Projection-free Methods and Their Applications

Guanghui Lan

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology, USA

Boris Polyak's Russian Seminar on Optimization
Moscow Institute of Physics and Technology (MIPT)
July 29, 2020

The H. Milton Stewart School of Industrial and Systems Engineering

- Background and motivation
  - Novelty Detection
  - Matrix Completion
  - Intensity Modulated Radiation Therapy (IMRT)
- Conditional Gradient over Simple Constraints
  - Classical Conditional Gradient and Recent Advancements
  - Applications
- Conditional Gradient over Function Constraints
  - Constraint-extrapolated CG
  - Constraint-extrapolated and Dual-regularized CG
  - Experimental Results for IMRT
- Summary

Background and Motivation

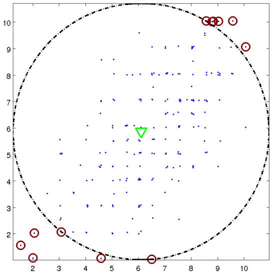
- Increasing interest in applying optimization to
  - Statistics, machine learning, artificial intelligence
  - Engineering design, finance, healthcare, ...
- New challenges in designing solution methods:
  - High dimensionality: millions of variables or more
  - Sparse solutions: statistically or physically meaningful
  - Low or moderate accuracy is acceptable
  - Simplicity: easy to implement

Motivating Problem I: Novelty Detection

Goal: find the boundary between inlier and outlier.

- Find the smallest ball with center c and radius r such that it includes all inlier data points.

$$\begin{aligned} \min \quad & r \\ \text{s.t.} \quad & (x^i - c)^T (x^i - c) \leq r, i = 1, \ldots, m. \end{aligned}$$



Dual Formulation of Novelty Detection

Quadratic optimization over a simplex.

$$\begin{aligned} \min \quad & \left\{ g(\alpha) := \textstyle\sum_{i=1}^m \alpha_i {x^i}^T x^i - \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j {x^i}^T x^j \right\} \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^m \alpha_i = 1, \alpha_i \geq 0. \end{aligned}$$

How to derive it?

- Lagrangian function:
  $L(r, c, \alpha) = r + \sum_{i=1}^m \alpha_i [(x^i - c)^T (x^i - c) - r)]$.
- Set the partial derivatives of L w.r.t. r and c to 0

  $$\frac{\partial L}{\partial r} = 1 - \textstyle\sum_{i=1}^m \alpha_i = 0 \Rightarrow \sum_{i=1}^m \alpha_i = 1,$$

  $$\frac{\partial L}{\partial c} = \textstyle\sum_{i=1}^m \alpha_i (-2x^i + 2c) = 0 \Rightarrow c = \sum_{i=1}^m \alpha_i x^i.$$

  Using these relations to simplify $\max_{\alpha \geq 0} \min_{r,c} L(r, c, \alpha)$.

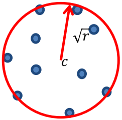
Challenges of Novelty Detection

- High dimension: a collection of 10^6 or more data points implies 10^6 or more dual variables.
- Sparse solution: many data points will be inside the circle

$$(x^i - c)^T (x^i - c) < r,$$

$$\alpha_i = 0.$$

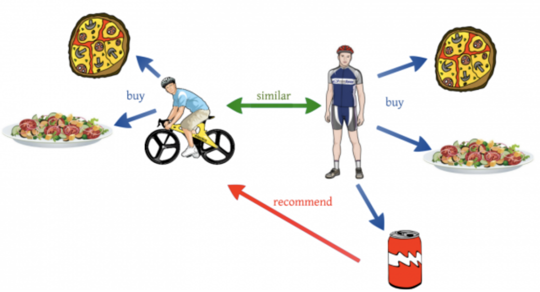
  Very few data points are on the boundary



- Do not seek highly accurate solutions due to inherent data uncertainty.

Motivating Problem II: Recommendation Systems

Assumption: people like things similar to other things they like, and things that are liked by other people with similar taste.



