Optimization for Data Science

Cristian Perez Jensen

March 11, 2024

Contents

1 Risk minimization 1
1.1 Algorithms in data science 1
1.2 Empirical and expected risk 1
1.3 The map of learning 3
2 Theory of convex functions 4
2.1 Mathematical background 4
2.2 Convex sets 4
2.3 Convex functions 5
2.4 Minimizing convex functions 9
2.5 Convex programming 10
3 Gradient descent 14
3.1 Vanilla analysis 14
3.2 Lipschitz convex functions 15
3.3 Smooth functions 16
3.4 Strongly convex functions 18

List of symbols

$\doteq$	Equality by definition
$\mathbb{R}$	Set of real numbers
$f : A \to B$	Function f that maps elements of set A to elements of set B
$\boldsymbol{v} \in \mathbb{R}^n$	n-dimensional vector
$\boldsymbol{M} \in \mathbb{R}^{m \times n}$	$m \times n$ matrix
$\boldsymbol{M}^\top$	Transpose of matrix $\boldsymbol{M}$
$\boldsymbol{M}^{-1}$	Inverse of matrix $\boldsymbol{M}$
$\det(\boldsymbol{M})$	Determinant of $\boldsymbol{M}$
$\frac{\mathrm{d}}{\mathrm{d}x} f(x)$	Ordinary derivative of $f(x)$ w.r.t. x at point $x \in \mathbb{R}$
$\frac{\partial}{\partial x} f(\boldsymbol{x})$	Partial derivative of $f(\boldsymbol{x})$ w.r.t. x at point $\boldsymbol{x} \in \mathbb{R}^n$
$\boldsymbol{\nabla} f(\boldsymbol{x}) \in \mathbb{R}^n$	Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at point $\boldsymbol{x} \in \mathbb{R}^n$
$\boldsymbol{D} f(\boldsymbol{x}) \in \mathbb{R}^{n \times m}$	Jacobian of $f : \mathbb{R}^n \to \mathbb{R}^m$ at point $\boldsymbol{x} \in \mathbb{R}^n$
$\boldsymbol{\nabla}^2 f(\boldsymbol{x}) \in \mathbb{R}^{n \times n}$	Hessian of $f : \mathbb{R}^n \to \mathbb{R}$ at point $\boldsymbol{x} \in \mathbb{R}^n$

1 Risk minimization

1.1 Algorithms in data science

In classical algorithm theory, an optimization problem solves a well-defined problem. For example, Kruskal's algorithm computes the minimum spanning tree of a graph. In data science, it is not as well-defined. The starting point is a learning problem, and the optimization typically happens on training data. However, even a perfect result may fail to solve the learning problem, which is a failure of the model in which the optimization algorithm was applied, rather than the optimization algorithm itself.

1.2 Empirical and expected risk

Typically, we have a data source $\mathcal{X}$, equipped with an unknown probability distribution. However, we do have access to independent samples $X_1, \dots, X_n \sim \mathcal{X}$. The goal is to "explain" $\mathcal{X}$ through these samples. More specifically, we have a class $\mathcal{H}$ of hypotheses (possible explanations of $\mathcal{X}$). The goal is then to select the hypothesis $H \in \mathcal{H}$ that best explains $\mathcal{X}$, which we measure by a *risk* (or loss) function $\ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R}$.

The expected risk can be computed by

$$\ell(H) \doteq \mathbb{E}_{\mathcal{X}}[\ell(H, X)].$$

The best explanation is the *expected risk minimizer*,

$$H^\star = \operatorname*{argmin}_{H \in \mathcal{H}} \ell(H).$$

However, since we do not have access to the distribution over $\mathcal{X}$, we cannot compute $\ell(H)$ or $H^\star$. Thus, we need to compute the *probably approximately correct* (PAC) hypothesis. This means that given $\delta, \epsilon > 0$, we want to produce, with probability $1 - \delta$, a hypothesis $\tilde{H} \in \mathcal{H}$ such that

$$\ell(\tilde{H}) \leq \inf_{H \in \mathcal{H}} \ell(H) + \epsilon,$$

meaning that with high probability, we approximately solve the expected risk minimization problem.

However, we can still not compute this, thus we must use our training data to compute the *empirical risk*,

$$\ell_n(H) = \frac{1}{n} \sum_{i=1}^{n} \ell(H, X_i).$$

This is a random variable, because it depends on the training data, which are all random variables, distributed according to the probability distribution of $\mathcal{X}$.

Lemma 1 (Weak law of large numbers)**.** Let $H \in \mathcal{H}$ be a hypothesis. For any $\delta, \epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$,

$$|\ell_n(H) - \ell(H)| \leq \epsilon,$$

with probability at least $1 - \delta$.

Given $n \in \mathbb{N}$ and training data $X_1, \ldots, X_n$ from $\mathcal{X}$, we want to produce a hypothesis $\tilde{H}_n$ such that

$$\ell_n(\tilde{H}_n) \leq \inf_{H \in \mathcal{H}} \ell_n(H) + \epsilon.$$

In an ideal world, $\tilde{H}_n$ is also almost the best explanation for the data source $\mathcal{X}$,

$$\ell(\tilde{H}_n) \leq \inf_{H \in \mathcal{H}} \ell(H) + \epsilon.$$

Remark. Note that the weak law of large numbers can only be applied to a *fixed* hypothesis, but not to the data-dependent hypothesis $\tilde{H}_n$. Thus, we are not always in an ideal world scenario.

A sufficient condition for an ideal world scenario is that the weak law of large numbers uniformly holds for all hypotheses with high probability. This leads us to the following theorem.

Theorem 2. Assume that for any $\delta, \epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$,

$$\sup_{H \in \mathcal{H}} |\ell_n(H) - \ell(H)| \leq \epsilon,$$

with probability at least $1 - \delta$. Then, for $n \geq n_0$, an approximate empirical risk minimizer $\tilde{H}_n$ is PAC for expected risk minimization, meaning that it satisfies

$$\ell(\tilde{H}_n) \leq \inf_{H \in \mathcal{H}} \ell(H) + 3\epsilon,$$

with probability at least $1 - \delta$.

Proof.

$$\begin{aligned}
\ell(\tilde{H}_n) &\leq \ell_n(\tilde{H}_n) + \epsilon && \text{Follows from } \sup_{H \in \mathcal{H}} |\ell_n(H) - \ell(H)| \leq \epsilon. \\
&\leq \inf_{H \in \mathcal{H}} \ell_n(H) + 2\epsilon && \tilde{H}_n \text{ is an almost best explanation of the training data.} \\
&\leq \inf_{H \in \mathcal{H}} \ell(H) + 3\epsilon, && \text{Follows from } \sup_{H \in \mathcal{H}} |\ell_n(H) - \ell(H)| \leq \epsilon.
\end{aligned}$$

with probability at least $1 - \delta$. ■

It turns out that the assumption made by Theorem 2 holds in many relevant cases, but it is not always true.

In this course, we will not learn how to pick the theory ($\mathcal{H}$ and ℓ), but rather how to solve the optimization problems that arise in empirical risk minimization after the theory has been chosen.

