Introduction to Optimization

Yangyang Xu

Mathematical Sciences, RPI

Ubiquitousness of optimization

- Optimization in everyday life: shortest path, least time, most efficient ...
- Optimization problems in many areas including operations research, management, engineering, finance, and so on.
- In recent years, optimization becomes particularly popular in machine learning, statistics, signal processing, and various data sciences.

Basic formulation

In general, an optimization problem can be formulated as

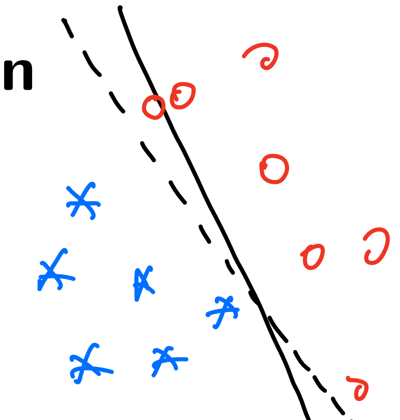
$$\underset{\mathbf{x}}{\text{minimize}}\, f(\mathbf{x}), \quad \text{subject to } \mathbf{x} \in \mathcal{X}.$$

- f is called the objective function
- $\mathbf{x}$ is called the variable
- $\mathcal{X}$ is called a feasible set. In this course, we assume

  $$\mathcal{X} = \left\{\mathbf{x} \in \mathcal{X}_{\mathsf{smpl}} : g_i(\mathbf{x}) \le 0, \forall i \in [m],\, h_i(\mathbf{x}) = 0, \forall i \in [\ell]\right\},$$

  where $\mathcal{X}_{\mathsf{smpl}}$ is a simple set such as a box constraint set.
- $\min_{\mathbf{x}} f(\mathbf{x})$ is equivalent to $\max_{\mathbf{x}} -f(\mathbf{x})$

Example I: Logistic regression



Problem setting:

- given a set of training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ with $y_i \in \{+1, -1\}$
- Assume the conditional probability of the label y_i based on sample $\mathbf{x}_i$ takes the form of

$$\mathrm{Prob}(y_i \mid \mathbf{x}_i, \mathbf{w}, b) = \frac{\exp[y_i(\mathbf{w}^\top \mathbf{x}_i + b)]}{1 + \exp[y_i(\mathbf{w}^\top \mathbf{x}_i + b)]}, \forall i \in [N].$$

Goal:

- to learn a hyperplane $\{\mathbf{x} : \mathbf{w}^\top \mathbf{x} + b = 0\}$ to separate the dataset.
- for a new data point $\mathbf{x}^{\text{new}}$, it can be classified based on the margin $\mathbf{w}^\top \mathbf{x}^{\text{new}} + b$.

Example I: Logistic regression

Optimization model:

- Assume independent identical distribution
- The likelihood function is

$$\mathcal{L} = \Pi_{i=1}^{N} \frac{\exp[y_i(\mathbf{w}^\top \mathbf{x}_i + b)]}{1 + \exp[y_i(\mathbf{w}^\top \mathbf{x}_i + b)]},$$

- Maximizing likelihood is equivalent to

$$\underset{\mathbf{w}, b}{\text{minimize}} \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + \exp[-y_i(\mathbf{w}^\top \mathbf{x}_i + b)]\right). \tag{LogReg}$$

$$\max_{w,b} \prod_{i=1}^{N} \frac{\exp[y_i(w^T x_i)+b]}{1+\exp[y_i(w^T x_i)+b]}$$

$$\Leftrightarrow \max_{w,b} \log\left(\prod_{i=1}^{N} \frac{\exp[y_i(w^T x_i)+b]}{1+\exp[y_i(w^T x_i)+b]}\right)$$

$$\log\left(\prod_{i=1}^{N} \frac{\exp[y_i(w^T x_i)+b]}{1+\exp[y_i(w^T x_i)+b]}\right)$$

$$= \sum_{i=1}^{N} \log \frac{\exp[y_i(w^T x_i)+b]}{1+\exp[y_i(w^T x_i)+b]}$$

$$= \sum_{i=1}^{N} \left(\log \exp[y_i(w^T x_i)+b] - \log\left(1+\exp[y_i(w^T x_i)+b]\right)\right)$$

$$= \sum_{i=1}^{N} \left(y_i(w^T x_i+b) - \log\left(1+\exp[y_i(w^T x_i)+b]\right)\right)$$

