Short Course
Robust and Sparse Optimization
Part I: Robust Optimization

Laurent El Ghaoui

EECS and IEOR Departments
UC Berkeley

PGMO Course, Fondation Mathématique Jacques Hadamard

Dec. 4, 2013

Outline

Outline

Convex optimization models

"Nominal" optimization problem:

$$\min_x f_0(x) \; : \; f_i(x) \leq 0, \quad i = 1, \ldots, m$$

f_0, f_i's are convex.

- Includes many problems arising in decision making, statistics.
- Efficient (polynomial-time) algorithms.
- Convex relaxations for non-convex problems.

Uncertainties are a pain!!

In practice, problem data is uncertain:

- *Estimation* errors affect problem parameters.
- *Implementation* errors affect the decision taken.

Uncertainties often lead to highly unstable solutions, or much degraded realized performance.

These problems are compounded in problems with multiple decision periods.

Robust counterpart

"Nominal" optimization problem:

$$\min_x f_0(x) \ : \ f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

Robust counterpart:

$$\min_x \max_{u \in \mathcal{U}} f_0(x, u) \ : \ \forall\, u \in \mathcal{U}, \quad f_i(x, u) \leq 0, \quad i = 1, \ldots, m$$

- functions f_i now depend on a second variable u, the "uncertainty", which is constrained to lie in given set $\mathcal{U}$.
- Inherits convexity from nominal. Very tractable in some practically relevant cases.
- Complexity is high in general, but there are systematic ways to get relaxations.

Robust chance counterpart

(Assume for simplicity there are no constraints)

$$\min_{x} \max_{p \in \mathcal{P}} \mathbf{E}_p f_0(x, u).$$

- Uncertainty is now random, obeys distribution p.
- Distribution p is only known to belong to a class $\mathcal{P}$ (*e.g.*, unimodal, given first and second moments).
- Complexity is high in general, but there are systematic ways to get relaxations.
- Rich variety of related models, including Value-at-Risk constraints.

In this lecture: our main goal is to introduce some important concepts in robust optimization, *e.g.* robust counterparts, affine recourse, distributional robustness.

Uncertainty models

Nominal problem:

$$\min_x \; c^T x \;:\; a_i^T x \leq b_i, \quad i = 1, \ldots, m.$$

We assume that $a_i = \hat{a}_i + \rho u_i$, where

- $\hat{a}_i$'s are the nominal coefficients.
- u_i's are the uncertain vectors, with $u_i \in \mathcal{U}_i$ but otherwise unknown.
- $\rho \geq 0$ is a measure of uncertainty.

Assumption that uncertainties affect each constraint independently is done without loss of generality.

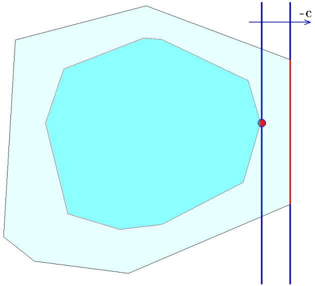
Robust counterpart

Robust counterpart:

$$\min_x \; c^T x \;:\; \forall u_i \in \mathcal{U}_i, \;\; (\hat{a}_i + \rho u_i)^T x \leq b_i, \;\; i = 1, \ldots, m.$$

Solution may be hard, but becomes easy when:

- $\mathcal{U}_i$ are polytopic, given by their vertices (“scenarios”);
- $\mathcal{U}_i$’s are “simple” sets such as ellipsoids, boxes, LMI sets, etc.
- Complexity governed by the support functions of sets $\mathcal{U}_i$.



Robust LP with ellipsoidal uncertainty.

Chance constraints

Simple case

Consider an LP, and assume one of the constraints is $a^T x \leq b$, where $x \in \mathbf{R}^n$ is the decision variable.

If a is random, we can often deal with the chance constraint

$$\text{Prob}\left\{a^T x \leq b\right\} \geq 1 - \epsilon$$

easily. For example, if a is Gaussian with mean $\hat{a}$ and covariance matrix Γ, above is equivalent to

$$\hat{a}^T x + \kappa(\epsilon) \|\Gamma^{1/2} x\|_2 \leq b,$$

where $\kappa(\cdot)$ is a known function that is positive when $\epsilon < 0.5$.

More complicated chance constraints

Often, the random variable enters quadratically in the constraint. This happens for example when x includes affine recourse, and a depends linearly on some random variables.

We are led to consider

$$\text{Prob}\left\{(u,1)^T W(u,1) > 0\right\} \leq \epsilon$$

where W depends *affinely* on the decision variables. Above is hard, even in the Gaussian case.

Distributional robustness

Consider instead

$$\sup_{p \in \mathcal{P}} \mathbf{Prob}_p \left\{ (u,1)^T W (u,1) > 0 \right\} \leq \epsilon$$

where the sup is taken with respect to all distributions p in a specific class $\mathcal{P}$, specifying *e.g.*:

- Moments.
- Symmetry, unimodality.