1.3 *The map of learning*

The map of learning can be seen in Figure 1. It plots the empirical risk ($\ell_n(H_n)$, training data) against the expected risk ($\ell(H_n)$, estimated by a validation set). We can only be in the area denoted by "empirical risk minimization", because

$$\begin{aligned}\ell_n(\tilde{H}_n) &\leq \inf_{H\in\mathcal{H}} \ell_n + \epsilon \\ &\leq \ell_n(\tilde{H}) + \epsilon \\ &\leq \ell(\tilde{H}) + 2\epsilon \\ &\leq \ell(\tilde{H}_n) + e\epsilon.\end{aligned}$$

A model is overfit when we have low empirical risk, while having high expected risk. This means that the explanation quality on the data source is much worse than on the training data. The main reason for this is that the theory ($\mathcal{H}$ and ℓ) is so complex that it can explain almost anything.

A model is underfit when we have high empirical risk, while having high expected risk. This means that we neither explain the training data nor the data source. The main reason for this is that the theory is too simple to capture the nature of the data.

The model is learning when we have low empirical risk and low expected risk. This means that the training was successful. Generalization occurs when the expected risk is close to the empirical risk. Note that this does not mean that the explanation is good, since any "blind explanation" will generalize well due to the weak law of large numbers. Ideally, we want generalization and learning.

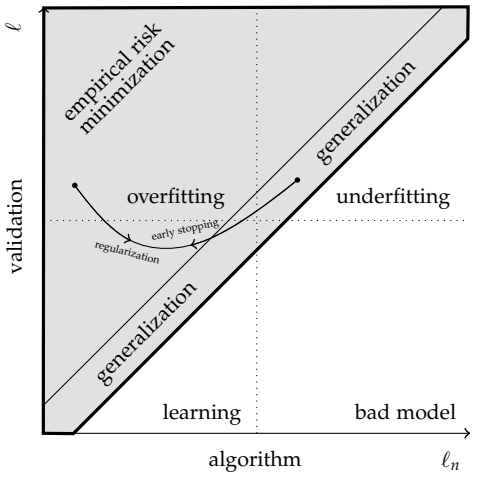


Figure 1. The map of learning.

2 *Theory of convex functions*

2.1 *Mathematical background*

Theorem 3 (Cauchy-Schwarz inequality)**.** Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, then

$$|\boldsymbol{u}^\top \boldsymbol{v}| \leq \|\boldsymbol{u}\|\|\boldsymbol{v}\|.$$

For non-zero vectors, this is equivalent to

$$-1 \leq \frac{\boldsymbol{u}^\top \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|} \leq 1.$$

Definition 4 (Spectral norm)**.** Let $\boldsymbol{A}$ be an $m \times d$ matrix, then

$$\|\boldsymbol{A}\| \doteq \max_{\boldsymbol{v} \in \mathbb{R}^d, \boldsymbol{v} \neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \max_{\|\boldsymbol{v}\|=1} \|\boldsymbol{A}\boldsymbol{v}\|.$$

Intuitively, this means that $\|\boldsymbol{A}\|$ is the largest factor by which a vector can be stretched in length under the mapping $\boldsymbol{v} \mapsto \boldsymbol{A}\boldsymbol{v}$.

Theorem 5 (Mean value theorem)**.** Let $a < b$ be real numbers, and let $h : [a, b] \to \mathbb{R}$ be a continuous function that is differentiable on (a, b). Then, there exists $c \in (a, b)$ such that

$$h'(c) = \frac{h(b) - h(a)}{b - a}.$$

Geometrically, this means that between a and b, there is a tangent to the graph of h that has the same slope; see Figure 2.

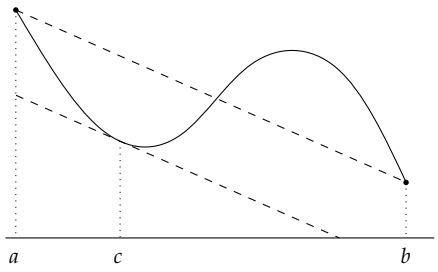


Figure 2. Illustration of the mean value theorem.

Theorem 6 (Fundamental theorem of calculus)**.** Let $a < b$ be real numbers, and let $h : \mathrm{dom}(h) \to \mathbb{R}$ be a differentiable function on an open domain $\mathrm{dom}(h) \supset [a, b]$, and such that h' is continuous on $[a, b]$. Then,

$$h(b) - h(a) = \int_a^b h'(t)dt.$$

2.2 *Convex sets*

Definition 7 (Convex set)**.** A set $\mathcal{C} \subseteq \mathbb{R}^d$ is convex if the line segment between any two points of $\mathcal{C}$ lies in $\mathcal{C}$. *I.e.*, if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ and any λ with $0 \leq \lambda \leq 1$, we have

$$\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y} \in \mathcal{C}.$$

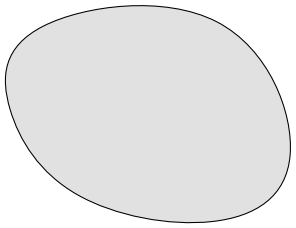
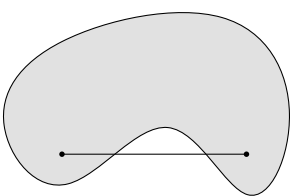


Figure 3. Example of a convex set in $\mathbb{R}^2$.

Observation. Let $\mathcal{C}_i, i \in I$ be convex sets, where I is a (possibly infinite)



index set. Then,

$$\mathcal{C} = \bigcap_{i \in I} \mathcal{C}_i,$$

is a convex set.

2.3 *Convex functions*

Definition 8 (Convexity)**.** A function $f : \mathrm{dom}(f) \to \mathbb{R}$ is convex if (i) $\mathrm{dom}(f)$ is convex and (ii) for all $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$ and all $\lambda \in [0,1]$, we have

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$

Geometrically, this condition means that the line segment connecting the two points $(\boldsymbol{x}, f(\boldsymbol{x}))$ and $(\boldsymbol{y}, f(\boldsymbol{y})) \in \mathbb{R}^{d+1}$ lies pointwise above the graph of f; see Figure 5.

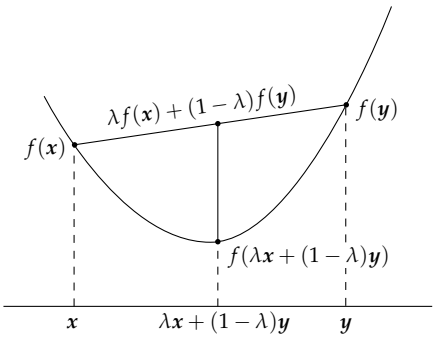


Figure 5. Illustration of the classic definition of convexity.

