A Brisk Overview of Convex Optimization

Looking at Figures That Took Too Long to Make.

Conner DiPaolo

Harvey Mudd College / Yelp

August, 2016

Table of Contents

Convex Optimization

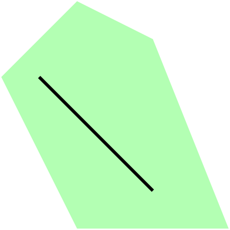
1. Core behind techniques in machine learning, signal processing, operations research, etc.
2. Can be very applied or very theoretical
3. If you need more resources check out *Convex Optimization* by Boyd and Vandenberghe[1].

[1] `http://stanford.edu/~boyd/cvxbook/`

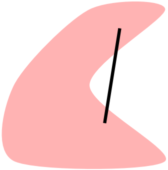
Table of Contents

Convex Sets

Convex Set



Non-Convex Set



Convex Sets - an Intuitive Definition

A set C is *convex* if, given any two points in that set, every point on the line segment between those two points is also in C.

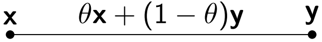
Convex Sets

Definition (Convex Combination / Line Segment)

A convex combination of two points **x** and **y** from an affine space is

$$\theta \mathbf{x} + (1-\theta)\mathbf{y},$$

where $0 \leq \theta \leq 1$.



Convex Sets - an Actual Definition

A set C is *convex* if, given any two points in that set, every point on the line segment between those two points is also in C. Mathematically, we have

Definition (Convex Set)

A set C is convex if, given $\mathbf{x}, \mathbf{y} \in C$, every convex combination of $\mathbf{x}$ and $\mathbf{y}$ still lies in C. That is,

$$\theta\mathbf{x} + (1-\theta)\mathbf{y} \in C. \qquad (0 \leq \theta \leq 1)$$

for all $\mathbf{x}$ and $\mathbf{y}$ in C.

Convex Sets

Example (The Space of Probability Distributions)

Let $\mathcal{P}$ be the space of continuous probability distributions over $\mathbb{R}^n$. That is, every element of $\mathcal{P}$ defines a unique probability density function $\mathbb{P}(x) \geq 0$ such that

$$\int_{\mathbb{R}^n} \mathbb{P}(x) dx = 1.$$

$\mathcal{P}$ is convex.

Convex Sets

Proof.

Let f and h be valid probability distributions from $\mathcal{P}$. That is,

$$f, h \in \left\{ \mathbb{P}(x) : \int_{\mathbb{R}^n} \mathbb{P}(x) dx = 1 \text{ and } \mathbb{P}(x) \geq 0 \right\}.$$

Now let $0 \leq \theta \leq 1$. Then

$$\theta f(x) + (1-\theta) h(x) \geq 0$$

as a positive combination of positive functions. Similarly,

$$\begin{aligned}\int_{\mathbb{R}^n} \left[\theta f(x) + (1-\theta) h(x)\right] dx &= \theta \int_{\mathbb{R}^n} f(x) dx + (1-\theta) \int_{\mathbb{R}^n} h(x) dx \\ &= \theta + (1-\theta) = 1.\end{aligned}$$

$\blacksquare$

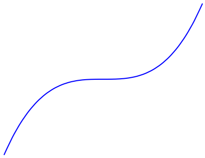
Table of Contents

Convex Functions

Convex Function

Concave Function

Neither



Convex Function

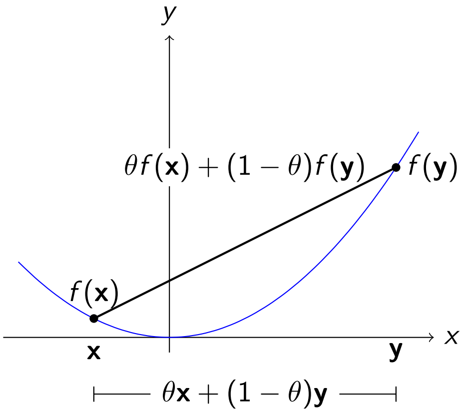
Definition (Convex Function)

A function $f : C \mapsto \mathbb{R}$ is *convex* if its domain C is convex and, for $0 \leq \theta \leq 1$, given any $\mathbf{x}, \mathbf{y} \in C$,

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}).$$

That is, a line segment drawn between two function values lies above the function.

Convex Function



Second Order Conditions

Theorem

A twice differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if and only if it's Hessian is positive semi-definite:

$$\nabla^2 f \succeq 0.$$

Concave and Convex Functions

Example

$f(x) = \log(x) : \mathbb{R}_{++} \mapsto \mathbb{R}$ is concave.

Proof.