$$\Leftrightarrow \min_{w,b} \frac{\sum_{i=1}^{N} \log\left(1+\exp[y_i(w^T x_i)+b]\right)}{N} - \frac{\sum_{i=1}^{N} y_i(w^T x_i+b)}{N}$$

$$\log\left(\prod_{i=1}^{N} \frac{\exp[y_i(w^T x_i)+b]}{1+\exp[y_i(w^T x_i)+b]}\right)$$

$$= \sum_{i=1}^{N} \log \frac{\exp[y_i(w^T x_i)+b]}{1+\exp[y_i(w^T x_i)+b]} = \sum_{i=1}^{N} \log \frac{1}{1+\exp[-y_i(w^T x_i+b)]}$$

Example II: Portfolio optimization

Problem setting:

- We have a unit of capital to invest on m assets
- Let x_i be the fraction of capital invested on the i-th asset and ξ_i be the (stochastic) return rate of the i-th asset
- The risk measured by variance of total return

$\mathbb{E}[x^\top \xi] = x^\top \mathbb{E}[\xi]$

$(x^\top(\xi - \mathbb{E}\xi))^2$

$= x^\top(\xi - \mathbb{E}\xi)(\xi - \mathbb{E}\xi)^\top x$

$$\begin{aligned}\mathrm{Var}\left(\mathbf{x}^\top \boldsymbol{\xi}\right) &= \mathbb{E}\left(\mathbf{x}^\top \boldsymbol{\xi} - \mathbb{E}[\mathbf{x}^\top \boldsymbol{\xi}]\right)^2 = \mathbb{E}\left(\mathbf{x}^\top(\boldsymbol{\xi} - \mathbb{E}\boldsymbol{\xi})\right)^2 \\ &= \mathbf{x}^\top \mathbb{E}\left[(\boldsymbol{\xi} - \mathbb{E}\boldsymbol{\xi})(\boldsymbol{\xi} - \mathbb{E}\boldsymbol{\xi})^\top\right]\mathbf{x} = \mathbf{x}^\top \boldsymbol{\Sigma} \mathbf{x}\end{aligned}$$

where $\mathbf{x} = (x_1; \ldots; x_m)$ and $\boldsymbol{\Sigma}$ is the covariance matrix

Goal:

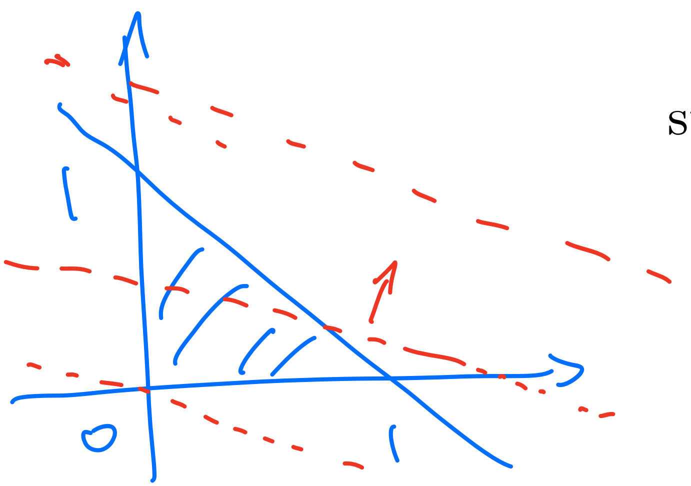
- To minimize risk subject to total unit capital and minimum expected return c

Example II: Portfolio optimization

Optimization model:

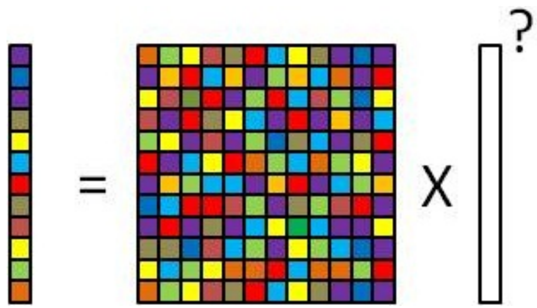
- variable: fraction vector of capital $\mathbf{x} = (x_1; \ldots; x_m)$
- risk function: $\frac{1}{2}\mathbf{x}^\top \mathbf{\Sigma}\mathbf{x}$
- constraint: total unit capital and minimum expected return c
- formulation:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \ & \frac{1}{2}\mathbf{x}^\top \mathbf{\Sigma}\mathbf{x} \\ \text{subject to} \ & \sum_{i=1}^{m} x_i \le 1, \\ & \mathbb{E}[\boldsymbol{\xi}^\top \mathbf{x}] \ge c, \\ & x_i \ge 0, \ \forall i \in [m]. \end{aligned} \tag{Portfolio}$$