Definition 9 (Epigraph)**.** The epigraph of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\mathrm{epi}(f) \doteq \left\{ (\boldsymbol{x}, \alpha) \in \mathbb{R}^{d+1} \;\middle|\; \boldsymbol{x} \in \mathrm{dom}(f), \alpha \geq f(\boldsymbol{x}) \right\}.$$

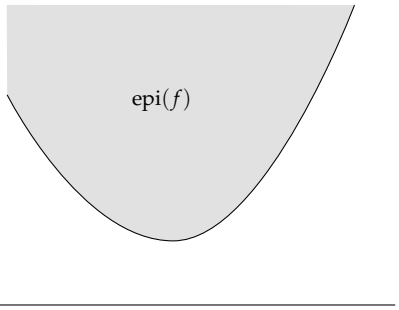


Figure 6. Epigraph of a convex function.

Observation. f is a convex function if and only if $\mathrm{epi}(f)$ is a convex set.

Lemma 10 (Jensen's inequality)**.** Let f be convex, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathrm{dom}(f)$, $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$ such that $\sum_{i=1}^m \lambda_i = 1$, then

$$f\left(\sum_{i=1}^m \lambda_i \boldsymbol{x}_i \right) \leq \sum_{i=1}^m \lambda_i f(\boldsymbol{x}_i).$$

Proof. We will prove Jensen's inequality by induction. The base case ($m = 2$) is true by definition of convexity. Let Jensen's inequality hold for $m = k$. Consider $m = k+1$, then

$$\begin{aligned}
f\left(\sum_{i=1}^{k+1} \lambda_i \boldsymbol{x}_i \right) &= f\left(\sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i + \lambda_{k+1} \boldsymbol{x}_{k+1} \right) \\
&= f\left((1-\lambda_{k+1}) \sum_{i=1}^{k} \frac{\lambda_i}{1-\lambda_{k+1}} \boldsymbol{x}_i + \lambda_{k+1} \boldsymbol{x}_{k+1} \right) \\
&\leq (1-\lambda_{k+1}) f\left(\sum_{i=1}^{k} \frac{\lambda_i}{1-\lambda_{k+1}} \boldsymbol{x}_i \right) + \lambda_{k+1} f(\boldsymbol{x}_{k+1}) && \text{By definition of convexity.} \\
&\leq (1-\lambda_{k+1}) \sum_{i=1}^{k} \frac{\lambda_i}{1-\lambda_{k+1}} f(\boldsymbol{x}_i) + \lambda_{k+1} f(\boldsymbol{x}_{k+1}) && \text{Induction assumption, } \textstyle\sum_{i=1}^{k} \frac{\lambda_i}{1-\lambda_{k+1}} = 1. \\
&= \sum_{i=1}^{k+1} \lambda_i f(\boldsymbol{x}_i).
\end{aligned}$$

Since it holds for the base case and the induction step, Jensen's inequality must hold for all m. ■

Lemma 11. Let f be convex and suppose that $\mathrm{dom}(f) \subseteq \mathbb{R}^d$ is open, then f is continuous.

Definition 12 (Differentiable functions)**.** Let $f : \mathrm{dom}(f) \to \mathbb{R}^m$ where $\mathrm{dom}(f) \subseteq \mathbb{R}^d$ is open. f is called differentiable at $\boldsymbol{x} \in \mathrm{dom}(f)$ if there exists an $m \times d$ matrix $\boldsymbol{A}$ and an error function $r : \mathbb{R}^d \to \mathbb{R}^m$ defined around $\mathbf{0} \in \mathbb{R}^d$ such that for all $\boldsymbol{y}$ in some neighborhood of $\boldsymbol{x}$,

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}) + r(\boldsymbol{y} - \boldsymbol{x}),$$

where $\lim_{v \to \mathbf{0}} \|r(v)\| / \|v\| = \mathbf{0}$ (error function r is sublinear around $\mathbf{0}$). $\boldsymbol{A}$ is unique and called the Jacobian matrix of f at $\boldsymbol{x}$.

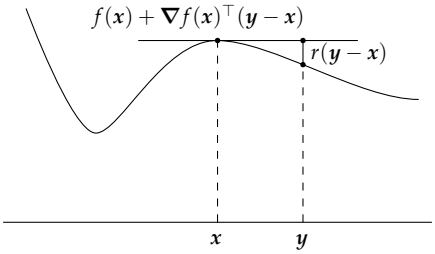


Figure 7. Graph of the affine function $f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})$ is a tangent hyperplane to the graph of f at $(\boldsymbol{x}, f(\boldsymbol{x}))$.

Lemma 13 (First-order convexity)**.** Suppose that $\mathrm{dom}(f)$ is open and that f is differentiable. In particular, the gradient

$$\nabla f(\boldsymbol{x}) \doteq \left[\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_d} \right]$$

exists at every point $\boldsymbol{x} \in \mathrm{dom}(f)$.

Then, f is convex if and only if (i) $\mathrm{dom}(f)$ is convex and (ii)

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f) : f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}).$$

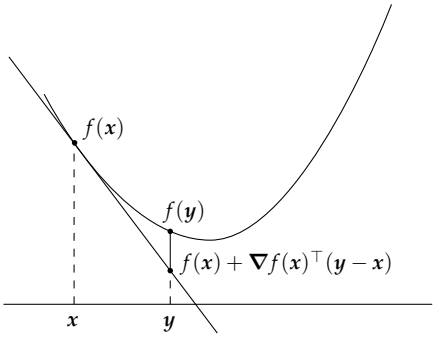


Figure 8. Illustration of the first-order characterization of convexity (Lemma 13).

Geometrically, this means that the graph is above all tangent hyperplanes; see Figure 8.

Proof. We will first prove that convexity implies that the first-order definition holds. Then, we will prove that the first-order definition implies convexity, making the definitions equivalent for differentiable functions f.

$\Rightarrow$: Suppose f is convex. Then, for all $t \in (0, 1)$,

$$\begin{aligned}
& f((1-t)\boldsymbol{x} + t\boldsymbol{y}) \leq (1-t)f(\boldsymbol{x}) + tf(\boldsymbol{y}) && \text{Definition of convexity} \\
\iff \quad & f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \leq f(\boldsymbol{x}) + t(f(\boldsymbol{y}) - f(\boldsymbol{x})) \\
\iff \quad & f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \frac{f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) - f(\boldsymbol{x})}{t} \\
& \qquad = f(\boldsymbol{x}) + \frac{\nabla f(\boldsymbol{x})^\top t(\boldsymbol{y} - \boldsymbol{x}) + r(t(\boldsymbol{y} - \boldsymbol{x}))}{t} \\
& \qquad = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \underbrace{\frac{r(t(\boldsymbol{y} - \boldsymbol{x}))}{t}}_{\to 0 \text{ for } t \to 0}.
\end{aligned}$$

$\Leftarrow$: Define $z \doteq \lambda x + (1-\lambda)y \in \mathrm{dom}(f)$ for $\lambda \in [0,1]$ by convexity of $\mathrm{dom}(f)$. Then, we have the following inequalities,

$$\begin{aligned} f(x) &\geq f(z) + \nabla f(z)^\top (x - z) \\ \lambda f(x) &\geq \lambda\Big(f(z) + \nabla f(z)^\top (x - z)\Big) \\ f(y) &\geq f(z) + \nabla f(z)^\top (y - z) \\ (1-\lambda) f(y) &\geq (1-\lambda)\Big(f(z) + \nabla f(z)^\top (y - z)\Big). \end{aligned}$$