$f''(x) = -\frac{1}{x^2} < 0$, so f is *strictly concave*. ■

Example

$f(x) = \exp(x) : \mathbb{R} \mapsto \mathbb{R}_{++}$ is convex.

Proof.

$f''(x) = e^x > 0$, so f is *strictly convex*. ■

Convex Function

Example

The function $f : \mathbb{R}^n \mapsto \mathbb{R}$ defined by

$$f(\mathbf{x}) = e^{\mathbf{x}^\top \mathbf{x}} = e^{\mathbf{x}_1^2 + \cdots + \mathbf{x}_n^2}$$

is convex.

Proof.

Each element of the Hessian

$$\nabla^2 f_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} = 4\mathbf{x}_i \mathbf{x}_j e^{\mathbf{x}^\top \mathbf{x}}$$
$$\nabla^2 f = 4\mathbf{x}\mathbf{x}^\top e^{\mathbf{x}^\top \mathbf{x}}.$$

Is this positive semi-definite? Consider $\mathbf{z} \in \mathbb{R}^n$. Then

$$\mathbf{z}^\top \nabla^2 f(x) \mathbf{z} = \mathbf{z}^\top 4\mathbf{x}\mathbf{x}^\top e^{\mathbf{x}^\top \mathbf{x}} \mathbf{z} = 4(\mathbf{z}^\top \mathbf{x})^2 e^{\mathbf{x}^\top \mathbf{x}} \geq 0$$

■

Convex Function

Example (Quadratic Functions)

The function $f : \mathbb{R}^n \mapsto \mathbb{R}$ defined by

$$f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$$

is convex if $A \succeq 0$.

Proof.

We know

$$\nabla^2 f = 2A.$$

Similarly we know that $2A \succeq 0$ if and only if $A \succeq 0$. Thus by the second order conditions f is convex if and only if $A \succeq 0$.
We can also show that f is concave if $A \preceq 0$. ∎

Affine Composition

Theorem

Given a convex $f : \mathbb{R}^m \mapsto \mathbb{R}$*, any matrix* $A \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$*, and* $\mathbf{b} \in \mathbb{R}^m$,

$$g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$$

is convex.

Proof.

$$\nabla^2 g = A^\top \nabla^2 f(A\mathbf{x} + \mathbf{b})A.$$

Then, for any $\mathbf{z} \in \mathbb{R}^n$,

$$\mathbf{z}^\top \nabla^2 g \mathbf{z} = \mathbf{z}^\top A^\top \nabla^2 f(A\mathbf{x} + \mathbf{b})A\mathbf{z} = (A\mathbf{z})^\top \nabla^2 f(A\mathbf{x} + \mathbf{b})(A\mathbf{z}) \geq 0$$

by the convexity assumption, so g is convex. ∎

Table of Contents

Optimization Problems

An optimization problem is a problem of the form

$$\begin{aligned} \text{minimize: } & f_0(\mathbf{x}) \\ \text{subj. to: } & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, p \end{aligned}$$

The goal of an optimization problem, as you might be able to guess, is to optimize f_0 where $\mathbf{x}$ is in the problem domain domain

$$\mathcal{D} = \bigcap_{i=0}^{m} \textbf{dom } f_i \, \cap \, \bigcap_{i=1}^{p} \textbf{dom } h_i,$$

satisfying the problem constraints.

Convex Optimization Problems

An optimization problem of the form

$$
\begin{aligned}
\text{minimize: } & f_0(\mathbf{x}) \\
\text{subj. to: } & f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\
& A\mathbf{x} = \mathbf{b}
\end{aligned}
$$

is convex if f_i for all $i = 0, \dots, m$ are convex.

Global Optimality of Convex Optimization Problems

Theorem (Global Optimality)

Given a convex optimization problem, and local optimum $\mathbf{x}^\star$ *such that*

$$f_0(\mathbf{x}) = \inf\{f_0(\mathbf{z}) : \mathbf{z} \textit{ feasible and } ||\mathbf{z} - \mathbf{x}||_2 \leq R\}$$

for some $R > 0$*.* $\mathbf{x}^\star$ *is the global optimum.*

Proof.

