Nonconvex? NP!

(No Problem!)

Lecturer: Ryan Tibshirani
Convex Optimization 10-725/36-725

Last time: integer programming

Given convex function f, convex set C, $J \subseteq \{1, \ldots n\}$, an integer program is a problem of the form

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & x \in C \\ & x_j \in \mathbb{Z}, \; j \in J \end{aligned}$$

IPs are like twisted cousin of convex optimization. Much harder to solve, but there is a huge literature on the topic. Key ideas:

- Lower and upper bounds
- Branch and bound method
- Cutting plane method

Application to modern statistical problems is growing, and exciting. E.g., least median of squares regression, best subset selection

My soap box

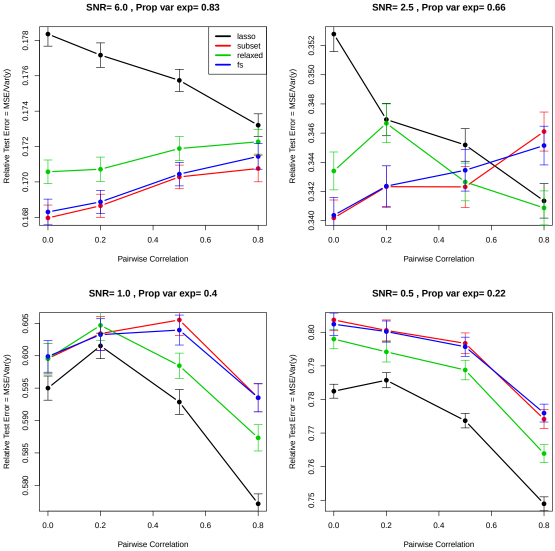
Recently, it's been said that statisticians should pay more attention to IPs. I agree

But just because we can solve a nonconvex problem by formulating it as IP, doesn't mean we should prefer its solution over that from related convex program

- The lasso is not a heuristic for best subset selection
- It's an estimator, with its own properties
- An ℓ_1 penalty shrinks coefficients (unlike ℓ_0); this can hurt or help, depending on the situation
- Even if we could always solve best subset selection efficiently, it would be unwise to think that we should always prefer it

Optimizers should be more aware of the bias-variance tradeoff

We (Hastie, Tibshirani x 2) are putting together some experiments to make this point salient. Preview:



SNR= 6.0 , Prop var exp= 0.83

SNR= 2.5 , Prop var exp= 0.66

SNR= 1.0 , Prop var exp= 0.4

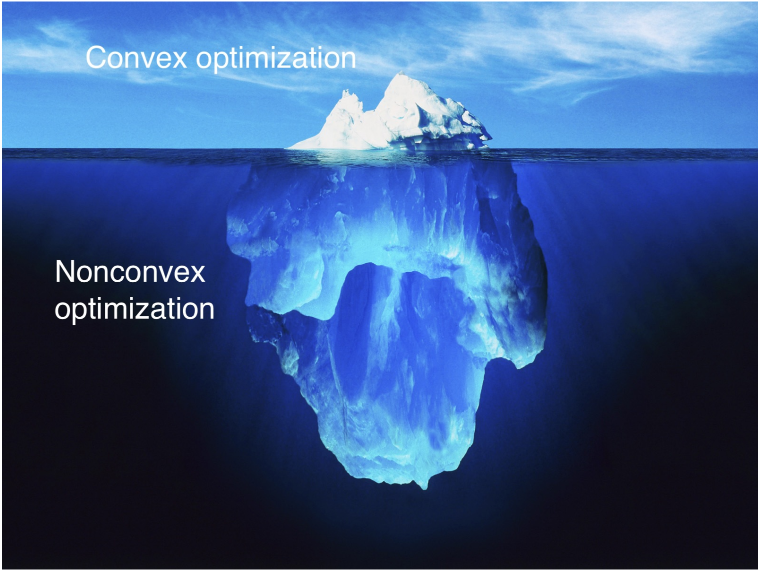
SNR= 0.5 , Prop var exp= 0.22

Outline

Today:

- Convex versus nonconvex?
- Classical nonconvex problems
- Eigen problems
- Graph problems
- Nonconvex proximal operators
- Discrete problems
- Infinite-dimensional problems
- Statistical problems
- Miscellaneous

Beyond the tip?



Some takeaway points

- If possible, formulate task in terms of convex optimization — typically easier to solve, easier to analyze
- Nonconvex does not necessarily mean nonscientific! However, statistically, it can often mean high(er) variance
- This is true both intrinsically, and because we can rarely solve nonconvex problems (to global optimality)
- In more cases than you might expect, nonconvex problems can be solved exactly (to global optimality)

What does it mean for a problem to be nonconvex?

Consider a generic optimization problem:

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \ i = 1, \ldots m \\ & h_j(x) = 0, \ j = 1, \ldots r \end{aligned}$$

This is a convex problem if f, g_i, $i = 1, \ldots m$ are convex, and h_j, $j = 1, \ldots r$ are affine

A nonconvex problem is one of this form, where not all conditions are met on the functions

But trivial modifications of convex problems can lead to nonconvex formulations ... so we really just consider nonconvex problems that are not trivially equivalent to convex ones

What does it mean to solve a nonconvex problem?