Fact: when $\mathcal{P}$ is the set of distributions having zero mean and unit covariance, the condition $\sup_{p \in \mathcal{P}} P_{\mathrm{wc}} \leq \epsilon$ is equivalent to the LMI in M, v:

$$\mathbf{Tr}\, M \leq \epsilon v, \quad M \succeq 0, \quad M \succeq vJ + W,$$

where J is all zero but a 1 in the bottom-right entry.

Example

Transaction costs In many financial decision problems, the transaction costs can be modeled with

$$T(x,u) = \|A(x)u + b(x)\|_1,$$

for appropriate affine $A(\cdot), b(\cdot)$.

Example:

$$\sum_{t=1}^{T} |x_{t+1} - x_t|$$

with decision variable x_t an affine function of u.

This leads to consider quantities such as

$$\max_{u \sim (0,I)} \mathbf{E}\, T(x,u)$$

where $u \sim (0, I)$ refers to distributions with zero mean and unit covariance matrices.

A useful result

For given $m \times d$ matrix A and d-vector b, define

$$\phi := \max_{u \sim (0,I)} \mathbf{E} \, \|Au + b\|_1$$

Let a_i denote the i-th row of A ($1 \leq i \leq m$). Then

$$\frac{2}{\pi}\psi \leq \phi \leq \psi,$$

where

$$\psi := \sum_{i=1}^{m} \left\| \begin{pmatrix} a_i \\ b_i \end{pmatrix} \right\|_2 .$$

Note: ψ is convex in A, b, which allows to minimize it if A, b are affine in the decision variables.

Outline

Basic idea

Nominal LP:

$$\min_x \; c^T x \; : \; Ax \leq b.$$

We assume that A, b are affected by uncertainty in affine fashion. We assume uncertainty is available to "known by" some decision variables (*e.g.*, price revealed as time unfolds).

We seek an affinely adjusted robust solution (*i.e.*, a linear feedback).

Example

Nominal LP:

$$\min_x \; c^T x \; : \; Ax \le b.$$

Assume that

- Right-hand side b is subject to uncertainty, $b(u) = \hat{b} + Bu$ with $u \in \mathcal{U}$.
- Decision variable can depend on (parts of) u: $x(u) = \hat{x} + Xu$.

Model information on u available to $x(\cdot)$ as $X \in \mathcal{X}$.

Affinely Adjustable Robust counterpart (AARC):

$$\min_{\hat{x}, X \in \mathcal{X}} \; \max_{u \in \mathcal{U}} \; c^T x(u) \; : \; \forall \, u \in \mathcal{U}, \;\; Ax(u) \le b(u).$$

Above is tractable (provided $\mathcal{U}$ is).

Example

Assume $\mathcal{U} = [-\rho, \rho]^m$, we obtain the AARC

$$\min_{\hat{x}, X \in \mathcal{X}} c^T \hat{x} - \rho \|c^T X\|_1 \;:\; A\hat{x} + \rho s \leq \hat{b}, \quad s_i \geq \|e_i^T (AX - B)\|_1, \quad i = 1, \ldots, m.$$

We recover the "pure" robust counterpart with $X = 0$.

Case with coefficient uncertainty

Approach can be extended to cases when A, c are also uncertain.

- AARC is usually not tractable.
- Efficient approximations via SDP.

A short-term financing problem

(From [8].)

Variables :

- Balance on the credit line x_i for month $i = 1, 2, 3, 4, 5$.
- Amount y_i of commercial paper issued ($i = 1, 2, 3$).
- Excess funds z_i for month $i = 1, 2, 3, 4, 5$.
- z_6, the company's wealth in June.

With these variables we have to meet certain cash-flow requirements each month.

Decision problem:

$$\text{maximize } z_6 \text{ subject to } \begin{cases} \text{Bounds on variables,} \\ \text{Cash-flow balance equations.} \end{cases}$$

Linear programming formulation

$$
\begin{aligned}
\max_{x,y,z} \quad & z_6 \\
\text{s.t.} \quad & x_1 + y_1 - z_1 = 150, \\
& x_2 + y_2 - 1.01x_1 + 1.003z_1 - z_2 = 100, \\
& x_3 + y_3 - 1.01x_2 + 1.003z_2 - z_3 = -200, \\
& x_4 - 1.02y_1 - 1.01x_3 + 1.003z_3 - z_4 = 200, \\
& x_5 - 1.02y_2 - 1.01x_4 + 1.003z_4 - z_5 = -50, \\
& -1.02y_3 - 1.01x_5 + 1.003z_5 - z_6 = -300, \\
& 0 \leq x \leq 100, \quad y \geq 0, \quad z \geq 0.
\end{aligned}
$$

The right-hand side contains the liabilities that we must meet.

