Notes are taken from Tibshirani's lecture notes: https://www.stat.cmu.edu/~ryantibs/convexopt-F13/scribes/

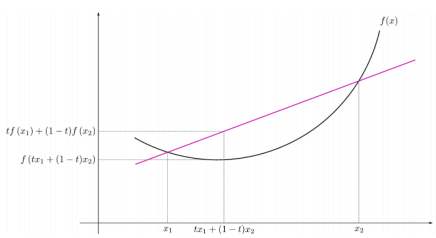
## Convex Function

**Definition 4.28** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if* $\mathrm{dom} f$ *is a convex set and if for all $x, y \in \mathrm{dom} f$, and $\theta$ with $0 \le \theta \le 1$, we have*

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y).$$

**Definition 4.29** *A function $f$ is strictly convex if whenever $x \ne y$, and $0 < \theta < 1$, strict inequality holds, that is, we have*

$$f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y).$$




f(y) >= f(x) + \grad f(x)^T (y-x)

## Strong Convexity

**Definition 4.32** *A differentiable function $f$ is called $m-$strongly convex if $m > 0$ and*

$$(\nabla f(x) - \nabla f(y))^T (x-y) \ge m \|x-y\|_2^2, \forall x, y \in \mathrm{dom} f$$

*An equivalent condition is*

$$f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{m}{2} \|y-x\|_2^2, \forall x, y \in \mathrm{dom} f$$

It is not necessary for a function to be differentiable. We could have the definition without gradient.

**Definition 4.33** *A function $f$ is called $m-$strongly convex if $m > 0$ and for $0 \le t \le 1$*

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) - \frac{1}{2} mt(1-t) \|x-y\|_2^2, \forall x, y \in \mathrm{dom} f$$

If the function is twice continuously differentiable, we could have the definition with Hessian matrix.

**Definition 4.34** *$f$ is called $m-$stronly convex if $m > 0$ and*

$$\nabla^2 f(x) \ge mI, \forall x, y \in \mathrm{dom} f$$

A strongly convex function is also strictly convex, but not vice-versa.

## Extended-Value Extension

**Definition 4.37** $\tilde{f} : \mathbb{R}^n \to \mathbb{R} \bigcup \{\infty\}$ *is extended-value extension of* $f$*:*

$$\tilde{f}(x) = \begin{cases} f(x) & x \in \mathrm{dom} f \\ \infty & x \notin \mathrm{dom} f \end{cases}$$

The extension $\tilde{f}$ is defined on all $\mathbb{R}^n$, and takes values in $\mathbb{R} \bigcup \{\infty\}$. This does not change its convexity

**Theorem 4.38** $f$ *is convex*

$\Leftrightarrow \tilde{f}$ *is convex*

$\Leftrightarrow \tilde{f}(\theta x + (1-\theta)y) \leq \theta \tilde{f}(x) + (1-\theta)\tilde{f}(y), 0 \leq \theta \leq 1$

## Properties of Convex Functions

Let $f$ be a differentiable function, $\mathrm{dom} f$ is open and convex, then we have

$f$ is convex $\Leftrightarrow f(y) \geq f(x) + \nabla f(x)^T (y - x)$

[redacted] The inequality states that for a convex function, the first-order Taylor approximation is a global underestimator of the function. Conversely, if the first-order Taylor approximation of a function is always a global underestimator of the function, then the function is convex.

Let $f$ be twice differentiable, $\mathrm{dom} f$ is open, then we have

$f$ is convex $\Leftrightarrow \nabla^2 f(x) \geq 0, \forall x \in \mathrm{dom} f$

If $\nabla^2 f(x) > 0, \forall x \in \mathrm{dom} f$, $f$ is strictly convex. The converse is not true.

For example, the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^4$ is strictly convex but has zero second derivative at $x = 0$

## Gradient Descent

Recall that we have $f : \mathbb{R}^n \to \mathbb{R}$, convex and differentiable. We want to solve

$$\min_{x \in \mathbb{R}^n} f(x)$$

i.e, to find $x^\star$ such that $f(x^\star)$=min $f(x)$ .

**Gradient descent:** choose initial $x^{(0)} \in \mathbb{R}^n$ , repeat :

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), k = 1, 2, 3, ...$$

Stop at some point(When to stop is quite dependent on what problems you are looking at).

