Convex Sets and Functions

Theo Diamandis

February 11, 2023

We care about convex optimization problems because (for the most part) we can solve them[1]. Nonconvex problems are often solved on a case-by-case basis, as there are no general-purpose algorithms. Here, we develop tools to answer the question, "Is it convex?". We look at the basic properties of convex sets and functions and review important examples, largely following [BV04, §2-3]. Our goal is to develop a calculus of convex functions that will allow us to prove complicated functions are convex (or concave) by constructing them from a set of simple functions with known convexity and convexity-preserving composition rules. This idea forms the basis of disciplined convex programming (DCP), which we will use heavily in the remainder of the course.

1 Convex Sets

We call a set S *convex* if for any two points $x, y \in S$, the line segment between x and y lies in S. Equivalently, we require that

$$\lambda x + (1-\lambda)y \in S \tag{1}$$

for all x and y in S and $\lambda \in [0,1]$. We call any point x of the form

$$x = \lambda_1 x_1 + \cdots + \lambda_k x_k$$

with $\lambda_1 + \cdots + \lambda_k = 1$ and $\lambda_i \geq 0$ a *convex combination* of $x_1, \ldots, x_k$. The *convex hull* of a set S, denoted $\mathbf{conv}(S)$, is the set of all convex combination of points in S. The convex hull of a set S is always convex.

[1] From Richard Feynman's Lectures on Physics: "Finally, we make some remarks on why linear systems are so important. The answer is simple: because we can solve them! So most of the time we solve linear problems. Second (and most important), it turns out that the fundamental laws of physics are often linear." If you substitute "convex optimization" for "linear systems" and "optimization problems in practice" for "the fundamental laws of physics," I think the same sentiment holds.

Convex cones. One important class of convex sets are the convex cones. Most convex optimization solvers deal with problems in conic form (*i.e.,* the constraint sets are all convex cones). We call a set C a *cone* if for every $x \in C$ and $\lambda \geq 0$, we have that $\lambda x \in C$. A *convex cone* is a cone that is also convex. Equivalently, this means that for any x_1, x_2 in C and λ_1, $\lambda_2 \geq 0$, we have

$$\lambda_1 x_1 + \lambda_2 x_2 \in C.$$

Visually, this sweeps out a pie slice. We define a *conic combination* and *conic hull* analogously to the convex case.

1.1 Simple examples

The following sets are some simple examples of convex sets, many of which we will encounter often throughout the remainder of the course.

Hyperplans and halfspaces. A *hyperplane* is a set of the form $\{x \mid a^T x = b\}$, where $a \neq 0$. A *halfspace* is a set of the form $\{x \mid a^T x \leq b\}$, where $a \neq 0$. For both of these sets, we refer to a as the normal vector. Hyperplanes are both convex and affine (every *linear combination*[2] of points in the set also lies inside the set), and halfspaces are convex. These facts can be verified directly with some algebra.

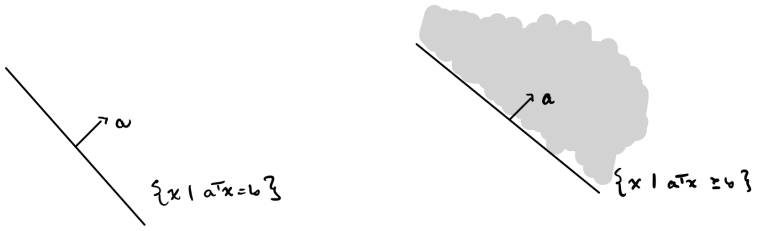

Figure 1: A hyperplane (left) and halfspace (right).

Norm balls and norm cones. Recall that a *norm* is a function $f : \mathbf{R}^n \to \mathbf{R}$ that satisfies

- $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$
- $\|tx\| = |t|\|x\|$ for $t \in \mathbf{R}$ (homogeneity of degree 1)
- $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)

A *norm ball* with center x_c and radius r is the set $\{x \mid \|x - x_c\| \leq r\}$. A special case of this is a Euclidean ball, where we take $\|\cdot\|$ to be the ℓ_2 norm, denoted by $\|\cdot\|_2$. Another special case is an ellipsoid, where we take the norm $\|x\|_P = \sqrt{x^T P x}$ for some positive definite matrix P. A *norm cone* is a set of the form $\{(x, t) \mid \|x\| \leq t\}$. The Euclidean norm cone is called the second-order cone, which is an important set for convex optimization solvers.

