Chapter 6: Optimization for Data Science
Optimization in Machine Learning and Statistics

TANN Chantara

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia

October 7, 2022

Table of Contents

Maximum likelihood estimation

Distribution estimation: Estimate a probability density $p(y)$ of a random variable from observed data

Parametric distribution estimation: Choose from a family of densities $p_\beta(y)$ parametrized in β

Maximum likelihood estimation: Observations y_i for $i = 1, \ldots, m$. Assume the values are iid samples from $p_\beta(\cdot)$. Then, the likelihood to observe y_i, $i = 1, \ldots, m$ is

$$\ell(\beta) = \prod_{i=1}^{m} p_\beta(y_i).$$

The parameters most likely to have generated the observations are found by solving $\max_\beta \ell(\beta)$ or, equivalently, $\max_\beta \log \ell(\beta)$.

$$L(\beta) = \log \ell(\beta) = \sum_{i=1}^{m} \log(p_\beta(y_i))$$

is the log-likelihood function.

Linear measurement model

$$y_i = x_i^\top \beta + v_i$$

- (y_i, x_i) observations
- β unknown parameters
- $v_i \sim p(\cdot)$ noise

The maximum likelihood (ML) estimate is an optimal solution of

$$\max_{\beta} L(\beta) = \max_{\beta} \sum_{i=1}^{m} \log(p(y_i - x_i^\top \beta))$$

Example: Gaussian noise: $p(v) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{v^2}{2\sigma^2}} \quad (\sigma > 0)$

$$\Longrightarrow L(\beta) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|y - X\beta\|_2^2,$$

where $X = [x_1, \ldots, x_m]^\top$ and $y = [y_1, \ldots, y_m]^\top$

Linear measurement model

$$y_i = x_i^\top \beta + v_i$$

- (y_i, x_i) observations
- β unknown parameters
- $v_i \sim p(\cdot)$ noise

The maximum likelihood (ML) estimate is an optimal solution of

$$\max_\beta L(\beta) = \max_\beta \sum_{i=1}^{m} \log(p(y_i - x_i^\top \beta))$$

Example: Laplacian noise: $p(v) = \frac{1}{2a} e^{-\frac{|v|}{a}} \quad (a > 0)$

$$\implies L(\beta) = -m \log(2a) - \frac{1}{a} \|y - X\beta\|_1,$$

where $X = [x_1, \ldots, x_m]^\top$ and $y = [y_1, \ldots, y_m]^\top$

Linear measurement model

$$y_i = x_i^\top \beta + v_i$$

- (y_i, x_i) observations
- β unknown parameters
- $v_i \sim p(\cdot)$ noise

The maximum likelihood (ML) estimate is an optimal solution of

$$\max_\beta L(\beta) = \max_\beta \sum_{i=1}^{m} \log(p(y_i - x_i^\top \beta))$$

Example: Uniform noise: $p(v) = \begin{cases} \frac{1}{2a}, & \text{if } v \in [-a, a], \quad (a > 0) \\ 0 & \text{else} \end{cases}$

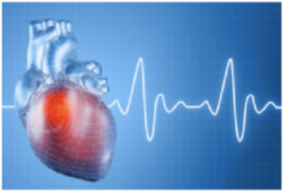
$$\implies L(\beta) = \begin{cases} -m \log(2a), & \text{if } \|y - X\beta\|_\infty \leq a \\ -\infty, & \text{else} \end{cases}$$

where $X = [x_1, \ldots, x_m]^\top$ and $y = [y_1, \ldots, y_m]^\top$

Logistic regression

Predicting the probability of a heart attach based on

- age
- height
- weight
- blood pressure
- cholesterol level
- etc.



$$\text{Label:} \quad y = \begin{cases} +1 & \text{person } x \text{ has a heart attack} \\ -1 & \text{person } x \text{ is healthy} \end{cases}$$

$$\text{Model:} \quad p_\theta(y|x) = \frac{1}{1 + \exp(-y \cdot \theta^\top x)} \qquad \theta \text{ unknown parameter}$$

Logistic regression (cont'd)

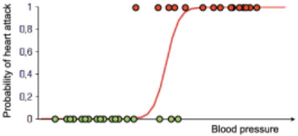
- Training data: $\{(x_i, y_i)\}_{i=1}^m$
- Log-likelihood function:

$$L(\theta) = \log \prod_{i=1}^{m} (1+\exp(-y_i \cdot \theta^\top x_i))^{-1} = -\sum_{i=1}^{m} \log(1+\exp(-y_i \cdot \theta^\top x_i))$$