Nonconvex problems can have local minima, i.e., there can exist a feasible x such that

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

but x is still not globally optimal. (Note: we proved that this could not happen for convex problems)

Hence by solving a nonconvex problem, we mean finding the global minimizer

We also implicitly mean doing it efficiently, i.e., in polynomial time

Addendum

This is really about putting together a list of interesting problems, that are suprisingly tractable ... so there will be exceptions about nonconvexity and/or requiring exact global optima

(Also, I'm sure that there are many more examples out there that I'm missing, so I invite you to give me ideas / contribute!)

Classical nonconvex problems

Linear-fractional programs

A linear-fractional program is of the form

$$\begin{aligned} \min_{x} \quad & \frac{c^T x + d}{e^T x + f} \\ \text{subject to} \quad & Gx \leq h, \; e^T x + f > 0 \\ & Ax = b \end{aligned}$$

This is nonconvex (but quasiconvex). Provided that this problem is feasible, it is in fact equivalent to the linear program

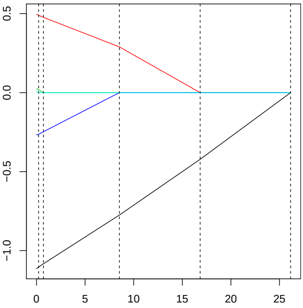
$$\begin{aligned} \min_{y,z} \quad & c^T y + dz \\ \text{subject to} \quad & Gy - hz \leq 0, \; z \geq 0 \\ & Ay - bz = 0, \; e^T y + fz = 1 \end{aligned}$$

The link between the two problems is the transformation

$$y = \frac{x}{e^T x + f}, \quad z = \frac{1}{e^T x + f}$$

The proof of their equivalence is simple; e.g., see B & V Chapter 4

Linear-fractional problems show up in the study of solutions paths for some common statistical estimation problems



E.g., knots in the lasso path (values of λ at which coefficient becomes nonzero) are optimal values of linear-fractional programs

See Taylor et al. (2013), "Inference in adaptive regression via the Kac-Rice formula"

Geometric programs

A monomial is a function $f : \mathbb{R}^n_{++} \to \mathbb{R}$ 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 of the form

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 1,\ i = 1, \ldots m \\ & h_j(x) = 1,\ j = 1, \ldots r \end{aligned}$$

where f, g_i, $i = 1, \ldots m$ are posynomials and h_j, $j = 1, \ldots r$ are monomials. This is nonconvex

This is equivalent to a convex problem, via a simple transformation. 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

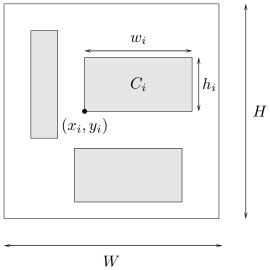
$$\gamma (e^{y_1})^{a_1} (e^{y_2})^{a_2} \cdots (e^{y_n})^{a_n} = e^{a^T y + b}$$

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

$$\begin{array}{ll} \displaystyle\min_{y} & \log\left(\displaystyle\sum_{k=1}^{p_0} e^{a_{0k}^T y + b_{0k}}\right) \\ \text{subject to} & \log\left(\displaystyle\sum_{k=1}^{p_i} e^{a_{ik}^T y + b_{ik}}\right) \leq 0,\ i = 1, \ldots m \\ & c_j^T y + d_j = 0,\ j = 1, \ldots r \end{array}$$

This is convex, recalling the convexity of soft max functions

Many interesting problems are geometric programs; see Boyd et al. (2007), "A tutorial on geometric programming", and also Chapters 4.5 and 8.8 of B & V book. Example floor planning program:



$$
\begin{aligned}
\min_{\substack{W,H,\\ x,y,w,h}} \quad & WH \\
\text{subject to} \quad & 0 \leq x_i \leq W,\ i = 1, \dots n \\
& 0 \leq y_i \leq H,\ i = 1, \dots n \\
& x_i + w_i \leq x_j,\ (i,j) \in \mathcal{L} \\
& y_i + h_i \leq y_j,\ (i,j) \in \mathcal{B} \\
& w_i h_i = C_i,\ i = 1, \dots n
\end{aligned}
$$

(Extension: Sra and Hosseini (2013), "Geometric optimization on positive definite matrices with application to elliptically contoured distributions")

Problems with two quadratic functions

Consider a problem involving two quadratics

$$\begin{array}{ll} \displaystyle\min_{x} & x^T A_0 x + 2b_0^T x + c_0 \\ \text{subject to} & x^T A_1 x + 2b_1^T x + c_1 \leq 0 \end{array}$$

Here A_0, A_1 need not be positive definite, so this is nonconvex. The dual problem can be cast as

$$\begin{array}{ll} \displaystyle\max_{u,v} & u \\ \text{subject to} & \left[\begin{array}{cc} A_0 + vA_1 & b_0 + vb_1 \\ (b_0 + vb_1)^T & c_0 + vc_1 - u \end{array}\right] \succeq 0,\ v \geq 0 \end{array}$$

Dual is convex (as always), and strong duality holds. See Appendix B of B & V, and also Beck and Eldar (2006), "Strong duality in nonconvex quadratic optimization with two quadratic constraints"