## Coordinate Descent

Similar but coordinate-by-coordinate by picking the coordinate with maximum gradient

## Step Size

### Fixed

### Backtracking Line Search

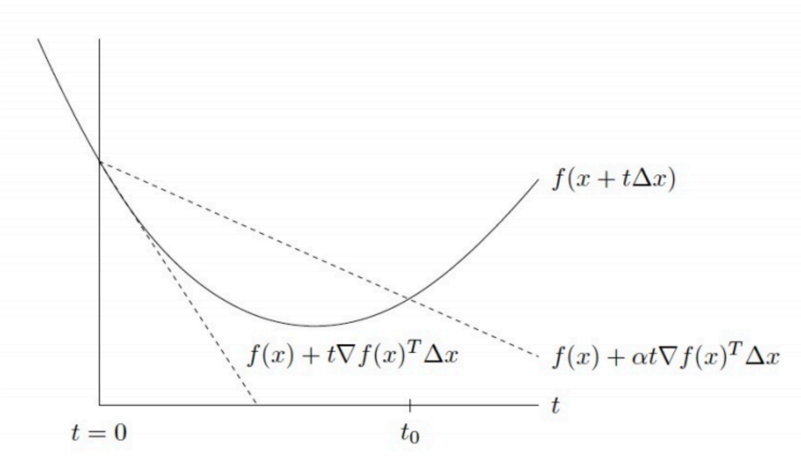
Starting with a maximum candidate step size value $\alpha_0 > 0$, using search control parameters $\tau \in (0,1)$ and $c \in (0,1)$, the backtracking line search algorithm can be expressed as follows:

1. Set $t = -c\,m$ and iteration counter $j = 0$.
2. Until the condition is satisfied that $f(\mathbf{x}) - f(\mathbf{x} + \alpha_j\,\mathbf{p}) \geq \alpha_j\,t$, repeatedly increment $j$ and set $\alpha_j = \tau\,\alpha_{j-1}$.
3. Return $\alpha_j$ as the solution.

In other words, reduce $\alpha_0$ by a factor of $\tau$ in each iteration until the Armijo–Goldstein condition is fulfilled.

Define the local slope of the function of $\alpha$ along the search direction $\mathbf{p}$ as $m = \nabla f(\mathbf{x})^{\mathrm{T}}\,\mathbf{p}$. It is assumed that $\mathbf{p}$ is a unit vector in a direction in which some local decrease is possible, i.e., it is assumed that $m < 0$.

Based on a selected control parameter $c \in (0,1)$, the Armijo–Goldstein condition tests whether a step-wise movement from a current position $\mathbf{x}$ to a modified position $\mathbf{x} + \alpha\,\mathbf{p}$ achieves an adequately corresponding decrease in the objective function. The condition is fulfilled if $f(\mathbf{x} + \alpha\,\mathbf{p}) \leq f(\mathbf{x}) + \alpha\,c\,m$.




### Exact Line Search

At each iteration, do the best we can along the direction of the gradient,

$$t = \underset{s\geq 0}{\operatorname{argmin}} f(x - s\nabla f(x)).$$

Usually, it is not possible to do this minimization exactly.

Approximations to exact line search are often not much more efficient than backtracking, and it's not worth it.

# Proof (first inequality is Lagrange form of Taylor's theorem)

**Theorem 6.1** *Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and that its gradient is Lipschitz continuous with constant $L > 0$, i.e. we have that $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$ for any $x, y$. Then if we run gradient descent for $k$ iterations with a fixed step size $t \leq 1/L$, it will yield a solution $f^{(k)}$ which satisfies*

$$f(x^{(k)}) - f(x^*) \leq \frac{\|x^{(0)} - x^*\|_2^2}{2tk}, \tag{6.1}$$

where $f(x^*)$ is the optimal value. Intuitively, this means that gradient descent is guaranteed to converge and that it converges with rate $O(1/k)$.

**Proof:** Our assumption that $\nabla f$ is Lipschitz continuous with constant $L$ implies that $\nabla^2 f(x) \preceq LI$, or equivalently that $\nabla^2 f(x) - LI$ is a negative semidefinite matrix. Using this fact, we can perform a quadratic expansion of $f$ around $f(x)$ and obtain the following inequality:

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}\nabla^2 f(x)\|y - x\|_2^2 \\ &\leq f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}L\|y - x\|_2^2 \end{aligned}$$