[2]A linear combination of points x and y is the set of points $\alpha x + \beta y$ where α, $\beta \in \mathbf{R}$.

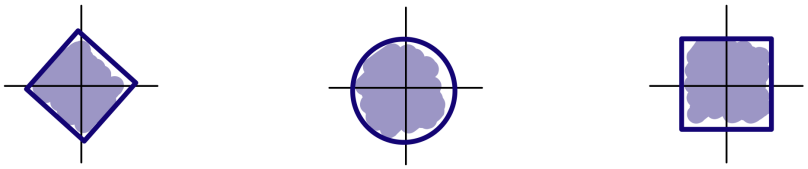

Figure 2: The ℓ_1 (left), ℓ_2 (center), and ℓ_∞ (right) norm balls centered at the origin.

Polyhedra and polytopes. A polyhedron is the solution of finitely many linear inequalities and equalities:

$$\{x \mid Ax = b, Cx \leq d\}.$$

We take the inequality to be componentwise. Linear programming problems are precisely those in which we minimize or maximize a linear objective over a polyhedral set. Note that a bounded polyhedron is sometimes refered to as a polytope (but some authors use the opposite convention). Important special cases are the nonnegative orthant, $\{x \mid x \geq 0\}$, and the probability simplex, $\{x \mid x \geq 0, \sum_i x_i = 1\}$.

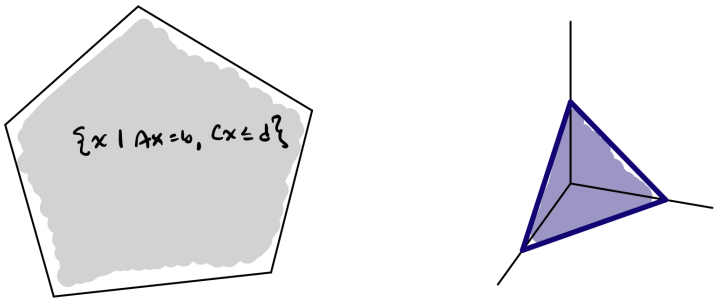

Figure 3: A polyhedron (left) and the probability simplex (right).

1.2 Operations that preserve convexity

Sometimes, we can establish convexity by directly applying the definition (1). Often, it is easier to build up a set from primitive convex sets and operations that preserve convexity. In this constructive calculus of convex sets, we can think of sets corresponding to expression trees: each leaf of the tree is a primitive set, and each interior node is a convexity-preserving operation. This calculus allows us to easily verify complicated sets are convex even when we do not have a simple description of the set. Some operations that preserve convexity are below.

Intersection. The intersection of any number (possibly infinite) of convex sets is convex. A simple example is that a polyhedron is the intersection of finitely many halfspaces. A more

complex example is the positive semidefinite cone, which is the set of positive semidefinite matrices, *i.e.,* all symmetric matrices X such that $z^T X z \geq 0$ for all vectors z. We denote the set of symmetric matrices by $\mathbf{S}^n$ and the set of positive semidefinite matrices by $\mathbf{S}^n_+$. While convexity can be proved directly from the definition, a slick proof is to recognize that

$$\mathbf{S}^n_+ = \bigcap_{z \neq 0} \{X \in \mathbf{S}^n \mid z^T X z \geq 0\},$$

which is the intersection of an infinite number of halfspaces parameterized by the vector z. (The function $z^T X z$ is linear in X for a fixed z). In fact, the converse is also true: every closed convex set is a intersection of halfspaces (usually infinitely many).