Matrix notation

Inequality form

$$\max_{v} \ c^T v \ : \ Av \geq b, \ \ Cv \leq d, \ \ v \geq 0.$$

Here:

- $v = (x, y, z)$ is the decision variable.
- A, b describe the liability constraints, b is the liability vector.
- C, d describe the upper bounds on x ($x \leq 100$).
- c is a vector such that $c^T v = z_6$. (That is, c is all 0's except 1 in the last entry.)

Note: we have, without loss of generality, replaced the equalities $Av = b$ by inequalities ($Av \geq b$) (Why is it safe to do so?)

CVX syntax

A short-term financing problem

Assume *A*, *b* have been defined in `matlab`'s workspace:

```
cvx_begin
    variable x(5,1);
    variable y(3,1);
    variable z(6,1);
    maximize( z(6) );
    A*[x; y; z] >= b;
    x <= 100; x >= 0;
    y >= 0; z >= 0;
cvx_end
```

Solution: (terminal wealth in red)

month	x	y	z
1	0.0000	0	0.0000
2	22.5813	0	0.0000
3	0.0000	0	351.9442
4	0.0000	150.0000	0.0000
5	28.9671	77.4187	0.0000
6	0	174.7512	92.4969

Uncertainty model

We now assume that the liability vector b is subject to uncertainty.

Specifically: at each time t, the liability is

$$b(t) = \hat{b}(t)(1 + 0.06u(t) + 0.02u(t-1))$$

for some values $u(t)$, $u(t-1)$ that are only known to be in $[-1, 1]$.
Here $\hat{b}$ contains the *nominal* liability values.

- This models *relative* errors in $b(t)$.
- The liability at time t depends on noise at time $t, t-1$.
- We set $\hat{b}$ to be the vector in the previous example.

Uncertainty model

Matrix form

The liability vector is now a function of uncertainty:

$$b(u) = \hat{b} + Bu, \quad \|u\|_\infty \leq 1,$$

where $u \in \mathbf{R}^6$, and

$$B = \mathbf{diag}(\hat{b}) \begin{pmatrix} 0.06 & 0 & 0 & 0 & 0 & 0 \\ 0.02 & 0.06 & 0 & 0 & 0 & 0 \\ 0 & 0.02 & 0.06 & 0 & 0 & 0 \\ 0 & 0 & 0.02 & 0.06 & 0 & 0 \\ 0 & 0 & 0 & 0.02 & 0.06 & 0 \\ 0 & 0 & 0 & 0 & 0.02 & 0.06 \end{pmatrix}.$$

Robust counterpart

$$\begin{array}{ll} \max\limits_{x,y,z} & c^T v \\ \text{s.t.} & \forall\, u,\ \ \|u\|_\infty \le 1 \ :\ Av \ge \hat{b} + Bu, \\ & v = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \ge 0, \ \ x \le 100. \end{array}$$

Robust counterpart

A single constraint

Let us examine a particular constraint (at time $t = 1, \dots, 6$):

$$a(t)^T v \geq \hat{b}(t) + B(t)^T u,$$

where $B(t)$ is the t-th row of B.

The robust counterpart obtains when insisting that the above remains true for *every* value of u permitted by the model (*i.e.*, $\|u\|_\infty \leq 1$).

The constraint becomes

$$\forall\, u, \quad \|u\|_\infty \leq 1 \;:\; a(t)^T v \geq \hat{b}(t) + B(t)^T u.$$

Useful fact

Fact: the condition on $z \in \mathbf{R}^p$, $\beta \in \mathbf{R}$

$$\forall\, u, \quad \|u\|_\infty \leq 1 \;:\; z^T u \leq \beta,$$

which is the same as:

$$\beta \geq \max_{u \,:\, |u_i| \leq 1,\; i=1,\ldots,p} z^T u,$$

is equivalent to

$$\|z\|_1 \leq \beta.$$

Proof: The scalar case ($p = 1$) is trivial. Then, exploit decoupling:

$$\max_{u \,:\, |u_i| \leq 1,\; i=1,\ldots,p} \sum_{i=1}^{p} u_i z_i = \sum_{i=1}^{p} \max_{u_i \,:\, |u_i| \leq 1} u_i z_i.$$

Each maximum is attained by $u_i = \mathbf{sign}(z_i)$, $i = 1, \ldots, p$.

Robust counterpart

Equivalent form of a single constraint

In our problem, the cash-flow matching constraint at time t is

$$\forall\, u, \quad \|u\|_\infty \leq 1 \ : \ a(t)^T v \geq b(t) + B(t)^T u.$$

Apply previous result and obtain the equivalent form:

$$a(t)^T v \geq \hat{b}(t) + \|B(t)\|_1$$

Recall $B(t)$ stands for the t-th row of matrix B.

Robust counterpart

All constraints together

Robust counterpart of cash-flow matching constraints:

$$\forall\ u, \quad \|u\|_\infty \leq 1 \ : \ Av \geq b(u).$$

Equivalent form:

$$Av \geq \hat{b} + |B|\mathbf{1},$$

where $|B|$ is the matrix of absolute values in B, and $\mathbf{1}$ is the vector of 1's. (Hence, $|B|\mathbf{1}$ is a vector with the l_1-norm of the rows of B in each entry.)

