Convex Optimization M2

Lecture 8

Applications

Outline

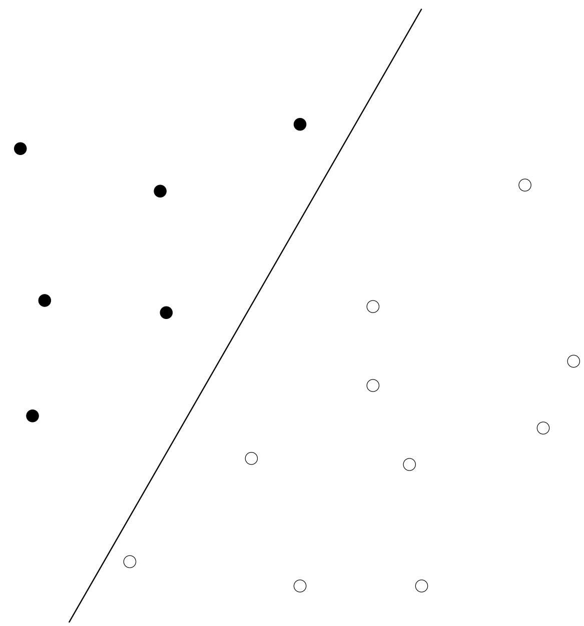
- Geometrical problems
- Approximation problems
- Combinatorial optimization
- Statistics

Geometrical problems

Linear discrimination

separate two sets of points $\{x_1, \ldots, x_N\}$, $\{y_1, \ldots, y_M\}$ by a hyperplane:

$$a^T x_i + b > 0, \quad i = 1, \ldots, N, \qquad a^T y_i + b < 0, \quad i = 1, \ldots, M$$



homogeneous in a, b, hence equivalent to

$$a^T x_i + b \geq 1, \quad i = 1, \ldots, N, \qquad a^T y_i + b \leq -1, \quad i = 1, \ldots, M$$

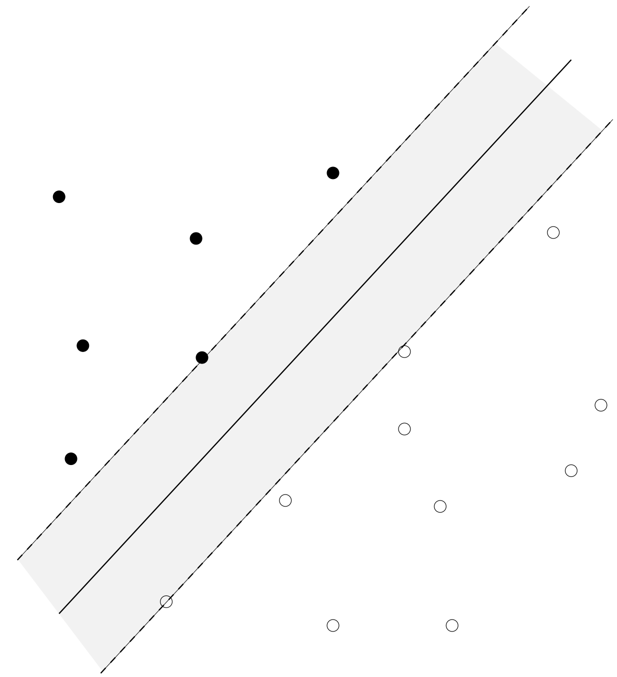
a set of linear inequalities in a, b

Robust linear discrimination

(Euclidean) distance between hyperplanes

$$\begin{aligned} \mathcal{H}_1 &= \{z \mid a^T z + b = 1\} \\ \mathcal{H}_2 &= \{z \mid a^T z + b = -1\} \end{aligned}$$

is $\mathbf{dist}(\mathcal{H}_1, \mathcal{H}_2) = 2/\|a\|_2$



to separate two sets of points by maximum margin,

$$\begin{array}{ll} \mbox{minimize} & (1/2)\|a\|_2 \\ \mbox{subject to} & a^T x_i + b \geq 1, \quad i = 1, \ldots, N \\ & a^T y_i + b \leq -1, \quad i = 1, \ldots, M \end{array} \tag{1}$$

(after squaring objective) a QP in a, b

Lagrange dual of maximum margin separation problem

$$\begin{array}{ll} \text{maximize} & \mathbf{1}^T\lambda + \mathbf{1}^T\mu \\ \text{subject to} & 2\left\|\sum_{i=1}^N \lambda_i x_i - \sum_{i=1}^M \mu_i y_i\right\|_2 \leq 1 \\ & \mathbf{1}^T\lambda = \mathbf{1}^T\mu, \quad \lambda \succeq 0, \quad \mu \succeq 0 \end{array} \tag{2}$$

from duality, optimal value is inverse of maximum margin of separation

interpretation

- change variables to $\theta_i = \lambda_i/\mathbf{1}^T\lambda$, $\gamma_i = \mu_i/\mathbf{1}^T\mu$, $t = 1/(\mathbf{1}^T\lambda + \mathbf{1}^T\mu)$
- invert objective to minimize $1/(\mathbf{1}^T\lambda + \mathbf{1}^T\mu) = t$

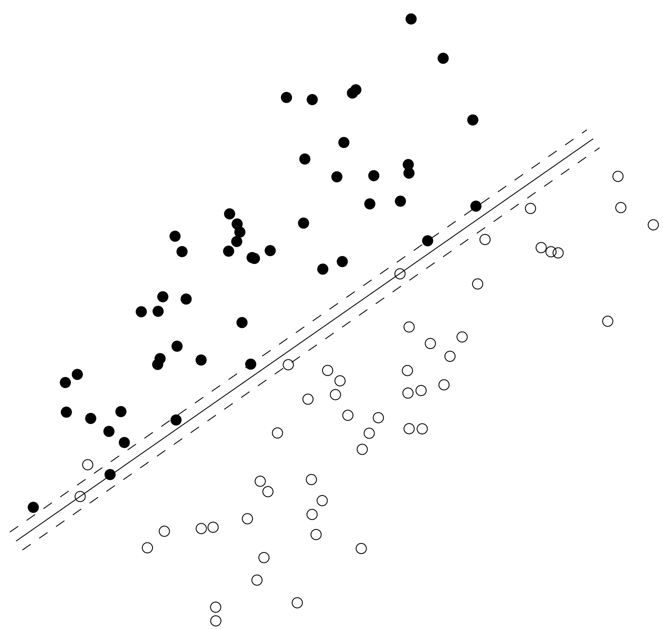
$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \left\|\sum_{i=1}^N \theta_i x_i - \sum_{i=1}^M \gamma_i y_i\right\|_2 \leq t \\ & \theta \succeq 0, \quad \mathbf{1}^T\theta = 1, \quad \gamma \succeq 0, \quad \mathbf{1}^T\gamma = 1 \end{array}$$

optimal value is distance between convex hulls

Approximate linear separation of non-separable sets

$$
\begin{array}{ll}
\text{minimize} & \mathbf{1}^T u + \mathbf{1}^T v \\
\text{subject to} & a^T x_i + b \geq 1 - u_i, \quad i = 1, \ldots, N \\
& a^T y_i + b \leq -1 + v_i, \quad i = 1, \ldots, M \\
& u \succeq 0, \quad v \succeq 0
\end{array}
$$

- an LP in a, b, u, v
- at optimum, $u_i = \max\{0, 1 - a^T x_i - b\}$, $v_i = \max\{0, 1 + a^T y_i + b\}$
- can be interpreted as a heuristic for minimizing #misclassified points