(Boyd) Suppose local optimality but not global optimality, with some feasible $\mathbf{y}$ such that $f_0(\mathbf{y}) < f_0(\mathbf{x})$. Then $||\mathbf{y} - \mathbf{x}||_2 > R$ because otherwise $f_0(\mathbf{x}) \leq f_0(\mathbf{y})$. Consider a point $\mathbf{z}$ given by

$$\mathbf{z} = (1-\theta)\mathbf{x} + \theta\mathbf{y} \quad \text{and} \quad \theta = \frac{R}{2||\mathbf{y} - \mathbf{x}||_2}.$$

Then $||\mathbf{x} - \mathbf{z}||_2 = R/2 < R$. By convexity of the feasible set, $\mathbf{z}$ is feasible. By the convexity of the objective function f_0 we have

$$f_0(\mathbf{z}) \leq (1-\theta) f_0(\mathbf{x}) + \theta f_0(\mathbf{y}) < f_0(\mathbf{x})$$

Common Problems - Linear Programs

$$
\begin{aligned}
\text{minimize: } & \mathbf{c}^\top \mathbf{x} \\
\text{subj. to: } & G\mathbf{x} \preceq \mathbf{h}, \\
& A\mathbf{x} = \mathbf{b}
\end{aligned}
$$

Common Problems - Quadratic Programs

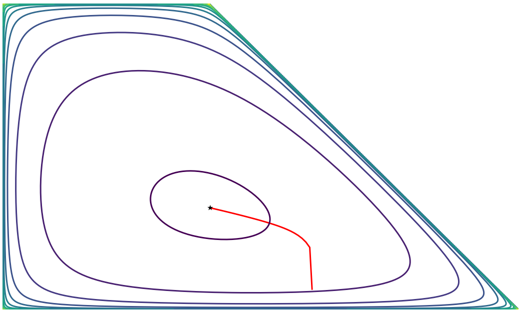
$$
\begin{aligned}
\text{minimize: } & \frac{1}{2}\mathbf{x}^\top P\mathbf{x} + \mathbf{q}^\top \mathbf{x} + \mathbf{r} \\
\text{subj. to: } & G\mathbf{x} \preceq \mathbf{h} \\
& A\mathbf{x} = \mathbf{b}
\end{aligned}
$$

with $P \in \mathbb{S}_+^n$

Example Problem - Analytic Centering

Want to find some sort of 'center' of a convex polygon (polytope) $A\mathbf{x} \preceq \mathbf{b}$:

$$\text{minimize: } f_0(\mathbf{x}) = -\sum_i \log(\mathbf{b}_i - a_i^\top \mathbf{x})$$



Example Problem - Least Squares Linear Regression

We have an overdetermined (tall A) system $A\mathbf{x} = \mathbf{b}$ we want to solve, and we want to find the "best" solution by minimizing

$f = ||A\mathbf{x} - \mathbf{b}||_2 = (A\mathbf{x} - \mathbf{b})^\top (A\mathbf{x} - \mathbf{b}) = \mathbf{x}^\top A^\top A\mathbf{x} - 2\mathbf{x}^\top A^\top \mathbf{b} + \mathbf{b}^\top \mathbf{b}.$

At optimality $\nabla f = 0$, so we can find

$$\nabla f = 2A^\top A\mathbf{x} - 2A^\top \mathbf{b} = \mathbf{0},$$

so

$$\mathbf{x}^\star = (A^\top A)^{-1} A^\top \mathbf{b}.$$

Example Problem - Logistic Regression

Have a dataset $\{\mathbf{x}_i, y_i\}_1^m$ with $y \in \{0, 1\}$. We want to predict

$$\hat{y} = \mathbb{P}(y_i = 1 | \mathbf{x}_i; \boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^\top \mathbf{x}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \mathbf{x})}.$$

Optimization problem: maximize the (log) likelihood of our data given the parameters $\boldsymbol{\theta}$:

$$\begin{aligned} \text{maximize:} \;\; \log \mathbb{P}(\mathcal{D} | \boldsymbol{\theta}) &= \sum_{i=1}^{m} y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}) + (1 - y_i) \log(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x})) \\ &= \sum_{i=1}^{m} \ell_i \end{aligned}$$

Example Problem - Logistic Regression

Is this convex? Note $\sigma'(x) = \sigma(x)\,[1 - \sigma(x)]$. Then

$$\begin{aligned}
\nabla \ell &= y\mathbf{x}\frac{\sigma(\boldsymbol{\theta}^\top \mathbf{x})\left[1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x})\right]}{\sigma(\boldsymbol{\theta}^\top \mathbf{x})} - (1-y)\mathbf{x}\frac{\sigma(\boldsymbol{\theta}^\top \mathbf{x})\left[1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x})\right]}{1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x})} \\
&= y\mathbf{x}\left[1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x})\right] - (1-y)\mathbf{x}\sigma(\boldsymbol{\theta}^\top \mathbf{x}) = \left[y - \sigma(\boldsymbol{\theta}^T \mathbf{x})\right]\mathbf{x}
\end{aligned}$$

and then

$$\nabla^2 \ell = -\sigma(\boldsymbol{\theta}^\top \mathbf{x})\left[1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x})\right]\mathbf{x}\mathbf{x}^\top \preceq 0$$

