Projection-free Methods and Their Applications

Guanghui Lan
Department of Industrial Engineering, Georgia Institute of Technology

July 2020

Typed by Yuelin Li

1 Background

The increasing interest in applying optimization to Statistics, machine learning, artificial intelligence, and engineering design, finance and healthcare...
New challenges in designing solution methods include:
(1) high dimensionality: millions of variables or more
(2) sparse solutions: statistically or physically meaningful
(3) low or moderate accuracy is acceptable
(4) simplicity: easy to implement

1.1 Problem 1: Novelty Detection

The goal is to find the boundary between inlier and outlier. Finding the smallest ball with center c and radius r such that it includes all inlier data points.

$$\min r$$

s.t.

$$(x^i - c)^T (x^i - c) \leq r \tag{1}$$

$i = 1, ..., m$. Quadratic optimization over a simplex.

$$\min g(\alpha) := \Sigma_{i=1}^m \alpha_i x^{iT} x^i - \Sigma_{i=1}^m \Sigma_{j=1}^m \alpha_i \alpha_j x^{iT} x^j \tag{2}$$

s.t. $\Sigma_{i=1}^m \alpha_i = 1$, $\alpha \geq 0$. We derive it through Lagrangian function:

$$L(r, c, \alpha) = r + \Sigma_{i=1}^m \alpha_i [(x^i - c)^T (x^i - c) - r] \tag{3}$$

Then set the partial derivatives of L w.r.t. r and c to 0:

$$\frac{\partial L}{\partial r} = 1 - \Sigma_{i=1}^m \alpha_i = 0 \implies \Sigma_{i=1}^m \alpha_i = 1 \tag{4}$$

$$\frac{\partial L}{\partial c} = \Sigma_{i=1}^m \alpha_i (-2x^i + 2c) = 0 \implies c = \Sigma_{i=1}^m \alpha_i x^i \tag{5}$$

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

The challenges of Novelty Detection is high dimension data, i.e. a collection of 10^6 or more data points implies 10^6 or more dual variables; a sparse solution s.t. many data points will be inside the circle $(x^i - c)^T (x^i - c) \leq r$, $\alpha_i = 0$, and very few data points on the boundary. Therefore, we do not seek highly accurate silutions due to inherent data uncertainty.

1.2 Problem 2: Recommendation Systems

We assume that people like things similar to other things they like, and things that are liked by other people with similar taste. 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. Let $M_{u,i}$ be a rating on movie i by user u. Then we need to estimate unrated movies.
Let A denotes the index set of given ratings, $||X||_* := \Sigma_{i=1}^{min(m,n)}\sigma_i(X)$ denotes the nuclear norm of X,

$$\min \Sigma_{(u,i)\in A}(M_{u,i} - X_{(u,i)})^2 : ||X||_* \leq r \tag{6}$$

High dimensionality: $X \in \mathbb{R}^{5,000,000*17,770}$
Sparsity: a small number of nonzero singular values ($\sigma_i(X)$)
Usually, we cannot afford full singular value decomposition.

1.3 Problem 3: Intensity Modulated Radiation Therapy (IMRT)

Each year approximately 1.7 millions people are diagnosed with cancer and more than half may benefit from IMRT. Patients irradiated by a linear accelerator from several angles denoted by A, and each structure s of the patient discretized into small voxels 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.

Let K_a be the set of allowed apertures determined by the position of the left and right leaves in beam angle a. The rectangle grid in each angle has m rows and n columns, the number of possible apertures in each angle is $(\frac{n(n-1)}{2})^m$. We use $x^{k,a}$ that comprised of binary decision variables $x^{k,a}_{(i,j)}$, to describe the shape of aperture $k \in K_a$.

$$x^{k,a}_{(i,j)} = \begin{cases} 1 & \text{if (i,j) is effective} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

To determine the influence rate $y^{k,z}$ 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** Let $D_{(i,j)v}$ in $Gray(Gy)$ be the dose received by voxel v from beamlet (i,j) at unit intensity, and Zv denote dose absorbed by a given voxel, where

$$z_v = \Sigma_{a \in A} \Sigma_{k \in K_a} \Sigma_{i=1}^{m} \Sigma_{j=1}^{n} R D_{(i,j)v} x_{(i,j)}^{k,a} y^{k,a} \tag{8}$$