Affine function. An affine function has the form $f(x) = Ax + b$ for some $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$. Suppose the set $S \subseteq \mathbf{R}^n$ is convex. Then the *image* of S under f,

$$f(S) = \{f(x) \mid x \in S\},$$

and the *inverse image* of S under f,

$$f^{-1}(S) = \{x \mid f(x) \in S\},$$

are both convex. The inverse image is defined even if the function f is not invertible. Some simple examples of affine functions are scaling, translation, rotation, and projection. A more complex example is the hyperbolic cone

$$\{x \mid x^T P x \leq (c^T x)^2, c^T x \geq 0\},$$

where $P \in \mathbf{S}^n_+$ and $c \in \mathbf{R}^n$. This set is convex since it is the inverse image of the second-order cone,

$$\{x \mid x^T x \leq t^2, t \geq 0\},$$

under the affine function $f(x) = (P^{1/2}x, c^T x)$. Similarly, ellipsoids can be written as the image of the Euclidean ball under an affine function:

$$\{y \mid (y - x_c) P^{-1} (y - x_c) \leq 1\} = f\left(\{x \mid \|x\|_2 \leq 1\}\right),$$

where $f(x) = P^{1/2}x + x_c$.

Perspective function. Recall that the perspective function $f : \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}^n$ is $f(x,t) = x/t$. If $S \subseteq \mathbf{dom}\, f$ is convex ($\mathbf{dom}\, f$ indicates the domain where f is defined), then the set

$$f(S) = \{f(x,t) \mid (x,t) \in S\}$$

is also convex. An interpretation is that a convex object viewed through a pin-hole camera yields a convex image. More generally, this result holds for linear fractional functions $f(x) = (Ax + b)/(c^T x + d)$. This construction may seem esoteric at first, but it comes up in surprisingly many applications. As one example, consider the conditional probability

$$f_{ij} = \mathbb{P}(u = i \mid v = j) = \frac{p_{ij}}{\sum_{k=1}^n p_{kj}}.$$

This is a linear fractional function. Thus, if C is a convex set of joint probabilities for (u, v), then the set of conditional probabilities of u given v is also convex.

1.3 Separating and supporting hyperplanes

We briefly cover arguably the most important theorem in convex analysis: the *separating hyperplane theorem*. This theorem states that if C and D are nonempty disjoint convex sets, then there exists a hyperplane that separates C and D. In other words, for nonempty convex C and D, such that $C \cap D = \emptyset$, there exist $a \neq 0$ and b such that $a^T x \leq b$ for all $x \in C$ and $a^T x \geq b$ for all $x \in D$. The hyperplane $\{x \mid a^T x = b\}$ is called the *separating hyperplane* for the sets C and D. The machine learning interpretation of this theorem is that there exists a linear classifier that distinguishes between disjoint convex sets.

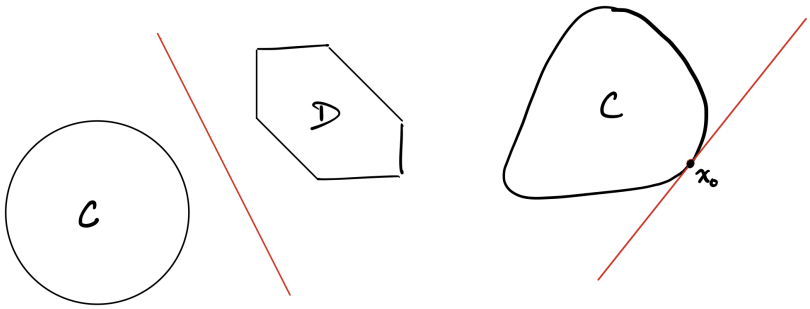

Figure 4: A seperating hyperplane (left) and a supporting hyperplane (right).

The supporting hyperplane theorem directly follows. Suppose C is a convex set and x_0 is a point in the boundary of C. Then apply the separating hyperplane theorem to $S_1 = \{x_0\}$ and S_2, the interior of C. This gives a hyperplane that is tangent to C at x_0 and defines a halfspace that contains C. (Recall the connection between convex sets and the intersection of halfspaces.)

