Convexity II: Optimization Basics

Ryan Tibshirani
Convex Optimization 10-725

See supplements for reviews of

- basic multivariate calculus
- basic linear algebra
Last time: convex sets and functions

“Convex calculus” makes it easy to check convexity. Tools:

- Definitions of convex sets and functions, classic examples
- Key properties (e.g., first- and second-order characterizations for functions)
- Operations that preserve convexity (e.g., affine composition)

E.g., is \( \max \left\{ \log \left( \frac{1}{(a^T x + b)^7} \right), \|Ax + b\|_1^5 \right\} \) convex?
Today:

- Optimization terminology
- Properties and first-order optimality
- Equivalent transformations
Optimization terminology

Reminder: a convex optimization problem (or program) is

$$\min_{x \in D} f(x)$$
$$\text{subject to } g_i(x) \leq 0, \quad i = 1, \ldots, m$$
$$Ax = b$$

where $f$ and $g_i$, $i = 1, \ldots, m$ are all convex, and the optimization domain is $D = \text{dom}(f) \cap \bigcap_{i=1}^m \text{dom}(g_i)$ (often we do not write $D$)

- $f$ is called criterion or objective function
- $g_i$ is called inequality constraint function
- If $x \in D$, $g_i(x) \leq 0$, $i = 1, \ldots, m$, and $Ax = b$ then $x$ is called a feasible point
- The minimum of $f(x)$ over all feasible points $x$ is called the optimal value, written $f^*$
• If \( x \) is feasible and \( f(x) = f^* \), then \( x \) is called optimal; also called a solution, or a minimizer\(^1\)

• If \( x \) is feasible and \( f(x) \leq f^* + \epsilon \), then \( x \) is called \( \epsilon \)-suboptimal

• If \( x \) is feasible and \( g_i(x) = 0 \), then we say \( g_i \) is active at \( x \)

• Convex minimization can be reposed as concave maximization

\[
\begin{align*}
\min_x f(x) & \quad \text{max}_x & -f(x) \\
\text{subject to} & \quad g_i(x) \leq 0, & \quad i = 1, \ldots, m \\
& \quad Ax = b & \quad \iff \quad \text{subject to} \quad g_i(x) \leq 0, & \quad i = 1, \ldots, m \\
& \quad Ax = b
\end{align*}
\]

Both are called convex optimization problems

\(^1\)Note: a convex optimization problem need not have solutions, i.e., need not attain its minimum, but we will not be careful about this
Solution set

Let $X_{\text{opt}}$ be the set of all solutions of convex problem, written

$$X_{\text{opt}} = \arg\min \quad f(x)$$
subject to $g_i(x) \leq 0$, $i = 1, \ldots, m$

$Ax = b$

Key property: $X_{\text{opt}}$ is a **convex set**

Proof: use definitions. If $x, y$ are solutions, then for $0 \leq t \leq 1$,

- $g_i(tx + (1 - t)y) \leq tg_i(x) + (1 - t)g_i(y) \leq 0$
- $A(tx + (1 - t)y) = tAx + (1 - t)Ay = b$
- $f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y) = f^*$

Therefore $tx + (1 - t)y$ is also a solution

Another key property: if $f$ is strictly convex, then the solution is **unique**, i.e., $X_{\text{opt}}$ contains one element
Example: lasso

Given \( y \in \mathbb{R}^n, \ X \in \mathbb{R}^{n \times p} \), consider the lasso problem:

\[
\min_{\beta} \quad \|y - X \beta\|_2^2 \\
\text{subject to} \quad \|\beta\|_1 \leq s
\]

Is this convex? What is the criterion function? The inequality and equality constraints? Feasible set? Is the solution unique, when:

- \( n \geq p \) and \( X \) has full column rank?
- \( p > n \) ("high-dimensional" case)?