Now let's plug in the gradient descent update by letting $y = x^+ = x - t\nabla f(x)$. We then get:

$$\begin{aligned} f(x^+) &\leq f(x) + \nabla f(x)^T (x^+ - x) + \tfrac{1}{2}L\|x^+ - x\|_2^2 \\ &= f(x) + \nabla f(x)^T (x - t\nabla f(x) - x) + \tfrac{1}{2}L\|x - t\nabla f(x) - x\|_2^2 \\ &= f(x) - \nabla f(x)^T t\nabla f(x) + \tfrac{1}{2}L\|t\nabla f(x)\|_2^2 \\ &= f(x) - t\|\nabla f(x)\|_2^2 + \tfrac{1}{2}Lt^2\|\nabla f(x)\|_2^2 \\ &= f(x) - \left(1 - \tfrac{1}{2}Lt\right) t\|\nabla f(x)\|_2^2 \end{aligned} \tag{6.2}$$

Using $t \leq 1/L$, we know that $-(1 - \frac{1}{2}Lt) = \frac{1}{2}Lt - 1 \leq \frac{1}{2}L(1/L) - 1 = \frac{1}{2} - 1 = -\frac{1}{2}$. Plugging this in to (??) 6.2 we can conclude the following:

$$f(x^+) \leq f(x) - \tfrac{1}{2}t\|\nabla f(x)\|_2^2 \tag{6.3}$$

Since $\frac{1}{2}t\|\nabla f(x)\|_2^2$ will always be positive unless $\nabla f(x) = 0$, this inequality implies that the objective function value strictly decreases with each iteration of gradient descent until it reaches the optimal value $f(x) = f(x^*)$. Note that this convergence result only holds when we choose $t$ to be small enough, i.e. $t \leq 1/L$. This explains why we observe in practice that gradient descent diverges when the step size is too large.

Next, we can bound $f(x^+)$, the objective value at the next iteration, in terms of $f(x^*)$, the optimal objective value. Since $f$ is convex, we can write

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^T (x^* - x) \\ f(x) &\leq f(x^*) + \nabla f(x)^T (x - x^*) \end{aligned}$$

where the first inequality yields the second through simple rearrangement of terms. Plugging this in to (??) 6.3 we obtain:

$$\begin{aligned} f(x^+) &\leq f(x^*) + \nabla f(x)^T (x - x^*) - \frac{t}{2}\|\nabla f(x)\|_2^2 \\ f(x^+) - f(x^*) &\leq \frac{1}{2t}\left(2t\nabla f(x)^T (x - x^*) - t^2\|\nabla f(x)\|_2^2\right) \\ f(x^+) - f(x^*) &\leq \frac{1}{2t}\left(2t\nabla f(x)^T (x - x^*) - t^2\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right) \\ f(x^+) - f(x^*) &\leq \frac{1}{2t}\left(\|x - x^*\|_2^2 - \|x - t\nabla f(x) - x^*\|_2^2\right) \end{aligned} \tag{6.4}$$

where the final inequality is obtained by observing that expanding the square of $\|x - t\nabla f(x) - x^*\|_2^2$ yields $\|x - x^*\|_2^2 - 2t\nabla f(x)^T (x - x^*) + t^2\|\nabla f(x)\|_2^2$. Notice that by definition we have $x^+ = x - t\nabla f(x)$. Plugging this in to (??) 6.4 yields:

$$f(x^+) - f(x^*) \leq \frac{1}{2t}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\right) \tag{6.5}$$

This inequality holds for $x^+$ on every iteration of gradient descent. Summing over iterations, we get:

$$\begin{aligned} \sum_{i=1}^{k} f(x^{(i)}) - f(x^*) &\leq \sum_{i=1}^{k} \frac{1}{2t}\left(\|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2\right) \\ &= \frac{1}{2t}\left(\|x^{(0)} - x^*\|_2^2 - \|x^{(k)} - x^*\|_2^2\right) \\ &\leq \frac{1}{2t}\left(\|x^{(0)} - x^*\|_2^2\right) \end{aligned} \tag{6.6}$$

where the summation on the right-hand side disappears because it is a telescoping sum. Finally, using the fact that $f$ decreasing on every iteration, we can conclude that

$$\begin{aligned} f(x^{(k)}) - f(x^*) &\leq \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) \\ &\leq \frac{\|x^{(0)} - x^*\|_2^2}{2tk} \end{aligned} \tag{6.7}$$

where in the final step, we plug in (??) 6.6 to get the inequality from (??) 6.1 that we were trying to prove. ■