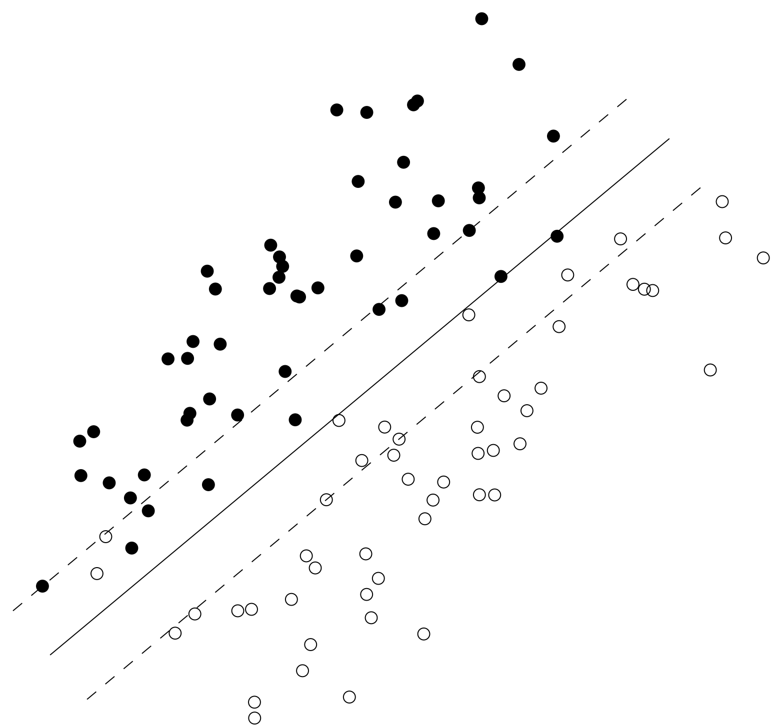


Support vector classifier

$$\begin{array}{ll} \text{minimize} & \|a\|_2 + \gamma(\mathbf{1}^T u + \mathbf{1}^T v) \\ \text{subject to} & a^T x_i + b \geq 1 - u_i, \quad i = 1, \ldots, N \\ & a^T y_i + b \leq -1 + v_i, \quad i = 1, \ldots, M \\ & u \succeq 0, \quad v \succeq 0 \end{array}$$

produces point on trade-off curve between inverse of margin $2/\|a\|_2$ and classification error, measured by total slack $\mathbf{1}^T u + \mathbf{1}^T v$

same example as previous page, with $\gamma = 0.1$:



Support Vector Machines: Duality

Given m data points $x_i \in \mathbb{R}^n$ with labels $y_i \in \{-1, 1\}$.

- The maximum margin classification problem can be written

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|_2^2 + C\mathbf{1}^T z \\ \text{subject to} & y_i(w^T x_i) \geq 1 - z_i, \quad i = 1, \ldots, m \\ & z \geq 0 \end{array}$$

  in the variables w, $z \in \mathbb{R}^n$, with parameter $C > 0$.

- We can set $w = (w, \mathbf{1})$ and increase the problem dimension by 1. So we can assume w.l.o.g. $b = 0$ in the classifier $w^T x_i + b$.
- The Lagrangian is written

$$L(w, z, \alpha) = \frac{1}{2}\|w\|_2^2 + C\mathbf{1}^T z + \sum_{i=1}^{m} \alpha_i (1 - z_i - y_i w^T x_i)$$

  with dual variable $\alpha \in \mathbb{R}^m_+$.

Support Vector Machines: Duality

- The Lagrangian can be rewritten

$$L(w,z,\alpha)=\frac{1}{2}\left(\left\|w-\sum_{i=1}^{m}\alpha_i y_i x_i\right\|_2^2-\left\|\sum_{i=1}^{m}\alpha_i y_i x_i\right\|_2^2\right)+(C\mathbf{1}-\alpha)^T z+\mathbf{1}^T\alpha$$

with dual variable $\alpha \in \mathbb{R}^n_+$.

- Minimizing in (w,z) we form the dual problem

$$\begin{array}{ll}\text{maximize} & -\frac{1}{2}\left\|\sum_{i=1}^{m}\alpha_i y_i x_i\right\|_2^2+\mathbf{1}^T\alpha\\ \text{subject to} & 0\leq\alpha\leq C\end{array}$$

- At the optimum, we must have

$$w=\sum_{i=1}^{m}\alpha_i y_i x_i \quad \text{and} \quad \alpha_i = C \text{ if } z_i>0$$

(this is the representer theorem).

Support Vector Machines: the kernel trick

- If we write X the data matrix with columns x_i, the dual can be rewritten

$$\begin{array}{ll} \text{maximize} & -\frac{1}{2}\alpha^T \,\mathbf{diag}(y) X^T X \,\mathbf{diag}(y)\alpha + \mathbf{1}^T\alpha \\ \text{subject to} & 0 \leq \alpha \leq C \end{array}$$

- This means that the data only appears in the dual through the gram matrix

$$K = X^T X$$

  which is called the **kernel** matrix.

- In particular, the original **dimension n does not appear in the dual.** SVM complexity only grows with the number of samples.
- In particular, the x_i are allowed to be infinite dimensional.
- The only requirement on K is that $K \succeq 0$.

Approximation problems

Norm approximation

$$\text{minimize} \quad \|Ax - b\|$$

($A \in \mathbb{R}^{m \times n}$ with $m \geq n$, $\|\cdot\|$ is a norm on $\mathbb{R}^m$)

interpretations of solution $x^\star = \mathop{\rm argmin}_x \|Ax - b\|$:

- **geometric**: $Ax^\star$ is point in $\mathcal{R}(A)$ closest to b
- **estimation**: linear measurement model

$$y = Ax + v$$

  y are measurements, x is unknown, v is measurement error

  given $y = b$, best guess of x is $x^\star$
- **optimal design**: x are design variables (input), Ax is result (output)

  $x^\star$ is design that best approximates desired result b

examples

- least-squares approximation ($\|\cdot\|_2$): solution satisfies normal equations

$$A^T Ax = A^T b$$

($x^\star = (A^T A)^{-1} A^T b$ if $\mathbf{Rank}\, A = n$)

- Chebyshev approximation ($\|\cdot\|_\infty$): can be solved as an LP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & -t\mathbf{1} \preceq Ax - b \preceq t\mathbf{1} \end{array}$$

- sum of absolute residuals approximation ($\|\cdot\|_1$): can be solved as an LP

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T y \\ \text{subject to} & -y \preceq Ax - b \preceq y \end{array}$$

Penalty function approximation

$$\begin{array}{ll} \text{minimize} & \phi(r_1) + \cdots + \phi(r_m) \\ \text{subject to} & r = Ax - b \end{array}$$

($A \in \mathbb{R}^{m \times n}$, $\phi : \mathbb{R} \to \mathbb{R}$ is a convex penalty function)

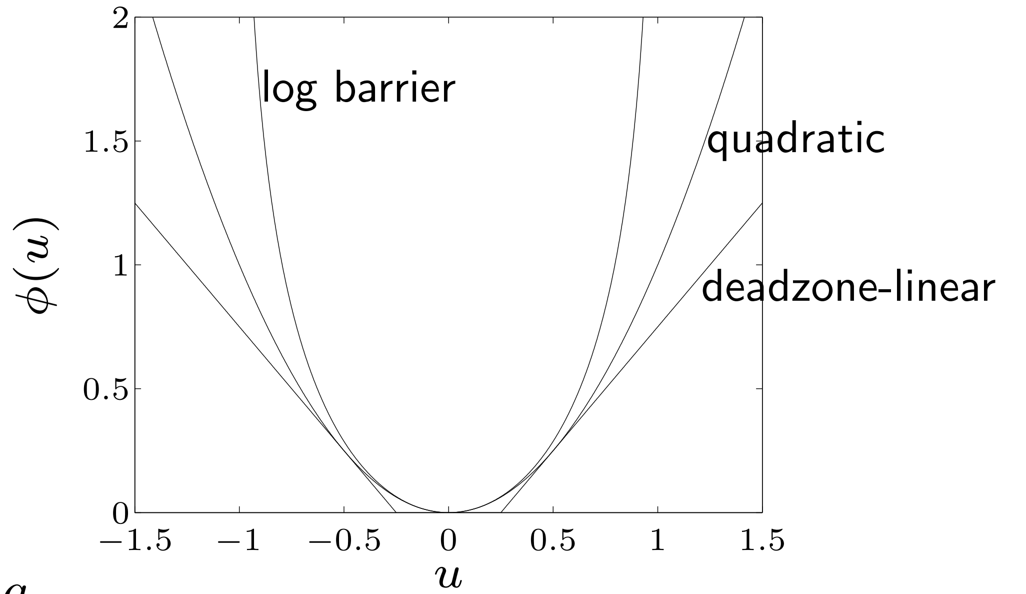
examples

- quadratic: $\phi(u) = u^2$
- deadzone-linear with width a:

$$\phi(u) = \max\{0, |u| - a\}$$

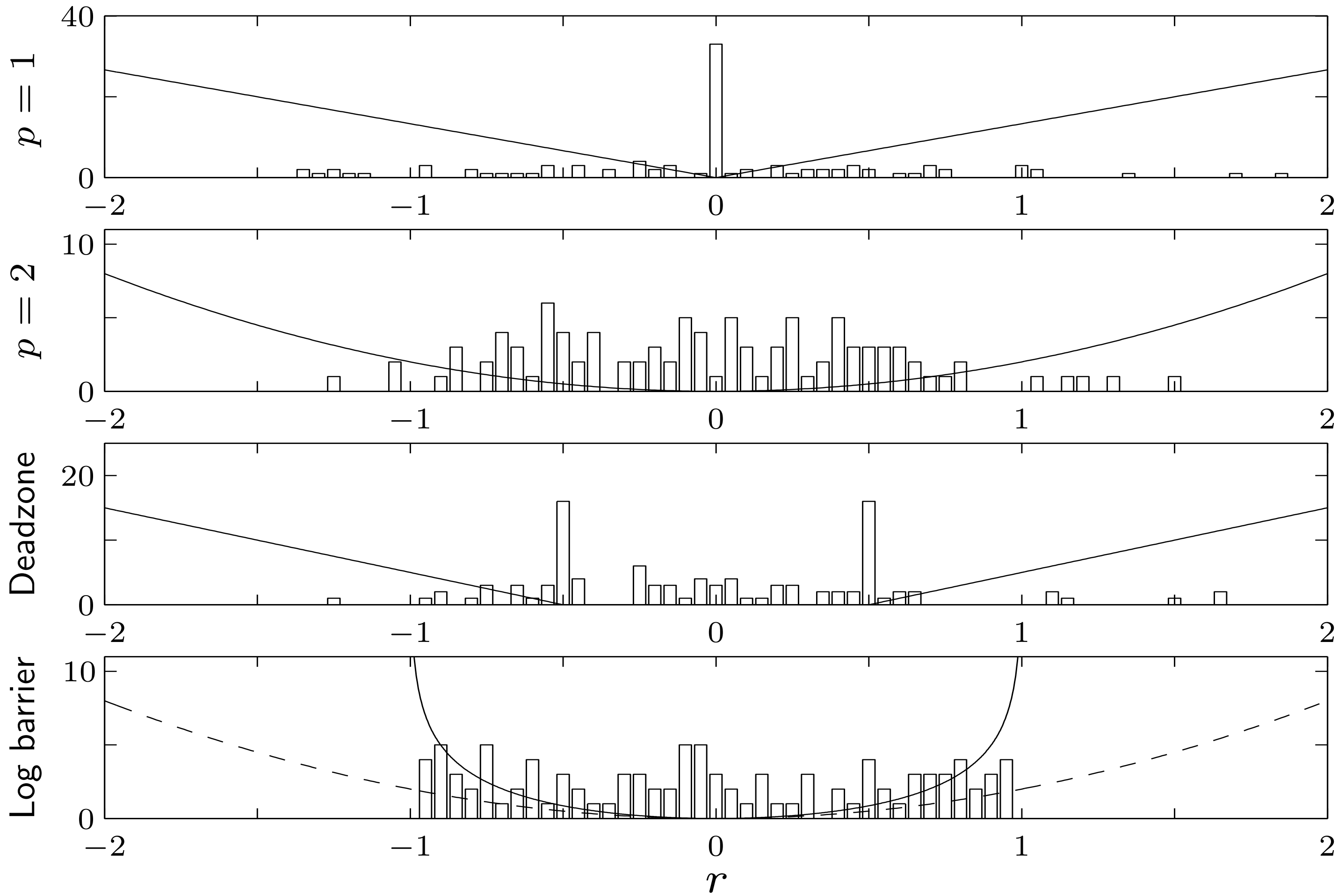
- log-barrier with limit a:

$$\phi(u) = \begin{cases} -a^2 \log(1 - (u/a)^2) & |u| < a \\ \infty & \text{otherwise} \end{cases}$$



example ($m = 100$, $n = 30$): histogram of residuals for penalties

$$\phi(u) = |u|, \quad \phi(u) = u^2, \quad \phi(u) = \max\{0, |u| - a\}, \quad \phi(u) = -\log(1 - u^2)$$

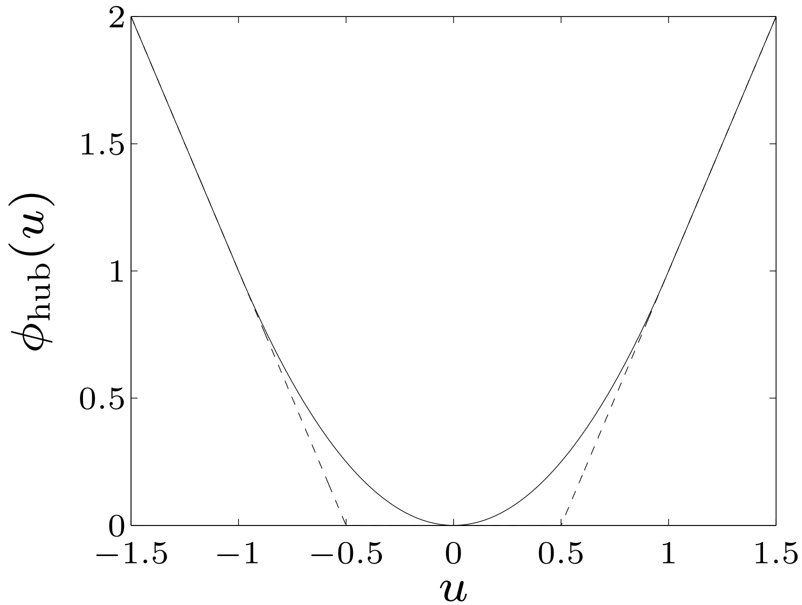


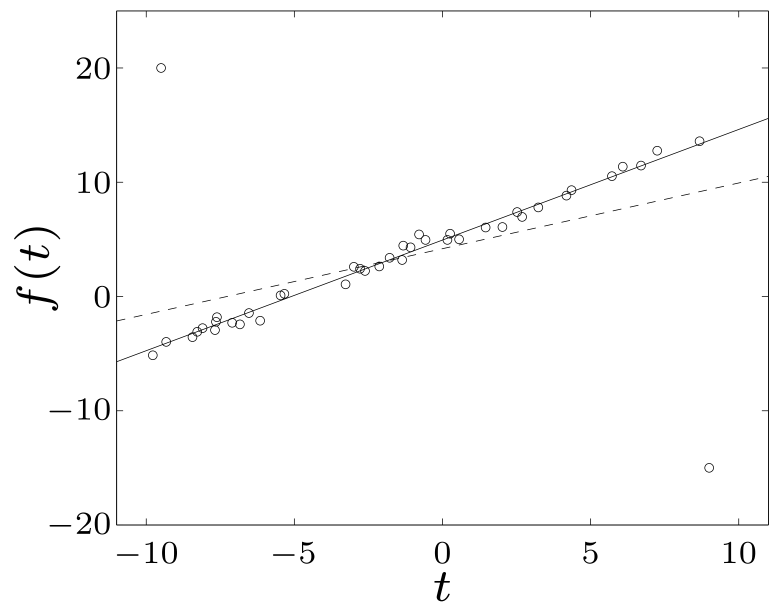
shape of penalty function has large effect on distribution of residuals

Huber penalty function (with parameter M)

$$\phi_{\text{hub}}(u) = \begin{cases} u^2 & |u| \leq M \\ M(2|u| - M) & |u| > M \end{cases}$$

linear growth for large u makes approximation less sensitive to outliers





- left: Huber penalty for $M = 1$
- right: affine function $f(t) = \alpha + \beta t$ fitted to 42 points t_i, y_i (circles) using quadratic (dashed) and Huber (solid) penalty

Combinatorial problems

Nonconvex Problems

Nonconvexity makes problems **essentially untractable**...