This implies the following,

$$\begin{aligned} \lambda f(x) + (1-\lambda) f(y) &\geq \lambda\Big(f(z) + \nabla f(z)^\top (x - z)\Big) + (1-\lambda)\Big(f(z) + \nabla f(z)^\top (y - z)\Big) \\ &= f(z) + \lambda \nabla f(z)^\top (x - z) + (1-\lambda)\nabla f(z)^\top (y - z) \\ &= f(z) + \nabla f(z)^\top (\lambda(x - z) + (1-\lambda)(y - z)) \\ &= f(z) + \nabla f(z)^\top (\lambda x + (1-\lambda)y - (\lambda z + (1-\lambda)z)) \\ &= f(z) + \nabla f(z)^\top (z - z) \\ &= f(\lambda x + (1-\lambda)y). \end{aligned}$$

■

Lemma 14 (First-order convexity alternative)**.** Suppose that $\mathrm{dom}(f)$ is open and that f is differentiable. Then, f is convex if and only if (i) $\mathrm{dom}(f)$ is convex and (ii)

$$\forall x, y \in \mathrm{dom}(f) : (\nabla f(y) - \nabla f(x))^\top (y - x) \geq 0.$$

Geometrically, this means that the gradient is monotonic.

Proof. We will first prove that it holds from left to right, and then from right to left.

$\Rightarrow$: If f is convex, the first-order characterization of convexity (Lemma 13) yields the following,

$$\begin{aligned} f(x) &\geq f(y) + \nabla f(y)^\top (x - y) \\ f(y) &\geq f(x) + \nabla f(x)^\top (y - x), \end{aligned}$$

for all $x, y \in \mathrm{dom}(f)$. Adding up these two inequalities yields

$$\begin{aligned} & f(x) + f(y) \geq f(y) + \nabla f(y)^\top (x - y) + f(x) + \nabla f(x)^\top (y - x) \\ \iff \quad & 0 \geq \nabla f(y)^\top (x - y) + \nabla f(x)^\top (y - x) \\ & \quad = (\nabla f(y) - \nabla f(x))^\top (x - y) \\ \iff \quad & 0 \leq (\nabla f(y) - \nabla f(x))^\top (y - x). \end{aligned}$$

$\Leftarrow$: Define $z \doteq (1-t)x + ty \in \mathrm{dom}(f)$ for $x, y \in \mathrm{dom}(f), t \in (0,1)$ by convexity of $\mathrm{dom}(f)$. Observe that $z = x + t(y - x)$. Then, we have the

following inequality, according to the monotonicity of the gradient,

$$\begin{aligned}(\nabla f(z) - \nabla f(x))^\top (z - x) &\geq 0 \\ (\nabla f(x + t(y - x)) - \nabla f(x))^\top (x + t(y - x) - x) &\geq 0 \\ (\nabla f(x + t(y - x)) - \nabla f(x))^\top (y - x) &\geq 0. \qquad \text{Divide by } t\end{aligned}$$

Let $h(t) = f(x + t(y - x))$, then $h'(t) = \nabla f(x + t(y - x))^\top (y - x)$. Hence, we can rewrite the inequality as the following,

$$h'(t) \geq \nabla f(x)^\top (y - x).$$

By the mean value theorem, there exists $c \in (0, 1)$ such that $h'(c) = h(1) - h(0)$. *I.e.*,

$$h'(c) = f(y) - f(x).$$

Thus, we can rewrite the inequality to the following,

$$\begin{aligned}f(y) &= f(x) + h'(c) \\ &\geq f(x) + \nabla f(x)^\top (y - x),\end{aligned}$$

which is the first-order characterization of convexity (Lemma 13).

■

Lemma 15 (Second-order convexity)**.** Suppose that $\mathrm{dom}(f)$ is open and that f is twice differentiable. In particular, the Hessian

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_d^2} \end{bmatrix}$$

exists at every point $x \in \mathrm{dom}(f)$ and is symmetric.

Then, f is convex if and only if (i) $\mathrm{dom}(f)$ is convex and (ii) for all $x \in \mathrm{dom}(f)$, we have

$$\nabla^2 f(x) \succeq 0.$$

I.e., $\nabla^2 f(x)$ is positive semidefinite (M is positive semidefinite if $x^\top M x \geq 0$ for all x and $x^\top M x > 0$ for all $x \neq \mathbf{0}$).

Geometrically, this means that f has non-negative curvature everywhere. *I.e.*, the growth rate should be growing.

Observation (Operations that preserve convexity)**.** Let $f_1, \ldots, f_m$ be convex functions, $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$, then

$$f \doteq \max_{i=1}^{m} f_i,$$

and

$$f \doteq \sum_{i=1}^{m} \lambda_i f_i.$$

are convex on $\mathrm{dom}(f) \doteq \bigcap_{i=1}^{m} \mathrm{dom}(f_i)$.

Let f be a convex function with $\mathrm{dom}(f) \subseteq \mathbb{R}^d, g : \mathbb{R}^m \to \mathbb{R}^d$ an affine function, *i.e.* $g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ for some $\boldsymbol{A} \in \mathbb{R}^{d\times m}, \boldsymbol{b} \in \mathbb{R}^d$. Then, the function $f \circ g$ is convex on $\mathrm{dom}(f \circ g) \doteq \{\boldsymbol{x} \in R^m \mid g(\boldsymbol{x}) \in \mathrm{dom}(f)\}$.

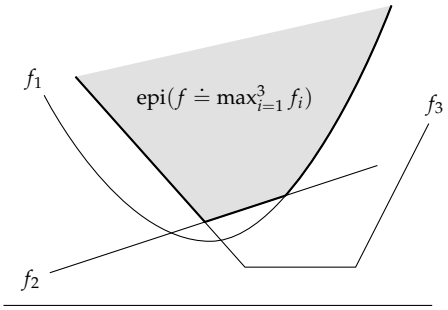


Figure 9. The maximum operator over m convex functions is a convex function. As can be seen, the epigraph of f is convex.

2.4 *Minimizing convex functions*

Definition 16 (Local minimum)**.** A local minimum of $f : \mathrm{dom}(f) \to \mathbb{R}$ is a point $\boldsymbol{x}$ such that there exists $\epsilon > 0$ with

$$f(\boldsymbol{x}) \leq f(\boldsymbol{y}) \;\; \forall \boldsymbol{y} \in \mathrm{dom}(f) \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{x}\| < \epsilon.$$

Remark. This definition of local minima means that in some small neighborhood around $\boldsymbol{x}$, $\boldsymbol{x}$ is the smallest point.

Lemma 17. Let $\boldsymbol{x}^\star$ be a local minimum of a convex function $f : \mathrm{dom}(f) \to \mathbb{R}$. Then, $\boldsymbol{x}^\star$ is a global minimum, meaning that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{y}) \;\; \forall \boldsymbol{y} \in \mathrm{dom}(f)$.

