Random Convex Programs with L_1-Regularization

Sparsity and Generalization

Campi & Carè, SIAM J. Contr. Optim, 50(5), 3532-3357, 2013.

D2 Zhicheng Zhang

Division of Operations Research

November 16, 2021

Random Convex Program

Random Convex Program

Consider a standard random convex program (min-max)

$$\text{RCP} \qquad \min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \max_{i=1,\cdots,N} L(x, \delta^{(i)}),$$

where $\delta^{(i)}$, $i = 1, \cdots, N$ are N scenarios sampled from Δ in an i.i.d. fashion according to probability $\mathbb{P}$.

♣ RCP is equivalent to a epigraphic form

$$\min_{L \in \mathbb{R}, x \in \mathcal{X}} L \quad \textit{subject to} \quad L(x, \delta^{(i)}) \leq L, \ i = 1, \cdots, N.$$

The optimization is to take worst-case minimization w.r.t. scenarios $\delta^{(i)}$.

Generalization Property (ϵ-level Performance Robustness)

There is a set Δ_ϵ with $\mathbb{P}\{\Delta_\epsilon\} \geq 1 - \epsilon$ such that $\max_{\delta \in \Delta_\epsilon} {}_{L(x_N^*, \delta)} \leq L_N^*$. where $L_N^* = max_{i=1,\cdots,N} \ L(x, \delta^{(i)})$, and x_N^* is the optimal solution of RCP.

Random Convex Program with L_1-Regularization

$$L_1\text{-RCP} \qquad \min_{x\in\mathcal{X}\in\mathbb{R}^d} \max_{i=1,\cdots,N} L(x,\delta^{(i)}) \quad \textit{subject to} \quad \|Ax-b\|_1 \le r,$$

where $A \in \mathbb{R}^{p\times d}$, $b \in \mathbb{R}^p$, $\|\cdot\|_1$ is the L_1 norm (e.g., $\|z\|_1 = \sum_{j=1}^{p} |z_j|$), and $r \in \mathbb{R}$ is the constraining parameter to tune the level of sparsity.

♣ L_1-RCP is equivalent to a epigraphic form

$$\min_{L\in\mathbb{R}, x\in\bar{\mathcal{X}}} L \quad \textit{subject to} \quad L(x,\delta^{(i)}) \le L, \; i = 1,\cdots,N,$$

where $\bar{\mathcal{X}} = \{x \in \mathcal{X} : \|Ax-b\|_1 \le r\}$.

- Reduce the effective dimension of decision variable (i.e., $\dim(x) = d$).

Assumption 1 (Convexity)

Function $L(x,\delta)$ is convex in x, while it has an arbitrary dependence on δ, and the optimization domain $\mathcal{X}$ is a convex and closed set.

Examples (lasso and basalt column constraints)

Example (lasso constraint)

Letting $A = I$ and $b = 0$, then the generalized constraint in L_1-RCP reduces to the following lasso constraint

$$\|x\|_1 \leq r$$

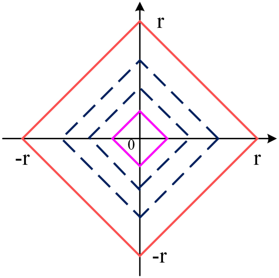
Example (basalt column constraint)

Letting A be a total variation matrix and $b = 0$, then the generalized constraint in L_1-RCP reduces to the following basalt column constraint

$$\begin{bmatrix} 1 & -1 & 0 & 0 & \cdots \\ 0 & 1 & -1 & 0 & \cdots \\ & & \vdots & & \\ \cdots & 0 & 0 & 1 & -1 \\ -1 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{d-1} \\ x_d \end{bmatrix} \implies \left\| \begin{bmatrix} x_2 - x_1 \\ x_3 - x_2 \\ \vdots \\ x_d - x_{d-1} \\ x_d - x_1 \end{bmatrix} \right\|_1 \leq r$$

- Moderate the number of jumps or switches for piecewise functions.