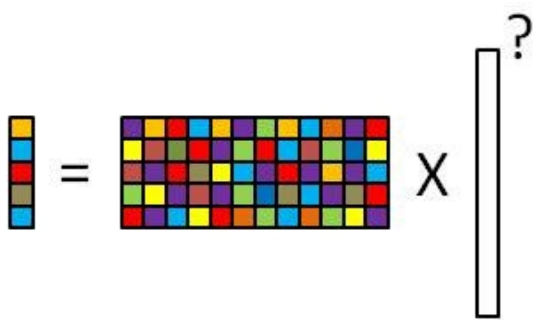


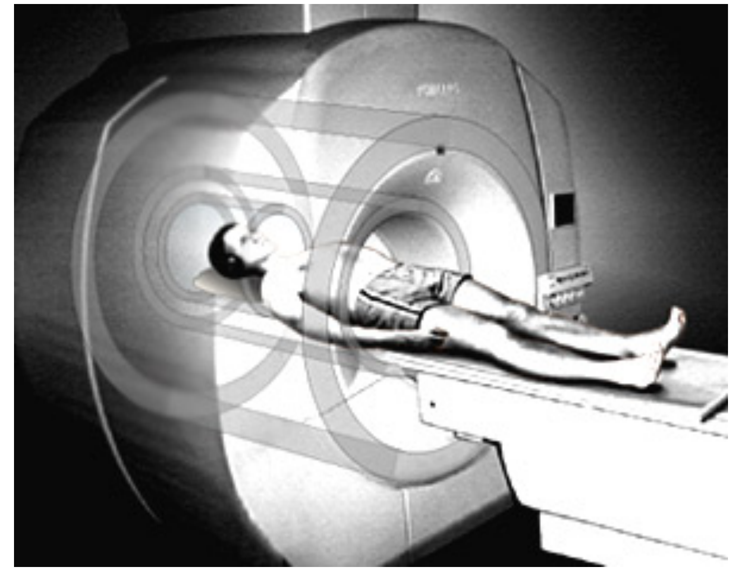
Example III: compressed sensing MRI[1]

- Classic technique



- Compressed sensing





- Classic technique: about 8 minutes
- Compressed sensing: about 80 seconds

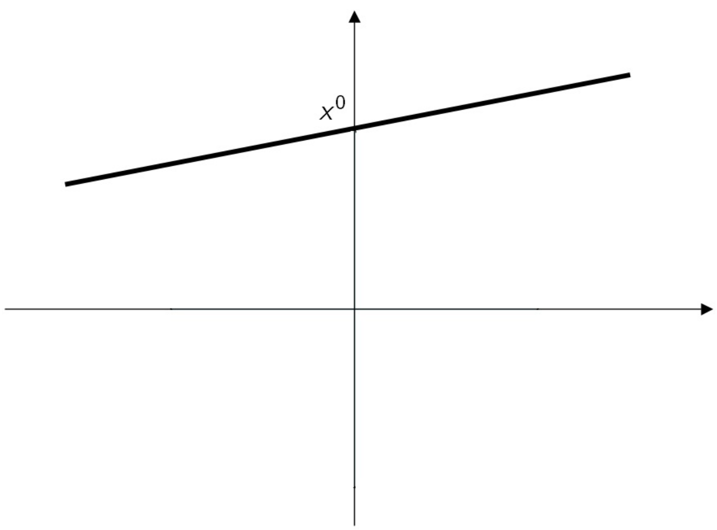
[1] Donoho'06, Candès et al.'06

How to recover the signal

- find the sparsest solution

$$\begin{cases} \min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_0 \\ \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b} \end{cases}$$

