On the Minimization Over Sparse Symmetric Sets: Projections, Optimality Conditions and Algorithms

Amir Beck

Technion - Israel Institute of Technology
Haifa, Israel

Based on joint work with Nadav Hallak and Yakov Vaisbourd

OPT2014: Workshop on Optimization for Machine Learning (NIPS 2014), Montreal, December 12, 2014

Problem Formulation

The sparse optimization problem

$$(P) \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in C_s \cap B, \end{array}$$

(1) f continuously differentiable

(1) B is closed and convex

(2) $C_s = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_0 \leq s\}$

Difficulties:

(a) $C_s \cap B$ non-convex

(b) $C_s \cap B$ induces a combinatorial constraint

No global optimality conditions, "solution" methods are heuristic in nature.

Example - Compressed Sensing

Linear CS Recover a sparse signal $\mathbf{x}$ with a sampling matrix $\mathbf{A}$ and a measure $\mathbf{b}$.

$$(CS) \qquad \begin{array}{ll} \min & \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \\ \text{s.t.} & \mathbf{x} \in C_s \cap \mathbb{R}^n \end{array}$$

Literature - CS

- Linear:
  1. **Conditions for reconstruction:** RIP (Candes and Tao '05), SRIP (Beck and Teboulle '10), spark (Donoho and Elad '03; Gorodnitsky and Rao '97), mutual coherence (Donoho et al. '03; Donoho and Huo '99; Mallat and Zhang '93)
  2. **Reviews:** Bruckstein et al. '09, Davenport et al. '11, Tropp and Wright '10.
  3. **Iterative algorithms:** IHT (Blumensath and Davis '08, '09, '12; Beck and Teboulle '10), CoSaMP (Needell and Tropp '09)
- Nonlinear:
  1. **Phase retrieval:** Shechtman et al. '13; Ohlsson and Eldar '13; Eldar and Mendelson '13; Eldar et al. '13; Hurt. '89
  2. **Nonlinear:** optimality conditions (Beck and Eldar '13), GraSP (Bahmani et al. '13)

Example - Sparse Index Tracking

Sparse Index Tracking Track an index $\mathbf{b}$ with at most s assets, with return matrix $\mathbf{A}$.

$$(IT) \quad \begin{array}{ll} \min & \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \\ \text{s.t.} & \mathbf{x} \in C_s \cap \Delta_n \end{array}$$

Example: Finance - track the S&P500 with a small number of assets

Takeda et al '12

Example - Sparse Principal Component Analysis

Sparse Principal Component Analysis Find the first principal eigenvector of a matrix $\mathbf{A}$.

$$(PCA) \quad \begin{array}{ll} \max & \mathbf{x}^T \mathbf{A} \mathbf{x} \\ \text{s.t.} & \mathbf{x} \in C_s \cap \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}\|_2 \leq 1\} \end{array}$$

Example: Finance - identify the group which explains most of the variance in the S&P500

Sample of works:

Moghaddam, Weiss, Avidan 06', d'Aspremont, Bach, El-Ghaoui 08', d'Aspremont, El-Ghaoui, Jordan Lanckriet 07',

recent review: Luss and Teboulle '13

Objectives

The sparse optimization problem

$$(P) \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in C_s \cap B, \end{array}$$

B closed and convex

Main Objectives:

- Define necessary **optimality conditions**
- Develop corresponding **algorithms**
- Establish **hierarchy** between algorithms and conditions.

The case $B = \mathbb{R}^n$: Beck, Eldar 13'

Objectives

The sparse optimization problem

$$(P) \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in C_s \cap B, \end{array}$$

B closed and convex

Main Objectives:

- Define necessary **optimality conditions**
- Develop corresponding **algorithms**
- Establish **hierarchy** between algorithms and conditions.

The case $B = \mathbb{R}^n$: Beck, Eldar 13'
However, we will also need to study and compute Orthogonal Projections on $B \cap C_s$.

Recap of Necessary First Order Opt. Conditions over Convex Sets: Stationarity

$$(*)\ \min\{f(\mathbf{x}) : \mathbf{x} \in S\},$$

S closed and convex, f continuously differentiable.

Equivalent Definitions of Stationarity: $\mathbf{x}^*$ stationary point iff

Projection Form

$$\mathbf{x}^* = P_S\left(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\right)$$

for some $L > 0$

Variational Form

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \forall \mathbf{x} \in S$$

Recap of Necessary First Order Opt. Conditions over Convex Sets: Stationarity

$$(*)\ \min\{f(\mathbf{x}) : \mathbf{x} \in S\},$$

S closed and convex, f continuously differentiable.

Equivalent Definitions of Stationarity: $\mathbf{x}^*$ stationary point iff

Projection Form

$$\mathbf{x}^* = P_S\left(\mathbf{x}^* - \frac{1}{L}\nabla f(\mathbf{x}^*)\right)$$

for some $L > 0$

Variational Form

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \forall \mathbf{x} \in S$$

- conditions are equivalent $\Rightarrow$ independent of L
- most algorithms that use first order information converge to stat. points.
- condition relies on the properties/computatbility of $P_S(\cdot)$

$$P_S(\mathbf{y}) = \operatorname{argmin}\{\|\mathbf{y} - \mathbf{x}\|^2 : \mathbf{y} \in S\}.$$

Why Study Orthogonal Projections?

$$\mathrm{P}_{C_s \cap B}(\mathbf{x}) = \operatorname{argmin}\left\{\|\mathbf{z}-\mathbf{x}\|_2^2 : \mathbf{z} \in C_s \cap B\right\}$$

To define optimality conditions, we need to

- compute and analyze properties of the orthogonal projection $P_{C_s \cap B}$.

Computing $P_{C_s \cap B}$ is in general a difficult task, but in fact tractable under symmetry assumptions on B

Why Study Orthogonal Projections?

$$\mathrm{P}_{C_s \cap B}(\mathbf{x}) = \operatorname{argmin}\left\{\|\mathbf{z}-\mathbf{x}\|_2^2 : \mathbf{z} \in C_s \cap B\right\}$$

To define optimality conditions, we need to

- compute and analyze properties of the orthogonal projection $P_{C_s \cap B}$.