That is, we simply replace b by its worst-case value $\hat{b} + |B|\mathbf{1}$ in the original model.

Robust counterpart
Formulation

Original robust counterpart:

$$\max_{x,y,z} \quad c^T v$$
$$\text{s.t.} \quad \forall\, u, \quad \|u\|_\infty \leq 1 \;:\; Av \geq \hat{b} + Bu,$$
$$v = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \geq 0, \quad x \leq 100.$$

After our derivations, simply replace b by $\hat{b} + |B|\mathbf{1}$:

$$\max_{x,y,z} \quad c^T v$$
$$Av \geq \hat{b} + |B|^T \mathbf{1},$$
$$v = \begin{pmatrix} x \\ y \\ z \end{pmatrix} \geq 0, \quad x \leq 100.$$

Robust counterpart

Results

(Terminal wealth in red)

Nominal solution:

month	x	y	z
1	0.0000	0	0.0000
2	22.0916	0	0.0000
3	0.0000	0	351.9442
4	0.0000	150.0000	0.0000
5	29.4665	77.9084	0.0000
6	0	174.2567	92.4969

Robust solution:

month	x	y	z
1	0.0000	0	0.0001
2	33.5287	0	0.0000
3	0.0000	0	377.0489
4	0.0000	159.0000	0.0000
5	33.9808	75.4714	0.0000
6	0	224.9128	17.2684

Comments

The results delivered by the robust solution seem substantially worse (terminal wealth goes from 92.5 to 17.3) .

However the terminal wealth delivered by the robust solution has to be compared to the one delivered by the nominal policy *when uncertainty is present* .

With uncertainty, the nominal policy is actually a complete failure: the cash-flow requirements are simply not met!

Robust counterpart to example problem

Interpretation

The policy implied by the robust counterpart is quite simple in this problem.

Simply replace all the nominal values of the liability vector by their extreme ones, and solve the original problem with these new (worse) values.

Can we do better?

Recourse approach

We can do better by exploiting the fact that this problem involves multiple decision periods, and assuming that the uncertainty is revealed to us as time goes on.

For example, we may assume that at each time t the values of $(u(1), \ldots, u(t-1))$ are available to the decision maker.

This lets us make the decision vectors (strictly causal) functions of u.

Recourse approach

Linear recourse

We assume that the decision variables are strictly causal *linear* functions:

$$x(u) = x + Xu, \quad y(u) = y + Yu, \quad z(u) = z + Zu.$$

where X, Y, Z are a strictly lower-triangular matrices. That way, $x(t), y(t), z(t)$ depend only on $(u(1), \ldots, u(t-1))$. Here, the variables are x, y, z *and* X, Y, Z.

We can write the vector of variables $v = (x, y, z)$ as a function of u:

$$v(u) = v + Vu, \quad V = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

where v is a vector, and V is a matrix. (Both are variables, and V is constrained by the triangular structure of X, Y, Z.)

Recourse approach

Linear recourse

Problem with linear recourse:

$$\begin{array}{ll} \max\limits_{x,y,z} & \min\limits_{u \,:\, \|u\|_\infty \le 1} \; c^T(v + Vu) \\ \text{s.t.} & X, Y, Z \text{ strictly lower-triangular,} \\ & \forall\, u, \;\; \|u\|_\infty \le 1 : \quad A(v + Vu) \ge \hat{b} + Bu \\ & \qquad\qquad\qquad\qquad\quad v + Vu \ge 0, \;\; x + Xu \le 100. \end{array}$$

- Note that we consider the worst-case objective.
- We make sure that the constraints remain valid for every admissible u, including the sign constraints on the variables themselves.
- We can apply the same technique as before to get to a tractable solution . . .

Recourse approach

Linear recourse, tractable form

$$\max_{x,y,z,X,Y,Z} \quad c^T v - \|V^T c\|_1$$
$$\text{s.t.} \quad Av \geq \hat{b} + |B - AV|^T \mathbf{1},$$
$$v \geq |V|\mathbf{1}, \quad x + |X|\mathbf{1} \leq 100,$$
$$X, Y, Z \text{ stricly lower triangular},$$
$$v = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad V = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}.$$

Problem is still an LP!

- The presence of V makes it easier to satisfy the cash-flow constraints.
- But it has a negative impact on the other constraints (bounds on variables), and the objective.
- With $V = 0$ we recover the robust solution seen before—information about the uncertainty can only help.

