Course Notes for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt

Email: hardt+ee227c@berkeley.edu

Graduate Instructor: Max Simchowitz

Email: msimchow+ee227c@berkeley.edu

January 27, 2018

Abstract

These are course notes for EE227C (Spring 2018): Convex Optimization and Approximation, taught at UC Berkeley. For further information, see the course page at:

https://ee227c.github.io/.

List of contributors

Acknowledgments

These notes closely follow an earlier course by Ben Recht.

Contents

1 Lecture 1: Convexity

This lecture provides the most important facts about convex sets and convex functions that we'll heavily make use of. These are often simple consequences of Taylor's theorem.

1.1 Convex sets

Definition 1.1 (Convex set)**.** A set $K \subseteq \mathbb{R}^n$ is *convex* if it the line segment between any two points in K is also contained in K. Formally, for all $x, y \in K$ and all scalars $\gamma \in [0, 1]$ we have $\gamma x + (1-\gamma)y \in K$.

Theorem 1.2 (Separation Theorem)**.** *Let $C, K \subseteq \mathbb{R}^n$ be convex sets with empty intersection $C \cap K = \emptyset$. Then there exists a point $a \in \mathbb{R}^n$ and a number $b \in \mathbb{R}$ such that*

1. *for all $x \in C$, we have $\langle a, x\rangle \geqslant b$.*
2. *for all $x \in K$, we have $\langle a, x\rangle \leqslant b$.*

If C and K are closed and at least one of them is bounded, then we can replace the inequalities by strict inequalities.

The case we're most concerned with is when both sets are compact (i.e., closed and bounded). We highlight its proof here.

Proof of Theorem 1.2 for compact sets. In this case, the Cartesian product $C \times K$ is also compact. Therefore, the distance function $\|x - y\|$ attains its minimum over $C \times K$. Taking p, q to be two points that achieve the minimum. A separating hyperplane is given by the hyperplane perpendicular to $q - p$ that passes through the midpoint between p and q. That is, $a = q - p$ and $b = (\langle a, q\rangle - \langle a, p\rangle)/2$. For the sake of contradiction, suppose there is a point r on this hyperplane contained in one of the two sets, say, C. Then the line segment from p to r is also contained in C by convexity. We can then find a point along the line segment that is close to q than p is, thus contradicting our assumption. ■

1.1.1 Notable convex sets

- Linear spaces $\{x \in \mathbb{R}^n \mid Ax = 0\}$ and halfspaces $\{x \in \mathbb{R}^n \mid \langle a, x\rangle \geqslant 0\}$
- Affine transformations of convex sets. If $K \subseteq \mathbb{R}^n$ is convex, so is $\{Ax + b \mid x \in K\}$ for any $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. In particular, affine subspaces and affine halfspaces are convex.
- Intersections of convex sets. In fact, every convex set is equivalent to the intersection of all affine halfspaces that contain it (a consequence of the separating hyperplane theorem).
- The cone of positive semidefinite matrices, denotes, $S^n_+ = \{A \in \mathbb{R}^{n\times n} \mid A \succeq 0\}$. Here we write $A \succeq 0$ to indicate that $x^\top Ax \geqslant 0$ for all $x \in \mathbb{R}^n$. The fact that S^n_+ is convex can be verified directly from the definition, but it also follows from what we already knew. Indeed, denoting by $S_n = \{A \in \mathbb{R}^{n\times n} \mid A^\top = A\}$ the set of all $n \times n$ symmetric matrices, we can write S^n_+ as an (infinite) intersection of halfspaces $S^n_+ = \bigcap_{x\in\mathbb{R}^n\setminus\{0\}}\{A \in S_n \mid x^\top Ax \geqslant 0\}$.
- See Boyd-Vandenberghe for lots of other examples.