Let $\underline{T}_v$ and $\overline{T}_v$ be pre-specified lower and upper dose thresholds for voxel v. Define

$$f(z) := \Sigma_{v \in V} \underline{w}_v [\underline{T}_v - z_v]_+^2 + \overline{w}_v [z_v - \overline{T}_v]_+^2 \tag{9}$$

where $[\cdot]_+$ denotes $\max(0, \cdot)$. Here, f acts as a surrogate for some clinicical criterion. Denote $\hat{D}_v^{k,a} := \Sigma_{i=1}^{m} \Sigma_{j=1}^{n} D_{(i,j)v} x_{i,j}^{k,a}$,

$$minf(z) := \frac{1}{N_v} \Sigma_{v \in V} \underline{w}_v [\underline{T}_v - z_v]_+^2 + \overline{w}_v [z_v - \overline{T}_v]_+^2 \tag{10}$$

$$z_v = \Sigma_{a \in A} \Sigma_{k \in K_a} R \hat{D}_v^{k,a} y^{k,a},$$

$$\Sigma_{a \in A} \Sigma_{k \in K_a} y^{k,a} \leq 1,$$

$$y^{k,a} \geq 0.$$

The challenges of IMRT lie in the huge-scale of data, i.e. the size of $y^{k,a}$ exponentially increases w.r.t. m; and the sparsity, i.e. smaller number of apertures or angles implies less radiation exposure. We also do not seek highly accurate solutions since f acts as surrogate for clinical criterion.

2 Conditional Gradient over Simple Constraints

One of the earliest methods initially developed by Frank and Wolfe for convex optimization:

$$\min f(x) \text{ s.t. } x \in X$$

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

Linear Optimiztion (LO) Oracle Minimizing a linear function over X is simple: for a given $p \in \mathbb{R}^n$, we can easilty compute a solution of $\min_{x \in X} (p, x)$.

Algorithm 1 Conditional Gradient Method

```
Input $x_0 \in X$, $\alpha_t = 2/(t+1)$
for $t = 0, ..., k$ do
    Compute gradient $\nabla f(x_t)$
    $y_t \in f(x_t) + (\nabla f(x_t), x - x_t)$
    $x_{t+1} = (1 - \alpha_t)x_t + \alpha_t y_t$
end for
```

(Note that α_t can be improved by a simple line search procedure.)

The features of conditional gradient include:
a) simple, i.e. no need to choose stepsize
b) projection-free: only need to solve a linear optimizatio problem, which is useful when the projection step is complicated
c) sparse solution: only one extreme point is added at each iteration
d) convergence: slower than accelerated projected gradient descent in general.

Convergence of conditional gradient

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

The number calls to linear optimization oracle is not improvable, but the number of gradient computations can be improved. [Explore the extensions to non-smooth problems...]

2.1 Application to novelty detection

$$\min g(\alpha) := \Sigma_{i=1}^{m} \alpha_i x^{i^T} x^i - \Sigma_{i=1}^{m} \Sigma_{j=1}^{m} \alpha_i \alpha_j x^{i^T} x^j \tag{11}$$

s.t. $\Sigma_{i=1}^{m} \alpha_i = 1$, $\alpha \geq 0$.
We find the most negative gradient component, set the corresponding coordinate of α to 1, and return the corresponding extreme point.
a) high dimensionality: complexity independent of dimension
b) sparsity: the number of nonzero elements bounded by $O(\frac{1}{\epsilon})$
c) accuracy: low-moderate
d) Easy to implement

2.2 Application to matrix completion

$$\min \Sigma_{(u,i)\in A}(M_{u,i} - X_{(u,i)})^2 : ||X||_* \leq r \tag{12}$$

We find the largest singular value of $\nabla f(X_t)$ and the correspinding singular vectors (u_t, v_t), return $ru_t v_t^T$
a) high dimensionality: complexity independent of dimension
b) sparsity: the number of nonzero elements bounded by $O(\frac{1}{\epsilon})$
c) accuracy: low-moderate
d) Implementation: no full singular value decomposition.

2.3 Application to IMRT

Gradient computation and linear optimization:
Given $y_t^{k,a}$, we 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 $x_i j^{k,a}$. To 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, and the value of $x_i j^{k,a}$ is fixed for the selected aperture. Note, no full gradient information is computed or stored.
Sparsity: the number of aperture is bounded by $O(\frac{1}{\epsilon})$.

2.4 Challenges

Even handling with models with simple constraints, the challenges remain. For example, in Matrix Completion, adding linear constraints will make the subproblem as hard as a general semidefinite program; in IMRT, we need to add different types of function constraints: ensure a small number of angles, risk averse constraints to avoid overdose (underdose) for normal (tumor) structures...need to develop new project-free methods for solving problems with general function constraints.