CVX syntax

```
cvx_begin
    variables x(5,1); y(3,1); z(6,1);
    variable X(5,6) lower_triangular;
    variable Y(3,6) lower_triangular;
    variable Z(6,6) lower_triangular;
    maximize( z(6)- c'*sum(abs(Z),2) );
    A*[x; y; z] >= b+sum(abs(A*[X; Y; Z]-B),2);
    x <= 100-sum(abs(X),2);
    x >= sum(abs(X),2);
    y >= sum(abs(Y),2);
    z >= sum(abs(Z),2);
    diag(X) == 0; diag(Y) == 0; diag(Z) == 0;
cvx_end
```

- Use `lower_triangular` to specify lower-triangular variables.
- Further use `diag(...) == 0` constraints to specify strictly lower-triangular structure.

Results

- *Nominal solution:* worst-case terminal wealth $-\infty$ (infeasible problem).
- *Robust solution:* worst-case terminal wealth 17.3.
- *Linear recourse solution:* worst-case terminal wealth 31.5.

Summary

- Recourse idea is the same as feedback: we use available information for "planning to re-plan".
- Robust optimization with linear recourse allows less conservative strategies in multi-period problems.
- The problem often reduces to an LP or QP, with substantially more variables (recourse matrices).
- Versions exist with chance constraints.

Dynamic programming

- Finite-state, discrete-time Markov decision process.
- Finite-horizon control problem: minimize expected cost.
- $a \in \mathcal{A}$ denote actions, $s \in \mathcal{S}$ states, and $c_t(s, a)$ the cost for action a in state s at time t.

Bellman recursion (value iteration):

$$v_t(s) = \min_{a \in \mathcal{A}} c_t(s, a) + p_t(a)^T v_{t+1}, \quad s \in \mathcal{S}$$

with $p_t(a)$ the transition probabilities at time t under action a.

Uncertainty on transition matrix

We assume that *at each stage*, "nature" picks a transition probability vector $p_t(a)$ in a given set $\mathcal{P}_t(a)$.

Robust counterpart: the robust control problem, with "nature" the adversary.

Robust Bellman recursion:

$$v_t(s) = \min_{a \in \mathcal{A}} c_t(s,a) + \max_{p \in \mathcal{P}_t(a)} p^T v_{t+1}, \quad s \in \mathcal{S}.$$

For a wide variety of sets $\mathcal{P}_t(a)$, inner problem very easy to solve.

Entropy uncertainty model

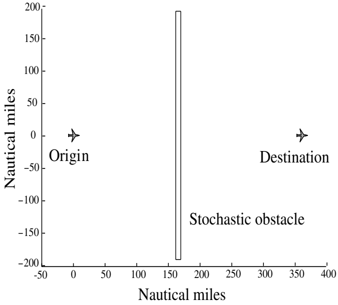
A natural way to model uncertainty in the transition matrices involves relative entropy bounds

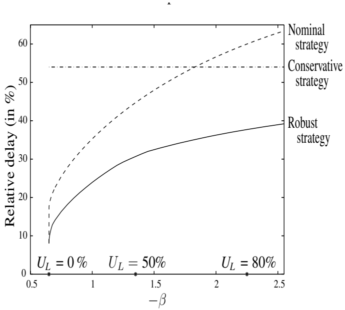
$$\mathcal{P} = \left\{ p \geq 0 \ : \ \sum_j p_j \log \frac{p_j}{q_j} \leq \beta, \ \sum_j p_j = 1 \right\}.$$

where $\beta > 0$ is a measure of uncertainty, and q is the nominal distribution.

The corresponding inner problem can be solved in $O(n)$ via bisection.

Example
Robust path planning





Outline

Supervised learning problems

Many supervised learning problems (*e.g.*, classification, regression) can be written as

$$\min_{w} \mathcal{L}(X^T w)$$

where $\mathcal{L}$ is convex, and X contains the data.

Penalty approach

Often, optimal value and solutions of optimization problems are sensitive to data.

A common approach to deal with sensitivity is via penalization, *e.g.*:

$$\min_x \mathcal{L}(X^T w) + \|Wx\|_2^2 \qquad (W = \text{weighting matrix}).$$

- How do we choose the penalty?
- Can we choose it in a way that reflects knowledge about problem structure, or how uncertainty affects data?
- Does it lead to better solutions from machine learning viewpoint?

Support Vector Machine

Support Vector Machine (SVM) classification problem:

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b))_+$$

- $Z := [z_1, \ldots, z_m] \in \mathbf{R}^{n \times m}$ contains the *data points* .
- $y \in \{-1, 1\}^m$ contain the *labels* .
- $x := (w, b)$ contains the *classifier parameters* , allowing to classify a new point z via the rule

$$y = \mathbf{sgn}(z^T w + b).$$

Robustness to data uncertainty

Assume the data matrix is only partially known, and address the robust optimization problem:

$$\min_{w,b} \max_{U \in \mathcal{U}} \sum_{i=1}^{m} (1 - y_i((z_i + u_i)^T w + b))_+,$$

where $U = [u_1, \ldots, u_m]$ and $\mathcal{U} \subseteq \mathbf{R}^{n \times m}$ is a set that describes additive uncertainty in the data matrix.