Pictorial interpretation



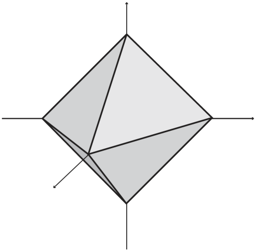


FIG. 2.1. *The lasso constraint.*

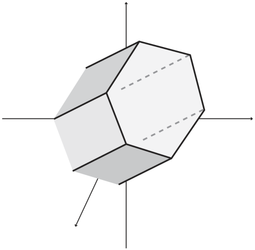


FIG. 2.2. *The basalt column constraint.*

- The contour of lasso is a diamond in $\mathbb{R}^2$.
- As r increases, search domain enlarges, optimal value ($\min_{x\in\mathbb{R}^d} \; L_N^*$) improves, optimal solution x_N^* loses generalization property.
- Recall some constrained optimization problems
  - $\min_x \|b - Ax\|_2 \quad \text{s.t. } \|x\|_0 \le r \qquad$ (Best subset selection)
    $\Leftrightarrow \quad \min_x \|x\|_0 \quad \text{s.t. } \|b - Ax\|_2 \le t$
  - $\min_x \|b - Ax\|_2 \quad \text{s.t. } \|x\|_1 \le r \qquad$ (Lasso)

q-dimensional subspace

- q is a user-chosen "complexity barrier" and satisfies $q < d$.
- Optimal q-dimensional subspace: $\mathcal{Z}^{opt}$:
  Set some rows of $Ax - b$ as zero (i.e., $a_h^\top - b_h = 0$), that is,

  $$\begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1d}x_d - b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2d}x_d - b_2 \\ \vdots \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pd}x_d - b_p \end{bmatrix} = \begin{bmatrix} a_1^\top - b_1 \\ a_2^\top - b_2 \\ \vdots \\ a_p^\top - b_p \end{bmatrix}$$

  so that

  $$\min_{x \in \mathcal{Z}^{opt} \cap \mathcal{X}} \max_{i=1,\cdots,N} L(x, \delta^{(i)}) \leq \min_{x \in \mathcal{Z} \cap \mathcal{X}} \max_{i=1,\cdots,N} L(x, \delta^{(i)})$$

- Two requirements for a suitable selection of q:
  - Guarantee adequate generalization properties;
  - Allow for a satisfactory optimal cost.

Algorithm (L_1-RCA)

Random convex algorithm with L_1-regularization (L_1–RCA).

(a) Let s be the dimension of the affine subspace of $\mathbb{R}^d$ identified by relation $Ax - b = 0$. Select an integer q with $s < q < d$.
Initialize $r = 0$.

(b) Let $x_N^*(r)$ be the optimal solution path of **L_1–RCP** as r is increased.
For all values of $r \geq 0$, evaluate which components of $Ax_N^*(r) - b$ are zero, and let $H(r)$ be the index set of the zero components of $Ax_N^*(r) - b$; thus, if, for example, the first two components of $Ax_N^*(r) - b$ are zero, we have $H(r) = \{1, 2\}$. Further, define $\mathcal{Z}(r) := \{x : a_h^T x - b_h = 0, h \in H(r)\}$, where $a_h^T x - b_h$ is the hth component of $Ax - b$; that is, $\mathcal{Z}(r)$ is the affine subspace of $\mathbb{R}^d$ preserving the null components of $Ax_N^*(r) - b$.

Set $\bar{r}$ to be the largest r such that $\dim(\mathcal{Z}(r)) = q$.

(c) Solve

$$\min_{x \in \mathcal{Z}(\bar{r}) \cap \mathcal{X}} \max_{i=1,\ldots,N} L(x, \delta^{(i)}),$$

and let x_N^* and L_N^* be the optimal solution and the optimal value of this problem.

Assumptions

Assumption 2 (Existence and Uniqueness)

W.p. 1 w.r.t. the multisample $\boldsymbol{\delta}$, any RCP considered here admits a unique solution.

Assumption 3

W.p. 1 w.r.t. the multisample $\boldsymbol{\delta}$, when function $m(r) = \dim(\mathcal{Z}(r))$ increases, it does so one unit at a time, that is, it does not have jumps up of 2 or more units, and $m(\infty) : \lim_{r\to\infty} \; m(r) = d$.