Handling convex equality constraints

Given convex f, g_i, $i = 1, \ldots m$, the problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \ i = 1, \ldots m \\ & h(x) = 0 \end{aligned}$$

is nonconvex when h is convex but not affine. A convex relaxation of this problem is

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \ i = 1, \ldots m \\ & h(x) \leq 0 \end{aligned}$$

If we can ensure that $h(x^*) = 0$ at any solution x^* of the above problem, then the two are equivalent

From B & V Exercises 4.6 and 4.58, e.g., consider the maximum utility problem

$$\begin{aligned} \max_{x,b} \quad & \sum_{t=1}^{T} \alpha_t u(x_t) \\ \text{subject to} \quad & b_{t+1} = b_t + f(b_t) - x_t, \; t = 1, \ldots T \\ & 0 \le x_t \le b_t, \; t = 1, \ldots T \end{aligned}$$

where $b_0 \ge 0$ is fixed. Interpretation: x_t is the amount spent of your total available money b_t at time t; concave function u gives utility, concave function f measures investment return

This is not a convex problem, because of the equality constraint; but can relax to

$$b_{t+1} \le b_t + f(b_t) - x_t, \;\; t = 0, \ldots T$$

without changing solution (think about throwing out money)

Convexifying constraint sets

Given nonconvex set C, consider the nonconvex problem

$$\min_x \; c^T x \;\; \text{subject to} \;\; x \in C$$

Due to linearity of objective, this is equivalent to convex problem

$$\min_x \; c^T x \;\; \text{subject to} \;\; x \in \text{conv}(C)$$

Proof: let $f^\star$ be optimal value in first problem, $x^\star$ be solution in second. Then $x^\star = \sum_i a_i x_i$ where $x_i \in C$, $a_i \geq 0$, and $\sum_i a_i = 1$. Note $f^\star \geq c^T x^\star = \sum_i a_i c^T x_i \geq \sum_i a_i f^\star = f^\star$. Thus all x_i must be optimal for first problem

But note that the convex problem is not necessarily easy! Could be very hard to even form $\text{conv}(C)$ (recall the cutting plane method for integer programming)

Eigen problems

Principal component analysis

Given a matrix $X \in \mathbb{R}^{n \times p}$, consider the nonconvex problem

$$\min_R \; \|X - R\|_F^2 \;\text{ subject to }\; \text{rank}(R) = k$$

for some fixed k. The solution here is given by the singular value decomposition of X: if $X = UDV^T$, then

$$\hat{R} = U_k D_k V_k^T,$$

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., $\hat{R}$ is the reconstruction of X from its first k principal components

This is often called the Eckart-Young Theorem, established in 1936, but was probably known even earlier — see Stewart (1992), "On the early history of the singular value decomposition"

Fantope

Another characterization of the SVD is via the following nonconvex problem, given $X \in \mathbb{R}^{n \times p}$:

$$\min_{Z \in \mathbb{S}^p} \|X - XZ\|_F^2 \;\; \text{subject to} \;\; \mathrm{rank}(Z) = k, \; Z \text{ projection}$$
$$\iff \max_{Z \in \mathbb{S}^p} \langle X^T X, Z \rangle \;\; \text{subject to} \;\; \mathrm{rank}(Z) = k, \; Z \text{ projection}$$

The solution here is $\hat{Z} = V_k V_k^T$, where the columns of $V_k \in \mathbb{R}^{p \times k}$ give the first k eigenvectors of $X^T X$

This is equivalent to a convex problem. Express constraint set C as

$$\begin{aligned} C &= \Big\{ Z \in \mathbb{S}^p : \mathrm{rank}(Z) = k, \; Z \text{ is a projection} \Big\} \\ &= \Big\{ Z \in \mathbb{S}^p : \lambda_i(Z) \in \{0, 1\}, \; i = 1, \ldots p, \; \mathrm{tr}(Z) = k \Big\} \end{aligned}$$

Now consider the convex hull $\mathcal{F}_k = \mathrm{conv}(C)$:

$$\begin{aligned}\mathcal{F}_k &= \Big\{ Z \in \mathbb{S}^p : \lambda_i(Z) \in [0,1],\ i = 1, \ldots p,\ \mathrm{tr}(Z) = k \Big\} \\ &= \Big\{ Z \in \mathbb{S}^p : 0 \preceq Z \preceq I,\ \mathrm{tr}(Z) = k \Big\}\end{aligned}$$

This is called the Fantope of order k. Further, the convex problem

$$\max_{Z \in \mathbb{S}^p} \langle X^T X, Z \rangle \quad \text{subject to} \quad Z \in \mathcal{F}_k$$

admits the same solution as the original one, i.e., $\hat{Z} = V_k V_k^T$

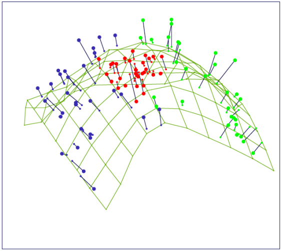
See Fan (1949), "On a theorem of Weyl conerning eigenvalues of linear transformations", and Overton and Womersley (1992), "On the sum of the largest eigenvalues of a symmetric matrix"