- Logloss function: $l(z) = \log(1 + \exp(-z))$
  - smooth overestimation of $\max\{-z, 0\}$
  - special case of a log-sum-exp function $\Longrightarrow$ convex

ML estimation $\Longleftrightarrow$ empirical logloss minimization

$$\min_\theta \frac{1}{m} \sum_{i=1}^{m} l(y_i \cdot \theta^\top x_i)$$



Covariance estimation for Gaussian variables

- $y \in \mathbb{R}^n$ is Gaussian with mean zero and covariance matrix R. Its density is

$$p_R(y) = \frac{1}{\sqrt{(2\pi)^n \det R}} e^{-\frac{1}{2} y^\top R^{-1} y}$$

- Log-likelihood function for observations y_i, $i = 1, \ldots, m$

$$L(R) = -\frac{mn}{2} \log(2\pi) - \frac{m}{2} \log(\det R) - \frac{m}{2} \mathrm{tr}(YR^{-1}),$$

where $Y = \frac{1}{m} \sum_{i=1}^{m} y_i y_i^\top$ is the sample covariance matrix

Note: This log-likelihood function is not concave

Covariance estimation for Gaussian variables (cont'd)

- Information matrix: $S = R^{-1} \quad (S \succ 0 \iff R \succ 0)$
- Using S instead of R as the parameter, we find

$$L(S) = -\frac{mn}{2}\log(2\pi) + \frac{m}{2}\log(\det S) - \frac{m}{2}\mathrm{tr}(YS),$$

which is concave

- The ML estimate of S (and thus R) is found by solving

$$\begin{cases} \max & \log(\det S) - \mathrm{tr}(YS) \\ \text{s.t.} & S \in \mathcal{S}, \end{cases}$$

where $\mathcal{S}$ contains all constraints that capture prior structural information

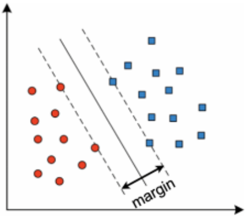
- If $\mathcal{S} = \mathbb{S}^n_{++}$, then $S = Y^{-1} \iff R = Y$ at optimality
(Recall that $\nabla \log(\det S) = S^{-1}$)

Support Vector Machines (SVM)

Classification problem: Given labelled data pairs (x_i, y_i), $i = 1, \dots, m$, where $x_i \in \mathbb{R}^d$ are the features (e.g., age, blood pressure, ...) and $y_i \in \{1, -1\}$ are the labels (e.g., healthy vs. heart attack, red vs. blue,...)

Goal: Predict the label of a new feature $x \in \mathbb{R}^n$

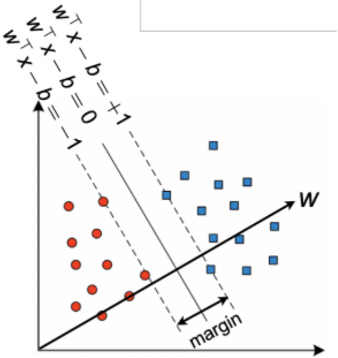
Idea: Find a hyperplane that separates the " blue" and "red" points with maximum margin. Predict labels of new points x depending on which side of the hyperplane they fall



Hard margin SVM

Hyperplane: $w^\top x - b = 0, w \neq 0$

We require: $w_i^\top x - b \geq 1 \quad \forall i$ with $y_i = 1$ (blue)
$w_i^\top x - b \leq -1 \ \forall i$ with $y_i = -1$ (red)



Margin $= \frac{2}{\|w\|_2}$

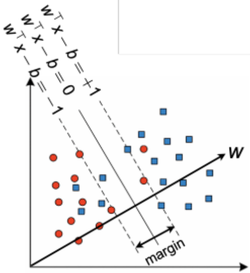
Hyperplane is found via the QP

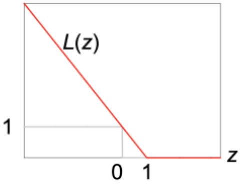
$$\begin{cases} \min_{w,b} & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} & y_i(w^\top x_i - b) \geq 1 \ \forall i \end{cases}$$

Soft margin SVM

What if the blue and red points are not linearly separable?

$$\min_{w,b} \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i(w^\top x_i - b)\}}_{\text{empirical Hinge loss}} + \underbrace{\frac{\rho}{2}\|w\|_2^2}_{\text{regularization term}}$$