Computing $P_{C_s \cap B}$ is in general a difficult task, but in fact tractable under symmetry assumptions on B

Revised Layout:

Projections, Optimality Conditions, Algorithms

Projection Onto Symmetric Sets

Definitions - Basics

Σ_n = permutation group of $[n]$

$\mathbf{x}^\sigma$ = reordering of $\mathbf{x}$ according to $\sigma \in \Sigma_n$,

$$(\mathbf{x}^\sigma)_i = x_{\sigma(i)}.$$

Example (permutation)

$\mathbf{x} = \begin{pmatrix} 5 & 4 & 6 \end{pmatrix}^T$, and

$$\sigma(1) = 3, \sigma(2) = 1, \sigma(3) = 2,$$

then

$$\mathbf{x}^\sigma = \begin{pmatrix} 6 & 5 & 4 \end{pmatrix}^T.$$

Definitions - sorting permutations

- $\sigma \in \Sigma_n$ is a **sorting permutation** of $\mathbf{x}$ if

$$x_{\sigma(1)} \geq x_{\sigma(2)} \geq \cdots \geq x_{\sigma(n-1)} \geq x_{\sigma(n)}$$

- $\tilde{\Sigma}(\mathbf{x})$ is the set of all the sorting permutations of $\mathbf{x}$

Example (sorting permutation)

$\mathbf{x} = \begin{pmatrix} 7 & 9 & 8 & 9 \end{pmatrix}^T$, and

$$\sigma(1) = 2, \sigma(2) = 4, \sigma(3) = 3, \sigma(4) = 1,$$

then $\mathbf{x}^\sigma = \begin{pmatrix} 9 & 9 & 8 & 7 \end{pmatrix}^T$

Also $\tilde{\sigma} \in \tilde{\Sigma}(\mathbf{x})$ where $\tilde{\sigma}(1) = 4, \tilde{\sigma}(2) = 2, \tilde{\sigma}(3) = 3, \tilde{\sigma}(4) = 1$.

Definitions - Type-1 Symmetric Set

- D is a **type-1 symmetric set** if

$$\mathbf{x} \in D \Rightarrow \mathbf{x}^{\sigma} \in D$$

$\sigma \in \Sigma_n$

set	description	type-1	nonneg. type-1	type-2
Δ_n'[1]	unit sum	✓		
$[\ell, u]^n (\ell < u)$	box	✓		

[1] $\Delta_n' = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{1}^T\mathbf{x} = 1\}$

Definitions - Nonnegative Type-1 Symmetric Set

- D is **nonnegative** if $\forall \mathbf{x} \in D$, $\mathbf{x} \geq \mathbf{0}$

set	description	type-1	nonneg. type-1	type-2
$\mathbb{R}^n_+$	nonnegative orthant	✓	✓	
Δ_n	unit simplex	✓	✓	

Definitions - Type-2 Symmetric Set

- D is a **type-2 symmetric set** if it is type-1 symmetric and

$$\mathbf{x} \in D, \mathbf{y} \in \{-1,1\}^n \Rightarrow \mathbf{x} \circ \mathbf{y} \equiv (x_i y_i)_{i=1}^n \in D$$

set	description	type-1	nonneg. type-1	type-2
$\mathbb{R}^n$	entire space	✓		✓
$B_p[0,1](p \geq 1)$	p-ball	✓		✓

Summary of symmetry properties of simple sets

set	desc.	type-1	non. t-1	type-2
$\mathbb{R}^n$	entire space	✓		✓
$\mathbb{R}^n_+$	nonnegative orthant	✓	✓	
Δ_n	unit simplex	✓	✓	
Δ_n'	unit sum	✓		
$B_p[0,1](p \geq 1)$	p-ball	✓		✓
$[\ell, u]^n(\ell < u)$	box	✓		

Symmetric Projection Monotonicity Lemma

Symmetric Projection Monotonicity Lemma. Let D be a type-1 symmetric set, $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{y} \in \mathrm{P}_D(\mathbf{x})$. Then

$$(y_i - y_j)(x_i - x_j) \geq 0$$

for any $i, j \in [n]$.

Order Preservation Property

Theorem. Let D be a type-1 symmetric set, and $\sigma \in \tilde{\Sigma}(\mathbf{x})$. Suppose that $P_D(\mathbf{x}) \neq \emptyset$.
Then

$$\exists \mathbf{y} \in P_D(\mathbf{x}) \text{ s.t. } \sigma \in \tilde{\Sigma}(\mathbf{y})$$

That is

$$\begin{aligned} \mathbf{x}_{\sigma(1)} &\geq \mathbf{x}_{\sigma(2)} \geq \cdots \geq \mathbf{x}_{\sigma(n)} \\ \mathbf{y}_{\sigma(1)} &\geq \mathbf{y}_{\sigma(2)} \geq \cdots \geq \mathbf{y}_{\sigma(n)} \end{aligned}$$

Example: $D = C_2, P_D((3,2,2,0)) = \{(3,2,0,0),(3,0,2,0)\}$.

Sparse Projection Onto Symmetric Sets

The Sparse Projection Problem

The sparse projection problem

Find an element in orthogonal projection of $\mathbf{x} \in \mathbb{R}^n$ onto $B \cap C_s$:

$$P_{C_s \cap B}(\mathbf{x}) = \operatorname{argmin} \left\{ \|\mathbf{z} - \mathbf{x}\|_2^2 : \mathbf{z} \in C_s \cap B \right\}$$

1. $B \cap C_s$ closed $\Rightarrow P_{C_s \cap B}(\mathbf{x}) \neq \emptyset$
2. $B \cap C_s$ nonconvex $\Rightarrow |P_{C_s \cap B}(\mathbf{x})| \geq 1$

A DIFFICULT NONCONVEX PROBLEM IN GENERAL

(known for $B = \mathbb{R}^n, \mathbb{R}^n_+, \Delta_n$)

Supports, Super Supports

Let $\mathbf{x} \in \mathbb{R}^n$, $s \in [n] = \{1, \ldots, n\}$.

1. **Support** of $\mathbf{x}$: $I_1(\mathbf{x}) \equiv \{i \in [n] \ : \ x_i \neq 0\}$.
2. **Super support** of $\mathbf{x}$: any set T s.t. $I_1(\mathbf{x}) \subseteq T$ and $|T| = s$.
3. $\mathbf{x}$ has **full support** if $\|\mathbf{x}\|_0 = |I_1(\mathbf{x})| = s$.
4. **Off-support** of $\mathbf{x}$: $I_0(\mathbf{x}) \equiv \{i \in [n] \ : \ x_i = 0\}$.