1.2 Convex functions

Definition 1.3 (Convex function)**.** A function $f\colon \Omega \to \mathbb{R}$ is *convex* if for all $x, y \in \Omega$ and all scalars $\gamma \in [0, 1]$ we have $f(\gamma x + (1-\gamma)y) \leqslant \gamma f(x) + (1-\gamma)f(y)$.

Jensen (1905) showed that for continuous functions, convexity follows from the "midpoint" condition that for all $x, y \in \Omega$,

$$f\left(\frac{x+y}{2}\right) \leqslant \frac{f(x)+f(y)}{2}\,.$$

This result sometimes simplifies the proof that a function is convex in cases where we already know that it's continuous.

Definition 1.4. The *epigraph* of a function $f\colon \Omega \to \mathbb{R}$ is defined as

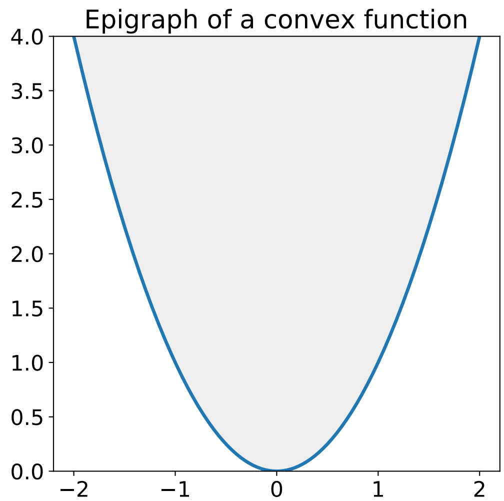
$$\mathrm{epi}(f) = \{(x,t) \mid f(x) \leqslant t\}\,.$$

Fact 1.5. *A function is convex if and only if its epigraph is convex.*

Convex functions enjoy the property that local minima are also global minima. Indeed, suppose that $x \in \Omega$ is a local minimum of $f\colon \Omega \to \mathbb{R}$ meaning that any point in a neighborhood around x has larger function value. Now, for every $y \in \Omega$, we can find a small enough γ such that

$$f(x) \leqslant f((1-\gamma)x + \gamma y) \leqslant (1-\gamma)f(x) + \gamma f(y)\,.$$

Therefore, $f(x) \leqslant f(y)$ and so x must be a global minimum.



1.2.1 First-order characterization

It is helpful to relate convexity to Taylor's theorem, which we recall now. We define the *gradient* of a differentiable function $f\colon \Omega \to \mathbb{R}$ at $x \in \Omega$ as the vector of partial derivatives

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_i}\right)_{i=1}^{n}.$$

We note the following simple fact that relates linear forms of the gradient to a one-dimensional derivative evaluated at 0. It's a consequence of the multivariate chain rule.

Fact 1.6. *Assume $f\colon \Omega \to \mathbb{R}$ is differentiable and let $x, y \in \Omega$. Then,*

$$\nabla f(x)^\top y = \left.\frac{\partial f(x+\gamma y)}{\partial \gamma}\right|_{\gamma=0}.$$

Taylor's theorem implies the following statement.

Proposition 1.7. *Assume $f\colon \Omega \to \mathbb{R}$ is continuously differentiable along the line segment between two points x and y. Then,*

$$f(y) = f(x) + \nabla f(x)^\top (y-x) + \int_0^1 (1-\gamma)\frac{\partial^2 f(x+\gamma(y-x))}{\partial \gamma^2}\,\mathrm{d}\gamma$$

Proof. Apply a second order Taylor's expansion to $g(\gamma) = f(x+\gamma(y-x))$ and apply Fact 1.6 to the first-order term. ■

Among differentiable functions, convexity is equivalent to the property that the first-order Taylor approximation provides a global lower bound on the function.