Proof. Proof by contradiction. Suppose there exists $\boldsymbol{y} \in \mathrm{dom}(f)$ such that $f(\boldsymbol{y}) < f(\boldsymbol{x}^\star)$. Define $\boldsymbol{y}' \doteq \lambda \boldsymbol{x}^\star + (1-\lambda)\boldsymbol{y}$ for $\lambda \in (0,1)$. From convexity, we get that $f(\boldsymbol{y}') < f(\boldsymbol{x}^\star)$, because $f(\boldsymbol{y}) < f(\boldsymbol{x}^\star)$. If we choose λ so close to 1 such that $\|\boldsymbol{y}' - \boldsymbol{x}^\star\| < \epsilon$ gives the inequality $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{y}')$. This yields a contradiction, thus there cannot exist a $\boldsymbol{y} \in \mathrm{dom}(f)$ such that $f(\boldsymbol{y}) < f(\boldsymbol{x}^\star)$, meaning that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathrm{dom}(f)$. ■

Lemma 18. Suppose that f is convex and differentiable over an open domain $\mathrm{dom}(f)$. Let $\boldsymbol{x} \in \mathrm{dom}(f)$. If $\boldsymbol{\nabla} f(\boldsymbol{x}) = \boldsymbol{0}$ (critical point), then $\boldsymbol{x}$ is a global minimum.

Proof. Suppose that $\boldsymbol{\nabla} f(\boldsymbol{x}) = \boldsymbol{0}$. According to the first-order characterization of convexity, we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \underbrace{\boldsymbol{\nabla} f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})}_{=0} = f(\boldsymbol{x})$$

for all $\boldsymbol{y} \in \mathrm{dom}(f)$, so $\boldsymbol{x}$ is a global minimum. ■

Definition 19 (Strict convexity)**.** A function $f : \mathrm{dom}(f) \to \mathbb{R}$ is strictly convex if (i) $\mathrm{dom}(f)$ is convex and (ii) for all $\boldsymbol{x} \neq \boldsymbol{y} \in \mathrm{dom}(f)$ and all $\lambda \in (0,1)$, we have

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) < \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$

An example of a strictly convex function can be found in Figure 5, while an example of a function that is not strictly convex can be found in Figure 10.

Lemma 20. Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be strictly convex. Then, f has *at most* one global minimum.

But, how do we know a minimizer exists? To be able to find this out, we must use the Weierstrass theorem.

Definition 21 (Sublevel set)**.** $f : \mathbb{R}^d \to \mathbb{R}, \alpha \in \mathbb{R}$. The set $f^{\leq \alpha} \doteq \{\boldsymbol{x} \in \mathbb{R}^d \mid f(\boldsymbol{x}) \leq \alpha\}$ is the α-sublevel set of f.

Theorem 22 (Weierstrass)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function, and suppose there is a non-empty and bounded sublevel set $f^{\leq \alpha}$. Then, f has a global minimum.

Proof. We know that f attains a minimum over the closed and bounded set $f^{\leq \alpha}$ at some $\boldsymbol{x}^\star$, because f is continuous. This $\boldsymbol{x}^\star$ is also a global minimum as it has value $f(\boldsymbol{x}^\star) \leq \alpha$, while any $\boldsymbol{x} \notin f^{\leq \alpha}$ has value $f(\boldsymbol{x}) > \alpha \geq f(\boldsymbol{x}^\star)$. ■

Since convex functions are continuous by Lemma 11, the Weierstrass theorem also applies to convex functions. Thus, to figure out whether a convex function has a minimizer, we need to find any non-empty and bounded sublevel set of this function.

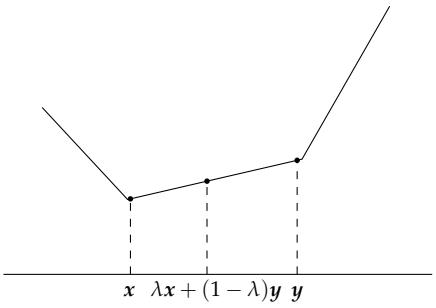


Figure 10. A non-strictly convex function.

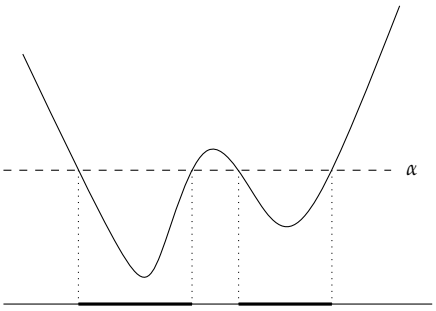


Figure 11. Illustration of a sublevel set of a non-convex function.

2.5 *Convex programming*

In standard form, optimization problems look like the following,

$$\begin{aligned} \text{minimize} \quad & f_0(\boldsymbol{x}) \\ \text{subject to} \quad & f_i(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, m \\ & h_j(\boldsymbol{x}) = 0, \quad j = 1, \ldots, p. \end{aligned}$$

The problem domain is then $\mathcal{D} = \bigcap_{i=0}^{m} \mathrm{dom}(f_i) \cap \bigcap_{j=1}^{p} \mathrm{dom}(h_j)$.

In a convex program, all f_i are convex functions, and all h_j are affine functions with domain $\mathbb{R}^d$.

Definition 23 (Lagrange dual function)**.** Given an optimization problem in standard form, its Lagrangian is the function $L : \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ given by

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{j=1}^{p} \nu_j h_j(\boldsymbol{x}).$$

The λ_i, ν_j are called Lagrange multipliers. The Lagrange dual function is the function $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R} \cup \{-\infty\}$ defined by

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x} \in \mathcal{D}} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

Only the $(\boldsymbol{\lambda}, \boldsymbol{\nu})$ pairs with $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) > -\infty$ are interesting.

Lemma 24 (Weak Lagrange duality)**.** Let $\boldsymbol{x}$ be a feasible solution ($f_i(\boldsymbol{x}) \leq 0$ for $i = 1, \dots m$ and $h_j(\boldsymbol{x}) = 0$ for $j = 1, \dots, p$). Let g be the Lagrange dual function and $\boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{\nu} \in \mathbb{R}^p$ such that $\boldsymbol{\lambda} \geq \mathbf{0}$. Then

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\boldsymbol{x}).$$

Proof.

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\boldsymbol{x}) + \underbrace{\sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x})}_{\leq 0} + \underbrace{\sum_{j=1}^{p} \nu_j h_j(\boldsymbol{x})}_{=0} \leq f_0(\boldsymbol{x}).$$

■

However, we want to know what the best lower bound is that we can get in this way. For this, we must choose $\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}$ such that $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is maximized. This can be phrased as another optimization problem, called the Lagrange dual,

$$\begin{aligned} \text{maximize} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned}$$

Observation. The Lagrange dual is a convex program, even if the primal is not.

