Foundations of Data Science, Fall 2020

7b. Optimisation I

Prof. Dan Olteanu

DaST Data•(Systems+Theory)

University of Zurich[UZH]

Oct 16, 2020

https://lms.uzh.ch/url/RepositoryEntry/16830890400

https://uzh.zoom.us/j/96690150974?pwd=cnZmMTduWUtCeWoxYW85Z3RMYnpTZz09

Solving Machine Learning Problems

Most machine learning methods can be cast as optimisation problems.

- So far: Closed-form solutions
  e.g., minimisation of least squares and ridge regression objectives
- Most interesting learning problems do not admit closed-form solutions :(

Two approaches to solving the problems beyond closed-form solutions:

1. Frame the objective of the ML problem as a mathematical problem

   Use existing blackbox solver for such problems

   When objectives can be formulated as convex optimisation problems

2. Gradient-based optimisation methods

   They are not blackbox: optimisation hyper-parameters affect performance

A Crash Course in Optimisation

Today:

- Convex optimisation

Next time:

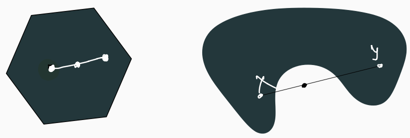
- Recap: Gradients, Hessians
- Gradient Descent
- Stochastic Gradient Descent
- Constrained optimisation

Most machine learning packages, e.g., scikit-learn, tensorflow, octave, torch, have optimisation methods readily implemented.

You need to understand the basics of optimisation to use them effectively.

Convex Sets

A set $C \subseteq \mathbb{R}^D$ is convex if for any $\mathbf{x}, \mathbf{y} \in C$ and $\lambda \in [0, 1]$, it holds $\lambda\,\mathbf{x} + (1-\lambda)\,\mathbf{y} \in C$



Examples of Convex Sets

- Set $\mathbb{R}^D$

  $\lambda\,\mathbf{x} + (1-\lambda)\,\mathbf{y} \in \mathbb{R}^D$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$

- Intersections of convex sets

  Given convex sets $C_1, \ldots, C_n$, the set $\bigcap_{i=1}^n C_i$ is convex

- Norm balls

  For any L-norm $||\cdot||$, the set $B = \{\mathbf{x} \in \mathbb{R}^D : ||\mathbf{x}|| \le 1\}$ is convex

- Polyhedra

  $$A \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} \overbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}^{x} \le \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

  Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$, the polyhedron $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\,\mathbf{x} \le \mathbf{b}\}$ is convex

- Positive semidefinite cone: The set of positive semi-definite matrices

Showing the Set of PSD Matrices is Convex

$A \in \mathbb{R}^{D \times D}$ is PSD : $\forall x \in \mathbb{R}^D$: $x^T A x \geq 0$.

To show: Any combination $\lambda A + (1-\lambda) B \in \underline{S_+^D}$ if $A, B \in S_+^D$

PSD cone.

$$x^T(\lambda A + (1-\lambda)B)x = x^T \lambda A x + x^T(1-\lambda)Bx$$

$$= \underbrace{\lambda}_{\geq 0} \cdot \underbrace{x^T A x}_{\geq 0} + \underbrace{(1-\lambda)}_{\geq 0} \cdot \underbrace{x^T B x}_{\geq 0}$$

$\lambda \in [0,1]$.

≥ 0.

Showing the Norm Balls Form Convex Sets

$B = \{ x \in \mathbb{R}^D \mid \|x\| \le 1 \}$

Take $x, y \in B$ and $\lambda \in [0,1]$.

To show: $\lambda x + (1-\lambda) \cdot y \in B$. triangle inequality.