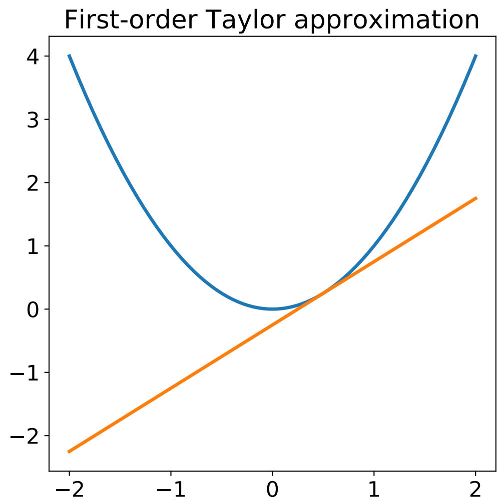


Figure 1: Taylor approximation of $f(x) = x^2$ at 0.5.

Proposition 1.8. *Assume $f\colon \Omega \to \mathbb{R}$ is differentiable. Then, f is convex if and only if for all $x, y \in \Omega$ we have*

$$f(y) \geqslant f(x) + \nabla f(x)^\top (y - x)\,. \tag{1}$$

Proof. First, suppose f is convex, then by definition

$$\begin{aligned} f(y) &\geqslant \frac{f((1-\gamma)x + \gamma y) - (1-\gamma) f(x)}{\gamma} \\ &\geqslant f(x) + \frac{f(x + \gamma(y - x)) - f(x)}{\gamma} \\ &\to f(x) + \nabla f(x)^\top (y - x) \quad \text{as } \gamma \to 0 && \text{(by Fact 1.6.)} \end{aligned}$$

On the other hand, fix two points $x, y \in \Omega$ and $\gamma \in [0, 1]$. Putting $z = \gamma x + (1 - \gamma) y$ we get from applying Equation 1 twice,

$$f(x) \geqslant f(z) + \nabla f(z)^\top (x - z) \quad \text{and} \quad f(y) \geqslant f(z) + \nabla f(z)^\top (y - z)$$

Adding these inequalities scaled by γ and $(1 - \gamma)$, respectively, we get $\gamma f(x) + (1 - \gamma) f(y) \geqslant f(z)$, which establishes convexity. ■

A direct consequence of Proposition 1.8 is that if $\nabla f(x) = 0$ vanishes at a point x, then x must be a global minimizer of f.

Remark 1.9 (Subgradients)**.** *Of course, not all convex functions are differentiable. The absolute value $f(x) = |x|$, for example, is convex but not differentiable at 0. Nonetheless, for every x, we can find a vector g such that*

$$f(y) \geqslant f(x) + g^\top (y - x)\,.$$

Such a vector is called a subgradient *of f at x. The existence of subgradients is often sufficient for optimization.*

1.2.2 Second-order characterization

We define the *Hessian* matrix of $f\colon \Omega \to \mathbb{R}$ at a point $x \in \Omega$ as the matrix of second order partial derivatives:

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j\in[n]}.$$

Schwarz's theorem implies that the Hessian at a point x is symmetric provided that f has continuous second partial derivatives in an open set around x.

In analogy with Fact 1.6, we can relate quadratic forms in the Hessian matrix to one-dimensional derivatives using the chain rule.

Fact 1.10. *Assume that $f\colon \Omega \to \mathbb{R}$ is twice differentiable along the line segment from x to y. Then,*

$$y^\top \nabla^2 f(x+\gamma y) y = \frac{\partial^2 f(x+\gamma y)}{\partial \gamma^2}.$$

Proposition 1.11. *If f is twice continuously differentiable on its domain Ω, then f is convex if and only if $\nabla^2 f(x) \succeq 0$ for all $x \in \Omega$.*

Proof. Suppose f is convex and our goal is to show that the Hessian is positive semidefinite. Let $y = x + \alpha u$ for some arbitrary vector u and scalar α. Proposition 1.8 shows

$$f(y) - f(x) - \nabla f(x)^\top (y-x) \geqslant 0$$

Hence, by Proposition 1.7,

$$\begin{aligned}
0 &\leqslant \int_0^1 (1-\gamma)\frac{\partial^2 f(x+\gamma(y-x))}{\partial \gamma^2}\,\mathrm{d}\gamma \\
&= (1-\gamma)\frac{\partial^2 f(x+\gamma(y-x))}{\partial \gamma^2} \quad \text{for some } \gamma \in (0,1) \quad \text{(by the mean value theorem)} \\
&= (1-\gamma)(y-x)^\top \nabla^2 f(x+\gamma(y-x))(y-x)\,. \qquad \text{(by Fact 1.10)}
\end{aligned}$$