Theorem 25. Suppose that a convex program has a feasible solution $\tilde{\boldsymbol{x}}$ that in addition satisfies $f_i(\tilde{\boldsymbol{x}}) < 0, i = 1, \dots, m$ (Slater point). Then, the infimum value of the primal equals the supremum value of its Lagrange dual. Moreover, if the value is finite, it is attained by a feasible solution of the dual,

$$\inf f_0(\boldsymbol{x}) = \max g(\boldsymbol{\lambda}, \boldsymbol{\nu}).$$

A case of particular interest is that strong duality holds and the joint value is attained in both the primal and dual problem.[1] If all f_i and h_j are differentiable, then the Karush-Kuhn-Tucker (KKT) conditions provide necessary and, under convexity, also sufficient conditions for this case to occur.

[1] This would mean that $\min f_0(\boldsymbol{x}) = \max g(\boldsymbol{\lambda}, \boldsymbol{\nu})$.

Definition 26 (Zero duality gap)**.** Let $\tilde{\boldsymbol{x}}$ be feasible for the primal and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ feasible for the Lagrange dual. The primal and dual solutions $\tilde{\boldsymbol{x}}$ and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ are said to have zero duality gap if $f_0(\tilde{\boldsymbol{x}}) = g(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$.

If the primal and dual have zero duality gap, then $\min f_0(\boldsymbol{x}) = \max g(\boldsymbol{\lambda}, \boldsymbol{\nu})$.

The consequence of zero duality gap is the *master equation*,

$$\begin{aligned} f_0(\tilde{\boldsymbol{x}}) &= g(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \\ &= \inf_{\boldsymbol{x} \in \mathcal{D}} \left(f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i f_i(\boldsymbol{x}) + \sum_{j=1}^{p} \tilde{\nu}_j h_j(\boldsymbol{x}) \right) \\ &\leq f_0(\tilde{\boldsymbol{x}}) + \sum_{i=1}^{m} \underbrace{\tilde{\lambda}_i f_i(\tilde{\boldsymbol{x}})}_{\leq 0} + \sum_{j=1}^{p} \underbrace{\tilde{\nu}_j h_j(\tilde{\boldsymbol{x}})}_{=0} \\ &\leq f_0(\tilde{\boldsymbol{x}}), \end{aligned}$$

which means that all inequalities turn into equalities. Thus,

$$\tilde{\lambda}_i f_i(\tilde{\boldsymbol{x}}) = 0, \quad i = 1, \ldots, m,$$

which is called complementary slackness, because if $\tilde{\lambda}_i \neq 0$, then $f_i(\tilde{\boldsymbol{x}}) = 0$, and vice versa. Furthermore, if all f_i and h_j are differentiable, then

$$\boldsymbol{\nabla} f_0(\tilde{\boldsymbol{x}}) + \sum_{i=1}^{m} \tilde{\lambda}_i \boldsymbol{\nabla} f_i(\tilde{\boldsymbol{x}}) + \sum_{j=1}^{p} \tilde{\nu}_j \boldsymbol{\nabla} h_j(\tilde{\boldsymbol{x}}) = \mathbf{0},$$

which is called the vanishing Lagrangian gradient condition.

In summary, we get the following result.

Theorem 27 (Karush-Kuhn-Tucker necessary conditions)**.** Let $\tilde{\boldsymbol{x}}$ and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ be feasible solutions of the primal optimization problem and its Lagrange dual, respectively, with zero duality gap. If all f_i and h_j are differentiable, then

$$\tilde{\lambda}_i f_i(\tilde{\boldsymbol{x}}) = 0, \quad i = 1, \ldots, m$$

Complementary slackness

$$\boldsymbol{\nabla} f_0(\tilde{\boldsymbol{x}}) + \sum_{i=1}^{m} \tilde{\lambda}_i \boldsymbol{\nabla} f_i(\tilde{\boldsymbol{x}}) + \sum_{j=1}^{p} \tilde{\nu}_j \boldsymbol{\nabla} h_j(\tilde{\boldsymbol{x}}) = \mathbf{0}.$$

Vanishing Lagrangian gradient

If we have a convex program, then the KKT conditions are sufficient for zero duality gap. The motivation behind this is that they may be easier to solve for than solving the primal optimization problem. However, we cannot always count on the KTT conditions to be solvable, because they are only guaranteed if there are primal and dual solutions of zero duality

gap. But, if the primal has a Slater point[2], then the KTT conditions are equivalent to the existence of a primal minimizer.

[2] A point $\tilde{x}$ such that $f_i(\tilde{x}) < 0$ for all $i = 1, \ldots, m$.

3 *Gradient descent*

Gradient descent is an optimization algorithm that aims to find the global minimizer. We assume that $f : \mathbb{R}^d \to \mathbb{R}$ is convex, differentiable, and has a global minimizer $x^\star$. Then, the goal of gradient descent is, for $\epsilon > 0$, to find a $x \in \mathbb{R}^d$ such that

$$f(x) - f(x^\star) \leq \epsilon.$$

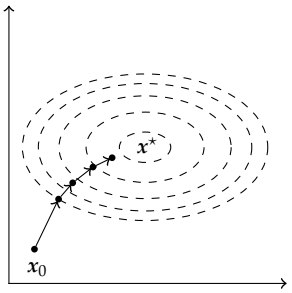


Figure 12. Gradient descent updates.

It works by first choosing a $x_0 \in \mathbb{R}^d$ and then iteratively updating this value by

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

for timesteps $t = 0, 1, \ldots$ and stepsize $\gamma > 0$.

Take a small step into the direction of the negative gradient to move toward the minimum.

3.1 *Vanilla analysis*

We want to be able to bound $f(x_t) - f(x^\star)$, which tells us how far off we are from the minimum at timestep t. For this, we can use the first-order characterization of convexity,

$$f(x^\star) \geq f(x_t) + \nabla f(x_t)^\top (x^\star - x_t) \iff f(x_t) - f(x^\star) \leq \nabla f(x_t)^\top (x_t - x^\star).$$

Using the gradient descent update rule, we get

$$\nabla f(x_t) = \frac{x_t - x_{t+1}}{\gamma},$$

which gives the following bound

$$\begin{aligned}
f(x_t) - f(x^\star) &\leq \nabla f(x_t)^\top (x_t - x^\star) \\
&= \frac{1}{\gamma}(x_t - x_{t+1})^\top (x_t - x^\star) \\
&= \frac{1}{2\gamma}\Big(\|x_t - x_{t+1}\|^2 + \|x_t - x^\star\|^2 - \|x_{t+1} - x^\star\|^2\Big) && 2v^\top w = \|v\|^2 + \|w\|^2 - \|v - w\|^2 \text{ (cosine theorem).} \\
&= \frac{\gamma}{2}\|\nabla f(x_t)\|^2 + \frac{1}{2\gamma}\Big(\|x_t - x^\star\|^2 - \|x_{t+1} - x^\star\|^2\Big). && x_t - x_{t+1} = \gamma \nabla f(x_t).
\end{aligned}$$