How do our answers change if we changed criterion to Huber loss:

\[
\sum_{i=1}^{n} \rho(y_i - x_i^T \beta), \quad \rho(z) = \begin{cases} 
\frac{1}{2}z^2 & |z| \leq \delta \\
\delta |z| - \frac{1}{2} \delta^2 & \text{else}
\end{cases}
\]
Example: support vector machines

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$ with rows $x_1, \ldots x_n$, consider the support vector machine or SVM problem:

$$
\min_{\beta, \beta_0, \xi} \quad \frac{1}{2} \|\beta\|_2^2 + C \sum_{i=1}^{n} \xi_i
$$

subject to $\xi_i \geq 0, \ i = 1, \ldots n$

$$
y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots n
$$

Is this convex? What is the criterion, constraints, feasible set? Is the solution $(\beta, \beta_0, \xi)$ unique? What if changed the criterion to

$$
\frac{1}{2} \|\beta\|_2^2 + \frac{1}{2} \beta_0^2 + C \sum_{i=1}^{n} \xi_i^{1.01}?
$$

For original criterion, what about $\beta$ component, at the solution?
Local minima are global minima

For a convex problem, a feasible point $x$ is called \textit{locally optimal} is there is some $R > 0$ such that

$$f(x) \leq f(y) \quad \text{for all feasible } y \text{ such that } \|x - y\|_2 \leq R$$

Reminder: for convex optimization problems, \textit{local optima are global optima}

Proof simply follows from definitions

\begin{figure}
\centering
\includegraphics[width=\textwidth]{convex_nonconvex.png}
\caption{Convex vs Nonconvex}
\end{figure}
Rewriting constraints

The optimization problem

$$\min_x f(x) \quad \text{subject to} \quad g_i(x) \leq 0, \quad i = 1, \ldots, m$$
$$Ax = b$$

can be rewritten as

$$\min_x f(x) \quad \text{subject to} \quad x \in C$$

where $C = \{x : g_i(x) \leq 0, \quad i = 1, \ldots, m, \quad Ax = b\}$, the feasible set. Hence the latter formulation is completely general.

With $I_C$ the indicator of $C$, we can write this in unconstrained form

$$\min_x f(x) + I_C(x)$$
First-order optimality condition

For a convex problem

\[
\min_x f(x) \text{ subject to } x \in C
\]

and differentiable \( f \), a feasible point \( x \) is optimal if and only if

\[
\nabla f(x)^T (y - x) \geq 0 \quad \text{for all } y \in C
\]

This is called the first-order condition for optimality

In words: all feasible directions from \( x \) are aligned with gradient \( \nabla f(x) \)

Important special case: if \( C = \mathbb{R}^n \) (unconstrained optimization), then optimality condition reduces to familiar \( \nabla f(x) = 0 \)
Example: quadratic minimization

Consider minimizing the quadratic function

\[ f(x) = \frac{1}{2} x^T Q x + b^T x + c \]

where \( Q \succeq 0 \). The first-order condition says that solution satisfies

\[ \nabla f(x) = Qx + b = 0 \]

- if \( Q \succ 0 \), then there is a unique solution \( x = -Q^{-1} b \)
- if \( Q \) is singular and \( b \notin \text{col}(Q) \), then there is no solution (i.e., \( \min_x f(x) = -\infty \))
- if \( Q \) is singular and \( b \in \text{col}(Q) \), then there are infinitely many solutions
  \[ x = -Q^+ b + z, \quad z \in \text{null}(Q) \]

where \( Q^+ \) is the pseudoinverse of \( Q \).
Example: equality-constrained minimization

Consider the equality-constrained convex problem:

\[ \min_x f(x) \quad \text{subject to} \quad Ax = b \]

with \( f \) differentiable. Let’s prove Lagrange multiplier optimality condition

\[ \nabla f(x) + A^T u = 0 \quad \text{for some} \quad u \]

According to first-order optimality, solution \( x \) satisfies \( Ax = b \) and

\[ \nabla f(x)^T (y - x) \geq 0 \quad \text{for all} \quad y \text{ such that} \quad Ay = b \]

This is equivalent to

\[ \nabla f(x)^T v = 0 \quad \text{for all} \quad v \in \text{null}(A) \]

Result follows because \( \text{null}(A)^\perp = \text{row}(A) \)
Example: projection onto a convex set