Plugging in our choice of y, this shows $0 \leqslant u^\top \nabla^2 f(x+\alpha\gamma u)u$. Letting α tend to zero establishes that $\nabla^2 f(x) \succeq 0$. (Note that γ generally depends on α but is always bounded by 1.)

Now, suppose the Hessian is positive semidefinite everywhere in Ω and our goal is to show that the function f is convex. Using the same derivation as above, we can see that the second-order error term in Taylor's theorem must be nonnegative. Hence, the first-oder approximation is a global lower bound and so the function f is convex by Proposition 1.8. ∎

1.3 Convex optimization

Much of this course will be about different ways of minimizing a convex function $f\colon \Omega \to \mathbb{R}$ over a convex domain Ω :

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

Convex optimization is not necessarily easy! For starters, convex sets do not necessarily enjoy compact descriptions. When solving computational problems involving convex sets, we need to worry about how to represent the convex set we're dealing with. Rather than asking for an explicit description of the set, we can instead require a computational abstraction that highlights essential operations that we can carry out. The Separation Theorem motivates an important computational abstraction called *separation oracle*.

Definition 1.12. A *separation oracle* for a convex set K is a device, which given any point $x \notin K$ returns a hyperplane separating x from K.

Another computational abstraction is a *first-order oracle* that given a point $x \in \Omega$ returns the gradient $\nabla f(x)$. Similarly, a *second-order oracle* returns $\nabla^2 f(x)$. A function value oracle or *zeroth-order oracle* only returns $f(x)$. First-order methods are algorithms that make do with a first-order oracle.

1.3.1 What is efficient?

Classical complexity theory typically quantifies the resource consumption (primarily running time or memory) of an algorithm in terms of the bit complexity of the input. This approach can be cumbersome in convex optimization and most textbooks shy away from it. Instead, it's customary in optimization to quantify the cost of the algorithm in terms of how often it accesses one of the oracles we mentioned.

The definition of "efficient" is not completely cut and dry in optimization. Typically, our goal is to show that an algorithm finds a solution x with $f(x) = \min_{x\in\Omega} f(x) + \epsilon$ for some additive error $\epsilon > 0$. The cost of the algorithm will depend on the target error. Highly practical algorithms often have a polynomial dependence on ϵ, such as $O(1/\epsilon)$ or even $O(1/\epsilon^2)$. Other algorithms achieve $O(\log(1/\epsilon))$ steps in theory, but are prohibitive in their actual computational cost. Technically, if we think of the parameter ϵ as being part of the input, it takes only $O(\log(1/\epsilon))$ bits to describe the error parameter. Therefore, an algorithm that depends more than logarithmically on $1/\epsilon$ may not be polynomial time algorithm in its input size.

In this course, we will make an attempt to highlight both the theoretical performance and practical appeal of an algorithm. Moreover, we will discuss other performance criteria such as robustness to noise. How well an algorithm performs is rarely decided by a single criterion, and usually depends on the application at hand.

2 Lecture 2: Gradient Method

In this lecture we encounter the fundamentally important *gradient method* and a few ways to analyze its convergence behavior. The goal here is to solve a problem of the form

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

where we'll make some additional assumptions on the function $f\colon \Omega \to \mathbb{R}$.

2.1 Gradient Descent

For a differentiable function f, the basic gradient method startingf from an initial point x_0 is defined by the iterative description

$$x_{t+1} = x_t - \eta \nabla f(x_t) \qquad\qquad (t \geqslant 1)$$

where η is the step size.