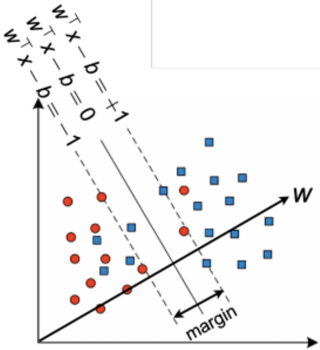
Hinge loss: $L(z) = \max\{0, 1 - z\}$

Signal: $z = y_i(w^\top x_i - b)$

- When $w^\top x_i - b$ and y_i have the same sign (i.e., y_i predicts the right class) and $|w^\top x_i - b| \geq 1$, then $L(z) = 0$

Soft margin SVM

What if the blue and red points are not linearly separable?



Replace Hinge loss of $i^t h$ sample with s_i

$$\begin{cases} \min_{w,b,s} & \frac{1}{m}\sum_{i=1}^{m} s_i + \frac{\rho}{2}\|w\|_2^2 \\ \text{s. t.} & y_i(w^\top x_i - b) + s_i \geq 1 \ \forall i \\ & s_i \geq 0 \ \forall i \end{cases}$$

Soft margin SVM

Primal QP:
$$\begin{cases} \min\limits_{w,b,s} & \frac{1}{m}\sum_{i=1}^m s_i + \frac{\rho}{2}\|w\|_2^2 \\ \text{s. t.} & y_i(w^\top x_i - b) + s_i \geq 1,\ s_i \geq 0,\ \forall i \end{cases}$$

Lagrangian:
$$L(w,b,s,\lambda,\gamma) = \frac{1}{m}\sum_{i=1}^m s_i + \frac{\rho}{2}\|w\|_2^2 - \sum_{i=1}^m \gamma_i s_i + \sum_{i=1}^m \lambda_i(1 - s_i - y_i(w^\top x_i - b))$$

Dual QP:
$$\begin{cases} \max\limits_{\lambda} & \sum_{i=1}^m \lambda_i - \frac{1}{2\rho}\sum_{i,j=1}^m \lambda_i\lambda_j y_i y_j x_i^\top x_j \\ \text{s. t.} & \sum_{i=1}^m y_i\lambda_i = 0, 0 \leq \lambda_i \leq \frac{1}{m},\ \forall i \end{cases}$$

KKT allows us to construct w and b from the dual solution.

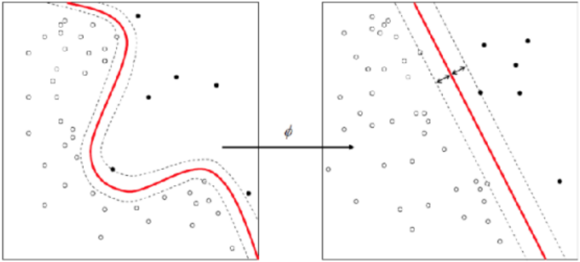
$$\nabla_w L(w,b,s,\lambda,\gamma) = 0 \qquad \Longrightarrow w = \frac{1}{\rho}\sum_{i=1}^m \lambda_i y_i x_i$$

$$\left.\begin{array}{l} \lambda_i(1 - s_i - y_i(w^\top x_i - b)) = 0 \\ (1/m - \lambda_i)s_i = 0 \end{array}\right\} \Longrightarrow \begin{cases} b = w^\top x_i - y_i \text{ for any } i \\ \text{with } 0 < \lambda_i < 1/m \end{cases}$$

Primal learns d and dual m parameters. Dual is easier if $m \ll d$

Kernel trick

- Improve classification via nonlinear separators.
- Use feature map $\phi : \mathbb{R}^d \to \mathbb{R}^D$ to lift the problem to a high-dimensional feature space $\mathbb{R}^D$, $D \gg d$



$x_1, \ldots, x_m$ $\qquad$ $x'_1 = \Phi(x_1), \ldots, x'_m = \Phi(x_m)$

Image source: Wikipedia

Nonlinear dimensionality reduction

SVM in high-dimensional space:

Primal QP:
$$\begin{cases} \min\limits_{w,b,s} & \frac{1}{m}\sum_{i=1}^{m} s_i + \frac{\rho}{2}\|w\|_2^2 \\ \text{s. t.} & y_i(w^\top \phi(x_i) - b) + s_i \geq 1,\ s_i \geq 0,\ \forall i \end{cases}$$