- sometimes the result of bad problem formulation
- however, often arises because of some natural limitation: fixed transaction costs, binary communications, ...

What can be done?... we will use convex optimization results to:

- find bounds on the optimal value, by **relaxation**
- get "good" feasible points via **randomization**

Nonconvex Problems

Focus here on a specific class of problems, general QCQPs, written

$$\begin{array}{ll}\text{minimize} & x^T P_0 x + q_0^T x + r_0 \\ \text{subject to} & x^T P_i x + q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m\end{array}$$

- if all P_i are p.s.d., this is a convex problem...
- so here, we suppose at least one P_i is not p.s.d.

Example: Boolean Least Squares

Boolean least-squares problem:

$$\begin{array}{ll} \text{minimize} & \|Ax-b\|^2 \\ \text{subject to} & x_i^2 = 1, \quad i = 1,\ldots,n \end{array}$$

- basic problem in digital communications
- could check all 2^n possible values of x ...
- an NP-hard problem, and very hard in practice
- many heuristics for approximate solution

Example: Partitioning Problem

Two-way partitioning problem described in §5.1.4 of the textbook

$$\begin{array}{ll} \text{minimize} & x^T W x \\ \text{subject to} & x_i^2 = 1, \quad i = 1, \ldots, n \end{array}$$

where $W \in \mathbf{S}^n$, with $W_{ii} = 0$.

- a feasible x corresponds to the partition

$$\{1, \ldots, n\} \;=\; \{i \mid x_i = -1\} \;\cup\; \{i \mid x_i = 1\}$$

- the matrix coefficient W_{ij} can be interpreted as the cost of having the elements i and j in the same partition.
- the objective is to find the partition with least total cost
- classic particular instance: MAXCUT $(W_{ij} \geq 0)$

Convex Relaxation

the original QCQP

$$\begin{array}{ll} \text{minimize} & x^T P_0 x + q_0^T x + r_0 \\ \text{subject to} & x^T P_i x + q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m \end{array}$$

can be rewritten

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(X P_0) + q_0^T x + r_0 \\ \text{subject to} & \mathbf{Tr}(X P_i) + q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m \\ & X \succeq x x^T \\ & \mathbf{Rank}(X) = 1 \end{array}$$

the only nonconvex constraint is now $\mathbf{Rank}(X) = 1$...

Convex Relaxation: Semidefinite Relaxation

- we can directly relax this last constraint, i.e. drop the nonconvex $\mathbf{Rank}(X) = 1$ to keep only $X \succeq xx^T$
- the resulting program gives a lower bound on the optimal value

$$\begin{array}{ll} \mbox{minimize} & \mathbf{Tr}(XP_0) + q_0^T x + r_0 \\ \mbox{subject to} & \mathbf{Tr}(XP_i) + q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m \\ & X \succeq xx^T \end{array}$$

Tricky. . . Can be improved?

Lagrangian Relaxation

Start from the original problem

$$\begin{array}{ll} \mbox{minimize} & x^TP_0x+q_0^Tx+r_0\\ \mbox{subject to} & x^TP_ix+q_i^Tx+r_i\leq 0,\quad i=1,\ldots,m \end{array}$$

form the Lagrangian

$$L(x,\lambda)=x^T\left(P_0+\sum_{i=1}^m\lambda_iP_i\right)x+\left(q_0+\sum_{i=1}^m\lambda_iq_i\right)^Tx+r_0+\sum_{i=1}^m\lambda_ir_i$$

in the variables $x\in\mathbb{R}^n$ and $\lambda\in\mathbb{R}^m_+$...

Lagrangian Relaxation: Lagrangian

the dual can be computed explicitly as an (unconstrained) quadratic minimization problem

$$\inf_{x\in\mathbb{R}} x^TPx + q^Tx + r = \begin{cases} r - \frac{1}{4}q^TP^\dagger q, & \text{if } P \succeq 0 \text{ and } q \in \mathcal{R}(P) \\ -\infty, & \text{otherwise} \end{cases}$$

so

$$\begin{aligned} \inf_x L(x,\lambda) = \quad & -\tfrac{1}{4}\left(q_0 + \textstyle\sum_{i=1}^m \lambda_i q_i\right)^T \left(P_0 + \textstyle\sum_{i=1}^m \lambda_i P_i\right)^\dagger \left(q_0 + \textstyle\sum_{i=1}^m \lambda_i q_i\right) \\ & + \textstyle\sum_{i=1}^m \lambda_i r_i + r_0 \end{aligned}$$

where we recognize a Schur complement...

Lagrangian Relaxation: Dual

the dual of the QCQP is then given by

$$\begin{array}{ll}\text{maximize} & \gamma + \sum_{i=1}^m \lambda_i r_i + r_0 \\ \text{subject to} & \left[\begin{array}{cc} \left(P_0 + \sum_{i=1}^m \lambda_i P_i\right) & \left(q_0 + \sum_{i=1}^m \lambda_i q_i\right)/2 \\ \left(q_0 + \sum_{i=1}^m \lambda_i q_i\right)^T/2 & -\gamma \end{array}\right] \succeq 0 \\ & \lambda_i \geq 0, \quad i = 1, \ldots, m \end{array}$$

which is a semidefinite program in the variable $\lambda \in \mathbb{R}^m$ and can be solved efficiently

Let us look at what happens when we use semidefinite duality to compute the dual of this last program (bidual of the original problem)...

Lagrangian Relaxation: Bidual

Taking the dual again, we get an SDP is given by

$$\begin{array}{ll}\mbox{minimize} & \mathbf{Tr}(XP_0) + q_0^T x + r_0 \\ \mbox{subject to} & \mathbf{Tr}(XP_i) + q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m \\ & \left[\begin{array}{cc} X & x^T \\ x & 1 \end{array}\right] \succeq 0 \end{array}$$

in the variables $X \in \mathbf{S}^n$ and $x \in \mathbb{R}^n$

- this is a convexification of the original program
- we have recovered the semidefinite relaxation in an "automatic" way

Lagrangian Relaxation: Boolean LS

An example: boolean lest squares

Using this technique, we can relax the original Boolean LS problem

$$\begin{array}{ll} \text{minimize} & \|Ax-b\|^2 \\ \text{subject to} & x_i^2 = 1, \quad i = 1,\ldots,n \end{array}$$

and relax it as an SDP

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(AX) + 2b^T Ax + b^T b \\ \text{subject to} & \left[\begin{array}{cc} X & x^T \\ x & 1 \end{array}\right] \succeq 0 \\ & X_{ii} = 1, \quad i = 1,\ldots,n, \end{array}$$

this program then produces a lower bound on the optimal value of the original Boolean LS program

Lagrangian Relaxation: Partitioning

the partitioning problem defined above is

$$\begin{array}{ll} \text{minimize} & x^T W x \\ \text{subject to} & x_i^2 = 1, \quad i = 1, \ldots, n \end{array}$$

the variable x disappears from the relaxation, which becomes:

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(WX) \\ \text{subject to} & X \succeq 0 \\ & X_{ii} = 1, \quad i = 1, \ldots, n \end{array}$$

Feasible points?