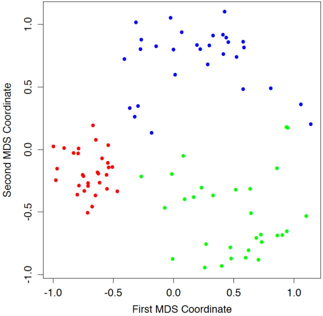
Sparse PCA extension: Vu et al. (2013), "Fantope projection and selection: near-optimal convex relaxation of sparse PCA"

Classical multidimensional scaling

Let $x_1, \ldots x_n \in \mathbb{R}^p$, and define similarities $S_{ij} = (x_i - \bar{x})^T (x_j - \bar{x})$. For fixed k, classical multidimensional scaling or MDS solves the nonconvex problem

$$\min_{z_1,\ldots z_n} \sum_{i,j=1}^{n} \Big(S_{ij} - (z_i - \bar{z})^T (z_j - \bar{z}) \Big)^2$$





From Hastie et al. (2009), "The elements of statistical learning"

Let S be the similarity matrix (entries $S_{ij} = (x_i - \bar{x})^T (x_j - \bar{x})$)

The classical MDS problem has an exact solution in terms of the eigendecomposition $S = UD^2U^T$:

$$\hat{z}_1, \ldots \hat{z}_n \text{ are the rows of } U_k D_k$$

where U_k is the first k columns of U, and D_k the first k diagonal entries of D

Note that other very similar forms of MDS are not convex, and not directly solveable, e.g., least squares scaling, with $d_{ij} = \|x_i - x_j\|_2$:

$$\min_{z_1, \ldots z_n} \sum_{i,j=1}^{n} \big(d_{ij} - \|z_i - z_j\|_2\big)^2$$

See Hastie et al. (2009), Chapter 14

Generalized eigenvalue problems

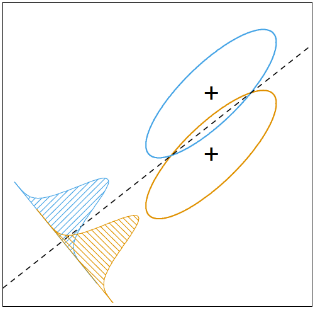
Given $B, W \in \mathbb{S}^p_{++}$, consider the nonconvex problem

$$\max_v \; \frac{v^T B v}{v^T W v}$$

This is a generalized eigenvalue problem, with exact solution given by the top eigenvector of $W^{-1}B$

This is important, e.g., in Fisher's discriminant analysis, where B is the between-class covariance matrix, and W the within-class covariance matrix

See Hastie et al. (2009), Chapter 4

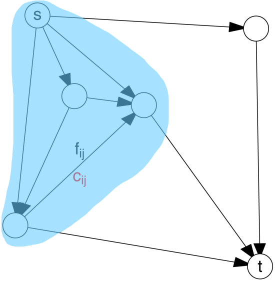


Graph problems

Min cut

Given a graph $G = (V, E)$ with $V = \{1, \ldots n\}$, two nodes $s, t \in V$, and costs $c_{ij} \geq 0$ on edges $(i, j) \in E$. Min cut problem:

$$\begin{aligned} \min_{b \in \mathbb{R}^{|E|}, x \in \mathbb{R}^{|V|}} \quad & \sum_{(i,j) \in E} b_{ij} c_{ij} \\ \text{subject to} \quad & b_{ij} \geq x_i - x_j \\ & b_{ij}, x_i, x_j \in \{0, 1\} \\ & \text{for all } i, j, \\ & x_s = 0, \; x_t = 1 \end{aligned}$$



Think of b_{ij} as the indicator that the edge (i, j) traverses the cut from s to t; think of x_i as an indicator that node i is grouped with t. This nonconvex problem can be solved exactly using max flow

A relaxation of min cut

$$\begin{aligned} \min_{b \in \mathbb{R}^{|E|},\, x \in \mathbb{R}^{|V|}} \quad & \sum_{(i,j) \in E} b_{ij} c_{ij} \\ \text{subject to} \quad & b_{ij} \geq x_i - x_j \ \text{ for all } i, j \\ & b \geq 0, \ x_s = 0, \ x_t = 1 \end{aligned}$$

This is an LP, and recall that it is the dual of the max flow LP:

$$\begin{aligned} \max_{f \in \mathbb{R}^{|E|}} \quad & \sum_{(s,j) \in E} f_{sj} \\ \text{subject to} \quad & f_{ij} \geq 0, \ f_{ij} \leq c_{ij} \ \text{ for all } (i,j) \in E \\ & \sum_{(i,k) \in E} f_{ik} = \sum_{(k,j) \in E} f_{kj} \ \text{ for all } k \in V \setminus \{s, t\} \end{aligned}$$

Max flow min cut theorem tells us that the relaxed min cut is tight

Max cut

Given an undirected graph $G = (V, E)$ with $V = \{1, \ldots n\}$, edge costs $c_{ij} \geq 0$, $(i, j) \in E$. Max cut problem:

$$\begin{array}{ll} \displaystyle\min_{v_1,\ldots,v_n \in \mathbb{R}} & \displaystyle\sum_{(i,j)\in E} c_{ij} \frac{(1 - v_i v_j)}{2} \\ \text{subject to} & v_i^2 = 1, \; i = 1, \ldots n \end{array}$$

