L}\right)^t \cdot ((f(\mathbf{x}^0) - f(\mathbf{x}^*))$$

In particular, if we want to guarantee that

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \leq \delta \cdot ((f(\mathbf{x}^0) - f(\mathbf{x}^*))$$

we just need to run the algorithm for $t = O\left(\frac{L}{\ell} \cdot \log \frac{1}{\delta}\right)$ steps.

4.3 Summary

If $f = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{b}^T\mathbf{x} + c$ is a quadratic strictly convex function, then setting $\epsilon_i = \frac{1}{2L}$ at each iteration (where L is the largest eigenvalue of A) guarantees convergence in time linear in the condition number of A and logarithmic in the desired quality of approximation. (The convergence is in terms of the distance of the point that we find from the optimum point.)

If f is an arbitrary strictly convex function, and ℓ and L are positive numbers such that the eigenvalues of the Hessian are always between ℓ and L, then setting $\epsilon_i = \frac{1}{L}$ guarantees convergence in time linear in the ratio L/ℓ and logarithmic in the desired approximation. (The convergence is in terms of difference between the cost function of the point that we find and the cost of the optimum point.)