Summing this up over the first T iterations, we get an upper bound on the summed error,

$$\begin{aligned}
\sum_{t=0}^{T-1} (f(x_t) - f(x^\star)) &\leq \sum_{t=0}^{T-1} \nabla f(x_t)^\top (x_t - x^\star) \\
&= \frac{\gamma}{2}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 + \frac{1}{2\gamma}\Big(\|x_0 - x^\star\|^2 - \|x_T - x^\star\|^2\Big) && \text{Telescoping sum.} \\
&\leq \frac{\gamma}{2}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 + \frac{1}{2\gamma}\|x_0 - x^\star\|^2. && \|x_T - x^\star\|^2 \geq 0.
\end{aligned}$$

We can also use this to get a bound on the average error,

$$\frac{1}{T}\sum_{t=0}^{T-1} (f(x_t) - f(x^\star)) \leq \frac{1}{T}\left(\frac{\gamma}{2}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 + \frac{1}{2\gamma}\|x_0 - x^\star\|^2\right).$$

3.2 *Lipschitz convex functions*

Theorem 28. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\boldsymbol{x}^\star$. Furthermore, suppose that $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\| \leq R$ and $\|\nabla f(\boldsymbol{x})\| \leq B$ for all $\boldsymbol{x}$. Choosing the stepsize,

$$\gamma \doteq \frac{R}{B\sqrt{T}},$$

gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) \leq \frac{RB}{\sqrt{T}}.$$

A function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous if there exists an M, such that $\forall x, y \in \mathbb{R}$:

$$\frac{|f(x) - f(y)|}{|x - y|} \leq M.$$

This holds if the gradient is bounded.

Proof. We can plug our bounds into the vanilla analysis,

$$\begin{aligned}\sum_{t=0}^{T-1}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) &\leq \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{1}{2\gamma}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2 \\ &\leq \frac{\gamma}{2}B^2T + \frac{1}{2\gamma}R^2 \\ &\doteq q(\gamma).\end{aligned}$$

Now, we want to choose γ such that the bound $q(\gamma)$ is minimized. We can compute the minimum by computing the derivative,

$$q'(\gamma) = \frac{1}{2}B^2T - \frac{R^2}{2\gamma^2},$$

and solving for $q'(\gamma) = 0$, which yields

$$\gamma = \frac{R}{B\sqrt{T}}.$$

Then, we can compute the bound by

$$q\left(\frac{R}{B\sqrt{T}}\right) = RB\sqrt{T}.$$

Dividing by T yields the result. ∎

We want to find out how many iterations we would need to ensure that the average error is bounded by ϵ. We can use Theorem 28 to compute a lower bound on the number of iterations,

$$\frac{RB}{\sqrt{T}} \leq \epsilon \implies T \geq \frac{R^2B^2}{\epsilon^2}.$$

So, the amount of iterations until convergence is of the order $\mathcal{O}(1/\epsilon^2)$. This means that we need at most $10000 \cdot R^2B^2$ iterations to achieve an error of 0.01.

3.3 *Smooth functions*

Definition 29 (Smooth functions)**.** Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be differentiable, $\mathcal{X} \subseteq \mathrm{dom}(f)$ convex, $L \in \mathbb{R}_+$. f is smooth (with parameter L) over $\mathcal{X}$ if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$:

$$\underbrace{f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})}_{\text{first-order convexity}} + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Remark. This definition does not require convexity.

Geometrically, this definition of smoothness means that the graph f is below a not too steep tangent paraboloid at $(\boldsymbol{x}, f(\boldsymbol{x}))$; see Figure 13. Furthermore, we can easily check if f is smooth with parameter L if the following function is convex,

$$g(\boldsymbol{x}) \doteq \frac{L}{2}\boldsymbol{x}^\top \boldsymbol{x} - f(\boldsymbol{x}),$$

over $\mathrm{dom}(g) \doteq \mathrm{dom}(f)$.

"Not too curved."

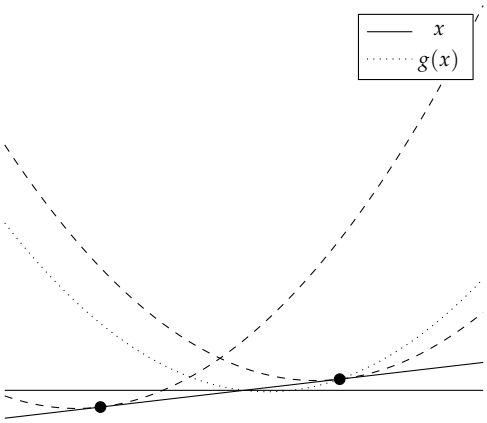


Figure 13. Plot of $f(x) = x$ with the tangent paraboloids at $x = -8,5$, showing smoothness with parameter $L = 1$. Furthermore, it shows that $g(x) = \frac{1}{2}x^2 - f(x)$ is convex.

Observation (Operations that preserve smoothness)**.** Let $f_1, \ldots, f_m$ be smooth with parameters $L_1, \ldots, L_m$ and let $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$. Then, the function $f \doteq \sum_{i=1}^m \lambda_i f_i$ is smooth with parameter $\sum_{i=1}^m \lambda_i L_i$.

Let f be smooth with parameter L and let $g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$. Then, the function $f \circ g$ is smooth with parameter $L\|\boldsymbol{A}\|^2$, where $\|\boldsymbol{A}\|$ is the spectral norm of $\boldsymbol{A}$.

Lemma 30. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, then f is smooth with parameter L if

$$\|\boldsymbol{\nabla} f(\boldsymbol{x}) - \boldsymbol{\nabla} f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|.$$

I.e., $\boldsymbol{\nabla} f$ is Lipschitz continuous.

Lemma 31 (Sufficient decrease)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L. With stepsize

$$\gamma \doteq \frac{1}{L},$$

gradient descent satisfies

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \frac{1}{2L}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2.$$

Sufficient decrease implies that, every step, we get closer to the minimum.

Proof.

$$
\begin{aligned}
f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) + \boldsymbol{\nabla} f(\boldsymbol{x}_t)^\top (\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 && \text{Smoothness.} \\
&= f(\boldsymbol{x}_t) + \boldsymbol{\nabla} f(\boldsymbol{x}_t)^\top \left(-\frac{\boldsymbol{\nabla} f(\boldsymbol{x}_t)}{L}\right) + \frac{L}{2}\left\|\frac{\boldsymbol{\nabla} f(\boldsymbol{x}_t)}{L}\right\|^2 && \text{GD: } \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \tfrac{1}{L}\boldsymbol{\nabla} f(\boldsymbol{x}_t). \\
&= f(\boldsymbol{x}_t) - \frac{1}{L}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 + \frac{1}{2L}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 \\
&= f(\boldsymbol{x}_t) - \frac{1}{2L}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2.
\end{aligned}
$$

■

Theorem 32. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\boldsymbol{x}^\star$. Furthermore, suppose that f is smooth with parameter L. Choosing stepsize

$$\gamma \doteq \frac{1}{L},$$

gradient descent yields

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2, \quad T > 0.$$

Proof. Due to sufficient decrease, we can bound the sum of squared gradients (which will be useful when bounding the vanilla analysis),

$$
\begin{aligned}
\frac{1}{2L}\sum_{t=0}^{T-1}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 &\leq \sum_{t=0}^{T-1} f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) && \text{SD: } f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \tfrac{1}{2L}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2. \\
&= f(\boldsymbol{x}_0) - f(\boldsymbol{x}_T). && \text{Telescoping sum.}
\end{aligned}
$$