Nonconvex, NP hard! SDP relaxation of Goemans and Williamson (1994), "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming":

$$\begin{array}{ll} \displaystyle\min_{v_1,\ldots,v_n \in \mathbb{R}^n} & \displaystyle\sum_{(i,j)\in E} c_{ij} \frac{(1 - v_i^T v_j)}{2} \\ \text{subject to} & \|v_i\|_2^2 = 1, \; i = 1, \ldots n \end{array}$$

Remarkable fact: with randomized rounding, $\mathbb{E}[f^\star_{\text{SDP}}] \geq 0.878 f^\star_{\text{OPT}}$

Nonconvex proximal operators

Hard-thresholding

One of the simplest nonconvex problems, given $y \in \mathbb{R}^n$:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \sum_{i=1}^{n} \lambda_i \cdot 1\{\beta_i \neq 0\}$$

Solution is given by hard-thresholding y,

$$\beta_i = \begin{cases} y_i & \text{if } y_i^2 > \lambda_i \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \ldots n$$

and can be seen by inspection. Special case of $\lambda_i = \lambda$, $i = 1, \ldots n$:

$$\min_{\beta} \|y - \beta\|_2^2 + \lambda \|\beta\|_0$$

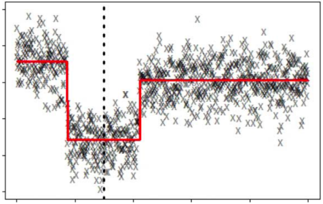
Compare to soft-thresholding, proximal operator for ℓ_1 norm. Also, note: changing the loss to $\|y - X\beta\|_2^2$ gives best subset selection, which is NP hard for general matrix X

Potts minimization

Consider 1d segmentation problem, also called Potts minimization:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_i)^2 + \lambda \sum_{i=1}^{n-1} 1\{\beta_i \neq \beta_{i+1}\}$$

Nonconvex, but solveable by dynamic programming, in two ways: Bellman (1961), "On the approximation of curves by line segments using dynamic programming", and Johnson (2013) "A dynamic programming algorithm for the fused lasso and L_0-segmentation"



Johnson: more efficient, Bellman: more general

Worst-case $O(n^2)$, but with practical performance more like $O(n)$

Tree-leaves projection

Given target $u \in \mathbb{R}^n$, tree g on $\mathbb{R}^n$, and label $y \in \{0, 1\}$, consider

$$\min_z \; \|u - z\|_2^2 + \lambda \cdot 1\{g(z) \neq y\}$$

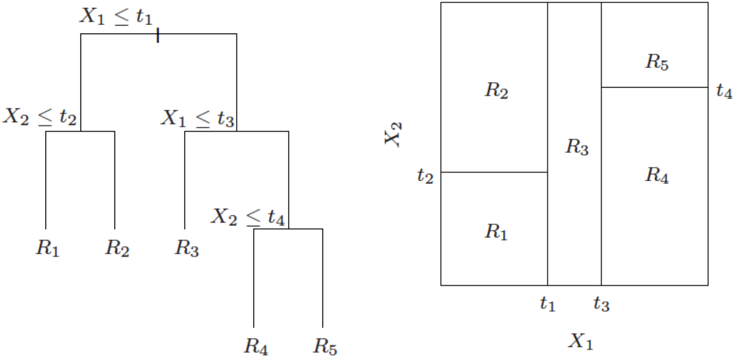
Interpretation: find z close to u, whose label under g is not unlike y. Argue directly that solution is either $\hat{z} = u$ or $\hat{z} = P_S(u)$, where

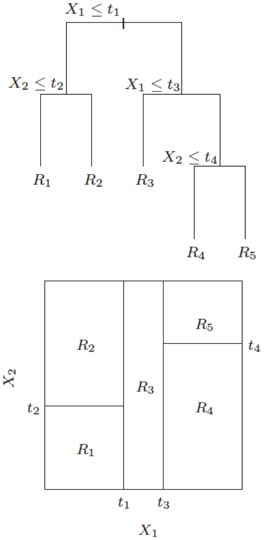
$$S = g^{-1}(1) = \{z : g(z) = y\}$$

the set of leaves of g assigned label y. We simply compute both options for $\hat{z}$ and compare costs. Therefore problem reduces to computing $P_S(y)$, the projection onto a set of tree leaves, a highly nonconvex set

(Subroutine of broader algorithm for nonconvex optimization; see Carreira-Perpinan and Wang (2012), "Distributed optimization of deeply nested systems")

The set S is a union of axis-aligned boxes; projection onto any one box is fast, $O(n)$ operations





To project onto S, could just scan through all boxes, and take the closest

Faster: decorate each node of tree with labels of its leaves, and bounding box. Perform depth-first search, pruning nodes:

- that do not contain a leaf labeled y, or
- whose bounding box is farther away than the current closest box

Discrete problems

Binary graph segmentation

Given $y \in \mathbb{R}^n$, undirected graph $G = (V, E)$ with $V = \{1, \ldots n\}$, consider binary graph segmentation:

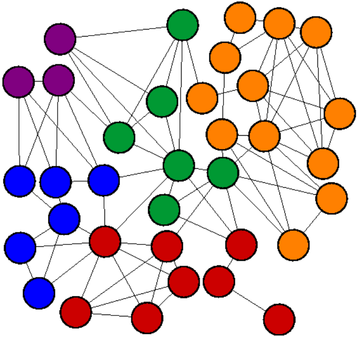
$$\min_{\beta \in \{0,1\}^n} \sum_{i=1}^n (y_i - \beta_i)^2 + \sum_{(i,j) \in E} \lambda_{ij} \cdot 1\{\beta_i \neq \beta_j\}$$

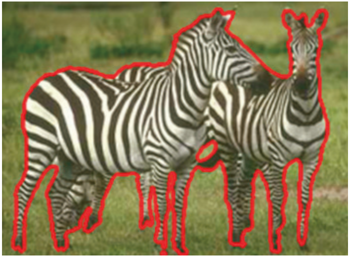
Nonconvex, but simple manipulation delivers the equivalent form

$$\max_{A \subseteq \{1,\ldots n\}} \sum_{i \in A} a_i + \sum_{j \in A^c} b_j - \sum_{(i,j) \in E,\, |A \cap \{i,j\}| = 1} \lambda_{ij}$$

which is a segmentation problem that can be solved exactly using min cut/max flow. E.g., Kleinberg and Tardos (2005), "Algorithm design", Chapter 7

Can apply recursively to get a verison of graph hierarchical clustering (divisive)





E.g., take the graph as a 2d grid for image segmentation (From `http://ailab.snu.ac.kr`)

Discrete Potts minimization

Given $y \in \mathbb{R}^n$, now consider discrete Potts minimization:

$$\min_{\beta \in \{b_1,\ldots b_k\}^n} \sum_{i=1}^n (y_i - \beta_i)^2 + \lambda \sum_{i=1}^{n-1} 1\{\beta_i \neq \beta_{i+1}\}$$

where $\{b_1, \ldots b_k\}$ is some fixed discrete set. This is nonconvex and can be efficiently solved using classical dynamic programming

Key insight is that the 1-dimensional structure allows us to exactly solve and store

$$\hat{\beta}_1(\beta_2) = \operatorname*{argmin}_{\beta_1 \in \{b_1,\ldots b_k\}} \underbrace{(y_1 - \beta_1)^2 + \lambda \cdot 1\{\beta_1 \neq \beta_2\}}_{f_1(\beta_1,\beta_2)}$$

$$\hat{\beta}_2(\beta_3) = \operatorname*{argmin}_{\beta_2 \in \{b_1,\ldots b_k\}} f_1\big(\hat{\beta}_1(\beta_2), \beta_2\big) + (y_2 - \beta_2)^2 + \lambda \cdot 1\{\beta_2 \neq \beta_3\}$$

$$\ldots$$

DP agorithm:

- Make a forward pass over $\beta_1, \ldots \beta_{n-1}$, keeping a look-up table; also keep a look-up table for the optimal partial criterion values $f_1, \ldots f_{n-1}$
- Solve exactly for β_n
- Make a backward pass $\beta_{n-1}, \ldots \beta_1$, reading off the look-up table

	b_1	b_2	$\ldots$	b_k
β_1				
β_2				
$\ldots$				
β_{n-1}				

	b_1	b_2	$\ldots$	b_k
f_1				
f_2				
$\ldots$				
f_{n-1}				

Requires $O(nk^2)$ operations

Nearly optimal K-means

Given data points $x_1, \ldots x_n \in \mathbb{R}^p$, the K-means problem is

$$\min_{c_1,\ldots c_K} \underbrace{\sum_{i=1}^n \min_{k=1,\ldots K} \|x_i - c_k\|_2^2}_{f(c_1,\ldots c_K)}$$

This is nonconvex, NP hard, and it is usually approximately solved using Lloyd's algorithm, run many times, with random starts

Careful choice of starting positions makes a big impact: if we run Lloyd's algorithm once, starting at $c_1 = s_1, \ldots c_K = s_K$, for special (random) $s_1, \ldots s_K$, then we get estimates $\hat{c}_1, \ldots \hat{c}_K$ with

$$\mathbb{E}\big[f(\hat{c}_1, \ldots \hat{c}_K)\big] \leq 8(\log k + 2) \cdot \min_{c_1,\ldots c_K \in \mathbb{R}^p} f(c_1, \ldots c_K)$$

See Arthur and Vassilvitskii (2007), "k-means++: The advantages of careful seeding". Their construction of $s_1, \ldots s_K$ is simple:

- Begin by choosing s_1 uniformly at random among $x_1, \ldots x_n$
- Compute squared distances

  $$d_i^2 = \|x_i - s_1\|_2^2$$

  for all points i not chosen, and choose s_2 by drawing from the remaining points, with probability weights $d_i^2 / \sum_j d_j^2$
- Recompute the squared distances as

  $$d_i^2 = \min\left\{\|x_i - s_1\|_2^2, \|x_i - s_2\|_2^2\right\}$$

  and choose s_3 according to the same recipe
- And so on, until all of $s_1, \ldots s_K$ are chosen

Infinite-dimensional problems

Smoothing splines

Given pairs $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, \ldots n$, smoothing spline solves

$$\min_f \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 + \lambda \int \big(f''(t)\big)^2 \, dt$$

Optimization domain: all functions f such that $\int (f''(t))^2 \, dt < \infty$, infinite-dimensional set