Measurement-wise, spherical uncertainty

Assume

$$\mathcal{U} = \{U = [u_1, \ldots, u_m] \in \mathbf{R}^{n \times m} \ : \ \|u_i\|_2 \leq \rho\},$$

where $\rho > 0$ is given.

Robust SVM reduces to

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b) + \rho \|w\|_2)_+.$$

Link with classical SVM

Classical SVM contains l_2-norm regularization term:

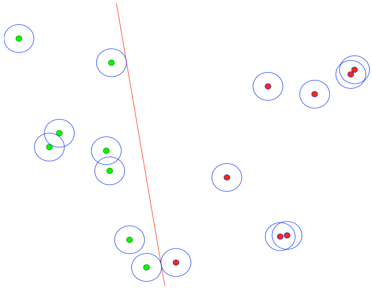
$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b))_+ + \lambda \|w\|_2^2.$$

where $\lambda > 0$ is a penalty parameter.

With spherical uncertainty, robust SVM is similar to classical SVM.

When data is separable, the two models are equivalent ...

Separable data



Maximally robust classifier for separable data, with spherical uncertainties around each data point. In this case, the robust counterpart reduces to the classical maximum-margin classifier problem.

Interval uncertainty

Assume

$$\mathcal{U} = \{U \in \mathbf{R}^{n \times m} \; : \; \forall (i,j), \;\; |U_{ij}| \leq \rho\},$$

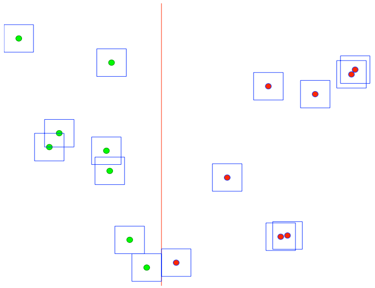
where $\rho > 0$ is given.

Robust SVM reduces to

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b) + \rho \|w\|_1)_+.$$

The l_1-norm term encourages sparsity, and may not regularize the solution.

Separable data



Maximally robust classifier for separable data, with box uncertainties around each data point. This uncertainty model encourages sparsity of the solution.

Other uncertainty models

We may generalize the approach to other uncertainty models, retaining tractability:

- “Measurement-wise” uncertainty models: perturbations affect each data point independent of each other.
- Other models couple the way uncertainties affect each measurement; for example we may control the number of errors across all the measurements.
- Norm-bound models allow for uncertainty of data matrix that is bounded in matrix norm.
- A whole theory, summarized next, is presented in [1].

Nominal problem

$$\min_{\theta \in \Theta} \mathcal{L}(Z^T \theta),$$

where

- $Z := [z_1, \ldots, z_m] \in \mathbf{R}^{n \times m}$ is the data matrix
- $\mathcal{L} : \mathbf{R}^m \to \mathbf{R}$ is a convex loss function
- Θ imposes "structure" (eg, sign) constraints on parameter vector θ

Loss function: assumptions

We assume that

$$\mathcal{L}(r) = \pi(\mathbf{abs}(P(r))),$$

where $\mathbf{abs}(\cdot)$ acts componentwise, $\pi : \mathbf{R}_+^m \to \mathbf{R}$ is a convex, monotone function on the non-negative orthant, and

$$P(r) = \begin{cases} r & (\text{"symmetric case"}) \\ r_+ & (\text{"asymmetric case"}) \end{cases}$$

with r_+ the vector with components $\max(r_i, 0)$, $i = 1, \ldots, m$.

Loss function: examples

- l_p-norm regression
- hinge loss
- Huber, Berhu loss

Robust counterpart

$$\min_{\theta \in \Theta} \max_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}(\mathbf{Z}^T \theta).$$

where $\mathcal{Z} \subseteq \mathbf{R}^{n \times m}$ is a set of the form

$$\mathcal{Z} = \{ Z + \Delta \ : \ \Delta \in \rho \mathcal{D}, \} ,$$

with $\rho \geq 0$ a measure of the size of the uncertainty, and $\mathcal{D} \subseteq \mathbf{R}^{l \times m}$ is given.

Generic analysis

For a given vector θ, we have

$$\max_{\mathbf{Z}\in\mathcal{Z}} \mathcal{L}(\mathbf{Z}^T\theta) = \max_{u} u^T Z^T\theta - \mathcal{L}^*(u) + \rho\phi_{\mathcal{D}}(uv^T),$$

where $\mathcal{L}^*$ is the conjugate of $\mathcal{L}$, and

$$\phi_{\mathcal{D}}(X) := \max_{\Delta\in\mathcal{D}} \langle X, \Delta\rangle$$

is the support function of $\mathcal{D}$.

Assumptions on uncertainty set $\mathcal{D}$

Separability condition: there exist two semi-norms ϕ, ψ such that

$$\phi_{\mathcal{D}}(uv^T) := \max_{\Delta \in \mathcal{D}} u^T \Delta v = \phi(u)\psi(v).$$

- Does not completely characterize (the support function of) the set $\mathcal{D}$
- Given ϕ, ψ, we can construct a set $\mathcal{D}_{\text{out}}$ that obeys condition
- The robust counterpart only depends on ϕ, ψ.