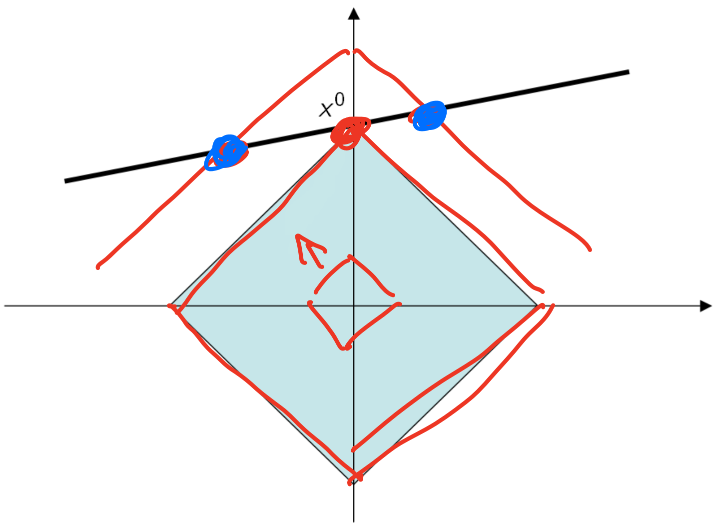
where $\|\mathbf{x}\|_0$ denotes #non-zeros in $\mathbf{x}$.



- ℓ_1 relaxation

$$\begin{cases} \min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_1 \\ \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b} \end{cases}$$

where $\|\mathbf{x}\|_1 \triangleq \sum_{i=1}^n |x_i|$.

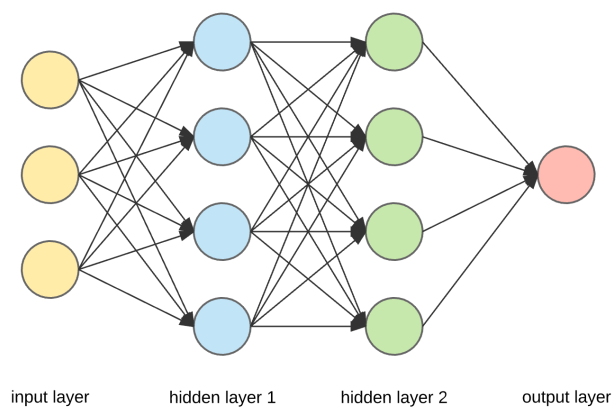


Example IV: Logistic regression with neural network approximation

Instead of hand pre-processing and directly applying LogReg on $\{\mathbf{x}_i\}$, auto-process the data first through a system (that is to be learned)

$$\underset{\mathbf{w},b,\boldsymbol{\theta}}{\text{minimize}} \frac{1}{N}\sum_{i=1}^{N} \log\left(1+\exp[-y_i(\mathbf{w}^\top f_{\boldsymbol{\theta}}(\mathbf{x}_i)+b)]\right)$$

- $f_{\boldsymbol{\theta}}$ a nonlinear transformation: extract features (maybe in other domain)
- currently popular: deep neural network $f_{\boldsymbol{\theta}}(\mathbf{x}) = f_{\boldsymbol{\theta}_d} \circ f_{\boldsymbol{\theta}_{d-1}} \circ \cdots \circ \underbrace{f_{\boldsymbol{\theta}_1}(\mathbf{x})}$



max(θ, x, 0)

Key questions

- Whether the minimization problem has a feasible and/or optimal solution
  - feasibility often assumed. But note checking feasibility can be hard
  - existence of optimal solution can be shown under mild conditions
- How to determine a candidate solution is optimal
  - by checking sufficient optimality conditions
- How to find an optimal solution (numerically and/or analytically)
  - analytically by formulating necessary optimality conditions and solving linear or nonlinear systems, e.g., $\min_{x,y}(x-1)^2+y^2, \text{ s.t. } x+y=2$
  - **Focus of this course:** numerically by moving iterate along feasible and descent directions

Outline of the rest

1. Concepts of numerical algorithm and convergence
2. Fundamentals of unconstrained optimization
3. Gradient type methods: steepest gradient descent, projected gradient, conjugate gradient, proximal gradient, and Nesterov's accelerated proximal gradient
4. Newton type methods: Newton's method, quasi-Newton method, and Gauss-Newton method
5. Derivative free methods: coordinate descent method
6. Theory of functional constrained optimization: Karush-Kuhn-Tucker (KKT) conditions and Lagrangian duality
7. Simplex method and Interior-point methods for linear programming
8. Ellipsoid method and cutting-plane methods
9. Penalty methods, barrier function methods, and augmented Lagrangian method
10. Alternating direction method of multipliers and applications