Can show that the solution $\hat{f}$ to the above problem is unique, and given by natural cubic spline, that has knots at $x_1, \ldots x_n$. (Proof: use integration by parts.) Hence we can parametrize by

$$f = \sum_{j=1}^n \theta_j \eta_j$$

where $\eta_1, \ldots \eta_n$ are natural cubic spline basis functions. Task now is to solve for coefficients $\theta \in \mathbb{R}^n$

Plugging in $f = \sum_{j=1}^n \theta_j \eta_j$, transform smoothing spline problem into finite-dimensional form:

$$\min_\theta \; \|y - N\theta\|_2^2 + \lambda \theta^T \Omega \theta$$

where $N_{ij} = \eta_j(x_i)$, and $\Omega_{ij} = \int \eta_i''(t)\, \eta_j''(t)\, dt$. The solution is explicitly given by

$$\hat{\theta} = (N^T N + \lambda\Omega)^{-1} N^T y$$

and fitted function is $\hat{f} = \sum_{j=1}^n \hat{\theta}_j \eta_j$. With proper choice of basis function (B-splines), calculation of $\hat{\theta}$ is $O(n)$

See Wahba (1990), “Splines models for observational data”; Green and Silverman (1994), “Nonparametric regression and generalized linear models”; Hastie et al. (2009), Chapter 5

Locally adaptive regression splines

Given same setup, (cubic) locally adaptive regression spline solves

$$\min_f \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 + \lambda \cdot \mathrm{TV}(f''')$$

Optimization domain: all functions f with $\mathrm{TV}(f''') < \infty$, which is again infinite-dimensional

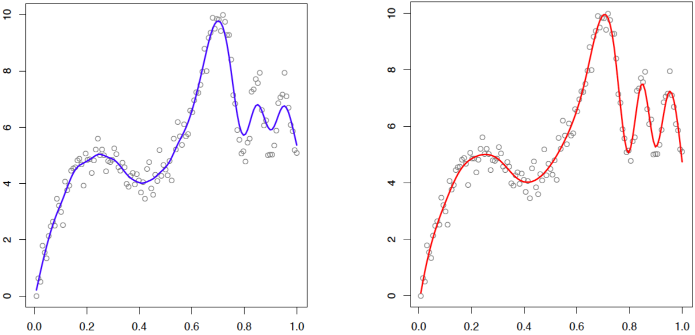
Similar to before, can show that the solution $\hat{f}$ to above problem is a cubic spline, but two key differences:

- Can have any number of knots $\leq n - 4$ (tuned by λ)
- Knots do not necessarily lie at input points $x_1, \ldots x_n$!

Details in Mammen and van de Geer (1997), "Locally adaptive regression splines". Summary: these are statistically more adaptive but computationally more challenging than smoothing splines

Smoothing spline
(easier to compute)

Locally adaptive spline
(more adaptive)



Finite-dimensional approximation is given in Mammen and van de Geer (1997), and a much faster approximation in Tibshirani (2014), “Adaptive piecewise polynomial estimation via trend filtering”

Reproducing kernel Hilbert spaces

Let $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{++}$ be a positive definite kernel, and $\mathcal{H}_\kappa$ the function space generated by (possibly infinite) linear combinations of functions $\kappa(\cdot, z)$, $z \in \mathbb{R}^d$. This is a reproducing kernel Hilbert space or RKHS, and is equipped with a norm $\|\cdot\|_{\mathcal{H}_\kappa}$

Given data $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \dots n$, consider the problem

$$\min_{f \in \mathcal{H}_\kappa} \sum_{i=1}^n \big(y_i - f(x_i)\big)^2 + \lambda \|f\|_{\mathcal{H}_\kappa}^2$$

This is infinite-dimensional, but by Mercer's Theorem, the solution satisfies $f = \sum_{j=1}^n \alpha_j K(\cdot, x_j)$. Letting $K \in \mathbb{R}^{n \times n}$ have elements $K_{ij} = \kappa(x_i, x_j)$, our problem reduces to

$$\min_\alpha \|y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha$$

(Beyond regression: same trick for kernel SVMs, kernel PCA, etc.)

Statistical problems

Sparse underdetermined linear systems

Suppose that $X \in \mathbb{R}^{n \times p}$ has unit normed columns, $\|X_i\|_2 = 1$, for $i = 1, \ldots n$. Given $y \in \mathbb{R}^n$, consider the problem of finding sparsest solution to linear system

$$\min_{\beta} \ \|\beta\|_0 \ \text{ subject to } \ X\beta = y$$

This is nonconvex and known to be NP hard, for a generic X. A natural convex relaxation is the ℓ_1 basis pursuit problem:

$$\min_{\beta} \ \|\beta\|_1 \ \text{ subject to } \ X\beta = y$$

It turns out that there is a deep connection between the two; we cite results from Donoho (2006), "For most large underdetermined systems of linear equations, the minimal ℓ_1 norm solution is also the sparsest solution"

As n, p grow large, $p > n$, there exists a threshold ρ (depending on the ratio p/n), such that for most matrices X, the following holds. If we solve the ℓ_1 problem and find a solution with:

- fewer than ρn nonzero components, then we must have found the unique solution of the ℓ_0 problem
- greater than ρn nonzero components, then there is no solution of the linear system with less than ρn nonzero components