WLOG, we can replace $\mathcal{D}$ by its convex hull.

Examples

- Largest singular value model: $\mathcal{D} = \{\Delta \ : \ \|\Delta\| \leq \rho\}$, with ϕ, ψ Euclidean norms.
- Any norm-bound model involving an induced norm (ϕ, ψ are then the norms dual to the norms involved).
- Measurement-wise uncertainty models, where each column of the perturbation matrix is bounded in norm, independently of the others, correspond to the case with $\psi(v) = \|v\|_1$.

Other examples

Bounded-error model: there are (at most K) errors affecting data

$$\mathcal{D} = \left\{ \Delta = [\lambda_1 \delta_1, \ldots, \lambda_m \delta_m] \in \mathbf{R}^{l \times m} \; : \; \begin{array}{c} \|\delta_i\| \leq 1, \quad i = 1, \ldots, m, \\ \sum_{i=1}^{m} \lambda_i \leq K, \quad \lambda \in \{0,1\}^m \end{array} \right\}.$$

for which $\phi(\cdot) = \| \cdot \|_*$, $\psi(v) =$ sum of the K largest magnitudes of the components of v.

Examples (follow'd)

- The set

$$\mathcal{D} = \left\{ \Delta = [\lambda_1 \delta_1, \ldots, \lambda_m \delta_m] \in \mathbf{R}^{l \times m} \; : \; \delta_i \in \{-1, 0, 1\}^l, \;\; \|\delta\|_1 \leq k \right\}$$

models measurement-wise uncertainty affecting Boolean data (we can impose $\delta_i \in \{x_i - 1, x_i\}$ to be more realistic)
In this case, we have $\psi(\cdot) = \|\cdot\|_1$ and

$$\phi(u) = \|u\|_{1,k} := \min_w \; k\|u - w\|_\infty + \|w\|_1.$$

Main result

For a given vector θ, we have

$$\min_{\theta} \max_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}(\mathbf{Z}^T\theta) = \min_{\theta,\kappa} \mathcal{L}_{\rm wc}(Z^T\theta, \kappa) \ : \ \kappa \geq \phi(U^T\theta)$$

where

$$\mathcal{L}(r, \kappa) := \max_{v} \ v^T r - \mathcal{L}^*(v) + \kappa\psi(v)$$

is the worst-case loss function of the robust problem.

Worst-case loss function

The tractability of the robust counterpart is directly linked to our ability to compute optimal solutions v^* for

$$\mathcal{L}(r, \kappa) = \max_{v} \; v^T r - \mathcal{L}^*(v) + \kappa \psi(v)$$

Dual representation (assume $\psi(\cdot) = \|\cdot\|$ is a norm):

$$\mathcal{L}(r, \kappa) = \max_{\xi} \; \mathcal{L}(r + \kappa \xi) \; : \; \|\xi\|_* \leq 1$$

When ψ is the Euclidean norm, robust regularization of $\mathcal{L}$ (see [10]).

Special cases

- When $\psi(\cdot) = \|\cdot\|_p$, $p = 1, \infty$, problem reduces to simple, tractable convex problem (assuming nominal problem is).
- For $p = 2$, problem can be reduced to such a simple form, for the hinge, l_q-norm and Huber loss functions.

Lasso

In particular, the least-squares problem with lasso penalty

$$\min_{\theta} \|X^T\theta - y\|_2 + \rho\|\theta\|_1$$

is the robust counterpart to a least-squares problem with uncertainty on X, with additive perturbation bounded in the norm

$$\|\Delta\|_{1,2} := \max_{1 \le i \le l} \sqrt{\sum_{j=1}^{n} \Delta_{ij}^2}.$$

Globalized robust counterpart

The robust counterpart is based on the worst-case value of the loss function assuming a bound on the data uncertainty ($Z \in \mathcal{Z}$):

$$\min_{\theta \in \Theta} \max_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}(\mathbf{Z}^T \theta).$$

The approach does not control the degradation of the loss outside the set $\mathcal{Z}$.

Globalized robust counterpart: formulation

In globalized robust counterpart, we fix a “rate” of degradation of the loss, which controls the amount of degradation of the loss as the data matrix Z goes “away from” the set $\mathcal{Z}$.
We seek to minimize τ, such that

$$\forall \Delta \ : \ \mathcal{L}((Z + \Delta)^T\theta) \leq \tau + \alpha\|\Delta\|,$$

where $\alpha > 0$ controls the rate of degradation, and $\| \cdot \|$ is a matrix norm.

Globalized robust counterpart

Examples

- For the SVM case, the globalized robust counterpart can be expressed as:

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b))_+ \ : \ \sqrt{m}\|\theta\|_2 \leq \alpha,$$

  which is a classical form of SVM.