Termination of L_1-RCA

For $r = 0$, $\|Ax_N^*(0) - b\|_1 = 0$ so that $Ax_N^*(0) - b = 0$ which entails that $m(0) = s$. Thus $m(r)$ goes from s to d, when it increases, it does so one unit at a time. Hence, an r exists where $m(r) = q$. Moreover, the sup $\bar{r}$ takes $m(\bar{r}) = q$. After $\bar{r}$ is determined in (b), then (c) generates x_N^* and L_N^* and terminates the ALGO.

Theory

THEOREM 3.2

For L_1-RCA algorithm, if it takes sample complexity

$$N \geq \frac{2}{\epsilon}\left[\ln\frac{1}{\beta} + q + (p-d+q)\ln\left(\frac{p\cdot e}{p-d+q}\right)\right].$$

Under Assumptions 1, 2, 3, for all multisample $\boldsymbol{\delta}$ with the exception of a set whose probability $\mathbb{P}^N$ is at most β:

There is a set Δ_ϵ with $\mathbb{P}\{\Delta_\epsilon\} \geq 1-\epsilon$ such that

$$\max_{\delta\in\Delta_\epsilon} L(x_N^*(\boldsymbol{\delta}), \delta) \leq L_N^*(\boldsymbol{\delta})$$

- This theorem is equivalent to a more general result, that is,

$$\binom{p}{d-q}\sum_{i=0}^{q}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i} \leq \beta.$$

- Special case: take $p = d$, then the term w.r.t N is as follows:

$$N \geq \frac{2}{\epsilon}\left[\ln\frac{1}{\beta} + q\left(1+\ln\frac{d\cdot e}{q}\right)\right].$$

Compared with previous results

TABLE 3.1

Values for N obtained using formula (3.15) (1st line in italic) and formula (3.8) (2nd through 9th lines); $\beta = 10^{-10}$, $p = d = 2000$.

	$\epsilon = 1\%$	$\epsilon = 2\%$	$\epsilon = 3\%$	$\epsilon = 4\%$	$\epsilon = 5\%$	$\epsilon = 6\%$	$\epsilon = 7\%$	$\epsilon = 8\%$	$\epsilon = 9\%$	$\epsilon = 10\%$
	229735	*114793*	*76478*	*57321*	*45826*	*38163*	*32689*	*28584*	*25390*	*22836*
$q = 1$	3403	1693	1123	838	668	554	472	411	364	325
$q = 2$	4427	2203	1462	1091	869	720	614	535	473	424
$q = 3$	5403	2689	1784	1332	1060	879	750	653	578	517
$q = 4$	6346	3158	2096	1564	1245	1033	881	767	679	608
$q = 5$	7264	3615	2399	1791	1426	1182	1009	878	777	696
$q = 10$	11594	5771	3829	2859	2276	1888	1610	1402	1240	1111
$q = 15$	15644	7786	5167	3858	3072	2548	2173	1893	1674	1500
$q = 20$	19506	9709	6443	4810	3831	3177	2711	2361	2088	1870

- Previous result: $\sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta.$
- For example: $\epsilon = 5\%$, $N_{old} = 45826$, $q = 10$, $N_{new} = 2276$.

[Campi & Garatti, *SIAM J. Optim.*, 2008]

Role of Probability $\mathbb{P}$, confidence β and risk ϵ

Role of Probability $\mathbb{P}$:

- Training samples $\{\delta^{(i)}\}_{i=1}^{N}$ are generated according to probability $\mathbb{P}$.
- Generalization property refers to sampling a new scenario δ again according to probability $\mathbb{P}$;
- Verify whether $L(x_N^*(\boldsymbol{\delta}), \delta) \leq L_N^*(\boldsymbol{\delta})$.

♠ QUES:

- What happens if the testing probability and the verification probability do not coincide ?
    - Ambiguity set (ACCP: ambiguous chance-constrained program)
    - Prohorov metric [Erdoğan & Iyengar, *Math. Program.*, 2006]
    - Wasserstein metric (DRO: distributionally robust optimization) [Esfahani & Kuhn, *Math. Program.*, 2017]

Role of confidence β and risk ϵ:

- For practical appeal of method, confidence β should take small.
- As scenarios N tends to infinity, the risk ϵ tends to zero.