Netflix Problem

- Netflix challenge: Netflix provides highly incomplete ratings from 0.5 million users for 17,770 movies.
- How to predict user ratings to recommend movies?

Matrix Completion

- Given a partially-observed noisy matrix M, we would like to approximately complete it.
- In Netflix problem,
  - $M_{u,i}$ is a rating on movie i by user u.
  - Need to estimate unrated movies.

movies

		3			
	3				5
users		5		4	
			2	5	
	1			2	

Low rank

A few factors explain most of the data $\longrightarrow$ low-rank approximation

How to exploit (approx.) low-rank structure in prediction?

Matrix Completion: Formulation and Challenges

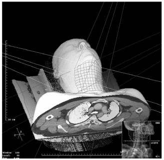
Let $\mathcal{A}$ denotes the index set of given ratings,
$\|X\|_* := \sum_{i=1}^{\min\{m,n\}} \sigma_i(X)$ denotes the nuclear norm of X.

$$\min_X \{\sum_{(u,i)\in\mathcal{A}} (M_{u,i} - X_{(u,i)})^2 : \|X\|_* \le r\}$$

- High dimensionality: $X \in \mathbb{R}^{5{,}000{,}000 \times 17{,}770}$.
- Sparsity: a small number of nonzero singular values ($\sigma_i(X)$).
- Do not seek highly accurate solutions.
- Usually cannot afford full singular value decomposition.

Motivating Example III: Intensity Modulated Radiation Therapy (IMRT)

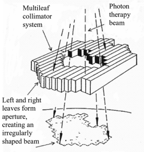
- Each year approximately 1.7 millions people are diagnosed with cancer and more than half may benefit from IMRT.



- Patient irradiated by a linear accelerator (Linac) from several *angles* denoted by A.
- Each structure s of the patient discretized into small *voxels* $\mathcal{V}$.

Definition of Aperture

- A beam in each angle, b_a, is decomposed into a rectangular grid of *beamlets*.
- A beamlet (i, j) is effective if it is not blocked by either the left, l_i, and right, r_i, leaves.
- An *aperture* is defined as the collection of effective beamlets.
- The motion of the leaves controls the set of effective beamlets and thus the shape of the aperture.



Intensity Modulated Radiation Therapy (IMRT)

- K_a: the set of allowed apertures determined by the position of the left and right leaves in beam angle a.
- The rectangular grid in each angle has m rows and n columns, the number of possible apertures in each angle: $\left(\frac{n(n-1)}{2}\right)^m$.
- Use $\mathbf{x}^{k,a}$, comprised of binary decision variables $x_{(i,j)}^{k,a}$, to describe the shape of aperture $k \in K_a$
  - $x_{(i,j)}^{k,a} = 1$ if beamlet (i,j) is effective, i.e., falling within the left and right leaves of row i,
  - otherwise $x_{(i,j)}^{k,a} = 0$.
- To determine the influence rate $y^{k,a}$ for aperture $k \in K_a$, which will be used to determine the dose intensity and the amount of radiation time from aperture k.

IMRT Problem Statement

- Dose received by voxel v from beamlet (i, j) at unit intensity is denoted by $D_{(i,j)v}$ in Gray(Gy).
- Dose absorbed by a given voxel: $z_v = \sum_{a \in \mathcal{A}} \sum_{k \in \mathcal{K}_a} \sum_{i=1}^{m} \sum_{j=1}^{n} RD_{(i,j)v}\, x_{ij}^{k,a} y^{k,a}$.
- Let $\underline{T}_v$ and $\overline{T}_v$ be pre-specified lower and upper dose thresholds for voxel v.
- Define $f(\mathbf{z}) := \sum_{v \in \mathcal{V}} \underline{w}_v \left[\underline{T}_v - z_v\right]_+^2 + \overline{w}_v \left[z_v - \overline{T}_v\right]_+^2$, where $[\cdot]_+$ denotes $\max\{0, \cdot\}$.
- f acts as a surrogate for some clinical criteria.