Consider projection onto convex set $C$:

$$
\min_x \|a - x\|_2^2 \text{ subject to } x \in C
$$

First-order optimality condition says that the solution $x$ satisfies

$$
\nabla f(x)^T (y - x) = (x - a)^T (y - x) \geq 0 \text{ for all } y \in C
$$

Equivalently, this says that

$$
a - x \in \mathcal{N}_C(x)
$$

where recall $\mathcal{N}_C(x)$ is the normal cone to $C$ at $x$
Partial optimization

Reminder: \( g(x) = \min_{y \in C} f(x, y) \) is convex in \( x \), provided that \( f \) is convex in \((x, y)\) and \( C \) is a convex set

Therefore we can always partially optimize a convex problem and retain convexity

E.g., if we decompose \( x = (x_1, x_2) \in \mathbb{R}^{n_1+n_2} \), then

\[
\begin{align*}
\min_{x_1, x_2} & \quad f(x_1, x_2) \\
\text{subject to} & \quad g_1(x_1) \leq 0, \\ & \quad g_2(x_2) \leq 0
\end{align*}
\]

\[\iff\]

\[
\begin{align*}
\min_{x_1} & \quad \tilde{f}(x_1) \\
\text{subject to} & \quad g_1(x_1) \leq 0
\end{align*}
\]

where \( \tilde{f}(x_1) = \min \{ f(x_1, x_2) : g_2(x_2) \leq 0 \} \). The right problem is convex if the left problem is
Example: hinge form of SVMs

Recall the SVM problem

\[
\min_{\beta, \beta_0, \xi} \quad \frac{1}{2} \|\beta\|_2^2 + C \sum_{i=1}^{n} \xi_i
\]

subject to \( \xi_i \geq 0, \ y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots, n \)

Rewrite the constraints as \( \xi_i \geq \max\{0, 1 - y_i(x_i^T \beta + \beta_0)\} \). Indeed we can argue that we have = at solution

Therefore plugging in for optimal \( \xi \) gives the hinge form of SVMs:

\[
\min_{\beta, \beta_0} \quad \frac{1}{2} \|\beta\|_2^2 + C \sum_{i=1}^{n} [1 - y_i(x_i^T \beta + \beta_0)]_+
\]

where \( a_+ = \max\{0, a\} \) is called the hinge function
Transformations and change of variables

If \( h : \mathbb{R} \rightarrow \mathbb{R} \) is a monotone increasing transformation, then

\[
\min_x f(x) \quad \text{subject to} \quad x \in C
\]

\[\iff\]

\[
\min_x h(f(x)) \quad \text{subject to} \quad x \in C
\]

Similarly, inequality or equality constraints can be transformed and yield equivalent optimization problems. Can use this to reveal the “hidden convexity” of a problem.

If \( \phi : \mathbb{R}^n \rightarrow \mathbb{R}^m \) is one-to-one, and its image covers feasible set \( C \), then we can change variables in an optimization problem:

\[
\min_x f(x) \quad \text{subject to} \quad x \in C
\]

\[\iff\]

\[
\min_y f(\phi(y)) \quad \text{subject to} \quad \phi(y) \in C
\]
Example: geometric programming

A monomial is a function \( f : \mathbb{R}_+^n \rightarrow \mathbb{R} \) of the form

\[
f(x) = \gamma x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}
\]

for \( \gamma > 0, a_1, \ldots a_n \in \mathbb{R} \). A posynomial is a sum of monomials,

\[
f(x) = \sum_{k=1}^{p} \gamma_k x_1^{a_{k1}} x_2^{a_{k2}} \cdots x_n^{a_{kn}}
\]

A geometric program is of the form

\[
\begin{align*}
\min_x & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 1, \ i = 1, \ldots m \\
& \quad h_j(x) = 1, \ j = 1, \ldots r
\end{align*}
\]

where \( f, g_i, i = 1, \ldots m \) are posynomials and \( h_j, j = 1, \ldots r \) are monomials. This is nonconvex
Given \( f(x) = \gamma x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n} \), let \( y_i = \log x_i \) and rewrite this as

\[
\gamma (e^{y_1})^{a_1} (e^{y_2})^{a_2} \cdots (e^{y_n})^{a_n} = e^{a^T y + b}
\]

for \( b = \log \gamma \). Also, a posynomial can be written as \( \sum_{k=1}^{p} e^{a_k^T y + b_k} \).