Example

$s = 3, n = 5$ and $\mathbf{x} = (-3, 4, 0, 0, 0)^T$

1. **Support**: $I_1(\mathbf{x}) = \{1, 2\}$
2. **Super support**: $T \in \{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}\}$
3. **Incomplete support**: $\|\mathbf{x}\|_0 < s$
4. **Off-support**: $I_0(\mathbf{x}) = \{3, 4, 5\}$

Restriction on Index Sets

$\mathbf{x} \in \mathbb{R}^n$, $T \subseteq [n]$ index set

1. $\mathbf{x}_T \in \mathbb{R}^{|T|}$ is the restriction of $\mathbf{x}$ to T
2. $\mathbf{U}_T$ is the submatrix of $\mathbf{I}_n$ constructed from the columns in T
3. $B_T = \{\mathbf{x} \in \mathbb{R}^{|T|} : \mathbf{U}_T\mathbf{x} \in B\}$ is **the restriction of B to T**
4. $\nabla_T f(\mathbf{x}) = \mathbf{U}_T^T \nabla f(\mathbf{x})$ is the restriction of $\nabla f(\mathbf{x})$ to T.

Example

$$\mathbf{x} = (8,7,6,5)^T \Rightarrow \mathbf{x}_{1,3} = (8,6)^T.$$

$$B = \{(x_1,x_2,x_3,x_4) : x_1 + 2x_2 + 3x_3 + 4x_4 = 1\} \Rightarrow B_{1,2} = \{(x_1,x_2)^T : x_1 + 2x_2 = 1\}.$$

$$f(\mathbf{x}) = x_1x_2 + x_2^2 + x_3^3 \Rightarrow \nabla_{\{1,3\}} f(\mathbf{x}) = (x_2, 3x_3^2)^T.$$

The Order Set

order set $S^{\sigma}_{[j_1,j_2]}$ For any permutation $\sigma \in \Sigma_n$, we define $S^{\sigma}_{[j_1,j_2]}$ as:

$$S^{\sigma}_{[j_1,j_2]} = \begin{cases} \{\sigma(j_1), \sigma(j_1+1), \ldots, \sigma(j_2)\} & 0 < j_1 \leq j_2 \leq n, \\ \emptyset & \text{otherwise.} \end{cases}$$

Example (order set)

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 3 & 2 \end{pmatrix}$$

$$S^{\sigma}_{[1,2]} = \{\sigma(1), \sigma(2)\} = \{4, 1\}, \; S^{\sigma}_{[3,4]} = \{\sigma(3), \sigma(4)\} = \{3, 2\}$$

Phases in Computing the Projection

To find $\mathbf{y} \in P_{C_s \cap B}(\mathbf{x})$:

(1) find its super support S

(2) Compute $\mathbf{y}_S = P_{B_S}(\mathbf{x}_S),\ \mathbf{y}_{S^c} = \mathbf{0}$

Naive approach: go over all possible $\binom{n}{s}$ super supports, compute the corresponding projections, and find the sparse projection vector. TOO EXPENSIVE

Type-1 Symmetric Sparse Projection Theorem

Type-1 Symmetric Projection Theorem B be a type-1 symmetric set, $\sigma \in \tilde{\Sigma}(\mathbf{x})$. Then $\exists \mathbf{y} \in P_{C_s \cap B}(\mathbf{x})$, $k \in \{0, \ldots, s\}$ for which

$$I_1(\mathbf{y}) \subseteq S^{\sigma}_{[1,k]} \cup S^{\sigma}_{[n+k-(s-1),n]}$$

Result: to find the super support, find $k \in \{0, \ldots, s\}$:

$$\begin{array}{ccccccccccc}
\mathbf{x}_{\sigma(1)} & \geq \cdots \geq & \mathbf{x}_{\sigma(k)} & \geq & \mathbf{x}_{\sigma(k+1)} & \geq & \cdots & \geq & \mathbf{x}_{\sigma(n+k-s)} & \geq \mathbf{x}_{\sigma(n+k-(s-1))} & \geq \cdots \geq \mathbf{x}_{\sigma(n)} \\
\mathbf{y}_{\sigma(1)} & \geq \cdots \geq & \mathbf{y}_{\sigma(k)} & \geq & 0 & \geq & \cdots & \geq & 0 & \geq \mathbf{y}_{\sigma(n+k-(s-1))} & \geq \cdots \geq \mathbf{y}_{\sigma(n)}
\end{array}$$

Type-1 Symmetric Sparse Projection Algorithm

Algorithm 1 Projection onto a type-1 symmetric sparse set

Input: $\mathbf{x} \in \mathbb{R}^n$. **Output:** $\mathbf{u} \in \mathrm{P}_{C_s \cap B}(\mathbf{x})$.

1. Find $\sigma \in \tilde{\Sigma}(\mathbf{x})$.
2. for any $k = s, s-1, \ldots, 0$ do:
   1. Set $T_k = S^{\sigma}_{[1,k]} \cup S^{\sigma}_{[n+k-(s-1),n]}$.
   2. Compute $\mathbf{g}_k = \mathrm{P}_{B_{T_k}}(\mathbf{x}_{T_k})$ and define $\mathbf{z}^k = \mathbf{U}_{T_k}\mathbf{g}_k$.
3. Return $\mathbf{u} = \operatorname{argmin}\{\|\mathbf{z} - \mathbf{x}\|^2 : \mathbf{z} \in \{\mathbf{z}^k : k = s, s-1, \ldots, 0\}\}$

Nonnegative Type-1 Symmetric Sparse Projection Theorem

Nonnegative Type-1 Symmetric Sparse Projection Theorem

B nonnegative type-1 symmetric, and $\sigma \in \tilde{\Sigma}(\mathbf{x})$. Then

$$\exists \mathbf{y} \in P_{C_s \cap B}(\mathbf{x}) \text{ s.t } I_1(\mathbf{y}) \subseteq S^{\sigma}_{[1,s]}$$

Example

$B = \Delta_4$, $s = 2$, $x = (1, 0.5, -0.5, -1)^T$, $S^{\sigma}_{[1,2]} = \{1, 2\}$,
$y = (0.75, 0.25, 0, 0)^T$.

The super support can be found by a simple sort operation

Type-2 Symmetric Sparse Projection Theorem

Type-2 Symmetric Sparse Projection Theorem B type-2 symmetric, and $\sigma \in \tilde{\Sigma}(|\mathbf{x}|)$. Then

$$\exists \mathbf{y} \in P_{C_s \cap B}(\mathbf{x}) \text{ s.t } I_1(\mathbf{y}) \subseteq S^{\sigma}_{[1,s]}$$

Example

$B = B_2[\mathbf{0}, 1]$, $s = 2$, $x = (3, 0.5, -0.7, -4)^T$, $S^{\sigma}_{[1,2]} = \{4, 1\}$, $y = (0.6, 0, 0, -0.8)^T$.