2 Convex Functions

Recall from the previous lecture that a function $f : \mathbf{R}^n \to \mathbf{R}$ is convex if its domain is a convex set and for all x, $y \in \mathbf{dom}\, f$ and all $\lambda \in [0, 1]$, we have

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y). \tag{2}$$

Geometrically, this mean that the line segment connecting the points $(x, f(x))$ and $(y, f(y))$ lies above the graph of f. In other words, the mixture of the endpoints is greater than the function evaluated at the mixture of the points. A function is strictly convex if (2) holds with strict inequality. A function is strongly convex with parameter $m > 0$ if $f - (m/2)\|x\|_2^2$ is convex. A function is *concave* if $-f$ is convex. It often simplifies the notation to extend a convex function to all of $\mathbf{R}^n$ by defining its value to be ∞ outside of the domain:

$$\tilde{f}(x) = \begin{cases} f(x) & x \in \mathbf{dom}\, f \\ \infty & \text{otherwise} \end{cases}$$

We call $\tilde{f}$ the *extended-value extension* of f. We will see some more general notions of convexity later in the course, which allow us to tackle broader classes of optimization problems.

First order condition for convexity. If f is differentiable, then f is convex if and only if for all x and y in $\mathbf{dom}\, f$, we have

$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

This condition says that the first order approximation (Taylor expansion) of f lies below the graph of f, *i.e.,* it is a global underestimator of f. In unconstrained convex optimization, this condition immediately tells us that if $\nabla f(x) = 0$, then x minimizes $f(x)$.

Second order condition for convexity. If f is twice differentiable, then f is convex if and only if for all x in $\mathbf{dom}\, f$, we have

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

For simple functions, this condition often provides an easy proof of convexity. For example, consider the least-squares function

$$f(x) = \|Ax - b\|_2^2 = x^T A^T A x - 2b^T A x + b^T b.$$

The Hessian is $\nabla^2 f(x) = 2A^T A$. Since $A^T A$ is positive semidefinite, f is convex.

Simple examples. For one-dimensional functions, we can easily check for convexity by looking at the graph of f. Some examples of convex functions where $x \in \mathbf{R}$ are

- Affine functions $f(x) = ax + b$ (these are both convex and concave).
- Exponential $f(x) = e^{ax}$ for any $a \in \mathbf{R}$.
- Powers $f(x) = x^p$ on $\mathbf{R}_{++}$ for $p \notin [0, 1]$ (it's concave for $p \in [0, 1]$).
- Powers of absolute value $f(x) = |x|^p$ for $p \geq 1$.
- Negative entropy $f(x) = -x \log x$ for $x \in \mathbf{R}_+$. ($f(x) = \log x$ is concave.)

Some examples of convex function with $x \in \mathbf{R}^n$ are

- Norms, *e.g.,* $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$, $\|x\|_1 = \sum_{i=1}^n |x_i|$, $\|x\|_\infty = \max_i |x_i|$.
- Max function $f(x) = \max\{x_1, \ldots, x_n\}$.
- Log-sum-exp $f(x) = \log \sum_i e^{x_i}$. (Sometimes called the softmax function.)
- Quadratic functions $f(x) = x^T P x + q^T x + r$ for $P \succeq 0$.

- Quadratic over linear function $f(x) = x^2/y$ for $y \in \mathbf{R}_{++}$.
- The indicator function of a convex set, $I_C(x) = \begin{cases} 0 & x \in C \\ \infty & \text{otherwise} \end{cases}$.

Some examples of concave functions are

- Min function $f(x) = \min\{x_1, \ldots, x_n\}$.
- Log determinant $f(x) = \log\det(X)$ for $X \succeq 0$.
- Geometric mean $f(x) = \left(\prod_{i=1}^n x_i\right)^{1/n}$.
- Log CDF of a Gaussian $f(x) = \log \Phi(x)$ for $x \in \mathbf{R}$.

Our aim will be to develop a calculus of convex functions using these atoms and convexity-preserving operations.

The epigraph. The epigraph of a function f is defined as

$$\mathbf{epi}\, f = \{(x,t) \in \mathbf{R}^n \times \mathbf{R} : f(x) \leq t\},$$

which is the set of all points above the function. This construction connects convex sets with convex functions. A function f is convex if and only if its epigraph is a convex set. In addition, the alpha sublevel sets of a convex function

$$S_\alpha = \{x \in \mathbf{R}^n : f(x) \leq \alpha\}$$

are convex sets. (The converse is not true.)