The first assumption that leads to a convergence analysis is that the gradients of the function aren't too big over the domain.

Definition 2.1 (L-Lipschitz)**.** A differentiable function f is said to be *L-Lipschitz* over the domain Ω if for all $x \in \Omega$, we have

$$\|\nabla f(x)\| \leqslant L\,.$$

Fact 2.2. *If a function f is L-Lipschitz, it implies that the difference between two points in the range is bounded,*

$$|f(x) - f(y)| \leqslant L\|x - y\|$$

How can we ensure that $x_{t+1} \in \Omega$? One natural approach is to "project" each iterate back onto the domain Ω.

Definition 2.3 (Projection)**.** The *projection* of a point x onto a set Ω is defined as

$$\Pi_\Omega(x) = \operatorname*{argmin}_{y \in \Omega} \|x - y\|$$

Example 2.4. A projection onto the Euclidean ball B_2 is just normalization:

$$\Pi_{B_2}(x) = \frac{x}{\|x\|}$$

A crucial property of projections is that when $x \in \Omega$, we have for any y (possibly outside Ω):

$$\|\Pi_\Omega(y) - x\|^2 \leqslant \|y - x\|^2$$

That is, the projection of y onto a convex set containing x is closer to x. See Figure 2 for a geometric picture.

Lemma 2.5.

$$\|\Pi_\Omega(y) - x\|^2 \leqslant \|y - x\|^2 - \|y - \Pi_\Omega(y)\|^2$$

Which follows from the Pythagorean theorem. Note that this lemma implies the above property.

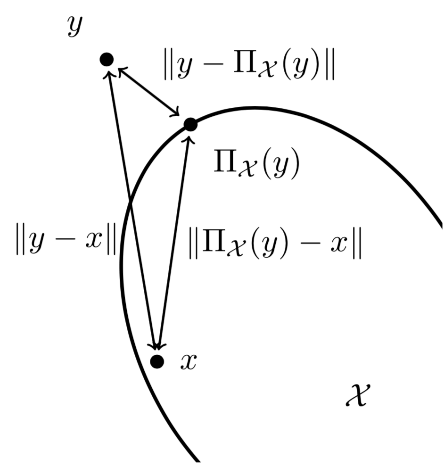


Figure 2: Projection of y onto set $\mathcal{X}$. (Figure taken from *S. Bubeck. Convex Optimization: Algorithms and Complexity)*

2.1.1 Modifying Gradient Descent with Projections

So now we can modify our original procedure to use two steps.

$$y_{t+1} = x_t - \eta \nabla f(x_t)$$

$$x_{t+1} = \Pi_\Omega(y_{t+1})$$

And we are guaranteed that $x_{t+1} \in \Omega$. Note that computing the projection may be the hardest part of your problem, as you are computing an argmin. However, there are convex sets for which we know explicitly how to compute the projection (see Example 2.4).

2.2 Convergence rate of gradient descent for Lipschitz functions

Theorem 2.6 (Projected Gradient Descent for L-Lipschitz Functions)**.** *Assume that function f is convex, differentiable, and closed with bounded gradients. Let L be the Lipchitz constant of f over the convex domain Ω. Let R be the upper bound on the distance $\|x_1 - x^*\|_2$ from the initial point x_1 to the optimal point $x^* = \arg\min_{x \in \Omega} f(x)$. Let t be the number of iterations of project gradient descent. If the learning rate η is set to $\eta = \frac{R}{L\sqrt{(t)}}$, then*

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f\left(x^*\right) \leqslant \frac{RL}{\sqrt{t}}.$$

This means that the difference between the functional value of the average point during the optimization process from the optimal value is bounded above by a constant proportional to $\frac{1}{\sqrt{t}}$.

Before proving the theorem, recall the "Fundamental Theorem of Optimization", which is that an inner product can be written as a sum of norms: $u^\top v = \frac{1}{2}(\|u\|^2 + \|v\|^2 - \|u - v\|^2$. This property can be seen by writing $\|u - v\|$ as $\|u - v\| = \|u\|^2 + \|v\|^2 - 2u^\top v$.