Table of Contents

Optimal Solutions

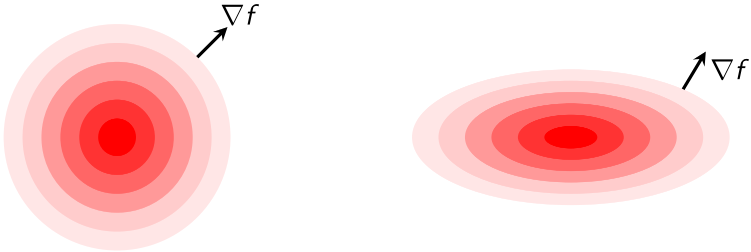
For convex unconstrained problems

$$\text{minimize: } f_0(\mathbf{x}) = f(\mathbf{x})$$

The optimal solution occurs when $\nabla f = \mathbf{0}$.

Gradient Descent

Dumb algorithm: start somewhere and take small steps down the hill.

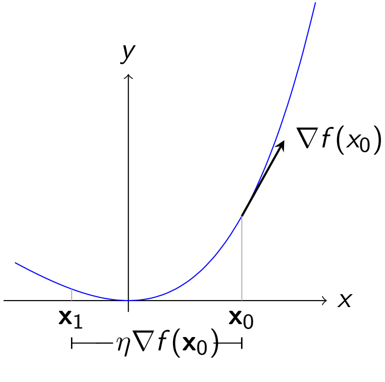


Gradient Descent

Gradient Descent
input : f, ∇f, $\eta(t)$, starting point $\mathbf{x}_0$, tolerance ϵ
output: optimal point $\mathbf{x}^\star$
$t \leftarrow 0$
while $||\nabla f||_2 \geq \epsilon$ **do**
    $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta(t)\nabla f(\mathbf{x}_t)$
    $t \leftarrow t + 1$
end
return $\mathbf{x}^\star = \mathbf{x}_{t+1}$

Newton's Method

Good algorithm: start somewhere, approximate f as a quadratic, and optimize that quadratic at every step



Newton's Method

Approximating f as a quadratic with it's Taylor Expansion:

$$f_{\mathbf{x}_0}(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

At optimality we have

$$\nabla f = \nabla f + \nabla^2 f(\mathbf{x} - \mathbf{x}_0) = 0,$$

so

$$\mathbf{x}^\star = \mathbf{x}_0 - \nabla^2 f^{-1} \nabla f.$$

Newton's Method

Newton's Method

input : f, ∇f, $\nabla^2 f$, starting point $\mathbf{x}_0$, tolerance ϵ

output: optimal point $\mathbf{x}^\star$

$t \leftarrow 0$

while $||\nabla f||_2 \geq \epsilon$ **do**

  solve $\nabla^2 f(\mathbf{x}_t)\mathbf{d}_t = \nabla f(\mathbf{x}_t)$ for $\mathbf{d}_t$

  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \mathbf{d}_t$

  $t \leftarrow t + 1$

end

return $\mathbf{x}^\star = \mathbf{x}_{t+1}$

Newton's Method is Good

$$
\begin{tikzcd}[column sep=huge, row sep=huge]
\text{minimize: } f(\mathbf{x}) \arrow[r, "\text{Linear Transformation}"] \arrow[d, "\text{Newton}"'] & \text{minimize: } f(T\mathbf{y}) \arrow[d] \\
\mathbf{x}_0, \mathbf{x}_1, \ldots \arrow[r] & T\mathbf{y}_0, T\mathbf{y}_1, \ldots
\end{tikzcd}
$$

Newton's Method is Scale Invariant

Proof.

(Induction) Assume $\mathbf{x}_t = A\mathbf{y}_t$. Recall at each step

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \nabla^2 f^{-1} \nabla f.$$

If we have $g(\mathbf{y}) = f(A\mathbf{y})$, then as before we have

$$\nabla g = A^T \nabla f(A\mathbf{y}) \qquad \text{and} \qquad \nabla^2 g = A^T \nabla^2 f(A\mathbf{y}) A$$

and then

$$\begin{aligned}
\mathbf{y}_{t+1} &= A^{-1}\mathbf{x}_t - (A^T \nabla^2 f(A\mathbf{x}_t) A)^{-1} A^T \nabla f(\mathbf{x}_t) \\
&= A^{-1}\mathbf{x}_t - A^{-1} \nabla^2 f(\mathbf{x}_t)^{-1} A^{-\top} A^T \nabla f(\mathbf{x}_t) \\
&= A^{-1} \left[\mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)\right] \\
A\mathbf{y}_{t+1} &= \mathbf{x}_{t+1}.
\end{aligned}$$

∎

Table of Contents