Dual QP:
$$\begin{cases} \max\limits_{\lambda} & \frac{1}{m}\lambda_i - \frac{1}{2\rho}\sum_{i,j=1}^{m} \lambda_i\lambda_j y_i y_j \phi(x_i)^\top \phi(x_j) \\ \text{s. t.} & \sum_{i=1}^{m} y_i\lambda_i = 0, 0 \leq \lambda_i \leq \frac{1}{m},\ \forall i \end{cases}$$

- The label of any new point x is predicted as

$$y = \text{sign}(w^\top \phi(x) - b) = \text{sign}\left(\frac{1}{\rho}\sum_{i=1}^{m} \lambda_i y_i \phi(x_i)^\top \phi(x) - b\right)$$

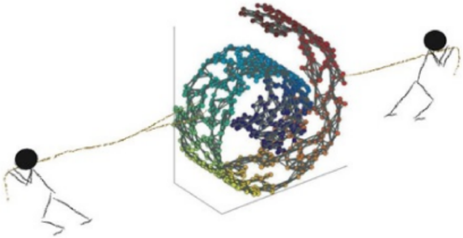
- Kernel function $K(x, x') = \phi(x)^\top \phi(x')$

The size of the dual QP is independent of the feature dimension D. Never evaluate inner products explicitly!

Nonlinear dimensionality reduction (cont'd)

Dimensionality reduction: Find meaningful low-dimensional structures hidden in high-dimensional observations (e.g., digital images, human genes, climate patterns etc.)

Example: Unwinding a Euro bill



The unwound Euro bill is flat ⇒ reduction from 3 to 2 dimensions

Nonlinear dimensionality reduction (cont'd)

Input: $y_i \in \mathbb{R}^d, \ i = 1, \ldots, m$

Construct a k-nearest neighbourhood graph $G = (V, E)$ with nodes $V = \{1, \ldots, m\}$ and edges E, where $(i, j) \in E$ if and only if y_i is among the k nearest neighbours of y_j

Example:

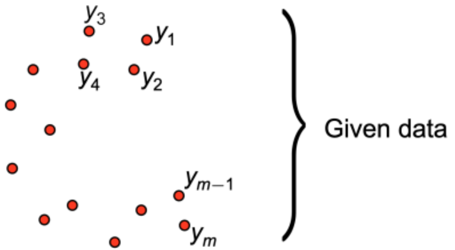


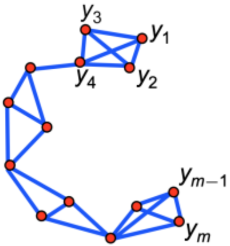
Figure:

Nonlinear dimensionality reduction (cont'd)

Input: $y_i \in \mathbb{R}^d,\ i = 1, \ldots, m$

Construct a k-nearest neighbourhood graph $G = (V, E)$ with nodes $V = \{1, \ldots, m\}$ and edges E, where $(i, j) \in E$ if and only if y_i is among the k nearest neighbours of y_j

Example:

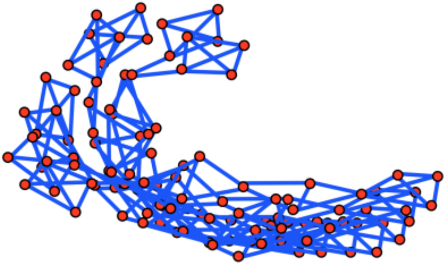


connect every node with its 3 nearest neighbors

Nonlinear dimensionality reduction (cont'd)

Idea: Spread out the data points as much as possible while keeping the distances between nearest neighbours fixed

Example:

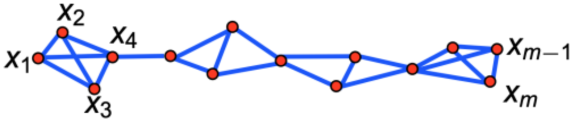


Nonlinear dimensionality reduction (cont'd)

Idea: Spread out the data points as much as possible while keeping the distances between nearest neighbours fixed

This can be done with and SDP !

Example: Let x_i, $i = 1, \ldots, m$ be the new positions of the data points (after spreading them out)



Nonlinear dimensionality reduction (cont'd)

Optimization problem for unfolding the kNN graph:

$$\begin{cases} \max\limits_{x} & \sum_{i=1}^{m} \|x_i\|_2^2 \\ \text{s.t.} & \sum_{i=1}^{m} x_i = 0 \\ & \|x_i - x_j\|_2^2 = \|y_i - y_j\|_2^2 \quad \forall (i,j) \in E \end{cases} \tag{1}$$

Maximize the variance of new positions. Require that mean is zero (eliminate translational degree of freedom) and require that distances between nearest neighbours are kept fixed.