Proof of Theorem 2.6 for compact sets. The proof begins by first bounding the difference in function values $f(x_s) - f(x^*)$.

$$
\begin{aligned}
f(x_s) - f(x^*) &\leqslant \nabla f(x_s)^\top (x_s - s^*) && (2)\\
&= \frac{1}{\eta}(x_s - y_{s+1})^\top (x_s - x^*) && (3)\\
&= \frac{1}{2\eta}\left(\|x_s - x^*\|^2 + \|x_s - y_{s+1}\|^2 - \|y_{s+1} - x^*\|^2\right) && (4)\\
&= \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(x_s)\|^2 && (5)\\
&\leqslant \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right) + \frac{\eta L^2}{2} && (6)\\
&\leqslant \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\right) + \frac{\eta L^2}{2} && (7)
\end{aligned}
$$

Equation 2 comes from the definition of convexity. Equation 3 comes from the update rule for projected gradient descent. Equation 4 comes from the "Fundamental Theorem of Optimization." Equation 5 comes from the update rule for projected gradient descent. Equation 6 is because f is L-Lipchitz. Equation 7 comes from Lemma 2.5.

Now, sum these differences from $s = 1$ to $s = t$:

$$
\begin{aligned}
\sum_{s=1}^{t} f(x_s) - f(x^*) &\leqslant \frac{1}{2\eta}\sum_{s=1}^{t}\left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\right) + \frac{\eta L^2 t}{2} && (8)\\
&= \frac{1}{2\eta}\left(\|x_1 - x^*\|^2 - \|x_t - x^*\|^2\right) + \frac{\eta L^2 t}{2} && (9)\\
&\leqslant \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta L^2 t}{2} && (10)\\
&\leqslant \frac{R^2}{2\eta} + \frac{\eta L^2 t}{2} && (11)
\end{aligned}
$$

Equation 9 is because Equation 8 is a telescoping sum. Equation 10 is because $\|x_t - x^*\|^2 \geqslant 0$. Equation 11 is by the assumption that $\|x_1 - x^*\|^2 \leqslant R^2$.

Then bound $f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f\left(x^*\right)$ by the above sum:

$$
\begin{aligned}
f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) &\leqslant \frac{1}{t}\sum_{s=1}^{t} f(x_s) && (12)\\
\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f\left(x^*\right) &\leqslant \frac{1}{t}\sum_{s=1}^{t} f(x_s) - f\left(x^*\right) && (13)\\
&\leqslant \frac{R^2}{2\eta} + \frac{\eta L^2 t}{2} && (14)
\end{aligned}
$$

Equation 12 is by convexity. $\frac{R^2}{2\eta} + \frac{\eta L^2 t}{2}$, the upper bound of the difference between $f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right)$ and $f(x^*)$ is minimized when η is set to be $\frac{RL}{\sqrt{t}}$.

■

2.3 Convergence rate for smooth functions

The next property we'll encounter is called *smoothness* and it often leads to stronger convergence guarantees.

Definition 2.7 (β-smoothness)**.** A continuously differentiable function f is β smooth if the gradient ∇f is β-Lipschitz, i.e

$$\|\nabla f(x) - \nabla f(y)\| \leqslant \beta \|x - y\|$$

Lemma 2.8. *Let f be a β-smooth function on $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$, one has*

$$|f(x) - f(y) - \nabla f(y)^\top (x - y)| \leqslant \frac{\beta}{2}\|x - y\|^2$$

Proof. First represent $f(x) - f(y)$ as an integral, apply Cauchy-Schwarz and then β-smoothness:

$$\begin{aligned}
&|f(x) - f(y) - \nabla f(y)^\top (x - y)| \\
&\qquad = \left| \int_0^1 \nabla f(y + t(x - y))^\top (x - y) dt - \nabla f(y)^\top (x - y) \right| \\
&\qquad \leqslant \int_0^1 \|\nabla f(y + t(x - y)) - \nabla f(y)\| \cdot \|x - y\| dt \\
&\qquad \leqslant \int_0^1 \beta t \|x - y\|^2 dt \\
&\qquad = \frac{\beta}{2}\|x - y\|^2
\end{aligned}$$

■

We also need the following lemma.