With this variable substitution, and after taking logs, a geometric program is equivalent to

\[
\min_{x} \log \left( \sum_{k=1}^{p_0} e^{a_{0k}^T y + b_{0k}} \right)
\]

subject to \( \log \left( \sum_{k=1}^{p_i} e^{a_{ik}^T y + b_{ik}} \right) \leq 0, \ i = 1, \ldots m \)

\[
c_j^T y + d_j = 0, \ j = 1, \ldots r
\]

This is convex, recalling the convexity of soft max functions.
Several interesting problems are geometric programs, e.g., floor planning:

\[ W \quad H \]
\[ h_i \quad w_i \quad (x_i, y_i) \]
\[ C_i \quad h_i \]

Figure 8.18 Floor planning problem. Non-overlapping rectangular cells are placed in a rectangle with width \( W \), height \( H \), and lower left corner at \((0, 0)\).

We also require that the cells do not overlap, except possibly on their boundaries: \( \text{int}(C_i \cap C_j) = \emptyset \) for \( i \neq j \). (It is also possible to require a positive minimum clearance between the cells.) The non-overlap constraint \( \text{int}(C_i \cap C_j) = \emptyset \) holds if and only if for \( i \neq j \),

- \( C_i \) is left of \( C_j \),
- \( C_i \) is right of \( C_j \),
- \( C_i \) is below \( C_j \),
- \( C_i \) is above \( C_j \).

These four geometric conditions correspond to the inequalities

\[
\begin{align*}
  x_i + w_i &\leq x_j, \\
  x_j + w_j &\leq x_i, \\
  y_i + h_i &\leq y_j, \\
  y_j + h_j &\leq y_i,
\end{align*}
\]

(8.32) at least one of which must hold for each \( i \neq j \). Note the combinatorial nature of these constraints: for each pair \( i \neq j \), at least one of the four inequalities above must hold.

8.8.1 Relative positioning constraints

The idea of relative positioning constraints is to specify, for each pair of cells, one of the four possible relative positioning conditions, i.e., left, right, above, or below. One simple method to specify these constraints is to give two relations on \( \{1, \ldots, N\} \): \( L \) (meaning 'left of') and \( B \) (meaning 'below'). We then impose the constraint that \( C_i \) is to the left of \( C_j \) if \((i, j) \in L\), and \( C_i \) is below \( C_j \) if \((i, j) \in B\).

This yields the constraints

\[
\begin{align*}
  x_i + w_i &\leq x_j \quad \text{for } (i, j) \in L, \\
  y_i + h_i &\leq y_j \quad \text{for } (i, j) \in B,
\end{align*}
\]

(8.33)

See Boyd et al. (2007), “A tutorial on geometric programming”, and also Chapter 8.8 of B & V book.
Eliminating equality constraints

Important special case of change of variables: eliminating equality constraints. Given the problem

\[
\begin{align*}
\min_x & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0, \quad i = 1, \ldots, m \\
& \quad Ax = b
\end{align*}
\]

we can always express any feasible point as \( x = My + x_0 \), where \( Ax_0 = b \) and \( \text{col}(M) = \text{null}(A) \). Hence the above is equivalent to

\[
\begin{align*}
\min_y & \quad f(My + x_0) \\
\text{subject to} & \quad g_i(My + x_0) \leq 0, \quad i = 1, \ldots, m
\end{align*}
\]

Note: this is fully general but not always a good idea (practically)
Introducing slack variables

Essentially opposite to eliminating equality contraints: introducing slack variables. Given the problem

\[
\begin{align*}
\min_{x} & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0, \ i = 1, \ldots, m \\
& \quad Ax = b
\end{align*}
\]

we can transform the inequality constraints via

\[
\begin{align*}
\min_{x,s} & \quad f(x) \\
\text{subject to} & \quad s_i \geq 0, \ i = 1, \ldots, m \\
& \quad g_i(x) + s_i = 0, \ i = 1, \ldots, m \\
& \quad Ax = b
\end{align*}
\]

Note: this is no longer convex unless \( g_i, \ i = 1, \ldots, n \) are affine
Relaxing nonaffine equalities

Given an optimization problem

\[
\min_x f(x) \text{ subject to } x \in C
\]

we can always take an enlarged constraint set \( \tilde{C} \supseteq C \) and consider

\[
\min_x f(x) \text{ subject to } x \in \tilde{C}
\]

This is called a relaxation and its optimal value is always smaller or equal to that of the original problem.