IMRT Treatment Planning: Basic Formulation

Denote $\hat{D}_v^{k,a} := \sum_{i=1}^m \sum_{j=1}^n D_{(i,j)v}\, x_{ij}^{k,a}$.

$$
\begin{aligned}
\min \quad & f(\mathbf{z}) := \frac{1}{N_v} \sum_{v\in\mathcal{V}} \underline{w}_v \left[\underline{T}_v - z_v\right]_+^2 + \overline{w}_v \left[z_v - \overline{T}_v\right]_+^2 \\
\text{s.t.} \quad & z_v = \sum_{a\in\mathcal{A}} \sum_{k\in\mathcal{K}_a} R \hat{D}_v^{k,a} y^{k,a}, \\
& \sum_{a\in\mathcal{A}} \sum_{k\in K_a} y^{k,a} \le 1, \\
& y^{k,a} \ge 0.
\end{aligned}
$$

IMRT: Challenges

- Huge-scale: the size of $y^{k,a}$ exponentially increases w.r.t. m.
  - 180×45^{10} decision variables for 180 angles and 10×10 grids.
- Sparsity: smaller number of apertures (or angles) implies less radiation exposure.
- Do not seek highly accurate solutions since f acts as surrogate for clinical criteria.

Conditional Gradient Method

- One of the earliest methods initially developed by Frank and Wolfe (1956) for convex optimization:

$$
\begin{aligned}
\min \quad & f(x) \\
\text{s.t.} \quad & x \in X.
\end{aligned}
$$

- $X \subseteq \mathbb{R}^n$ is a convex compact set.
- $f : X \to \mathbb{R}$ is smooth (differentiable with Lipschitz continuous gradients).

The Algorithm

Linear Optimization (LO) Oracle

Minimizing a linear function over X is simple: for a given $p \in \mathbb{R}^n$, we can easily compute a solution of $\min_{x \in X} \langle p, x \rangle$.

Algorithm 1 Conditional Gradient Method

Input: $x_0 \in X$, $\alpha_t = 2/(t+1)$[1].
for $t = 0, \ldots, k$ **do**
  Compute gradient $\nabla f(x_t)$.
  $y_t \in \mathrm{Argmin}_{x \in X} \{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle \}$.
  $x_{t+1} = (1 - \alpha_t) x_t + \alpha_t y_t$.
end for

[1] α_t can be improved by a simple line search procedure.

Projected Gradient vs Conditional Gradient

$$\min\{x^2 + y^2 : x + y = 1, x, y \geq 0\}, \; (x^*, y^*) = (1/2, 1/2).$$

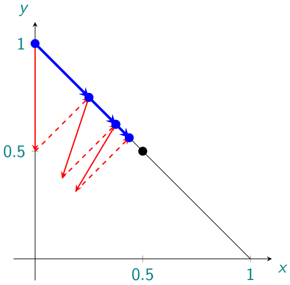


Figure: Left: Project gradient. Blue arrows: solution updates, red solid arrows: gradient descent moves, red dashed arrows: projection onto X. Right: Conditional gradient. Blue arrows: solution updates, red solid arrows: the updates (always extreme points), red dashed arrows: convex combination.

Features of Conditional Gradient

- Simple: no need to choose stepsize.
- Projection-free: only need to solve a linear optimization problem.
  - Useful if the projection step is complicated.
- Sparse solution: only one extreme point is added at each iteration.
- Convergence: slower than (accelerated) projected gradient descent in general.

Convergence of Conditional Gradient

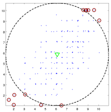
Theorem

Let $\epsilon > 0$ be given. The number of iterations performed by the conditional gradient method to find a solution $\bar{x} \in X$ s.t. $f(\bar{x}) - f^* \leq \epsilon$ is bound by $\mathcal{O}(1/\epsilon)$.

- The number calls to linear optimization oracle is not improvable (Jaggi 13, Lan 13, Guzman and Nemirovski 15).
- But the number of gradient computations can be improved (Lan and Zhou 14).
- Extensions to nonsmooth problems (Lan 13).
- Extensions to nonconvex problems (Jiang et. al. 16).
- See Chapter 7 of Lan 20 for more details.