- Lagrangian relaxations only provide lower bounds on the optimal value
- how can we compute good feasible points?
- can we measure how suboptimal this lower bound is?

Randomization

The original QCQP

$$
\begin{array}{ll}
\text{minimize} & x^T P_0 x + q_0^T x + r_0 \\
\text{subject to} & x^T P_i x + q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m
\end{array}
$$

was relaxed into

$$
\begin{array}{ll}
\text{minimize} & \mathbf{Tr}(XP_0) + q_0^T x + r_0 \\
\text{subject to} & \mathbf{Tr}(XP_i) + q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m \\
& \left[\begin{array}{cc} X & x^T \\ x & 1 \end{array} \right] \succeq 0
\end{array}
$$

- the last (Schur complement) constraint is equivalent to $X - xx^T \succeq 0$
- hence, if x and X are the solution to the relaxed program, then $X - xx^T$ is a covariance matrix...

Randomization

For the problem

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(XP_0) + q_0^T x + r_0 \\ \text{subject to} & \mathbf{Tr}(XP_i) + q_i^T x + r_i \leq 0, \quad i = 1, \ldots, m \\ & \left[\begin{array}{cc} X & x^T \\ x & 1 \end{array}\right] \succeq 0 \end{array}$$

- pick y as a Gaussian variable with $y \sim \mathcal{N}(x, X - xx^T)$
- y will solve the QCQP "on average" over this distribution

in other words, with $\mathbf{E}[yy^T] = \mathbf{E}[(y-x)(y-x)^T] + 2 \times 0 + xx^T = X$, which means $\mathbf{E}[y^T P y] = \mathbf{Tr}(PX)$ and the problem above becomes

$$\begin{array}{ll} \text{minimize} & \mathbf{E}[y^T P_0 y + q_0^T y + r_0] \\ \text{subject to} & \mathbf{E}[y^T P_i y + q_i^T y + r_i] \leq 0, \quad i = 1, \ldots, m \end{array}$$

a good feasible point can then be obtained by sampling enough y...

Bounds on suboptimality

- In certain particular cases, it is possible to get a hard bound on the gap between the optimal value and the relaxation result
- A classic example is that of the MAXCUT bound

The MAXCUT problem is a particular case of the partitioning problem

$$\begin{array}{ll} \text{maximize} & x^T W x \\ \text{subject to} & x_i^2 = 1, \quad i = 1, \ldots, n \end{array}$$

with $W \succeq 0$, its Lagrangian relaxation is computed as

$$\begin{array}{ll} \text{maximize} & \mathbf{Tr}(WX) \\ \text{subject to} & X \succeq 0 \\ & X_{ii} = 1, \quad i = 1, \ldots, n \end{array}$$

Bounds on suboptimality: MAXCUT

Let X be a solution to this program

- We look for a feasible point by sampling a normal distribution $\mathcal{N}(0, X)$
- We convert each sample point x to a feasible point by rounding it to the nearest value in $\{-1, 1\}$, i.e. taking

$$\hat{x} = \mathbf{sgn}(x)$$

Crucially, when $\hat{x}$ is sampled using that procedure, the expected value of the objective $\mathbf{E}[\hat{x}^T W \hat{x}]$ can be computed explicitly

$$\mathbf{E}[\hat{x}^T W \hat{x}] = \frac{2}{\pi} \sum_{i,j=1}^{n} W_{ij} \arcsin(X_{ij}) = \frac{2}{\pi} \mathbf{Tr}(W \arcsin(X))$$

Bounds on suboptimality: MAXCUT

- We are guaranteed to reach this expected value $\frac{2}{\pi}\mathbf{Tr}(W\arcsin(X))$ after sampling a few (feasible) points $\hat{x}$
- Hence we know that the optimal value OPT of the MAXCUT problem satisfies

$$\frac{2}{\pi}\mathbf{Tr}(W\arcsin(X)) \leq OPT \leq \mathbf{Tr}(WX)$$

- Furthermore, with

$$X \preceq \arcsin(X),$$

we can simplify (and relax) the above expression to get

$$\frac{2}{\pi}\mathbf{Tr}(WX) \leq OPT \leq \mathbf{Tr}(WX)$$

the procedure detailed above guarantees that we can find a feasible point at most $2/\pi$ suboptimal

Numerical Example: Boolean LS

Boolean least-squares problem

$$\begin{array}{ll} \text{minimize} & \|Ax-b\|^2 \\ \text{subject to} & x_i^2 = 1, \quad i = 1, \ldots, n \end{array}$$

with

$$\begin{aligned} \|Ax-b\|^2 &= x^T A^T A x - 2b^T A x + b^T b \\ &= \mathbf{Tr}\, A^T A X - 2b^T A^T x + b^T b \end{aligned}$$

where $X = xx^T$, hence can express BLS as

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}\, A^T A X - 2b^T A x + b^T b \\ \text{subject to} & X_{ii} = 1, \quad X \succeq xx^T, \quad \text{rank}(X) = 1 \end{array}$$

. . . still a very hard problem

SDP relaxation for BLS

using Lagrangian relaxation, remember:

$$X \succeq xx^T \iff \begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \succeq 0$$

we obtained the **SDP relaxation** (with variables X, x)

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}\, A^TAX - 2b^TA^Tx + b^Tb \\ \text{subject to} & X_{ii} = 1, \quad \begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \succeq 0 \end{array}$$

- optimal value of SDP gives **lower bound** for BLS
- if optimal matrix is rank one, we're done

Interpretation via randomization

- can think of variables X, x in SDP relaxation as defining a normal distribution $z \sim \mathcal{N}(x, X - xx^T)$, with $\mathbf{E}\, z_i^2 = 1$
- SDP objective is $\mathbf{E}\, \|Az - b\|^2$

suggests randomized method for BLS:

- find $X^{\rm opt}$, $x^{\rm opt}$, optimal for SDP relaxation
- generate z from $\mathcal{N}(x^{\rm opt}, X^{\rm opt} - x^{\rm opt}x^{{\rm opt}T})$
- take $x = \mathbf{sgn}(z)$ as approximate solution of BLS
  (can repeat many times and take best one)