The super support can be found by a simple sort operation

Unified Symmetric Projection Theorem

symmetry function The symmetry function $p : \mathbb{R}^n \to \mathbb{R}^n$:

$$p(\mathbf{x}) \equiv \begin{cases} \mathbf{x} & B \text{ is nonnegative type-1,} \\ |\mathbf{x}| & B \text{ is type-2 symmetric.} \end{cases}$$

Unified Symmetric Projection Theorem B be closed, convex, and a nonnegative type-1 or a type-2 symmetric. $\sigma \in \tilde{\Sigma}(p(\mathbf{x}))$.

Then

$$\exists \mathbf{y} \in P_{C_s \cap B}(\mathbf{x}) \text{ s.t. } I_1(\mathbf{y}) \subseteq S^{\sigma}_{[1,s]}$$

Unified Symmetric Sparse Projection Algorithm

Algorithm 2 Unified symmetric sparse projection algorithm

Input: $\mathbf{x} \in \mathbb{R}^n$.

Output: $\mathbf{u} \in P_{B \cap C_s}(\mathbf{x})$.

1. Compute $T = S^{\sigma}_{[1,s]}$ for $\sigma \in \tilde{\Sigma}(p(\mathbf{x}))$.
2. Return $\mathbf{u} = \mathbf{U}_T P_{B_T}(\mathbf{x}_T)$.

Super supports for sparse projection onto simple sets

B	$P_{B\cap C_s}(\mathbf{x})$	**super support set(s)**	B_T
$\mathbb{R}^n$	$\mathbf{U}_T\mathbf{x}_T$	$T = S^{\sigma}_{[1,s]}, \sigma \in \tilde{\Sigma}(\lvert\mathbf{x}\rvert)$	$B_T = \mathbb{R}^s$
$\mathbb{R}^n_+$	$\mathbf{U}_T[\mathbf{x}_T]_+$	$T = S^{\sigma}_{[1,s]}, \sigma \in \tilde{\Sigma}(\mathbf{x})$	$B_T = \mathbb{R}^s_+$
Δ_n	$\mathbf{U}_T P_{B_T}(\mathbf{x}_T)$	$T = S^{\sigma}_{[1,s]}, \sigma \in \tilde{\Sigma}(\mathbf{x})$	$B_T = \Delta_s$
$\Delta_n^{'}$	$\mathbf{U}_{T_k} P_{B_{T_k}}(\mathbf{x}_{T_k})$	$T_k = S^{\sigma}_{[1,k]} \cup S^{\sigma}_{[n+k-(s-1),n]}$ $k = 0,1,\ldots,s,\ \sigma \in \tilde{\Sigma}(\mathbf{x})$	$B_{T_k} = \Delta'_s$
$B_p^n[\mathbf{0},1]$ $(p \geq 1)$	$\mathbf{U}_T P_{B_T}(\mathbf{x}_T)$	$T = S^{\sigma}_{[1,s]}, \sigma \in \tilde{\Sigma}(\lvert\mathbf{x}\rvert)$	$B_T = B_p^s[\mathbf{0},1]$
$[\ell,u]^n$ $(\ell < u)$	$\mathbf{U}_{T_k} P_{B_{T_k}}(\mathbf{x}_{T_k})$	$T_k = S^{\sigma}_{[1,k]} \cup S^{\sigma}_{[n+k-(s-1),n]}$ $k = 0,1,\ldots,s,\ \sigma \in \tilde{\Sigma}(\mathbf{x})$	$B_{T_k} = [\ell,u]^s$

Optimality Conditions and Algorithms

Back to the Sparse Optimization Problem

The sparse optimization problem

$$(P) \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in C_s \cap B, \end{array}$$

- $C_s = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_0 \leq s\}$

Assumption

[A] $f : \mathbb{R}^n \to \mathbb{R}$ *is lower bounded, continuously differentiable.*

[B] B *is a closed and convex set.*

In some cases

[C] $f \in C^{1,1}_{L(f)}$.

Road Map of Optimality Conditions

$$(P) \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in C_s \cap B, \end{array}$$

- Basic Feasibility - "optimality" over the support.
- L-Stationarity - extension of stationarity over convex sets.
- CW-optimality

Basic Feasibility (BF)

- $\mathbf{x} \in C_s \cap B$ is a **basic feasible (BF) point** of (P) if for any super support set S of $\mathbf{x}$, for some $L > 0$:

$$\mathbf{x}_S = P_{B_S}\left(\mathbf{x}_S - \frac{1}{L}\nabla_S f(\mathbf{x})\right).$$

Remarks:

(a) If $|I_1(\mathbf{x})| = s$, then the only super support set is the support itself.

(b) If $|I_1(\mathbf{x})| = k < s$, then there are $\binom{n-k}{s-k}$ possible super supports.

(c) $\mathbf{x}$ is a **BF** $\Leftrightarrow$ $\mathbf{x}_T$ is stationary point of f over B_T for any super support T of $\mathbf{x}$

(d) Optimality $\Rightarrow$ BF

Basic Feasibility - Examples

B	BF conditions (full support)
$\mathbb{R}^n$	$\nabla_{I_1(\mathbf{x}^*)} f(\mathbf{x}^*) = \mathbf{0}$
$\mathbb{R}^n_+$	$\nabla_{I_1(\mathbf{x}^*)} f(\mathbf{x}^*) = \mathbf{0}$
Δ_n	$\exists \mu \in \mathbb{R} : \nabla_i f(\mathbf{x}^*) = \mu, i \in I_1(\mathbf{x}^*)$
Δ'_n	$\exists \mu \in \mathbb{R} : \nabla_i f(\mathbf{x}^*) = \mu, i \in I_1(\mathbf{x}^*)$
$B_2^n[\mathbf{0}, 1]$	$\nabla_{I_1(\mathbf{x}^*)} f(\mathbf{x}^*) = \mathbf{0}$ or $\|\mathbf{x}^*\| = 1$ and $\exists \lambda \leq 0$: $\nabla_{I_1(\mathbf{x}^*)} f(\mathbf{x}^*) = \lambda \mathbf{x}^*_{I_1(\mathbf{x}^*)}$
$[\ell, u]^n (\ell < u)$	$\frac{\partial f}{\partial x_i}(\mathbf{x}^*) \begin{cases} = 0 & \ell < x_i < u \\ \geq 0 & x_i = \ell \\ \leq 0 & x_i = u_i \end{cases}, i \in I_1(\mathbf{x}^*)$

What if $|I_1(\mathbf{x})| < s$?