Introduce Gram matrix $X \in \mathbb{S}_+^m$ with $X_{ij} = x_i^\top x_j$. We then have

- $\sum_{i=1}^{m} \|x_i\|_2^2 = \mathrm{tr}(X)$
- $\sum_{i=1}^{m} x_i = 0 \iff (\sum_{i=1}^{m} x_i)^\top (\sum_{j=1}^{m} x_j) = \sum_{i,j=1}^{m} X_{ij} = 0$
- $\|x_i - x_j\|_2^2 = X_{ii} - 2X_{ij} + X_{jj} = \|y_i - y_j\|_2^2 \quad \forall (i,j) \in E$

Nonlinear dimensionality reduction (cont'd)

Theorem: The "unfolding problem" (1) is equivalent to

$$\begin{cases} \max\limits_{X} & \mathrm{tr}(X) \\ \text{s. t.} & \sum_{i,j=1}^{m} X_{ij} = 0 \\ & X_{ii} - 2X_{ij} + X_{jj} = \|y_i - y_j\|_2^2 \quad \forall (i,j) \in E \\ & X \succeq 0, \mathrm{rank}(X) \leq d \end{cases} \tag{2}$$

Proof sketch: If x_i, $i = 1, \ldots, m$ is feasible in (1), then X defined via $X_{ij} = x_i^\top x_j$ is feasible in (2) with the same objective value. Note that

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_m^\top \end{pmatrix} (x_1 \ldots x_m) \succeq 0$$

Nonlinear dimensionality reduction (cont'd)

Theorem: The "unfolding problem" (1) is equivalent to

$$\begin{cases} \max\limits_{X} & \mathrm{tr}(X) \\ \text{s. t.} & \sum_{i,j=1}^{m} X_{ij} = 0 \\ & X_{ii} - 2X_{ij} + X_{jj} = \|y_i - y_j\|_2^2 \quad \forall (i,j) \in E \\ & X \succeq 0, \mathrm{rank}(X) \leq d \end{cases}$$

Proof sketch: If X is feasible in (2), then $X = RDR^\top$, where $D \in \mathbb{R}^{r\times r}$ is the diagonal matrix of all positive eigenvalues of X, the columns of $R \in \mathbb{R}^{m\times r}$ contain the corresponding orthonormal eigenvectors, and $r \leq \min\{d, m\}$ is the rank of X. Define x_i as the i^{th} row of $RD^{1/2}$. By construction, $X = (RD^{1/2})(RD^{1/2})^\top = (x_1 \dots x_m)^\top (x_1 \dots x_m)$ is the Gram matrix of the recovered x_i. Thus, the x_i are feasible in (1) and attain the same objective value as X in (2).

Nonlinear dimensionality reduction (cont'd)

The "unfolding problem" (1) is approximated by the SDP

$$\begin{cases} \max\limits_{X} & \mathrm{tr}(X) \\ \text{s. t.} & \sum_{i,j=1}^{m} X_{ij} = 0 \\ & X_{ii} - 2X_{ij} + X_{jj} = \|y_i - y_j\|_2^2 \quad \forall (i,j) \in E \\ & X \succeq 0, \cancel{\mathrm{rank}(X) \leq d} \end{cases} \tag{3}$$

How to recover a low-dimensional solution $x_i \in \mathbb{R}^r$, $i = 1, \ldots, m$ with $r \ll \min\{d, m\}$ from a solution X of the SDP (3)?

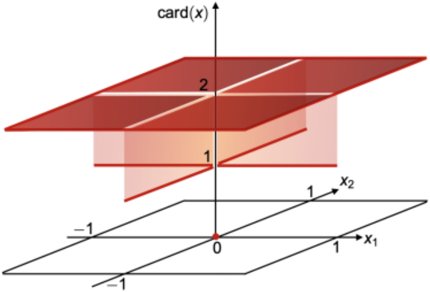
Heuristic approach: Let $D \in \mathbb{R}^{r \times r}$ be the diagonal matrix of the r largest eigenvalues of X, and let $R \in \mathbb{R}^{m \times r}$ be the matrix whose columns are the corresponding eigenvectors. Define x_i as the i^{th} row of $RD^{1/2}$. As small eigenvalues are ignored, we have

$$X \approx (RD^{1/2})(RD^{1/2})^\top = (x_1 \ldots x_m)^\top (x_1 \ldots x_m),$$

and thus the x_i are nearly feasible and optimal in (1).