$\| \lambda x + (1-\lambda) y \| \le \| \lambda x \| + \| (1-\lambda) y \|$

$= \lambda \underbrace{\|x\|}_{\le 1} + (1-\lambda) \underbrace{\|y\|}_{\le 1}$

$\le 1.$

Showing the Polyhedron is Convex + Example

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$, the polyhedron $P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\,\mathbf{x} \le \mathbf{b}\}$ is convex

Take $x, y \in P$: $A(\lambda x + (1-\lambda) y) = \lambda \underbrace{Ax}_{\le b} + (1-\lambda) \underbrace{Ay}_{\le b}$

$\le \lambda b + (1-\lambda) b$

$\le b.$

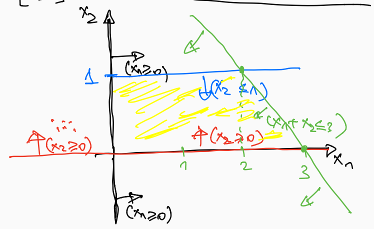
$$\underbrace{\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} \le \underbrace{\begin{bmatrix} 0 \\ 0 \\ 1 \\ 3 \end{bmatrix}}_{b}$$

$-x_1 \le 0 \Leftrightarrow x_1 \ge 0$

$-x_2 \le 0 \Leftrightarrow x_2 \ge 0$

$x_2 \le 1$

$x_1 + x_2 \le 3$

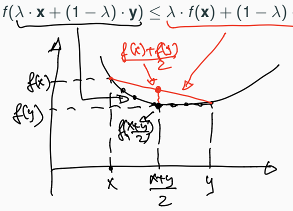


Convex Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ defined on a convex domain is convex if:

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ where f is defined and $0 \le \lambda \le 1$,

$$f(\lambda \cdot \mathbf{x} + (1-\lambda) \cdot \mathbf{y}) \le \lambda \cdot f(\mathbf{x}) + (1-\lambda) \cdot f(\mathbf{y})$$



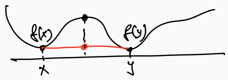
$\lambda = 0 \Rightarrow f(y) \le f(y)$

$\lambda = 1 \Rightarrow f(x) \le f(x)$

$\lambda = 0.5$

$f\left(\frac{x+y}{2}\right) \le \frac{f(x)+f(y)}{2}$

Not convex:



Examples of Convex Functions

- Affine functions: $f(\mathbf{x}) = \mathbf{b}^\top \mathbf{x} + c$
- Quadratic functions: $f(\mathbf{x}) = 1/2\, \mathbf{x}^\top \mathbf{A}\, \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$, where $\mathbf{A}$ is symmetric positive semidefinite
- Nonnegative weighted sums of convex functions: Given convex functions $f_1, \dots, f_n$ and $w_1, \dots, w_n \in \mathbb{R}_{\ge 0}$, the following is a convex function

$$f(\mathbf{x}) = \sum_{i=1}^{k} w_i \cdot f_i(\mathbf{x})$$

- Norms: $\| \cdot \|_p$ except $p = 0$ $\quad l_0 \doteq$ the number of non-zero entries in the vector.

Counterexample to convexity:

$x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Take $\theta \in (0,1)$:

$\| \theta x + (1-\theta) y \|_0 = \left\| \begin{bmatrix} \theta \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1-\theta \end{bmatrix} \right\|_0 = \left\| \begin{bmatrix} \theta \\ 1-\theta \end{bmatrix} \right\|_0 = 2$ $\Rightarrow$ not convex

$\overset{!}{\le}\ \theta \|x\|_0 + (1-\theta) \|y\|_0 = \theta \cdot 1 + (1-\theta) \cdot 1 = 1.$

Convex Optimisation

Given convex functions $f(\mathbf{x}), g_1(\mathbf{x}), \dots, g_m(\mathbf{x})$ and affine functions $h_1(\mathbf{x}), \dots h_n$,

a convex optimisation problem is of the form:

$$\begin{aligned} &\text{minimize } f(\mathbf{x}) \\ &\text{subject to } g_i(\mathbf{x}) \le 0 \qquad && i \in \{1, \dots, m\} \\ &\qquad\qquad h_j(\mathbf{x}) = 0 && j \in \{1, \dots, n\} \end{aligned}$$