Important special case: relaxing nonaffine equality constraints, i.e.,

\[ h_j(x) = 0, \quad j = 1, \ldots, r \]

where \( h_j, j = 1, \ldots, r \) are convex but nonaffine, are replaced with

\[ h_j(x) \leq 0, \quad j = 1, \ldots, r \]
Example: maximum utility problem

The maximum utility problem models investment/consumption:

\[
\max_{x, b} \sum_{t=1}^{T} \alpha_t u(x_t)
\]

subject to \( b_{t+1} = b_t + f(b_t) - x_t, \ t = 1, \ldots, T \)

\( 0 \leq x_t \leq b_t, \ t = 1, \ldots, T \)

Here \( b_t \) is the budget and \( x_t \) is the amount consumed at time \( t \); \( f \) is an investment return function, \( u \) utility function, both concave and increasing.

Is this a convex problem? What if we replace equality constraints with inequalities:

\( b_{t+1} \leq b_t + f(b_t) - x_t, \ t = 1, \ldots, T? \)
Example: principal components analysis

Given $X \in \mathbb{R}^{n \times p}$, consider the low rank approximation problem:

$$\min_{R} \|X - R\|_{F}^{2} \quad \text{subject to} \quad \text{rank}(R) = k$$

Here $\|A\|_{F}^{2} = \sum_{i=1}^{n} \sum_{j=1}^{p} A_{ij}^{2}$, the entrywise squared $\ell_2$ norm, and $\text{rank}(A)$ denotes the rank of $A$.

Also called principal components analysis or PCA problem. Given $X = UDV^{T}$, singular value decomposition or SVD, the solution is

$$R = U_{k}D_{k}V_{k}^{T}$$

where $U_{k}, V_{k}$ are the first $k$ columns of $U, V$ and $D_{k}$ is the first $k$ diagonal elements of $D$. I.e., $R$ is reconstruction of $X$ from its first $k$ principal components.
The PCA problem is not convex. Let’s recast it. First rewrite as

$$\min_{Z \in S^p} \|X - XZ\|_F^2 \text{ subject to } \text{rank}(Z) = k, \ Z \text{ is a projection}$$

$$\iff \max_{Z \in S^p} \text{tr}(SZ) \text{ subject to } \text{rank}(Z) = k, \ Z \text{ is a projection}$$

where \( S = X^TX \). Hence constraint set is the nonconvex set

$$C = \left\{ Z \in S^p : \lambda_i(Z) \in \{0, 1\}, \ i = 1, \ldots p, \ \text{tr}(Z) = k \right\}$$

where \( \lambda_i(Z) \), \( i = 1, \ldots n \) are the eigenvalues of \( Z \). Solution in this formulation is

$$Z = V_k V_k^T$$

where \( V_k \) gives first \( k \) columns of \( V \)
Now consider relaxing constraint set to $\mathcal{F}_k = \text{conv}(C)$, its convex hull. Note

$$\mathcal{F}_k = \{Z \in \mathbb{S}^p : \lambda_i(Z) \in [0, 1], i = 1, \ldots p, \text{tr}(Z) = k\}$$

$$= \{Z \in \mathbb{S}^p : 0 \preceq Z \preceq I, \text{tr}(Z) = k\}$$

This set is called the Fantope of order $k$. It is convex. Hence, the linear maximization over the Fantope, namely

$$\max_{Z \in \mathcal{F}_k} \text{tr}(SZ)$$

is a convex problem. Remarkably, this is equivalent to the original nonconvex PCA problem (admits the same solution)!

(Famous result: Fan (1949), “On a theorem of Weyl concerning eigenvalues of linear transformations”)
References and further reading