Basic Feasibility - Examples

B	BF conditions (full support)
$\mathbb{R}^n$	$\nabla_{I_1(\mathbf{x}^*)} f(\mathbf{x}^*) = \mathbf{0}$
$\mathbb{R}^n_+$	$\nabla_{I_1(\mathbf{x}^*)} f(\mathbf{x}^*) = \mathbf{0}$
Δ_n	$\exists \mu \in \mathbb{R} : \nabla_i f(\mathbf{x}^*) = \mu, i \in I_1(\mathbf{x}^*)$
Δ'_n	$\exists \mu \in \mathbb{R} : \nabla_i f(\mathbf{x}^*) = \mu, i \in I_1(\mathbf{x}^*)$
$B_2^n[\mathbf{0}, 1]$	$\nabla_{I_1(\mathbf{x}^*)} f(\mathbf{x}^*) = \mathbf{0}$ or $\|\mathbf{x}^*\| = 1$ and $\exists \lambda \leq 0$: $\nabla_{I_1(\mathbf{x}^*)} f(\mathbf{x}^*) = \lambda \mathbf{x}^*_{I_1(\mathbf{x}^*)}$
$[\ell, u]^n (\ell < u)$	$\frac{\partial f}{\partial x_i}(\mathbf{x}^*) \begin{cases} = 0 & \ell < x_i < u \\ \geq 0 & x_i = \ell \\ \leq 0 & x_i = u_i \end{cases}, i \in I_1(\mathbf{x}^*)$

What if $|I_1(\mathbf{x})| < s$? In all of the above cases, basic feasibility is exactly like stationarity over B

Characterization of BF points

Theorem. $\mathbf{x} \in C_s \cap B$ is a BF point if and only if

$$\mathbf{x}_T = \mathrm{P}_{B_T}\left(\mathbf{x}_T - \frac{1}{L}\nabla_T f(\mathbf{x})\right),$$

where

1. **Symmetry**: B is non-negative type-1 or type-2 symmetric
2. **Fill the support**: $i \in [n+1]$ s.t.
   $\left|S^{\sigma}_{[i,n]} \cup I_1(\mathbf{x})\right| = s \quad (\sigma \in \tilde{\Sigma}(-p(-\nabla f(\mathbf{x}))))$
3. **Into super support**: $T = I_1(\mathbf{x}) \cup S^{\sigma}_{[i,n]}$

Important: when $|I_1(\mathbf{x})| < s$ only one super support needs to be checked

Basic Feasible Search - Find a BF point

Algorithm 3 BAsic Feasible Search (BAFS)

Initialization: $\mathbf{x}^0 \in C_s \cap B, k = 0$.

Output: $\mathbf{u} \in C_s \cap B$ which is a basic feasible point.

1. Repeat
   1. $k \leftarrow k + 1$
   2. let $\sigma \in \tilde{\Sigma}\left(-p\left(-\nabla f(\mathbf{x}^k)\right)\right)$
   3. set $i \in \{1, \ldots, n+1\}$ such that $\left|S^{\sigma}_{[i,n]} \cup I_1(\mathbf{x}^k)\right| = s$
   4. set $T_k = I_1(\mathbf{x}^k) \cup S^{\sigma}_{[i,n]}$
   5. take $\mathbf{x}^k \in \operatorname{argmin}\{f(\mathbf{y}) : \mathbf{y} \in B, I_1(\mathbf{y}) \subseteq T_k\}$

   Until $f(\mathbf{x}^{k-1}) \leq f(\mathbf{x}^k)$
2. Set $\mathbf{u} = \mathbf{x}^{k-1}$

Finite Termination.

- requires the ability to minimize over the support set.

Road Map of Optimality Conditions

$$(P) \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in C_s \cap B, \end{array}$$

- Basic Feasibility - "optimality" over the support.
- L-Stationarity - **extension of stationarity over convex sets.**
- CW-optimality

L-Stationarity

Unfortunately, the variational form $\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \forall \mathbf{x} \in B \cap C_s$ is not a necessary optimality condition (in general...)

L-Stationarity

Unfortunately, the variational form $\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \forall \mathbf{x} \in B \cap C_s$ is not a necessary optimality condition (in general...)

Let $L > 0$. A vector $\mathbf{x} \in C_s \cap B$ is an **L-stationary point** of (P) if

$$\mathbf{x} \in P_{C_s \cap B}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right).$$

L-Stationarity in the Hierarchy:

1. L-Stationarity $\Rightarrow$ BF
2. If $f \in C^{1,1}_{L(f)}$, Optimality $\Rightarrow$ L-stationarity $\forall L > L(f)$

Condition depends on L, more restrictive as L gets smaller

Gradient Projection Method

$$\mathbf{x}^{k+1} \in P_{C_s \cap B}\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right)$$

- $B = \mathbb{R}^n \Rightarrow$ Iterative Hard Thresholding (IHT) method (Blumensath and Davis '08, '09, '12).
- Makes sense only when $f \in C^{1,1}$.
- Only guarantees convergence to an L-stationary point for $L > L(f)$.

L-stationarity characterization

Theorem. Let B be a nonnegative type-1 or type-2 symmetric set. Then a BF point $\mathbf{x}^*$ is an L-stationary point if and only if

$$\min_{i \in I_1(\mathbf{x}^*)} p(Lx_i^* - \nabla_i f(\mathbf{x}^*)) \geq \max_{j \in I_0(\mathbf{x}^*)} p(Lx_j^* - \nabla_j f(\mathbf{x}^*)),$$

Example ($B = \mathbb{R}^n$)

$B = \mathbb{R}^n$, and $\sigma \in \tilde{\Sigma}(|\mathbf{x}^*|)$. Then $\mathbf{x}^*$ is an L-stationary point of (P) if and only if[a]

$$|\nabla_i f(\mathbf{x}^*)| \begin{cases} \leq L|x^*_{\sigma(s)}| & \text{if } i \in I_0(\mathbf{x}^*), \\ = 0 & \text{if } i \in I_1(\mathbf{x}^*). \end{cases}$$

[a] Beck, A. & Eldar, Y. C., SIOPT, 2013

Back to the Variational Form of Stationarity

- In general, the variational form is not a necessary optimality conditions.
- However, when f is concave, it is in fact a necessary optimality condition.