Goal is to find an optimal value of a convex optimisation problem:

$$v^* = \min\{f(\mathbf{x}) : g_i(\mathbf{x}) \le 0, i \in \{1, \dots, m\}, h_i(\mathbf{x}) = 0, j \in \{0, \dots, n\}\}$$

Whenever $f(\mathbf{x}^*) = v^*$ then $\mathbf{x}^*$ is a (not necessarily unique) optimal point

- $v^* \overset{def}{=} +\infty$ for infeasible instances (feasible = fulfils all constraints g_i and h_j)
- $v^* \overset{def}{=} -\infty$ for unbounded instances (unbounded = the set of feasible instances has no infimum)

Local Optima are Global Optima for Convex Optimisation Problems

$\mathbf{x}$ is locally optimal if:

- $\mathbf{x}$ is feasible and
- There is $B > 0$ s.t. $f(\mathbf{x}) \le f(\mathbf{y})$ for all feasible $\mathbf{y}$ with $||\mathbf{x} - \mathbf{y}||_2 \le B$.

$\mathbf{x}$ is globally optimal if:

- $\mathbf{x}$ is feasible and
- $f(\mathbf{x}) \le f(\mathbf{y})$ for all feasible $\mathbf{y}$.

Theorem: For a convex optimisation problem, all locally optimal points are globally optimal.

Proof by contradiction. $\exists$ locally optimal point x that is not globally optimal: $\exists\, y \ne x$ s.t. $f(y) < f(x)$

x locally optimal $\Rightarrow \exists B$ s.t. $\forall$ feasible z $f(x) \le f(z)$ s.t. $\|x - z\|_2 \le B$.

Local Optima are Global Optima for Convex Optimisation Problems: Proof

Let $z = \lambda y + (1-\lambda) x$ where $\lambda = \frac{B}{2\|x-y\|_2}$.

To get $\lambda \in [0,1]$, we may choose B as we like.

With this λ, it holds that $\|x - z\|_2 \leq B$ since:

$$\|x - z\|_2 = \|x - (\lambda y + (1-\lambda)x)\|_2 = \|\lambda x - \lambda y\|_2$$

$$= \lambda \|x-y\|_2 = \frac{B}{\cancel{2\|x-y\|_2}} \cdot \cancel{\|x-y\|_2} = \frac{B}{2} \leq B.$$

Let us expand $f(z)$.

$$f(z) = f(\lambda y + (1-\lambda)x) \overset{f \text{ is convex}}{\leq} \lambda f(y) + (1-\lambda) f(x)$$

$$\overset{\text{hypothesis: } f(y) < f(x)}{<} \lambda f(x) + (1-\lambda) f(x)$$

$$= f(x).$$

We obtained: $f(z) < f(x)$. Contradiction (it means that x is not locally optimal).

Classes of Convex Optimisation Problems

Linear Programming:

$$\begin{aligned} &\text{minimize } \mathbf{c}^\top \mathbf{x} + d \\ &\text{subject to } \mathbf{A}\,\mathbf{x} \leq \mathbf{e} \\ &\qquad\qquad \mathbf{B}\,\mathbf{x} = \mathbf{f} \end{aligned}$$

Quadratically Constrained Quadratic Programming:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\mathbf{x}^\top \mathbf{B}\,\mathbf{x} + \mathbf{c}^\top \mathbf{x} + d \\ &\text{subject to } \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_i\,\mathbf{x} + \mathbf{r}_i^\top \mathbf{x} + s_i \leq 0 \qquad i \in \{1, \ldots, m\} \\ &\qquad\qquad \mathbf{A}\,\mathbf{x} = \mathbf{b} \end{aligned}$$

Semidefinite Programming:

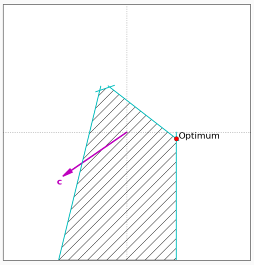
$$\begin{aligned} &\text{minimize } \mathrm{tr}(\mathbf{C}\,\mathbf{X}) \\ &\text{subject to } \mathrm{tr}(\mathbf{A}_i\,\mathbf{X}) = b_i \qquad i \in \{1, \ldots, m\} \\ &\qquad\qquad \mathbf{X} \text{ positive semidefinite} \end{aligned}$$

$\mathrm{tr}(\mathbf{A})$ is the trace of the matrix $\mathbf{A}$

Linear Programming

Looking for solutions $\mathbf{x} \in \mathbb{R}^n$ to the following optimisation problem

$$\begin{aligned} &\text{minimize } \mathbf{c}^\top \mathbf{x} + d \\ &\text{subject to } \mathbf{A}\,\mathbf{x} \leq \mathbf{e} \\ &\qquad\qquad \mathbf{B}\,\mathbf{x} = \mathbf{f} \end{aligned}$$



- No closed-form solution
- Efficient algorithms exist, both in theory and practice (for tens of thousands of variables)

Linear Model with Absolute Loss

Suppose we have data $(\mathbf{X}, \mathbf{y})$ and that we want to minimise the objective:

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} |\mathbf{x}_i^\top \mathbf{w} - y_i|$$

We would like to transform this optimisation problem into a linear program.

We introduce one ζ_i for each datapoint.

The linear program in the $D + N$ variables $w_1, \ldots, w_D, \zeta_1, \ldots, \zeta_N$

$$\text{minimize} \quad \sum_{i=1}^{N} \zeta_i$$

subject to:

$$\begin{aligned} \mathbf{w}^\top \mathbf{x}_i - y_i \leq \zeta_i, &\qquad i = 1, \ldots, N \\ y_i - \mathbf{w}^\top \mathbf{x}_i \leq \zeta_i, &\qquad i = 1, \ldots, N \end{aligned}$$

The solution to this linear program gives $\mathbf{w}$ that minimises the objective $\mathcal{L}$.

Recall: Likelihood of Linear Regression (Gaussian Noise Model)

Likelihood

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top(\mathbf{X}\mathbf{w} - \mathbf{y})\right)$$

Maximise Likelihood = Maximise Log-Likelihood ($\log : \mathbb{R}^+ \to \mathbb{R}$ is increasing)

$$\mathrm{LL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top(\mathbf{X}\mathbf{w} - \mathbf{y})$$

Maximise Log-Likelihood = Minimise Negative Log-Likelihood

$$\mathrm{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\top(\mathbf{X}\mathbf{w} - \mathbf{y})$$

$$= \underbrace{\frac{N}{2}\log(2\pi\sigma^2)}_{\text{constant}} + \frac{1}{2\sigma^2}\left(\underbrace{\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}}_{\mathbf{w}^\top\mathbf{B}\mathbf{w}} - \underbrace{2\mathbf{y}^\top\mathbf{X}\mathbf{w}}_{\mathbf{c}^\top\mathbf{w}} + \underbrace{\mathbf{y}^\top\mathbf{y}}_{\text{constant}}\right)$$

This is a convex quadratic optimisation problem with no constraints!

Minimising the Lasso Objective

For the Lasso objective, *i.e.*, linear model with ℓ_1-regularisation, we have

$$\mathcal{L}_{\text{lasso}}(\mathbf{w}) = \sum_{i=1}^{N}(\mathbf{w}^\top\mathbf{x}_i - y_i)^2 + \lambda\sum_{i=1}^{D}|w_i| = \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{y}^\top\mathbf{y} + \lambda\sum_{i=1}^{D}|w_i|$$

- Quadratic part of the loss function cannot be framed as linear programming
- Lasso regularisation does not allow for closed-form solutions
- Can be rephrased as quadratic programming problem
- Alternatively resort to general optimisation methods