Example: Minimax Regression

Minimax Regression

A signal $s(t)$ is obtained as the composition of 200 sinusoids,

$$s(t) = \sum_{j=1}^{200} \alpha_j \sin(jt), \;\; \hat{s}(t) = \sum_{j=1}^{200} x_j \sin(jt) \;\rightarrow\; \hat{s}(t) = \sum_{k=1}^{7} x_{j_k} \sin(j_k t)$$

- Take $\alpha_1 = \alpha_5 = \alpha_8 = \alpha_{45} = 0.2$, $\sum_{j \neq 1,5,8,45} \alpha_j = 1$ and $\sum_{j=1}^{200} \alpha_j = 1$.
- Gather $N = 332$ samples of $(t^{(i)}, s(t^{(i)}))$, where $t^{(i)} \sim U(-\pi, \pi)$.
- Select $q = 7$ nonzero coefficients x_{j_k} with frequencies $j_1 = 1$, $j_2 = 5$, $j_3 = 8$, $j_4 = 41$, $j_5 = 45$, $j_6 = 109$, $j_7 = 127$.

Consider L_1-RCP ALGO as follows

$$\min_{x \in \mathbb{R}^{200}} \max_{i=1,\cdots,332} \left| s(t^{(i)}) - \hat{s}(t^{(i)}) \right|, \quad \text{s.t. } \|x\|_1 \leq r,$$

$$\text{(reduced order)} \;\Rightarrow\; \min_{x_{j_1},\cdots,x_{j_7}} \max_{i=1,\cdots,332} \left| s(t^{(i)}) - \sum_{k=1}^{7} x_{j_k} \sin(x_{j_k} t^{(i)}) \right|,$$

The obtained optimal solutions are: $x_{j_1}^* = 0.1909$, $x_{j_2}^* = 0.1964$, $x_{j_3}^* = 0.2033$, $x_{j_4}^* = 0.0187$, $x_{j_5}^* = 0.2059$, $x_{j_6}^* = 0.0271$, $x_{j_7}^* = 0.0184$, and cost $L_{332}^* = 0.0649$.

Numerical Experiments

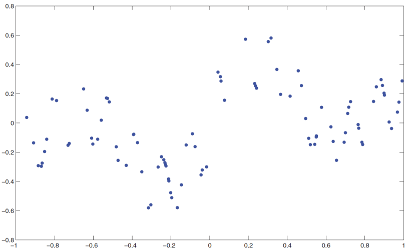


FIG. 4.1. *Samples* $(t^{(i)}, s(t^{(i)}))$.

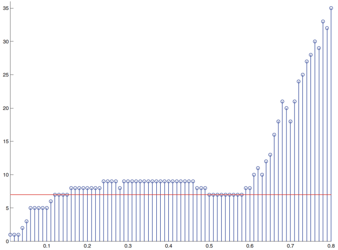


FIG. 4.2. *Number of nonzero coefficients for* $r \leq 0.8$*; the horizontal line is at level* $q = 7$.

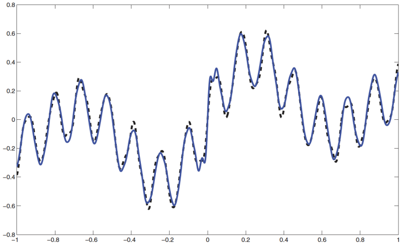


FIG. 4.3. *Signal* $s(t)$ *(solid line) and reduced order signal* $= \hat{s}^*_{332}(t)$ *(dashed line).*

A posterior evaluation

PROPOSITION 4.1

Let x_N^* be the solution obtained with L_1-RCA. Take

$$M \geq \frac{1}{\epsilon'} \ln \frac{1}{\beta'}$$

i.i.d. samples $\delta^{(N+1)}, \cdots, \delta^{(N+M)}$ distributed according to $\mathbb{P}$ and independent of $\delta^{(1)}, \cdots, \delta^{(N)}$ and let

$$L^* = \max_{i=N+1,\cdots,N+M} L(x_N^*, \delta^{(i)}).$$

Then, with confidence $1-\beta'$ w.r.t. the multisample $\delta^{(N+1)}, \cdots, \delta^{(N+M)}$, relation

$$L(x_N^*, \delta) \leq L^*$$

holds with probability at least $1-\epsilon'$ w.r.t. random choices of δ.