Example

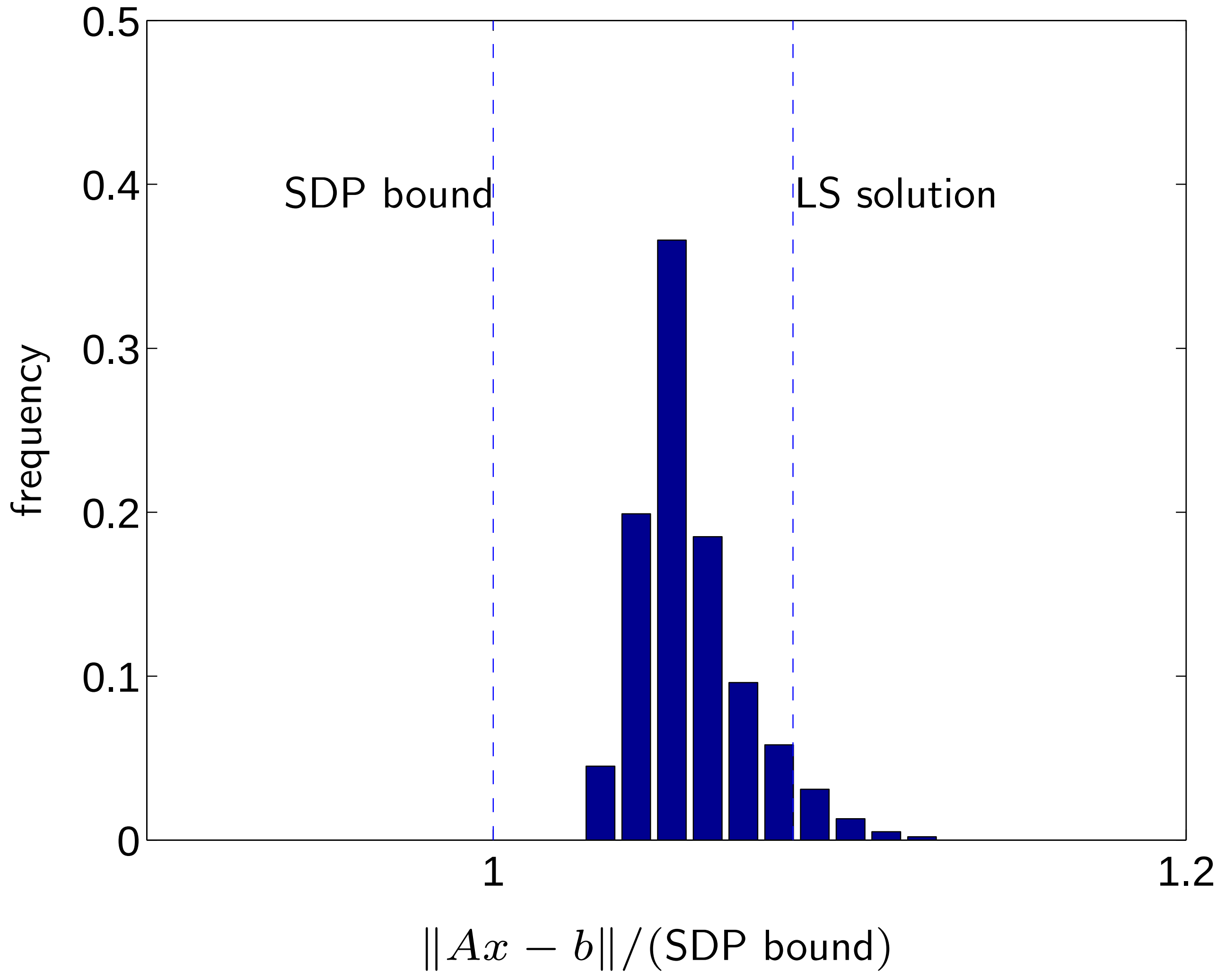
- (randomly chosen) parameters $A \in \mathbb{R}^{150 \times 100}$, $b \in \mathbb{R}^{150}$
- $x \in \mathbb{R}^{100}$, so feasible set has $2^{100} \approx 10^{30}$ points

LS approximate solution: minimize $\|Ax - b\|$ s.t. $\|x\|^2 = n$, then round yields objective 8.7% over SDP relaxation bound

randomized method: (using SDP optimal distribution)

- best of 20 samples: 3.1% over SDP bound
- best of 1000 samples: 2.6% over SDP bound

Example: Partitioning Problem



Example: Partitioning Problem

we go back now to the two-way partitioning problem considered in exercise 5.39 of the textbook:

$$\begin{array}{ll} \mbox{minimize} & x^TWx \\ \mbox{subject to} & x_i^2 = 1, \quad i = 1, \ldots, n \end{array}$$

the Lagrange dual of this problem is given by the SDP:

$$\begin{array}{ll} \mbox{maximize} & -\mathbf{1}^T\nu \\ \mbox{subject to} & W + \mathbf{diag}(\nu) \succeq 0 \end{array}$$

Example: Partitioning

the dual of this SDP is the new SDP

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}\, WX \\ \text{subject to} & X \succeq 0 \\ & X_{ii} = 1, \quad i = 1, \ldots, n \end{array}$$

the solution X^{opt} gives a lower bound on the optimal value p^{opt} of the partitioning problem

- solve this SDP to find X^{opt} (and the bound p^{opt})
- let v denote an eigenvector of X^{opt} associated with its largest eigenvalue
- now let

$$\hat{x} = \mathbf{sgn}(v)$$

the vector $\hat{x}$ is our guess for a good partition

Partitioning: Randomization

- we generate independent samples $x^{(1)}, \ldots, x^{(K)}$ from a normal distribution with zero mean and covariance $X^{\rm opt}$
- for each sample we consider the heuristic approximate solution

$$\hat{x}^{(k)} = \mathbf{sgn}(x^{(k)})$$

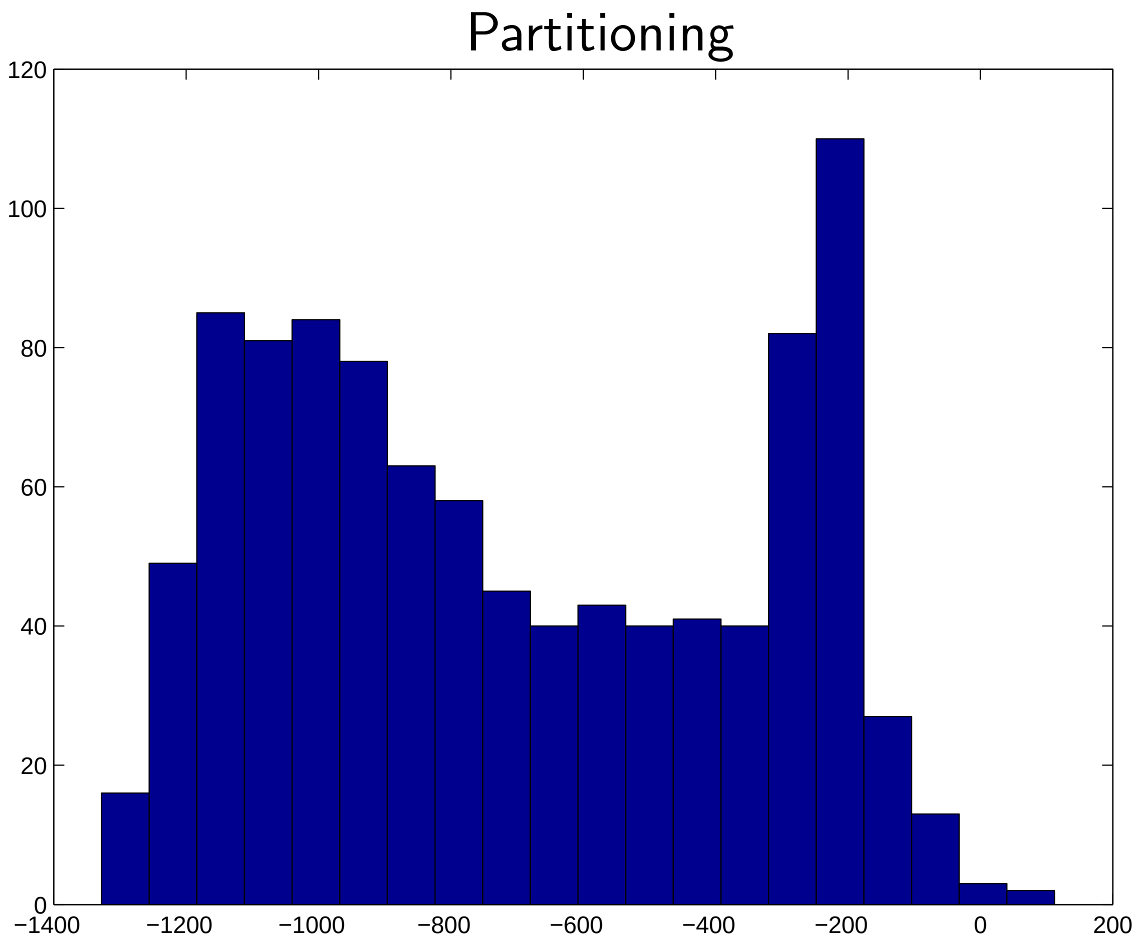
- we then take the one with lowest cost

We compare the performance of these methods on a randomly chosen problem

- the optimal SDP lower bound $p^{\rm opt}$ is equal to -1641
- the simple $sign(x)$ heuristic gives a partition with total cost -1280