Back to the Variational Form of Stationarity

- In general, the variational form is not a necessary optimality conditions.
- However, when f is concave, it is in fact a necessary optimality condition.

Theorem. Suppose that f is concave and cont. diff.. If $\mathbf{x}^*$ is an optimal solution of (P), then

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \quad \forall \mathbf{x} \in B \cap C_s.$$

(a direct consequence of Krein-Milman+attainment of opt. sol. at extreme points)

Back to the Variational Form of Stationarity

- In general, the variational form is not a necessary optimality conditions.
- However, when f is concave, it is in fact a necessary optimality condition.

Theorem. Suppose that f is concave and cont. diff.. If $\mathbf{x}^*$ is an optimal solution of (P), then

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \quad \forall \mathbf{x} \in B \cap C_s.$$

(a direct consequence of Krein-Milman+attainment of opt. sol. at extreme points)

We will call this type of stationarity co-stationarity

co-stationarity $\Rightarrow$ L-stationarity

Theorem. For any $L > 0$:

$$\mathbf{x}^* \text{ co-stationary } \Rightarrow \mathbf{x}^* \text{ is } L - \text{stationary}$$

co-stationarity $\Rightarrow$ L-stationarity

Theorem. For any $L > 0$:

$$\mathbf{x}^* \text{ co-stationary } \Rightarrow \mathbf{x}^* \text{ is } L - \text{stationary}$$

Consequence: for concave f, L-stationarity is a necessary optimality condition for any $L > 0$.

co-stationarity $\Rightarrow$ L-stationarity

Theorem. For any $L > 0$:

$$\mathbf{x}^* \text{ co-stationary } \Rightarrow \mathbf{x}^* \text{ is } L - \text{stationary}$$

Consequence: for concave f, L-stationarity is a necessary optimality condition for any $L > 0$.

sparse PCA $\max\{\mathbf{x}^T\mathbf{A}\mathbf{x} : \|\mathbf{x}\|^2 \leq 1, \|\mathbf{x}\|_0 \leq s\}$ $(\mathbf{A} \succeq \mathbf{0})$

- Most algorithms converge to a co-stationary point, e.g., conditional gradient method:

$$\mathbf{x}^{k+1} = \frac{\mathbf{y}^k}{\|\mathbf{y}^k\|_2}, \mathbf{y}^k = H_s(\mathbf{A}\mathbf{x}^k).$$

- Gradient projection is never employed (since L-stationarity is a weak condition).
- The above is only correct for concave f.

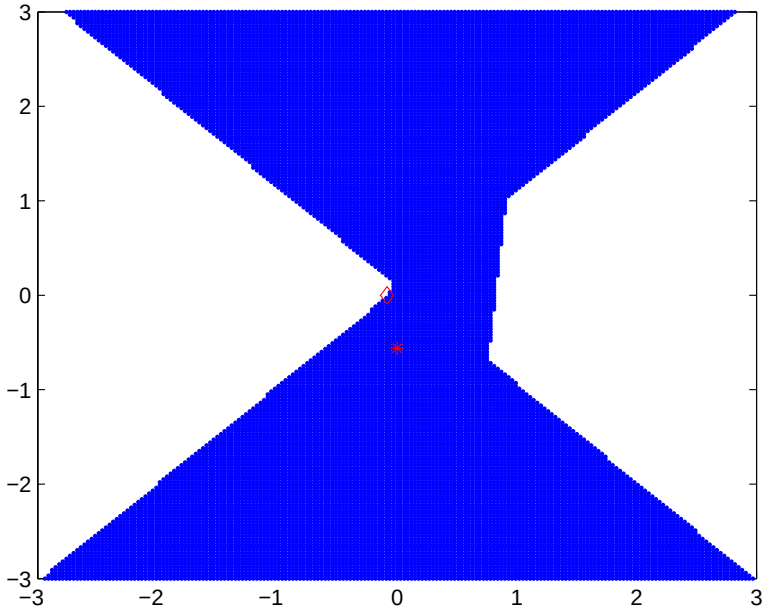
Back to L-stationarity - Example

$$\min\left\{f(x_1,x_2)\equiv 12x_1^2+20x_1x_2+32x_2^2 : \left\|(x_1;x_2)^T\right\|_0\leq 1\right\}$$

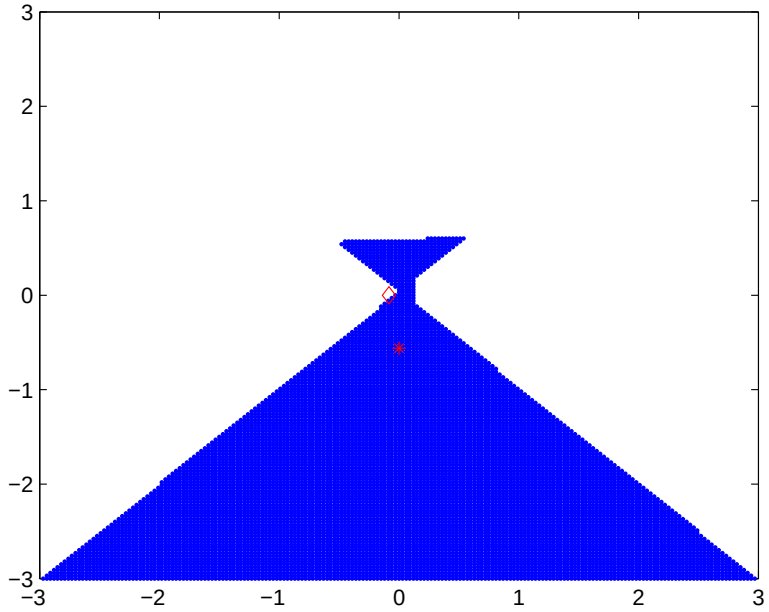
$$L(f)=48.3961$$

Two BF vectors: $(0,-9/16)$ - optimal solution. $(-1/12,0)$ - non-optimal, SL=196.

$L=250$



$L=500$



Main Questions

- Are there stronger (more restrictive) optimality conditions?
- Can we define algorithms that (1) do not depend on the Lipschitz property and (2) converge to "better" optimality conditions?

Main Questions

- Are there stronger (more restrictive) optimality conditions?
- Can we define algorithms that (1) do not depend on the Lipschitz property and (2) converge to "better" optimality conditions?