Application to Novelty Detection

$$\begin{array}{ll} \min & \left\{ g(\alpha) := \sum_{i=1}^{m} \alpha_i {x^i}^T x^i - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j {x^i}^T x^j \right\} \\ \text{s.t.} & \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0. \end{array}$$



- Linear optimization: $\min\{\langle \nabla g(\alpha_t), \alpha \rangle : \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0\}$:
    - Find the most negative gradient component, set the corresponding coordinate of α to 1.
    - Return the corresponding extreme point.
- High dimensionality: complexity independent of dimension.
- Sparsity: number of nonzero elements bounded by $\mathcal{O}(1/\epsilon)$.
- Accuracy: low/moderate.
- Convenient for implementation.

Application to Matrix Completion

$$\min \quad \left\{ f(X) := \textstyle\sum_{(u,i)\in\mathcal{A}} (M_{u,i} - X_{(u,i)})^2 \right\}$$
$$\text{s.t.} \quad \|X\|_* \leq r.$$

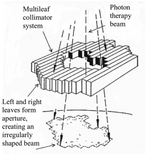
- Linear optimization: $\min\{\langle \nabla f(X_t), X\rangle : \|X\|_* \leq r\}$:
  - Find the largest singular value of $\nabla f(X_t)$ and the corresponding singular vectors (u_t, v_t),
  - Return $-r u_t v_t^T$.
- High dimensionality: complexity independent of dimension.
- Sparsity: rank bounded by $\mathcal{O}(1/\epsilon)$.
- Accuracy: low/moderate.
- Implementation: no full singular value decomposition.

Application to IMRT

$$\hat{D}_v^{k,a} := \sum_{i=1}^m \sum_{j=1}^n D_{(i,j)v}\, x_{ij}^{k,a}$$

$$\begin{aligned} \min \quad & f(\mathbf{z}) := \frac{1}{N_v} \sum_{v\in\mathcal{V}} \underline{w}_v \,[\underline{T}_v - z_v]_+^2 + \overline{w}_v\,[z_v - \overline{T}_v]_+^2 \\ \text{s.t.} \quad & z_v = \sum_{a\in\mathcal{A}} \sum_{k\in\mathcal{K}_a} R\hat{D}_v^{k,a} y^{k,a}, \\ & \sum_{a\in\mathcal{A}} \sum_{k\in K_a} y^{k,a} \le 1, \\ & y^{k,a} \ge 0. \end{aligned}$$

- $(\frac{n(n-1)}{2})^m$ apertures in each angle.
- 180×45^{10} for 180 angles and 10×10 grids.



Application to IMRT

Gradient computation and linear optimization:

- Given $y_t^{k,a}$ compute the gradient of f w.r.t. z.
- Apply chain rule to $y^{k,a}$, the magnitude of the gradient for each aperture will depend on the binary variables $x_{ij}^{k,a}$.
- Find the aperture with the most negative gradient component.
  - Examine the grid of each angle row by row,
  - Select the position of the leaves resulting in the smallest gradient along this row,
  - The value of $x_{ij}^{k,a}$ is fixed for the selected aperture.
- NO full gradient information is computed or stored.

Sparsity: the number of aperture will be bounded by $\mathcal{O}(1/\epsilon)$.

Remaining Challenges

Only handle models with simple constraints!

- Matrix completion: adding linear constraints will make the subproblem as hard as a general semidefinite program.
- IMRT: need to add different types of function constraints
  - Group sparsity constraints to ensure a small number of angles,
  - Risk averse constraints to avoid overdose (underdose) for normal (tumor) structures.
- To develop new projection-free methods for solving problems with general function constraints.