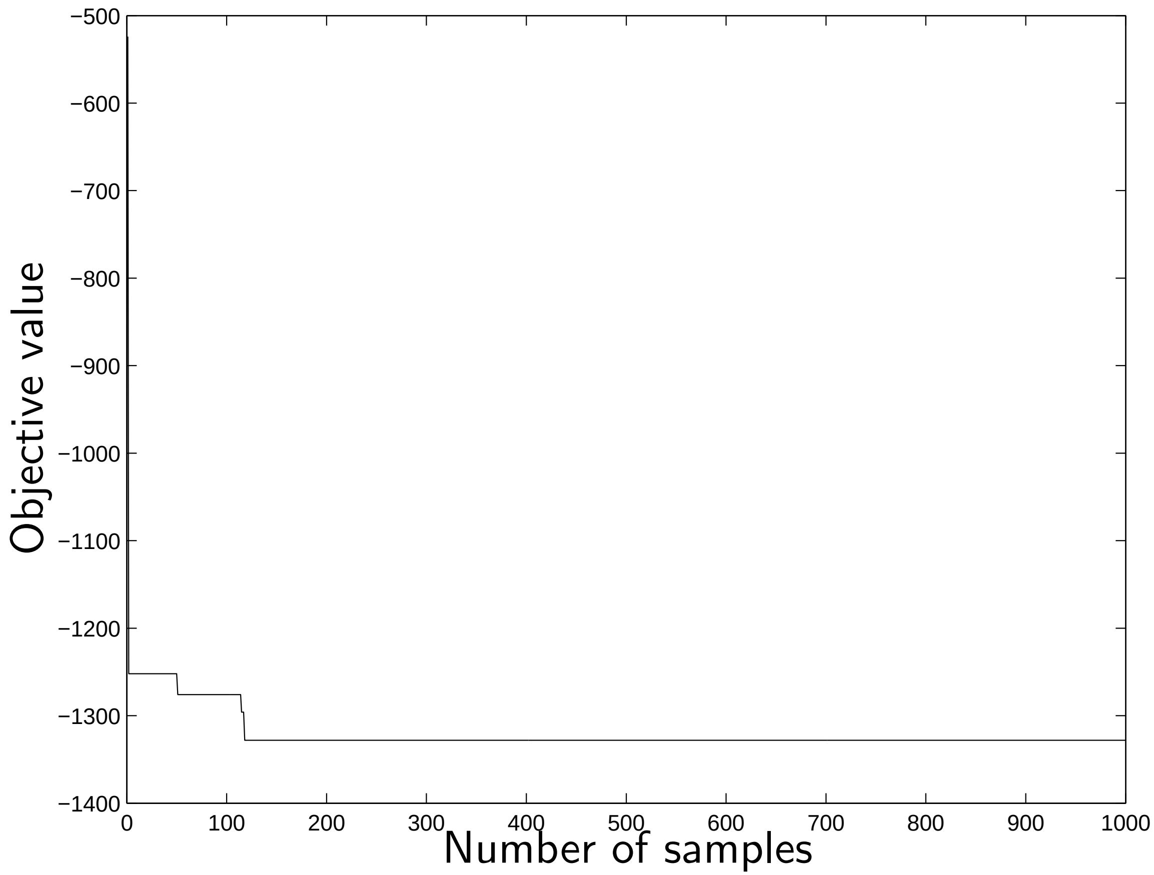
exactly what the optimal value is, we can't say; all we can say at this point is that it is between -1641 and -1280

Partitioning: Numerical Example



histogram of the objective obtained by the randomized heuristic, over 1000 samples: the minimum value reached here is -1328

Partitioning: Numerical Example



we're not sure what the optimal cost is, but now we know it's between -1641 and -1328

Applications in Statistics

Parametric distribution estimation

- distribution estimation problem: estimate probability density $p(y)$ of a random variable from observed values
- parametric distribution estimation: choose from a family of densities $p_x(y)$, indexed by a parameter x

maximum likelihood estimation

$$\text{maximize (over } x\text{)} \quad \log p_x(y)$$

- y is observed value
- $l(x) = \log p_x(y)$ is called log-likelihood function
- can add constraints $x \in C$ explicitly, or define $p_x(y) = 0$ for $x \notin C$
- a convex optimization problem if $\log p_x(y)$ is concave in x for fixed y

Linear measurements with IID noise

linear measurement model

$$y_i = a_i^T x + v_i, \quad i = 1, \ldots, m$$

- $x \in \mathbb{R}^n$ is vector of unknown parameters
- v_i is IID measurement noise, with density $p(z)$
- y_i is measurement: $y \in \mathbb{R}^m$ has density $p_x(y) = \prod_{i=1}^m p(y_i - a_i^T x)$

maximum likelihood estimate: any solution x of

$$\text{maximize} \quad l(x) = \sum_{i=1}^m \log p(y_i - a_i^T x)$$

(y is observed value)

examples

- Gaussian noise $\mathcal{N}(0, \sigma^2)$: $p(z) = (2\pi\sigma^2)^{-1/2} e^{-z^2/(2\sigma^2)}$,

$$l(x) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(a_i^T x - y_i)^2$$

  ML estimate is LS solution

- Laplacian noise: $p(z) = (1/(2a))e^{-|z|/a}$,

$$l(x) = -m\log(2a) - \frac{1}{a}\sum_{i=1}^{m}|a_i^T x - y_i|$$

  ML estimate is ℓ_1-norm solution

- uniform noise on $[-a, a]$:

$$l(x) = \begin{cases} -m\log(2a) & |a_i^T x - y_i| \leq a, \quad i = 1, \ldots, m \\ -\infty & \text{otherwise} \end{cases}$$