**Theorem 6.2** *Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and that its gradient is Lipschitz continuous with constant $L > 0$, i.e. we have that $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$ for any $x, y$. Then if we run gradient descent for $k$ iterations with step size $t_i$ chosen using backtracking line search on each iteration $i$, it will yield a solution $f^{(k)}$ which satisfies*

$$f(x^{(k)}) - f(x^*) \leq \frac{\|x^{(0)} - x^*\|_2^2}{2t_{\min}k}, \tag{6.8}$$

*where* $t_{\min} = \min\{1, \beta/L\}$

**Convex $f$.** From Theorem 6.1 ??, we know that the convergence rate of gradient descent with convex $f$ is $O(1/k)$, where $k$ is the number of iterations. This implies that in order to achieve a bound of $f(x^{(k)}) - f(x^*) \leq \epsilon$, we must run $O(1/\epsilon)$ iterations of gradient descent. This rate is referred to as "sub-linear convergence."

**Strongly convex $f$.** In contrast, if we assume that $f$ is strongly convex, we can show that gradient descent converges with rate $O(c^k)$ for $0 < c < 1$. This means that a bound of $f(x^{(k)}) - f(x^*) \leq \epsilon$ can be achieved using only $O(\log(1/\epsilon))$ iterations. This rate is typically called "linear convergence."

### 6.1.4 Pros and cons of gradient descent

The principal advantages and disadvantages of gradient descent are:

- Simple algorithm that is easy to implement and each iteration is cheap; just need to compute a gradient
- Can be very fast for smooth objective functions, i.e. well-conditioned and strongly convex
- However, it's often slow because many interesting problems are not strongly convex
- Cannot handle non-differentiable functions (biggest downside)

Subgradients

**Definition 6.3** *A subgradient of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at some point $x$ is any vector $g \in \mathbb{R}^n$ that achieves the same lower bound as the tangent line to $f$ at $x$, i.e. we have*

$$f(y) \geq f(x) + g^T(y - x) \qquad \forall\, x, y$$

The subgradient $g$ always exists for convex functions on the relative interior of their domain. Furthermore, if $f$ is differentiable at $x$, then there is a unique subgradient $g = \nabla f(x)$. Note that subgradients need not exist for nonconvex functions (for example, cubic functions do not have subgradients at their inflection points).

### 6.2.1 Examples of subgradients

**absolute value.** $f(x) = |x|$. Where $f$ is differentiable, the subgradient is identical to the gradient, $\text{sign}(x)$. At the point $x = 0$, the subgradient is any point in the range $[-1, 1]$ because any line passing through $x = 0$ with a slope in this range will lower bound the function.

**$\ell_2$ norm.** $f(x) = \|x\|_2$. For $x \neq 0$, $f$ is differentiable and the unique subgradient is given by $g = x/\|x\|_2$. For $x = 0$, the subgradient is any vector whose $\ell_2$ norm is at most 1. This holds because, by definition, in order for $g$ to be a subgradient of $f$ we must have that

$$f(y) = \|y\|_2 \geq f(x) + g^T(y - x) = g^T y \qquad \forall\ y.$$

In order for $\|y\|_2 \geq g^T y$ to hold, $g$ must have $\|g\|_2 \leq 1$.

**$\ell_1$ norm.** $f(x) = \|x\|_1$. Since $\|x\|_1 = \sum_{i=1}^{n} |x_i|$, we can consider each element $g_i$ of the subgradient separately. The result is very analogous to the subgradient of the absolute value function. For $x_i \neq 0$, $g_i = \text{sign}(g_i)$. For $x_i = 0$, $g_i isanypointin[-1, 1]$.

**maximum of two functions.** $f(x) = \max\{f_1(x), f_2(x)\}$, where $f_1$ and $f_2$ are convex and differentiable. Here we must consider three cases. First, if $f_1(x) > f_2(x)$, then $f(x) = f_1(x)$ and therefore there is a unique subgradient $g = \nabla f_1(x)$. Likewise, if $f_2(x) > f_1(x)$, then $f(x) = f_2(x)$ and $g = \nabla f_2(x)$. Finally, if $f_1(x) = f_2(x)$, then $f$ may not be differentiable at $x$ and the subgradient will be any point on the line segment that joints $\nabla f_1(x)$ and $\nabla f_2(x)$.