General convex optimization

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) := Ax - b = 0, \\ & h_i(x) \le 0, \quad i = 1, \ldots, d, \\ & x \in X. \end{aligned}$$

- $X \subseteq \mathbb{R}^n$ is a convex compact set.
- Only linear optimization over X, no projection.
- $f : X \to \mathbb{R}$ and $h_i : X \to \mathbb{R}$, $i = 1, \ldots, d$, are proper lower semicontinuous convex functions.
- $A : \mathbb{R}^n \to \mathbb{R}^m$ denotes a linear mapping.
- b: a given vector in $\mathbb{R}^m$.
- Denote $h(x) \equiv (h_1(x); \ldots, h_d(x))$.

Projection-free methods

- Naturally consider its saddle point reformulation:

  $\min_{x\in X} \max_{y\in\mathbb{R}^m, z\in\mathbb{R}^d_+} f(x) + \langle g(x), y\rangle + \langle h(x), z\rangle.$
- The smoothing CG method (Lan 13) is not applicable:
  - can only deal with linear coupling term $\langle g(x), y\rangle$ but not $\langle h(x), z\rangle$.
  - can not handle unbounded dual feasible set.
- The constrained extrapolation (ConEx) method by Boob, Deng and Lan (19):
  - Optimal complexity for solving a wide range of function constrained problems uniformly.
  - But require projections over X.

Constraint-extrapolated Conditional Gradient (CoexCG)

$$l_{h_i}(\bar{x},x) := h_i(\bar{x}) + \langle \nabla h_i(\bar{x}), x - \bar{x}\rangle, \; l_h(\bar{x},x) := (l_{h_1}(\bar{x},x); \ldots, l_{h_d}(\bar{x},x)).$$

Algorithm 2 CoexCG

for $k = 1$ **to** N **do**

$$\tilde{g}_k = g(p_{k-1}) + \lambda_k[g(p_{k-1}) - g(p_{k-2})],$$

$$\tilde{h}_k = l_h(x_{k-2}, p_{k-1}) + \lambda_k[l_h(x_{k-2}, p_{k-1}) - l_h(x_{k-3}, p_{k-2})],$$

$$q_k = \mathrm{argmin}_{y\in\mathbb{R}^m}\{\langle -\tilde{g}_k, y\rangle + \tfrac{\tau_k}{2}\|y - q_{k-1}\|_2^2\} = q_{k-1} + \tfrac{1}{\tau_k}\tilde{g}_k,$$

$$r_k = \mathrm{argmin}_{z\in\mathbb{R}^d_+}\{\langle -\tilde{h}_k, z\rangle + \tfrac{\tau_k}{2}\|z - r_{k-1}\|_2^2\} = [r_{k-1} + \tfrac{1}{\tau_k}\tilde{h}_k]_+,$$

$$p_k = \mathrm{argmin}_{x\in X}\{l_f(x_{k-1}, x) + \langle g(x), q_k\rangle + \langle l_h(x_{k-1}, x), r_k\rangle\},$$

$$x_k = (1-\alpha_k)x_{k-1} + \alpha_k p_k.$$

end for

Convergence of CoexCG

Theorem

Assume $\alpha_k = 2/(k+1)$, $\lambda_k = (k-1)/k$, *and* $\tau_k = N^{3/2}/k$ *in CoexCG. Let* $\epsilon > 0$ *be given. Assume that* f *and* h_i *are smooth convex functions. The total number of iterations performed by CoexCG before finding a point* $\bar{x} \in X$ *s.t.* $f(\bar{x}) - f(x^*) \leq \epsilon$ *and* $\|g(\bar{x})\|_2 + \|[h(\bar{x})]_+\|_2 \leq \epsilon$, *can be bounded by* $\mathcal{O}(1/\epsilon^2)$.

- This $\mathcal{O}(1/\epsilon^2)$ bound appears to be tight.
- If f or some h_i are structured nonsmooth (containing a bilinear saddle point), a similar $\mathcal{O}(1/\epsilon^2)$ can be attained.
- Requires to fix N a priori when setting algorithmic parameters: is it possible to improve?