The answer is YES

Road Map of Optimality Conditions

$$(P) \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in C_s \cap B, \end{array}$$

- Basic Feasibility - "optimality" over the support.
- L-Stationarity - extension of stationarity over convex sets.
- **CW-optimality**

Simple-CW optimality

Let B be nonnegative type-1 or type-2 symmetric.
A BF point $\mathbf{x}$ is a **simple-CW point** of (P) if

$$f(\mathbf{x}) \leq \begin{cases} f(\mathbf{x} - x_i e_i \pm x_i e_j) & \text{B type-2} \\ f(\mathbf{x} - x_i e_i + x_i e_j) & \text{B nonneg. type-1} \end{cases}$$

where

$$\begin{aligned} i &\in \operatorname*{argmin}_{\ell \in D(\mathbf{x})} \{p(-\nabla_\ell f(\mathbf{x}))\} \text{ with } D(\mathbf{x}) = \operatorname*{argmin}_{k \in I_1(\mathbf{x})} p(x_k) \\ j &\in \operatorname*{argmin}_{\ell \in I_0(\mathbf{x})} \{-p(-\nabla_\ell f(\mathbf{x}))\} \end{aligned}$$

Simple-CW optimality in the Hierarchy

Results: If B is a non-negative type-1 set or a type-2 symmetric

1. Optimality $\Rightarrow$ Simple-CW
2. If $f \in C^{1,1}_{L(f)}$, then Simple-CW $\Rightarrow$ $L_2(f)$-stationarity $\forall L \geq L_2(f)$

$L_2(f)$ - Lipschitz constant of ∇f restricted to two coordinates. Smaller than $L(f)$

Simple-CW is more restrictive than $L(f)$-stationarity

Local Lipschitz Constant

Under the Lipschitz assumption,

- For any $i \neq j$ there exists a $L_{i,j}(f)$ for which:

$$\|\nabla_{i,j} f(\mathbf{x}) - \nabla_{i,j} f(\mathbf{x}+\mathbf{d})\| \leq L_{i,j}(f)\|\mathbf{d}\|,$$

for any $\mathbf{d} \in \mathbb{R}^n$ satisfying $d_k = 0$ for any $k \neq i, j$.

Local Lipschitz Constant

Under the Lipschitz assumption,

- For any $i \neq j$ there exists a $L_{i,j}(f)$ for which:

$$\|\nabla_{i,j} f(\mathbf{x}) - \nabla_{i,j} f(\mathbf{x}+\mathbf{d})\| \leq L_{i,j}(f)\|\mathbf{d}\|,$$

  for any $\mathbf{d} \in \mathbb{R}^n$ satisfying $d_k = 0$ for any $k \neq i, j$.

- the Local Lipschitz constant is

$$L_2(f) = \max_{i \neq j} L_{i,j}(f)$$

Local Lipschitz Constant

Under the Lipschitz assumption,

- For any $i \neq j$ there exists a $L_{i,j}(f)$ for which:

$$\|\nabla_{i,j} f(\mathbf{x}) - \nabla_{i,j} f(\mathbf{x}+\mathbf{d})\| \leq L_{i,j}(f)\|\mathbf{d}\|,$$

  for any $\mathbf{d} \in \mathbb{R}^n$ satisfying $d_k = 0$ for any $k \neq i, j$.

- the Local Lipschitz constant is

$$L_2(f) = \max_{i \neq j} L_{i,j}(f)$$

- $L_2(f) \leq L(f)$.

Example: $f(\mathbf{x}) = \mathbf{x}^T\mathbf{Q}\mathbf{x} + 2\mathbf{b}^T\mathbf{x}$ with $\mathbf{Q}_n = \mathbf{I}_n + \mathbf{J}_n$ ($\mathbf{I}_n$ - identity, $\mathbf{J}_n$ - all ones)

Local Lipschitz Constant

Under the Lipschitz assumption,

- For any $i \neq j$ there exists a $L_{i,j}(f)$ for which:

$$\|\nabla_{i,j} f(\mathbf{x}) - \nabla_{i,j} f(\mathbf{x}+\mathbf{d})\| \leq L_{i,j}(f)\|\mathbf{d}\|,$$

  for any $\mathbf{d} \in \mathbb{R}^n$ satisfying $d_k = 0$ for any $k \neq i, j$.

- the Local Lipschitz constant is

$$L_2(f) = \max_{i \neq j} L_{i,j}(f)$$

- $L_2(f) \leq L(f)$.

Example: $f(\mathbf{x}) = \mathbf{x}^T\mathbf{Q}\mathbf{x} + 2\mathbf{b}^T\mathbf{x}$ with $\mathbf{Q}_n = \mathbf{I}_n + \mathbf{J}_n$ ($\mathbf{I}_n$ - identity, $\mathbf{J}_n$ - all ones)

$$L(f) = 2\lambda_{\max}(\mathbf{Q}_n) = 2(n+1)$$

On the other hand, $L_{i,j}(f) = 2\lambda_{\max}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = 6$

Local Lipschitz Constant

Under the Lipschitz assumption,

- For any $i \neq j$ there exists a $L_{i,j}(f)$ for which:

$$\|\nabla_{i,j} f(\mathbf{x}) - \nabla_{i,j} f(\mathbf{x}+\mathbf{d})\| \leq L_{i,j}(f)\|\mathbf{d}\|,$$

  for any $\mathbf{d} \in \mathbb{R}^n$ satisfying $d_k = 0$ for any $k \neq i, j$.

- the Local Lipschitz constant is

$$L_2(f) = \max_{i \neq j} L_{i,j}(f)$$

- $L_2(f) \leq L(f)$.

Example: $f(\mathbf{x}) = \mathbf{x}^T\mathbf{Q}\mathbf{x} + 2\mathbf{b}^T\mathbf{x}$ with $\mathbf{Q}_n = \mathbf{I}_n + \mathbf{J}_n$ ($\mathbf{I}_n$ - identity, $\mathbf{J}_n$ - all ones)

$$L(f) = 2\lambda_{\max}(\mathbf{Q}_n) = 2(n+1)$$

On the other hand, $L_{i,j}(f) = 2\lambda_{\max}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = 6$

We get: $L(f) = 2(n+1), L_2(f) = 6$

Zero-CW Optimality

Assumption: The function can be minimized over any super support, B nonnegative type-1 or type-2

$\mathbf{x}$ is called a **zero-CW optimal point** if

$$f(\mathbf{x}) \leq \min \{f(\mathbf{y}) : \mathbf{y} \in B, I_1(\mathbf{y}) \subseteq T\},$$

where $T = (I_1(\mathbf{x}) \cup \{j\}) \setminus \{i\}$ if $\|\mathbf{x}\|_0 = s$, otherwise additional (specific) indices are added

Full-CW Optimality

Assumption: The function can be minimized over any super support, B nonnegative type-1 or type-2

$\mathbf{x}$ is a **full-CW optimal point** if $\forall i \in I_1(\mathbf{x}), j \in I_0(\mathbf{x})$, it holds that

$$f(\mathbf{x}) \leq \min \{f(\mathbf{y}) : \mathbf{y} \in B, I_1(\mathbf{y}) \subseteq T_{i,j}\}.$$

In the non full-support case, additional indices are added to $T_{i,j}$.