- For l_p-norm regression with m data points, the globalized robust counterpart takes the form

$$\min_{\theta} \|X^T\theta - y\|_p \ : \ \kappa(m,p)\|\theta\|_2 \leq \alpha$$

  where $\kappa(m,1) = \sqrt{m}$, $\kappa(m,2) = \kappa(m,\infty) = 1$.

Chance constraints

Theory can address problems with "chance constraints"

$$\min_{\theta} \max_{p \in \mathcal{P}} \mathbf{E}_p \mathcal{L}(Z(\delta)^T \theta)$$

where δ follows distribution p, and $\mathcal{P}$ is a class of distributions

- Results are more limited, focused on upper bounds.
- Convex relaxations are available, but more expensive.
- Approach uses Bernstein approximations (Nemirovski & Ben-tal, 2006).

Robust regression with chance constraints: an example

$$\phi_p := \min_{\theta} \max_{x \sim (\hat{x}, X)} \mathbf{E}_x \|A(x)\theta - b(x)\|_p$$

- Regression variable is $\theta \in \mathbf{R}^n$
- $x \in \mathbf{R}^q$ is an uncertainty vector that enters affinely in the problem matrices: $[A(x), b(x)] = [A_0, b_0] + \sum_i x_i [A_i, b_i]$.
- The distribution of uncertainty vector x is unknown, except for its mean $\hat{x}$ and covariance X.
- Objective is worst-case (over distributions) expected value of l_p-norm residual ($p = 1, 2$).

Main result

(Assume $\hat{x} = 0$, $X = I$ WLOG)

For $p = 2$, the problem reduces to least-squares:

$$\phi_2^2 = \min_{\theta} \sum_{i=0}^{q} \|A_i\theta - b_i\|_2^2$$

For $p = 1$, we have $(2/\pi)\psi_1 \leq \phi_1 \leq \psi_1$, with

$$\psi_1 = \min_{\theta} \sum_{i=0}^{q} \|A_i\theta - b_i\|_2$$

Example: robust median

As a special case, consider the median problem:

$$\min_{\theta} \sum_{i=1}^{q} |\theta - x_i|$$

Now assume that vector x is random, with mean $\hat{x}$ and covariance X, and consider the robust version:

$$\phi_1 := \min_{\theta} \max_{x \sim (\hat{x}, X)} \mathbf{E}_x \sum_{i=1}^{q} |\theta - x_i|$$

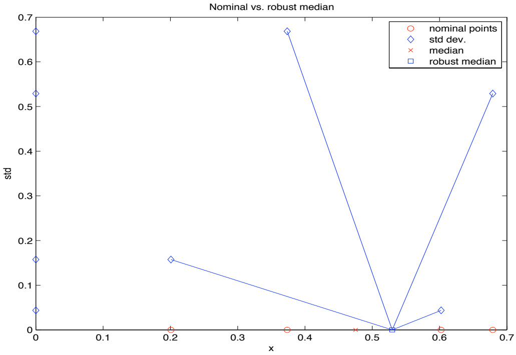
Approximate solution

We have $(2/\pi)\psi_1 \leq \phi_1 \leq \psi_1$, with

$$\psi_1 := \sum_{i=1}^{n} \sqrt{(\theta - \hat{x}_i)^2 + X_{ii}}$$

Amounts to find the minimum distance sum (a very simple SOCP).

Geometry of robust median problem



Outline

References I

[1] A. Bental, L. El Ghaoui, and A. Nemirovski.
Robust Optimization.
Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

[2] Dimitris Bertsimas, David B. Brown, and Constantine Caramanis.
Theory and applications of robust optimization.
SIAM Rev., 53(3):464–501, August 2011.

[3] Dimitris Bertsimas and Vineet Goyal.
On the power of robust solutions in two-stage stochastic and adaptive optimization problems.
Math. Oper. Res., 35(2):284–305, May 2010.

[4] S. Boyd and M. Grant.
The CVX optimization package, 2010.

[5] S. Boyd and L. Vandenberghe.
Convex Optimization.
Cambridge University Press, 2004.

[6] G. Calafiore and L. El Ghaoui.
Optimization Models.
Cambridge University Press, 2014.
To appear.

References II

[7] C. Caramanis, S. Mannor, and H. Xu.
Robust optimization in machine learning.
In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, chapter 14. MIT Press, 2011.

[8] G. Cornuejols and R. Tütüncü.
Optimization Methods in Finance.
Cambridge University Press, 2007.

[9] V. Guigues.
Robust production management.
Optimization Online, February 2011.
www.optimization-online.org/D8_HTML/2011/02/2935.html.

[10] Adrian S. Lewis and C. H. Jeffrey Pang.
Lipschitz behavior of the robust regularization.
SIAM J. Control Optim., 48(5):3080–3104, December 2009.

[11] Arnab Nilim and Laurent El Ghaoui.
Robust control of Markov decision processes with uncertain transition matrices.
Oper. Res., 53(5):780–798, September-October 2005.