Constraint-extrapolated and Dual-regularized Conditional Gradient (CoexDurCG)

Algorithm 3 CoexDurCG

The algorithm is the same as CoexCG except that
$q_k = \mathrm{argmin}_{y\in\mathbb{R}^m}\{\langle -\tilde{g}_k, y\rangle + \frac{\tau_k}{2}\|y - q_{k-1}\|_2^2 + \frac{\gamma_k}{2}\|y - q_0\|_2^2\}$,
$r_k = \mathrm{argmin}_{z\in\mathbb{R}^d_+}\{\langle -\tilde{h}_k, z\rangle + \frac{\tau_k}{2}\|z - r_{k-1}\|_2^2 + \frac{\gamma_k}{2}\|z - r_0\|_2^2\}$,
for some $\gamma_k \geq 0$.

- Set $\alpha_k = 2/(k+1)$, $\lambda_k = (k-1)/k$, $\tau_k = \sqrt{k}$, $\gamma_k = [(k+1)\sqrt{k+1} - k\sqrt{k}]/k$.
- CoexDurCG achieves similar convergence as CoexCG.
- Does not require N fixed in advance.
- Analysis is more involved.

Application to IMRT

Clinical criteria to avoid underdose (resp., overdose) for tumor (resp., healthy) structures.

- Usually specified as value at risk (VaR) constraints.
- "PTV56:V56$\geq$ 95%": the percentage of voxels in structure PTV56 that receive at least 56 Gy dose should be at least 95%.
- "PTV68: V74.8$\leq$ 10%": the percentage of voxels in structure PTV68 that receive more than 74.8 Gy dose should be at most 10%.
- Use Conditional Value at Risk (CVaR) as an approximation.

Group Sparsity

- In the basic IMRT formulation, the simplex constraint only results in a small number of apertures.
- In practice, a small number of angles is desired:
  - not necessary to rotate the patient often.
  - reduce the time for treatment and/or radiation exposure.
- $\sum_{a \in \mathcal{A}} \max_{k \in K_a} y^{k,a} \leq \Phi$ for some properly chosen $\Phi > 0$.
  - Intuitively, encourage the selection of apertures in those angles K_a that have already contained some nonzero elements of $y^{k,a}$, $k \in K_a$.

IMRT New Formulation

$$
\begin{aligned}
\min \quad & f(\mathbf{z}) := \tfrac{1}{N_v} \textstyle\sum_{v \in \mathcal{V}} \underline{w}_v \left[\underline{T}_v - z_v\right]_+^2 + \overline{w}_v \left[z_v - \overline{T}_v\right]_+^2 \\
\text{s.t.} \quad & z_v = \textstyle\sum_{a \in \mathcal{A}} \sum_{k \in \mathcal{K}_a} R \hat{D}_v^{k,a} y^{k,a}, \\
& -\tau_i + \tfrac{1}{p_i N_i} \textstyle\sum_{v \in S_i} [\tau_i - z_v]_+ \leq -b_i, \forall i \in \mathrm{UD}, \\
& \tau_i + \tfrac{1}{p_i N_i} \textstyle\sum_{v \in S_i} [z_v - \tau_i]_+ \leq b_i, \forall i \in \mathrm{OD}, \\
& \textstyle\sum_{a \in \mathcal{A}} \max_{k \in \mathcal{K}_a} y^{k,a} \leq \Phi, \\
& \textstyle\sum_{a \in \mathcal{A}} \sum_{k \in K_a} y^{k,a} \leq 1, \\
& y^{k,a} \geq 0, \tau_i \in [\underline{\tau}_i, \overline{\tau}_i],
\end{aligned}
$$

where OD and UD denote the set of overdose and underdose clinical criteria, respectively.

Implementation

- Smooth objective and (structured) nonsmooth constraints, $\mathcal{O}(1/\epsilon^2)$ iteration complexity.
- A similar procedure as before to simultaneously solve the linear optimization problem and compute the most negative combined gradients of objective and constraints.
- No need to compute full gradient or perform projection.
- Test instances: both randomly generated ones and real dataset.
- Goals:
  - to compare CoexCG with CoexDurCG,
  - to study the group sparsity,
  - to meet the clinical criteria.