  ML estimate is any x with $|a_i^T x - y_i| \leq a$

Logistic regression

random variable $y \in \{0, 1\}$ with distribution

$$p = \mathbf{Prob}(y = 1) = \frac{\exp(a^T u + b)}{1 + \exp(a^T u + b)}$$

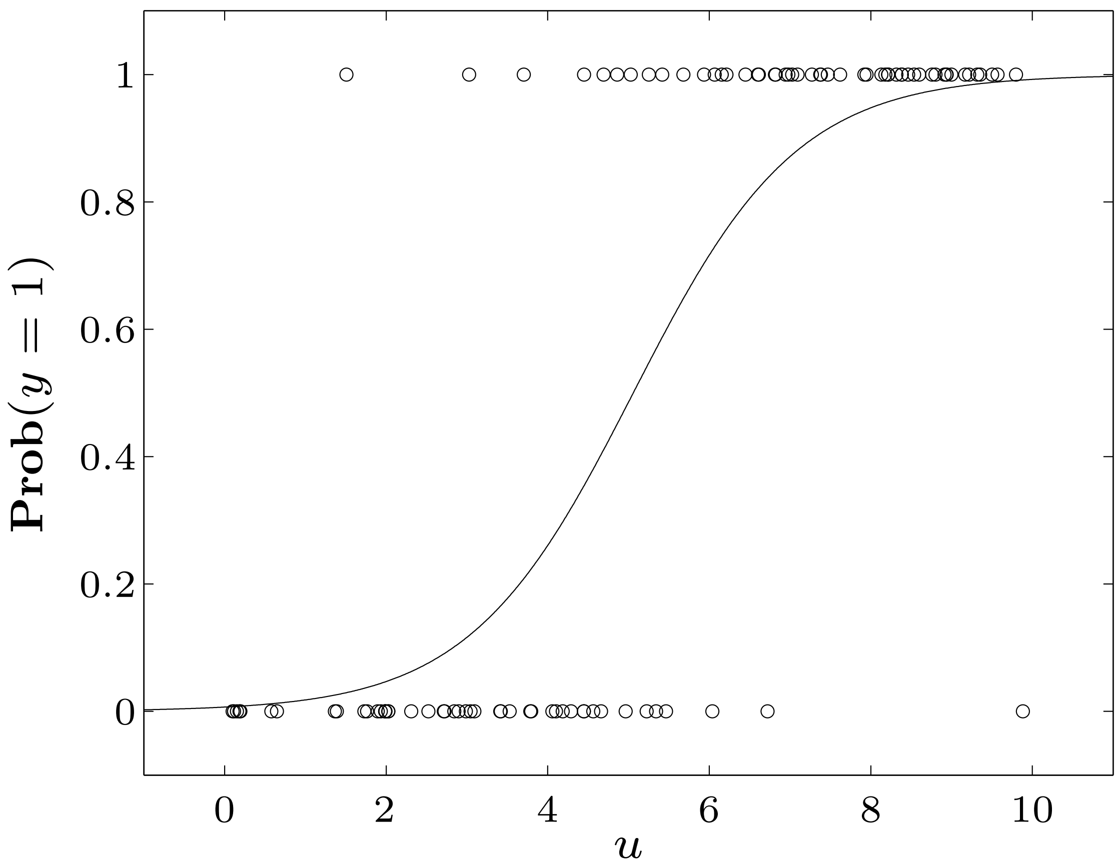
- a, b are parameters; $u \in \mathbb{R}^n$ are (observable) explanatory variables
- estimation problem: estimate a, b from m observations (u_i, y_i)

log-likelihood function (for $y_1 = \cdots = y_k = 1$, $y_{k+1} = \cdots = y_m = 0$):

$$\begin{aligned} l(a,b) &= \log\left(\prod_{i=1}^{k} \frac{\exp(a^T u_i + b)}{1 + \exp(a^T u_i + b)} \prod_{i=k+1}^{m} \frac{1}{1 + \exp(a^T u_i + b)}\right) \\ &= \sum_{i=1}^{k} (a^T u_i + b) - \sum_{i=1}^{m} \log(1 + \exp(a^T u_i + b)) \end{aligned}$$

concave in a, b

example ($n = 1$, $m = 50$ measurements)



- circles show 50 points (u_i, y_i)
- solid curve is ML estimate of $p = \exp(au + b)/(1 + \exp(au + b))$

Experiment design

m linear measurements $y_i = a_i^T x + w_i$, $i = 1, \ldots, m$ of unknown $x \in \mathbb{R}^n$

- measurement errors w_i are IID $\mathcal{N}(0,1)$
- ML (least-squares) estimate is

$$\hat{x} = \left(\sum_{i=1}^m a_i a_i^T\right)^{-1} \sum_{i=1}^m y_i a_i$$

- error $e = \hat{x} - x$ has zero mean and covariance

$$E = \mathbf{E}\, ee^T = \left(\sum_{i=1}^m a_i a_i^T\right)^{-1}$$

  confidence ellipsoids are given by $\{x \mid (x-\hat{x})^T E^{-1}(x-\hat{x}) \leq \beta\}$

experiment design: choose $a_i \in \{v_1, \ldots, v_p\}$ (a set of possible test vectors) to make E 'small'

vector optimization formulation

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}_+^n) & E = \left(\sum_{k=1}^p m_k v_k v_k^T\right)^{-1} \\ \text{subject to} & m_k \geq 0, \quad m_1 + \cdots + m_p = m \\ & m_k \in \mathbf{Z} \end{array}$$

- variables are m_k ($\#$ vectors a_i equal to v_k)
- difficult in general, due to integer constraint

relaxed experiment design

assume $m \gg p$, use $\lambda_k = m_k/m$ as (continuous) real variable

$$\begin{array}{ll} \text{minimize (w.r.t. } \mathbf{S}_+^n) & E = (1/m)\left(\sum_{k=1}^p \lambda_k v_k v_k^T\right)^{-1} \\ \text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1 \end{array}$$

- common scalarizations: minimize $\log\det E$, $\mathbf{Tr}\, E$, $\lambda_{\max}(E)$, ...
- can add other convex constraints, $e.g.$, bound experiment cost $c^T\lambda \leq B$

Experiment design

D-optimal design

$$
\begin{array}{ll}
\text{minimize} & \log\det\left(\sum_{k=1}^p \lambda_k v_k v_k^T\right)^{-1} \\
\text{subject to} & \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1
\end{array}
$$

interpretation: minimizes volume of confidence ellipsoids

dual problem

$$
\begin{array}{ll}
\text{maximize} & \log\det W + n\log n \\
\text{subject to} & v_k^T W v_k \le 1, \quad k = 1, \ldots, p
\end{array}
$$

interpretation: $\{x \mid x^T W x \le 1\}$ is minimum volume ellipsoid centered at origin, that includes all test vectors v_k

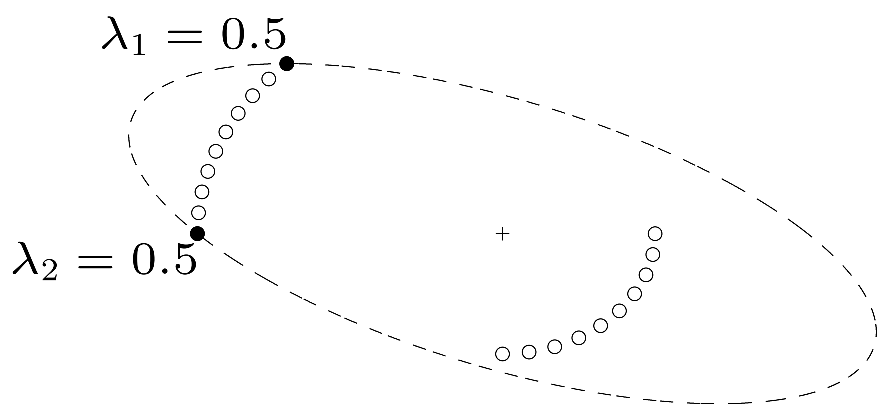
complementary slackness: for λ, W primal and dual optimal

$$
\lambda_k(1 - v_k^T W v_k) = 0, \quad k = 1, \ldots, p
$$

optimal experiment uses vectors v_k on boundary of ellipsoid defined by W

Experiment design

example $(p = 20)$



design uses two vectors, on boundary of ellipse defined by optimal W

Experiment design

Derivation of dual.

first reformulate primal problem with new variable X

$$\begin{array}{ll}\text{minimize} & \log\det X^{-1} \\ \text{subject to} & X = \sum_{k=1}^{p} \lambda_k v_k v_k^T, \quad \lambda \succeq 0, \quad \mathbf{1}^T\lambda = 1\end{array}$$

$$L(X,\lambda,Z,z,\nu) = \log\det X^{-1} + \mathbf{Tr}\left(Z\left(X - \sum_{k=1}^{p}\lambda_k v_k v_k^T\right)\right) - z^T\lambda + \nu(\mathbf{1}^T\lambda - 1)$$

- minimize over X by setting gradient to zero: $-X^{-1} + Z = 0$
- minimum over λ_k is $-\infty$ unless $-v_k^T Z v_k - z_k + \nu = 0$

Dual problem

$$\begin{array}{ll}\text{maximize} & n + \log\det Z - \nu \\ \text{subject to} & v_k^T Z v_k \le \nu, \quad k = 1,\ldots,p\end{array}$$

change variable $W = Z/\nu$, and optimize over ν to get dual of page 55.