3 Project-free methods

Consider saddle point reformulation:

$$\min_{x\in X} \max_{y\in\mathbb{R}^m, z\in\mathbb{R}^d_+} f(x) + \langle g(x), y\rangle + \langle h(x), z\rangle.$$

Previous study shows that the smoothing CG methods is not applicable bacause it can only deal with linear coupling term $\langle g(x), y\rangle$, but not $\langle h(x), z\rangle$. The constrained extrapolation (ConEx) method (by Boob, Deng and Lan) shows optimal complexity for solving a wide range of function constrained problems uniformly, but it requires projections over X.
Therefore, we introduce a new method **Constraint-extrapolated Conditional Gradient (CoexCG)**.

3.1 Constraint-extrapolated Conditional Gradient (CoexCG)

Algorithm 2 CoexCG

```
for $k = 1$ to $N$ do
    $\tilde{g}_k = g(p_{k-1}) + \lambda_k [g(p_{k-1}) - g(p_{k-2})]$
    $\tilde{h}_k = I_h(x_{k-2}, p_{k-1}) + \lambda_k [l_h(x_k - 2, p_{k-1}) - I_h(x_{k-3}, p_{k-2})]$
    $q_k = \text{argmin}_{y\in\mathbb{R}^m} \left\{\langle -\tilde{g}_k, y\rangle + \frac{\tau_k}{2}\|y - q_{k-1}\|_2^2\right\} = q_{k-1} + \frac{1}{\tau_k}\tilde{g}_k$
    $r_k = \text{argmin}_{z\in R^d} \left\{\left\langle -\tilde{h}_k, z\right\rangle + \frac{\tau_k}{2}\|z - r_{k-1}\|_2^2\right\} = \left[r_{k-1} + \frac{1}{\tau_k}\tilde{h}_k\right]_+$
    $p_k = \text{argmin}_{x\in X} \{I_f(x_{k-1}, x) + \langle g(x), q_k\rangle + \langle I_h(x_{k-1}, x), r_k\rangle\}$
    $x_k = (1 - \alpha_k) x_{k-1} + \alpha_k p_k$
end for
```

where the first four steps are simply sum of vectors, and the fifth step is solving a linear problem, the last step is convex optimization.

Theorem 2. 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. $f(\bar{x}) - f(x^*) \le \epsilon$ and $g(\bar{x})\|_2 + \|[h(\bar{x})] + \|_2 \le c$, 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. It requires to fix N a priori when setting algorithmic parameters. [1] [think about how to improve...]

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

Algorithm 3 CoexCG

```
for $k = 1$ to $N$ do
    $\tilde{g}_k = g(p_{k-1}) + \lambda_k [g(p_{k-1}) - g(p_{k-2})]$
    $\tilde{h}_k = I_h(x_{k-2}, p_{k-1}) + \lambda_k [l_h(x_k - 2, p_{k-1}) - I_h(x_{k-3}, p_{k-2})]$
    $q_k = \text{argmin}_{y \in \mathbb{R}^m} \left\{ \langle -\tilde{g}_k, y \rangle + \frac{\tau_k}{2} \|y - q_{k-1}\|_2^2 + \frac{\gamma_k}{2} \|y - q_0\|_2^2 \right\}$
    $r_k = \text{argmin}_{z \in \mathbb{R}^d} \left\{ \left\langle -\tilde{h}_k, z \right\rangle + \frac{\tau_k}{2} \|z - r_{k-1}\|_2^2 + \frac{\gamma_k}{2} \|z - r_0\|_2^2 \right\}$ for some $\gamma_k \geq 0$ .
    $p_k = \text{argmin}_{x \in X} \left\{ I_f(x_{k-1}, x) + \langle g(x), q_k \rangle + \langle I_h(x_{k-1}, x), r_k \rangle \right\}$
    $x_k = (1 - \alpha_k) x_{k-1} + \alpha_k p_k$
end for
```

Set $\alpha_k = 2/(k+1), \lambda_k = (k-1)/k, \tau_k = \sqrt{k}\ a_k = [(k+1)\sqrt{k+1} - k\sqrt{k}]/k$. The algorithm is the same as CoexCG except for q_k, r_k. Also it achieves similar convergence as CoexCG, and does not require N fixed in advance.

3.3 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 PTV 56 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%.
We use Conditional Value at Risk (CVaR) as an approximation. [2]
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 $y^{k,a}$ $k \in K_a$.

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

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

Implementation:
Smooth objective and (structured) nonsmooth constraints, $\mathcal{O}\left(1/\epsilon^2\right)$ 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.
There is no need to compute full gradient or perform projection. Test instances: both randomly generated ones and real dataset.The goals are set to compare CoexCG with CoexDurCG, study the group sparsity, and meet the clinical criteria.

References

[1] H. E. Romeijn G. Lan and Z. Zhou, *Conditional gradient methods for convex optimization with function constraints*, Communications in Contemporary Mathematics (2020).

[2] G. Lan, *First-order and stochastic optimization methods for machine learning*, Springer, 2020.