### 6.2.2 Subdifferential

**Definition 6.4** *The subdifferential of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ at some point $x$ is the set of all subgradients of $f$ at $x$, i.e. we say*

$$\partial f(x) = \{g \in \mathbb{R}^n : g \textit{ is a subgradient of } f \textit{ at } x\}$$

An important property of the subdifferential $\partial f(x)$ is that it is a closed and convex set, which holds even for nonconvex $f$. To verify this, suppose we have two subgradients $g_1, g_2 \in \partial f(x)$. We need to show that $g_0 = \alpha g_1 + (1 - \alpha) g_2$ is also in $\partial f(x)$ for arbitrary $\alpha$. If we write the following inequalities,

$$\alpha\Big(f(y) \geq f(x) + g_1^T(y - x)\Big)\alpha$$
$$(1-\alpha)\Big(f(y) \geq f(x) + g_2^T(y - x)\Big)(1 - \alpha),$$

which follow from the definition of subgradient applied to $g_1$ and $g_2$, we can add them together to yield $f(y) \geq f(x) + \alpha\, g_1^T(y - x) + (1 - \alpha)\, g_2^T(y - x) = g_0^T(y - x)$.

### 7.2.1 Subgradient method

For convex $f$, not necessarily differentiable, subgradient method finds the lowest value of the criterion by:

$$x^{(k)} = x^{(k-1)} - t_k g^{(k-1)}, \quad , k = 1, 2, 3, \cdots$$

where $g^{(k-1)}$ is any subgradient of $f$ at $x^{(k-1)}$. Note that it is not a decent method, that the next iterative doesn't always find the lower criterion. So we need to keep the best lowest criterion value at every iteration, i.e., $f(x^{(k)}_{\text{best}}) = \min_i f(x^{(i)})$.

### 7.2.2 Choosing the step size

i) Fixed step size: $t_k = t\ \ \forall k$.
However, for subgradient method, we do not typically chose fixed step size.

ii) Diminishing step size (Standard): choose $t_k$ that is square summable but not summable.

$$\sum_{k=1}^{\infty} t_k^2 < \infty, \ \sum_{k=1}^{\infty} = \infty.$$

Note that step sizes are all pre-defined, not adaptively computed during the optimization iteration.

### 7.2.3 Convergence analysis

i) Fixed step size: Suboptimal Convergence.
For convex, not differentiable function $f$, if the function itself is Lipschitz with constant $G$ such as,

$$|f(x) - f(y)| \le G\|x - y\|_2 \;\; \forall x, y$$

subgradient method using fixed step size t would give a point that is suboptimal such as,

$$\lim_{k\to\infty} f(x_{\text{best}}^{(k)}) \le f(x^*) + G^2 \frac{t}{2}.$$

In other words, the smaller the step size, the smaller the difference would be between the optimal and sub-optimal convergence.

ii) Diminishing step size that is square summable: Optimal Convergence.

$$\lim_{k\to\infty} f(x_{\text{best}}^{(k)}) = f(x^*).$$

Note that subgradient method is applicable to functions that may not look like Lipschitz, since the over the bounded set the function can be Lipschitz.

## Projection Method

Projected subgradient method can be used to minimize a convex function over a convex set C:

$$\min_{x\in C} f(x)$$

It is same as usual subgradient update except we project the solution back on to C every time so that at every iteration we move in the direction of the subgradient but still lies in the set C.

$$x^{(k)} = P_C(x^{(k-1)} - t_k g^{(k-1)}), \quad k = 1, 2, 3, \cdots$$

Alternative method:

$$\min_{x\in C} f(x) = \min_{x\in\mathbb{R}^n} f(x) + I_C(x)$$

**Examples for projection onto solution set C:**
i) $C = \{y : \;\; y_i \ge \;\; \forall i\} \Rightarrow [P_C(x)]_i = \max\{x_i, 0\}$.

### 7.2.7 Basic Pursuit Problem

We can use projected subgradient method to solve the basic pursuit problem:

$$\min_{\beta\in\mathbb{R}^p} \|\beta\|_1 \;\text{ s.t. }\; X\beta = y.$$

In this case, the solution set is $C = \{\beta : X\beta = y\}$.
The projection on to solution set C is $P_C(\beta) = \beta + X^T(XX^T)^{-1}(y - X\beta)$ as shown in example 2 above.
Projected subgradient method performs step

$$\begin{aligned}
\beta^{(k)} &= P_C(\beta^{(k-1)} - t_k g^{(k-1)}) \\
&= \beta^{(k-1)} - t_k g^{(k-1)} + X^(XX^T)^{-1}(y - X\beta^{(k-1)} + X t_k g^{(k-1)}) \\
&= \beta^{(k-1)} - (I - X^T(XX^T)^{-1}X) t_k g^{(k-1)}
\end{aligned}$$

Where, $g^{(k-1)} \in \partial\|\beta^{(k-1)}\|_1$.