Performance robustness cost ϵ_ℓ

Assessment of Robustness-loss curve

Let $\ell = q+1, \cdots, q+h$, $\alpha = p-d+q$, and

$$\epsilon_\ell = \frac{\ell}{N} + \frac{g-1+\sqrt{g^2+2(\ell-1)g}}{N}, \quad g = \ln\left[\frac{1}{\beta}\cdot\left(\frac{p\cdot e}{\alpha}\right)^\alpha\right],$$

where h is an arbitrary integer chosen by the user such that $q+h \leq N$. To easy notation, denote x^* as $x_N^*(\boldsymbol{\delta})$. Define

$$L^*_{\epsilon_\ell} = \max\{L \text{ such that } L(x^*, \delta^{(i)}) \geq L \text{ for } \ell \text{ scenarios } \delta^{(i)}\}.$$

Thus, $L^*_{\epsilon_\ell}$ are the values $L(x^*, \delta^{(i)})$ listed in decreasing order of magnitude.

- The first term of ϵ_ℓ is the *empirical probability* of the scenarios that greater than or equal to $L^*_{\epsilon_\ell}$.
- The second term $\frac{g+\sqrt{g^2+2(\ell-1)g}}{N}$ of ϵ_ℓ is the adjustment term accounting for the mismatch between empirical and real probability.

Results

THEOREM 5.1

The statement $L(x^*, \delta) \leq L^*_{\epsilon_\ell}$ holds with probability at least $1 - \epsilon_\ell$ is true simultaneously for all $\ell = q+1, \cdots, q+h$ with confidence $1 - h\beta$.

$$\Leftrightarrow \quad \mathbb{P}^N\{\boldsymbol{\delta} : \mathbb{P}\{L(x^*, \delta) > L^*_{\epsilon_\ell}\} > \epsilon_\ell\} \leq \binom{p}{d-q} \sum_{i=0}^{\ell-1} \binom{N}{i} \epsilon_\ell^i (1-\epsilon_\ell)^{N-i}.$$

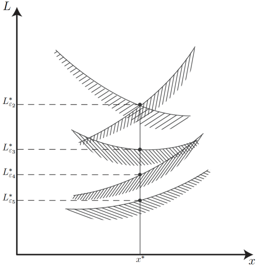


FIG. 5.1. *Visualization of $L^*_{\epsilon_\ell}$ for $q = 1$. Each constraint represents the region where $L(x, \delta^{(i)})$ for some $\delta^{(i)}$.*

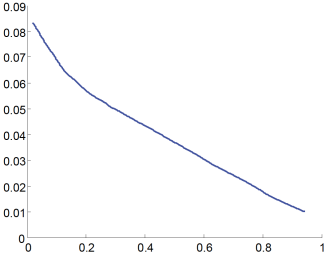


FIG. 5.3. *Robustness-loss curve: $L^*_{\epsilon_\ell}$ (vertical axis) vs. ϵ_ℓ (horizontal axis). ℓ is in the range* $8, \ldots, 6007$.

Conclusions

Take home messages:

- L_1-regularization shrinks the number of optimization variables.
- Induce a L_1-RCP algorithm.
- Enhance the generalization properties of the RCP.
- Perform a novel finite-sample guarantee:

$$\mathbb{P}^N\{\boldsymbol{\delta} : \mathbb{P}\{L(x^*, \delta) > L^*\} > \epsilon\} \leq \binom{p}{d-q} \sum_{i=0}^{q} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}.$$

- Does not require any knowledge of probability measure $\mathbb{P}$ (unknown).

Improvement

- Even use L_1-RCA, N scales as $\frac{1}{\epsilon} \cdot d$.
- Fast algorithm gives the form of sample complexity N as $\frac{1}{\epsilon} + d$.

[Carè, Garatti and Campi, *Operations Research*, 2014]