Lemma 2.9. *Let f be a β-smooth function, then for any $x, y \in \mathbb{R}^n$, one has*

$$f(x) - f(y) \leqslant \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|$$

Proof. Let $z = y - \frac{1}{\beta}(\nabla f(y) - \nabla f(x))$. Then one has

$$f(x) - f(y)$$

$$\begin{aligned}
&= f(x) - f(z) + f(z) - f(y) && (15)\\
&\leqslant \nabla f(x)^\top (x - z) + \nabla f(y)^\top (z - y) + \frac{\beta}{2}\|z - y\|^2 && (16)\\
&= \nabla f(x)^\top (x - y) + (\nabla f(x) - \nabla f(y))^\top (y - z) + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 && (17)\\
&= \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 && (18)
\end{aligned}$$

■

We will show that gradient descent with the update rule

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

attains a faster rate of convergence under the smoothness condition.

Theorem 2.10. *Let f be convex and β-smooth on $\mathbb{R}^n$ then gradient descent with $\eta = \frac{1}{\beta}$ satsifies*

$$f(x_t) - f(x^*) \leqslant \frac{2\beta\|x_1 - x^*\|^2}{t - 1}$$

To prove this we will need the following two lemmas.

Proof. By the update rule and lemma 2.8 we have

$$f(x_{s+1}) - f(x_s) \leqslant -\frac{1}{2\beta}\|\nabla f(x_s)\|^2$$

In particular, denoting $\delta_s = f(x_s) - f(x^*)$ this shows

$$\delta_{s+1} \leqslant \delta_s - \frac{1}{2\beta}\|\nabla f(x_s)\|^2$$

One also has by convexity

$$\delta_s \leqslant \nabla f(x)s)^\top (x_s - x^*) \leqslant \|x_s - x^*\| \cdot \|\nabla f(x_s)\|$$

We will prove that $\|x_s - x^*\|$ is decreasing with s, which with the two above displays will imply

$$\delta_{s+1} \leqslant \delta_s - \frac{1}{2\beta\|x_1 - x^*\|^2}\delta_s^2$$

We solve the recurrence as follows. Let $w = \frac{1}{2\beta\|x_1 - x^*\|^2}$, then

$$w\delta_s^2 + \delta_{s+1} \leqslant \delta_s \iff w\frac{\delta_s}{\delta_{s+1}} + \frac{1}{\delta_s} \leqslant \frac{1}{\delta_{s+1}} \implies \frac{1}{\delta_{s+1}} - \frac{1}{\delta_s} \geqslant w \implies \frac{1}{\delta_t} \geqslant w(t-1)$$

To finish the proof it remains to show that $\|x_s - x^*\|$ is decreasing with s. Using lemma 2.9 one immediately gets

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geqslant \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2$$

We use this and the fact that $\nabla f(x^*) = 0$

$$\|x_{s+1} - x^*\|^2 = \|x_s - \frac{1}{\beta}\nabla f(x_s) - x^*\|^2 \tag{19}$$

$$= \|x_s - x^*\|^2 - \frac{2}{\beta}\nabla f(x_s)^\top (x_s - x^*) + \frac{1}{\beta^2}\|\nabla f(x_s)\|^2 \tag{20}$$

$$\leqslant \|x_s - x^*\|^2 - \frac{1}{\beta^2}\|\nabla f(x_s)\|^2 \tag{21}$$

$$\leqslant \|x_s - x^*\|^2 \tag{22}$$

which concludes the proof. ■