Using the vanilla analysis, we can derive a bound on the average error,

$$
\begin{aligned}
\sum_{t=0}^{T-1}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) &\leq \frac{1}{2L}\sum_{t=0}^{T-1}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 + \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2 && \gamma \doteq \tfrac{1}{L}. \\
&\leq f(\boldsymbol{x}_0) - f(\boldsymbol{x}_T) + \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2 \\
\sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) &\leq \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2 \\
\frac{1}{T}\sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) &\leq \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2.
\end{aligned}
$$

From sufficient decrease, we know that the last iterate must be the best, thus

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \frac{1}{T}\sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)) \leq \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2.$$

■

Using this result, we can compute a bound on T to achieve an error smaller than ϵ,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2 \leq \frac{R^2 L}{2T} \leq \epsilon.$$

The error then becomes,

$$T \geq \frac{R^2 L}{2\epsilon},$$

which is on the order of $\mathcal{O}(1/\epsilon)$. This means that we need at most $50 \cdot R^2 L$ iterations for an error of 0.01 (as opposed to $10000 \cdot R^2 B^2$ in the Lipschitz case).

3.4 *Strongly convex functions*

Definition 33 (Strong convexity)**.** Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a convex and differentiable function, $\mathcal{X} \in \mathrm{dom}(f)$ convex and $\mu \in \mathbb{R}_+, \mu > 0$. f is called *strongly convex* with parameter μ over $\mathcal{X}$ if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{\nabla} f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2.$$

Geometrically, this means that, for any $\boldsymbol{x}$, the graph of f is above a not too flat tangent paraboloid at $(\boldsymbol{x}, f(\boldsymbol{x}))$; see Figure 14. Furthermore, we can easily check if f is strongly convex if and only if the following function is convex,

$$g(\boldsymbol{x}) \doteq f(\boldsymbol{x}) - \frac{\mu}{2}\boldsymbol{x}^\top \boldsymbol{x}.$$

By assuming that a function f is smooth and strongly convex, we can use a stronger lower bound to derive a bound on the error from the vanilla analysis, TODO

"Not too flat."

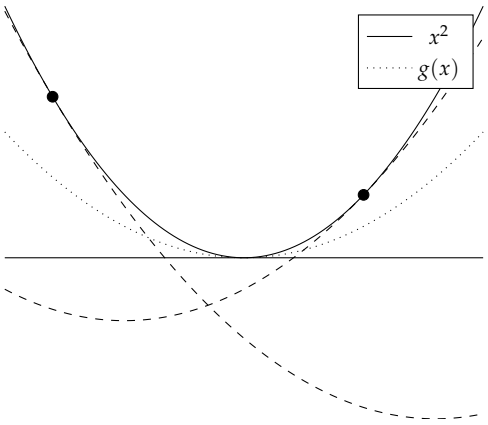


Figure 14. Plot of $f(x) = x^2$ with the tangent paraboloids at $x = -6, 4$, showing strong convexity with parameter $\mu = 1$. Furthermore, it shows $g(x) = f(x) - \frac{1}{2}x^2$, which is convex.

Theorem 34. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\boldsymbol{x}^\star$. Furthermore, suppose that f is smooth with parameter L and strongly convex with parameter $\mu > 0$. Choosing $\gamma \doteq \frac{1}{L}$, gradient descent with arbitrary $\boldsymbol{x}_0$ satisfies the following two properties,

1. Squared distances to $\boldsymbol{x}^\star$ are geometrically decreasing,

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2;$$

2. The absolute error after T iterations is exponentially small in T,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2.$$

Proof of 1. Using the vanilla analysis, we know the following,

$$\begin{aligned} f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) &\leq \boldsymbol{\nabla} f(\boldsymbol{x}_t)^\top (\boldsymbol{x}_t - \boldsymbol{x}^\star) \\ &= \frac{\gamma}{2}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 + \frac{1}{2\gamma}\Big(\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2\Big). \end{aligned}$$

But, using strong convexity, we have a stronger lower bound on the left hand side,

$$\begin{aligned} f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) &\leq \boldsymbol{\nabla} f(\boldsymbol{x}_t)^\top (\boldsymbol{x}_t - \boldsymbol{x}^\star) - \frac{\mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 && \text{Strong convexity.} \\ &= \frac{\gamma}{2}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 + \frac{1}{2\gamma}\Big(\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2\Big) - \frac{\mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2. \end{aligned}$$

This can be rewritten to the following bound,

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \underbrace{2\gamma(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) + \gamma^2\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2}_{\text{“noise“}} + (1 - \mu\gamma)\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2.$$

Now we need to show that the noise is non-positive,

$$\begin{aligned} 2\gamma(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) + \gamma^2\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 &= \frac{2}{L}(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_t)) + \frac{1}{L^2}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 && \gamma \doteq \tfrac{1}{L}. \\ &\leq \frac{2}{L}(f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t)) + \frac{1}{L^2}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 && \text{Sufficient decrease.} \\ &\leq -\frac{1}{L^2}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 + \frac{1}{L^2}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2 && \text{SD: } f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \tfrac{1}{2L}\|\boldsymbol{\nabla} f(\boldsymbol{x}_t)\|^2. \\ &= 0. \end{aligned}$$

Hence, the noise is non-positive, and we get the following,

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \Big(1 - \frac{\mu}{L}\Big)\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2.$$

■

Proof of 2. From (1), we know the following,

$$\|\boldsymbol{x}_T - \boldsymbol{x}^\star\|^2 \leq \Big(1 - \frac{\mu}{L}\Big)^T\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2.$$

Using smoothness, we can derive a bound on the final iterate error,

$$\begin{aligned} f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) &\leq \boldsymbol{\nabla} f(\boldsymbol{x}^\star)^\top (\boldsymbol{x}_T - \boldsymbol{x}^\star) + \frac{L}{2}\|\boldsymbol{x}_T - \boldsymbol{x}^\star\|^2 \\ &= \frac{L}{2}\|\boldsymbol{x}_T - \boldsymbol{x}^\star\|^2 && \boldsymbol{\nabla} f(\boldsymbol{x}^\star) = 0, \text{ because it is a stationary point.} \\ &\leq \frac{L}{2}\Big(1 - \frac{\mu}{L}\Big)^T\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2. && \text{Using (1).} \end{aligned}$$

■

From there, we can derive a lower bound on the number of iterations T to get an error of at most ϵ,

$$T \geq \frac{L}{\mu}\log\Big(\frac{R^2 L}{2\epsilon}\Big),$$

This means that we need $\frac{L}{\mu}\log(50 \cdot R^2 L)$ iterations for an error of at most 0.01, as opposed to $50 \cdot R^2 L$ in the smooth case. This bound only depends linearly on $\frac{\mu}{L}$, which might be very high.

References