Here "most" can be quantified precisely via a uniform probability measure over matrices X with unit norm columns

There is a large and fast-moving body of related literature. See Donoho et al. (2009), "Message-passing algorithms for compressed sensing" for a nice review

Exact low-rank matrix completion

Given a matrix $Y \in \mathbb{R}^{n \times n}$, partially observed, over a set of indices $\Omega \subseteq \{1, \ldots, n\}^2$. Consider the problem of finding the lowest-rank matrix matching Y on the observed set

$$\min_B \ \mathrm{rank}(B) \ \text{ subject to } \ B_{ij} = Y_{ij}, \ (i,j) \in \Omega$$

This is nonconvex. Natural convex relaxation:

$$\min_B \ \|B\|_{\mathrm{tr}} \ \text{ subject to } \ B_{ij} = Y_{ij}, \ (i,j) \in \Omega$$

Under some assumptions, it can be shown that the solution to the convex problem is exactly equal to the solution to the nonconvex problem, with high probability over the sampling model

See, e.g., Candes and Recht (2008), "Exact matrix completion via convex optimization", and many papers since

Miscellaneous

Curvature domination

Given convex f and nonconvex g, consider a problem

$$\min_x \; f(x) + g(x)$$

For convex h, we can always rewrite this problem as

$$\min_x \; \underbrace{f(x) + g(x) - h(x)}_{F(x)} + h(x)$$

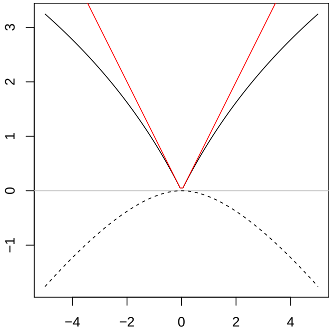
Sometimes, can choose h to make F smooth and strictly convex! How? Prove that $\nabla^2 f(x) \succ \nabla^2 (h-g)(x)$ for all x

This is apparently an old idea. See Parekh and Selesnick (2015), "Convex fused lasso denoising with non-convex regularization and its use for pulse detection" and references therein. (Related ideas on curvature domination from statistics: Zhang, Loh, Wainwright)

Setting of Parekh and Selesnick (2015) (simplified):

$$\min_{\beta} \; \frac{1}{2}\|y - \beta\|_2^2 + \lambda \sum_{i=1}^{n-1} \phi_a(\beta_i - \beta_{i+1})$$

where $\phi_a(x) = \frac{1}{a}\log(1 + a|x|)$, for $a > 0$. This uses a nonconvex segmentation penalty — whole problem appears to be nonconvex



Fact: $s_a(x) = \phi_a(x) - |x|$ is twice differentiable, strictly concave, with $-a \leq s_a''(x) \leq 0$ for all x

Rewrite problem (denoting by D the first difference operator) as

$$\min_{\beta} \; \underbrace{\frac{1}{2}\|y-\beta\|_2^2 + \lambda \sum_{i=1}^{n-1} \big(\phi_a([D\beta]_i) - |[D\beta]_i|\big)}_{F_a(\beta)} + \lambda\|D\beta\|_1$$

Now compute

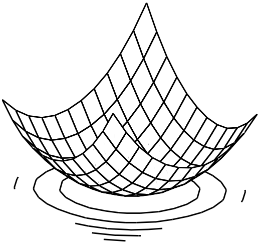
$$\nabla^2 F_a(\beta) = I + \lambda D^T S_a(D\beta) D$$

where $S_a(x) = \mathrm{diag}(s_a''(x_1), \ldots s_a''(x_{n-1}))$. Easy to check that

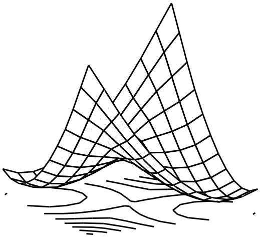
$$D^T S_a(D\beta) D \succeq -a D^T D \succeq -4a$$

using our previous fact, and $\lambda_{\max}(D^T D) = 4$. Thus $F_a(\beta)$ — and our whole problem — is strictly convex provided $1 - 4a\lambda > 0$, i.e., $a < 1/(4\lambda)$

From Parekh and Selesnick (2015): contour plots of criterion



a small enough



a too large

Gradient descent converges to minimizers

Given f twice continuously differentiable, $\mathrm{dom}(f) = \mathbb{R}^n$, and initial point $x^{(0)} \in \mathbb{R}^n$. Our old friend gradient descent, repeats:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \dots$$

When f has Lipschitz gradient, constant $L > 0$, but is nonconvex, can anything be said?

A very old problem. Remarkable new result from Lee et al. (2016), "Gradient descent red converges to minimizers": if step sizes are small enough $t_k \leq 1/L$, $k = 1, 2, 3, \dots$, and $x^{(0)}$ is drawn from any density over $\mathbb{R}^n$, then saddle points are unlikely limit points, i.e.,

$$\mathbb{P}\Big(\lim_{k \to \infty} x^{(k)} = \tilde{x}\Big) = 0, \quad \text{for any isolated strict saddle point } \tilde{x} \text{ of } f$$

Panageas and Piliouras (2016) have already relaxed isolated saddle point and global Lipschitz conditions