Jensen's inequality. Jensen's inequality is simply a generalization of (2) to expectations. It states that for a convex function f and random variable X,

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

It includes the standard definition as a special case: consider a random variable that takes value x with probability λ and y with probability $1 - \lambda$.

Verifying convexity. We just saw many ways to verify the convexity of a function. We can directly apply the definition. We can try to equivalently show that f restricted to any line is convex. We can show the Hessian is positive semidefinite. Or we can show that f is obtained by composing simple convex functions with operations that preserve convexity. For the remainder of the course, we will concentrate on the last technique, which forms the basis of *disciplined convex programming* (DCP).

2.1 Operations that preserve convexity

For the most part, we will identify convex functions by building them from simple functions with known convexity (or concavity) and calculus rules. Again, we can think of a convex function as an expression tree with leaves as atoms (simple functions) with known convexity and nodes as operations that modify convexity in known ways.

Simple rules. The following rules preserve convexity and can be directly verified:

- **Nonnegative scaling:** f convex, $\alpha \geq 0 \implies \alpha f$ convex.
- **Sum:** f, g convex $\implies f + g$ convex.
- **Affine composition:** f convex $\implies f(Ax + b)$ convex.

The properties above immediately imply that any regularized norm minimization problem, which has the form

$$\text{minimize} \quad \|Ax - b\| + \lambda\|x\|,$$

is convex for any choice of the norms and $\lambda \geq 0$.

Pointwise maximum and supremum. A less obvious rule is that if $f_1, \ldots, f_m$ are convex functions, then $f(x) = \max\{f_1(x), \ldots, f_m(x)\}$ is convex. Examples of convex functions built from this rule include piecewise linear functions of the form

$$f(x) = \max_{i=1,\ldots,m} (a_i^T x + b_i),$$

and the sum of the largest k components of a vector,

$$f(x) = x_{[1]} + x_{[2]} + \cdots + x_{[k]},$$

where $x_{[i]}$ is the ith largest component of x. The second example follows from considering the $\binom{n}{k}$ linear functions that order each unique set of k components of x as the first k components. The pointwise supremum is also convex: if $f(x, y)$ is convex in x for each $y \in \mathcal{A}$, then the function

$$g(x) = \sup_{y \in \mathcal{A}} f(x, y)$$

is convex. An example built from this rule is the maximum eigenvalue of a symmetric matrix. which we can write as

$$\lambda_{\max}(X) = \sup_{\|y\|_2 = 1} y^T X y.$$

This function is linear in X for each fixed y, so the pointwise supremum is convex.

Composition. Consider the function $f(x) = h(g(x)) = h(g_1(x), \ldots, g_k(x))$ for $g : \mathbf{R}^n \to \mathbf{R}^k$ and $h : \mathbf{R}^k \to \mathbf{R}$. We can show that f is convex if

- g_i convex, h convex, $\tilde{h}$ nondecreasing in each argument.
- g_i concave, h convex, $\tilde{h}$ nonincreasing in each argument.

Similarly, f is concave if

- g_i concave, h concave, $\tilde{h}$ nondecreasing in each argument.
- g_i convex, h concave, $\tilde{h}$ nonincreasing in each argument.

In the scalar case, we can directly prove this by examining the second derivative

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x).$$

When we can deduce the sign, we can deduce convexity. This rule forms the basis of disciplined convex programming (DCP). In fact, you can get away with only knowing this rule; we can use it to derive all the other rules. For example, since max is a convex, increasing function, the pointwise maximum of convex functions is convex. This rule also illustrates the importance of how we parse a function. For example, consider

$$f(x) = \log \sum_{i=1}^{m} \exp(g_i(x))$$

where g_i are convex. We cannot parse this from the inside; we have to parse it with h as the entire log-sum-exp function.