Random instances

Index	# of voxels	# of apertures	b_i & p_i
Ins. 1	4096	46080	[30,40,200] & [0.05,0.05,0.05]
Ins. 2	4096	46080	[40,50,100] & [0.01,0.01,0.05]
Ins. 3	4096	46080	[50,60,80] & [0.01,0.01,0.01]
Ins. 4	262144	737280	[40,50,100] & [0.01,0.01,0.05]
Ins. 5	262144	737280	[50,60,80] & [0.01,0.01,0.01]

Comparison of Algorithms

Index	N	CoexCG $f(x_N)$	CoexCG $\|h(x_N)\|$	CoexCG CPU(s)	CoexDurCG $f(x_N)$	CoexDurCG $\|h(x_N)\|$	CoexDurCG CPU(s)
Ins. 1	1	46.8723	1.7237e+03				
	100	0.0683	0.4234	34	0.0616	0.3705	33
	1000	0.0197	0.0319	323	0.0210	0.0219	327
Ins. 2	1	46.8723	1.7237e+03				
	100	0.0568	0.4424	33	0.0583	0.5002	34
	1000	0.0224	0.0426	327	0.0232	0.0334	339
Ins. 3	1	46.8723	1.7237e+03				
	100	0.0625	13.7567	33	0.0604	7.3929	33
	1000	0.0227	0.0514	332	0.0226	0.0193	332
Ins. 4	1	47.7099	8.7850e+03				
	100	0.4643	163.3043	1645	0.4643	163.3043	1645
	1000	0.0398	12.1765	17254	0.0398	12.1765	17356
Ins. 5	1	47.7099	8.7850e+03				
	100	0.4866	253.9389	1644	0.4581	206.9143	1637
	1000	0.0406	39.2051	17146	0.0417	38.6486	17607

Real Prostate Cancer Data

- $3,047,040$ number of voxels, over 2×10^{30} potential apertures.
- Clinical criteria:
  - PTV56: V56$\geq 95\%$
  - PTV68: V68$\geq 95\%$, V74.8$\leq 10\%$
  - Rectum: V30$\leq 80\%$, V50$\leq 50\%$, V65$\leq 25\%$
  - Bladder: V40$\leq 70\%$, V65$\leq 30\%$
  - Left femoral head: V50$\leq 1\%$
  - Right femoral head: V50$\leq 1\%$
- Our results satisfy all these criteria.
- Can reduce the number of angles from 39 to 3 without sacrificing much these criteria.

Dose Voxel Histogram (DVH)

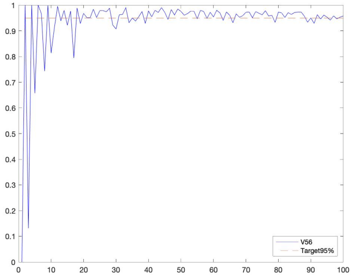


Figure: PTV56 with $\phi = 1$

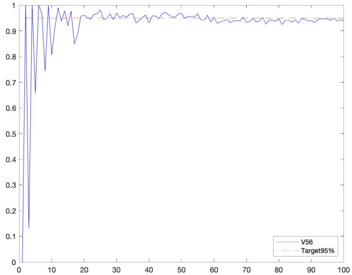


Figure: PTV56 with $\phi = 0.005$

Summary

- CG methods over simple sets can handle
  - high-dimensionality, sparsity, low/moderate accuracy
- CoexCG/CoexDurCG significantly extend these methods to
  - deal with affine, smooth/nonsmooth function constraints
- Wide range of applications
  - Novelty detection, recommendation systems, IMRT,
  - Structured SVM, SDP, ...

References

- G. Lan, First-order and Stochastic Optimization Methods for Machine Learning, Springer Nature, Switzerland AG, 2020 (Chapter 7),
- G. Lan, H. E. Romeijn, Z. Zhou, Conditional Gradient Methods for Convex Optimization with Function Constraints, arXiv 2007.00153, 2020,
- and many relevant references therein.