Cardinality

Definition: The cardinality $\mathrm{card}(x)$ of $x \in \mathbb{R}^n$ is the number of non-zero entries of x.



The cardinality function is non-convex !

Convex cardinality problems

A convex cardinality problem is a convex except for a single cardinality function in the objective or in the constraints.

Assume that $C \subset \mathbb{R}^n$ is a convex set, and $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function.

Convex minimum cardinality problem

$$\begin{cases} \min\limits_{x} & \operatorname{card}(x) \\ \text{s.t.} & x \in C \end{cases}$$

Convex problem with cardinality constraint:

$$\begin{cases} \min\limits_{x} & f(x) \\ \text{s.t.} & x \in C,\ \operatorname{card}(x) \leq k \end{cases}$$

Examples: Statistics

Regressor selection: Fit $b \in \mathbb{R}^m$ as a linear combination of k out of n possible columns of $A \in \mathbb{R}^{m \times n}$

$$\begin{cases} \min\limits_{x} & \|Ax - b\|_2 \\ \text{s. t.} & \operatorname{card}(x) \leq k \end{cases}$$

Linear classification with fewest errors: Replace the objective of the soft margin SVM with card(s)

$$\begin{cases} \min\limits_{w,b,s} & \operatorname{card}(s) \\ \text{s. t.} & y_i(w^\top x_i - b) + s_i \geq 1, \ s_i \geq 0, \ \forall i \end{cases}$$

Example: Minimum number of violations

Find $x \in C$ that violates as few of the following m convex inequalities as possible:

$$f_1(x) \leq 0, \ldots f_m(x) \leq 0$$

Such an x can be found by solving

$$\begin{cases} \min\limits_{x,t} & \mathrm{card}(t) \\ \text{s.t.} & x \in C,\ t \geq 0 \\ & f_i(x) \leq t_i \quad \forall i = 1, \ldots, m \end{cases}$$

Example: Sparse design

Find the sparsest design vector that satisfies a set of specifications

$$\begin{cases} \min\limits_{x} & \mathrm{card}(x) \\ \text{s.t.} & x \in C \end{cases}$$

Examples:

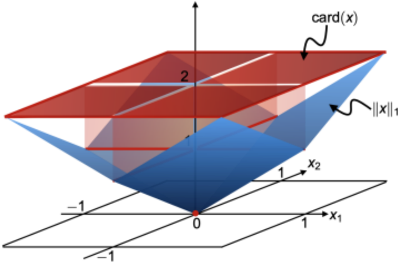
- finite impulse response filter design (zero entires reduce the required hardware)
- antenna array beamforming (zero entries correspond to unneeded antenna elements)
- truss design (zero entries correspond to unneeded bars)
- wire sizing (zero entries correspond to unneeded wires)

Exact solution of convex cardinality problems

- The decision $x \in \mathbb{R}^n$ has 2^n sparsity patterns (each component of x can be zero or nonzero)
- A convex cardinality problem can thus be solved exactly by solving 2^n convex problems (each enforcing a sparsity pattern)
- This may be practical for $n \leq 10$ but impractical for $n \geq 15$

ℓ_1-Norm heuristic

Replace $\text{card}(x)$ with $\gamma\|x\|_1$ or add a regularization term $\gamma\|x\|_1$ to the objective. Tune $\gamma > 0$ to achieve the desired sparsity



Note: $\|x\|_1$ is the convex envelope of $\text{card}(x)$ on $\{x : \|x\|_\infty \leq 1\}$

ℓ_1-Norm heuristic (cont'd)

Convex minimum cardinality problem:

$$\begin{cases} \min\limits_{x} & \mathrm{card}(x) \\ \text{s. t.} & x \in C \end{cases} \implies \begin{cases} \min\limits_{x} & \|x\|_1 \\ \text{s. t.} & x \in C \end{cases}$$

Convex problem with cardinality constraint:

$$\begin{cases} \min\limits_{x} & f(x) \\ \text{s. t.} & x \in C,\ \mathrm{card}(x) \le k \end{cases} \implies \begin{cases} \min\limits_{x} & f(x) \\ \text{s. t.} & x \in C,\ \|x\|_1 \le \beta \end{cases}$$

β can be tuned to ensure that $\mathrm{card}(x) \le k$.