Partial minimization. This rule will come up again when we examine duality theory. If $f(x, y)$ is convex in (x, y) and C is a convex set, then the function

$$g(x) = \inf_{y \in C} f(x, y)$$

is convex. This type of minimization also forms the basis of dynamic programming. Examples include partial minimization of a convex quadratic (related to the Schur complement) and the distance to a convex set

$$\mathbf{dist}(x, S) = \inf_{y \in S} \|x - y\|_2,$$

which is convex if S is convex.

Perspective. Like in the case of convex sets, the projective transformation on the epigraph of f preserves convexity. Specifically,

$$g(x,t) = tf(x/t)$$

is convex if f is convex. An important example is the relative entropy function

$$g(x,t) = -t\log(x/t) = t\log t - t\log x,$$

which is convex because $f(x) = -\log x$ is convex. From the convexity of g, we can establish the convexity of the *Kullback-Leiber divergence*

$$D_{\mathrm{KL}}(p,q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}.$$

The perspective transformation comes up often and in surprising places.

2.2 Disciplined Convex Programming (DCP)

DCP provides a constructive proof of convexity for a function, which is represented as an expression tree. Each leaf is a variable or a constant, and nodes are functions with known convexity, monotonicity, and sign properties. Convexity of the function is verified via composition rules[3]. Importantly, DCP verification is a sufficient but not necessary condition for convexity. It's easy to build a DCP parser from the information in this lecture. Tools like `Convex.jl` also include a *canonicalizer* that transforms a DCP into a form that can be accepted by a solver (or in the case of `Convex.jl`, into a form that is accepted by the modeling framework `JuMP.jl`[4]).

Examples. We consider two examples[5] of DCP analysis. You will work through several more on your homework. First we show that for $x < 1$, $y < 1$, the function

$$f(x,y) = \frac{(x-y)^2}{1 - \max\{x,y\}}$$

is convex. We begin by drawing out the expression tree.

- The leaves x, y, 1 are affine expressions.
- The function max is convex; the function $x - y$ is affine.
- The function $1 - \max$ is concave.
- The function u^2/v is convex, monotone decreasing in v for $v > 0$.

[3]Check out `https://dcp.stanford.edu/rules`

[4]Technically speaking, `Convex.jl` directly uses `MathOptInterface.jl`, the backend for `JuMP.jl`.

[5]These examples are taken from `https://web.stanford.edu/~boyd/papers/pdf/cvx_dcp.pdf`

- Hence, the function is convex.

As a second example, consider the function

$$f(x) = \sqrt{1 + x^2}.$$

Here, applying the composition rules directly does not work. However, if we recognize that

$$f(x) = \sqrt{1 + x^2} = \left\| \begin{bmatrix} 1 \\ x \end{bmatrix} \right\|_2,$$

then the function is clearly convex. Often, you will have to apply some creativity to find a suitable representation of the function for DCP verification.

From DCP to optimization. This lecture is necessarily somewhat dry for us to build up the machinery necessary to tackle convex optimization problems. Recall that a convex optimization problem has the form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b, \end{array}$$

where all the f_i's are convex. From the calculus developed above, we know that this problem is the minimization of a convex function over a convex set. DCP provides a means for a computer to verify that each of these functions is indeed convex and transform the problem into a form that can be solved by available convex optimization software. Importantly, for these problems, we have that

1. All locally optimal points are globally optimal.
2. An optimal point can be found efficiently.

An aside on tractability. It is important that a optimization problem is not only convex but also *tractable*. For example, consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & x \in \{0, 1\}^n. \end{array}$$

The constraints are clearly not convex, but we can rewrite this problem as

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & t \geq f(x), \quad \text{for all } x \in \{0, 1\}^n. \end{array}$$

This problem is a single variable convex optimization problem, but it has 2^n constraints, so it is not very useful. To write the problem, we have to evaluate f at every point in the set $\{0, 1\}^n$! Roughly speaking, we want the number of constraints m to be polynomial in n for tractability. Of course, a problem often has multiple equivalent representations and heuristics like convex relaxations can work very well in practice, which is why optimization can be somewhat of an art.

Acknowledgements

Thank you to John Drago and Hye Woong (Alex) Jeon for helpful edits.

References

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.