Hierarchy - Summary

Full-CW

$\Downarrow$

Zero-CW

$\Downarrow$

Simple-CW

$\Downarrow$

$L_2(f)$-Stationarity

$\Downarrow$

Basic Feasibility

Concave f

$\mathbf{x}$ is called CW-minimum point if $f(\mathbf{x}) \geq f(\mathbf{z})$ for any

$$S_2(\mathbf{x}) \equiv \{\mathbf{z} : \|\mathbf{z} - \mathbf{x}\|_0 \leq 2, \mathbf{z} \in C_s \cap B\}.$$

Concave f

$\mathbf{x}$ is called CW-minimum point if $f(\mathbf{x}) \geq f(\mathbf{z})$ for any

$$S_2(\mathbf{x}) \equiv \{\mathbf{z} : \|\mathbf{z} - \mathbf{x}\|_0 \leq 2, \mathbf{z} \in C_s \cap B\}.$$

CW-optimality is a necessary optimality conditions (of a combinatorial flavor...)

Theorem. If f is concave and $B = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\}$, then any CW-maximal point is a co-stationary point.

Quite difficult to prove since co-stationarity is a rather strong condition.

pit-prop data

13 variables measuring 180 properties of pitprops. 715 BF points, 28 co-stationary points, 3 CW-optimal points.

#	Support	CW	Value
1	{1,2,9,10}	*	2.937
2	{1,2,7,10}		2.883
3	{1,2,7,9}		2.859
4	{1,2,8,9}		2.797
5	{1,2,8,10}		2.759
6	{1,2,6,7}		2.697
7	{2,7,9,10}		2.696
8	{2,6,7,10}		2.592
9	{1,6,7,10}		2.587
10	{1,2,3,4}	*	2.563
11	{7,8,9,10}		2.549
12	{6,7,9,10}		2.522
13	{6,7,10,13}		2.459
14	{6,7,8,10}		2.444

#	Support	CW	Value
15	{5,6,7,10}		2.337
16	{7,8,10,12}		2.314
17	{7,8,10,13}		2.302
18	{5,6,7,13}		2.28
19	{3,4,6,7}		2.209
20	{4,5,6,7}		2.196
21	{7,10,12,13}		2.136
22	{3,4,8,12}		1.995
23	{3,4,10,12}		1.992
24	{3,10,11,12}		1.609
25	{3,5,12,13}		1.516
26	{1,5,12,13}		1.414
27	{2,5,12,13}		1.408
28	{3,5,11,13}		1.382

LS over sparse l_1-ball

$$\min\left\{\left\|\begin{pmatrix}1000 & 0 & 0 & 1\\ 0 & 1 & 0 & 1\\ 0 & 0 & 0.01 & 1\end{pmatrix}\mathbf{x}-\begin{pmatrix}3\\1\\9\end{pmatrix}\right\|_2^2 : \mathbf{x}\in C_2\cap B_1^4[\mathbf{0},1]\right\}.$$

support	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
values	0.003	0.003	0.002	0		
	0.997	0	0	0.910		
	0	0.997	0	0.090		
	0	0	0.998	0		
BF	✓	✓	✓	✓		
L(f)-stationary			✓	✓		
simple-CW			✓	✓		
zero-CW			✓			
full-CW			✓			

Zero-CW search method

Algorithm 4 Zero-CW search method (ZCWS)

Initialization: $\mathbf{x}^0 \in C_s \cap B$ a BF, $k = 0$.

Output: $\mathbf{u} \in C_s \cap B$ which is a zero-cw.

General Step ($k = 0, 1, 2, \ldots$)

1. $D(\mathbf{x}^k) = \underset{\ell \in I_1(\mathbf{x}^k)}{\operatorname{argmin}} \, p(x_\ell^k)$
2. $i \in \underset{\ell \in D(\mathbf{x}^k)}{\operatorname{argmin}} \{p(-\nabla_\ell f(\mathbf{x}^k))\}$
3. $j \in \underset{\ell \in I_0(\mathbf{x}^k)}{\operatorname{argmin}} \left\{-p(-\nabla_\ell f(\mathbf{x}^k))\right\}$

Zero-CW search method - Find a Zero-CW point

Algorithm 5 Zero-CW search method (ZCWS)

4. let $\sigma \in \tilde{\Sigma}(-p(-\nabla f(\mathbf{x}^k)))$ and let ℓ be such that

$$\left| \left(S^{\sigma}_{[\ell,n]} \cup I_1(\mathbf{x}^k) \cup \{j\} \right) \setminus \{i\} \right| = s$$

5. Define

$$T_k = \left(S^{\sigma}_{[\ell,n]} \cup I_1(\mathbf{x}^k) \cup \{j\} \right) \setminus \{i\}.$$

6. Set $\mathbf{x} \in \operatorname{argmin} \{ f(\mathbf{y}) : \mathbf{y} \in B, I_1(\mathbf{y}) \subseteq T_k \}$.
7. $\mathbf{x}^{k+1} = \mathrm{BFS}(\mathbf{x})$
8. If $f(\mathbf{x}^k) \leq f(\mathbf{x}^{k+1})$, then STOP and the output is $\mathbf{u} = \mathbf{x}^k$. Otherwise, $k \leftarrow k+1$ and go back to step 1.

ZCWS generates a points that is a zero-CW point, and this converges to "better" points than gradient projection/IHT.

Full-CW search method - Find a Full-CW point

Algorithm 6 Full-CW search method

Initialization: $\mathbf{x}^0 \in C_s \cap B$ - a basic feasible point, $k = 0$.

Output: $\mathbf{u} \in C_s \cap B$ which is a full-cw point.

General Step ($k = 0, 1, 2, \ldots$)

1. $\mathbf{x}^{k+1} = \mathrm{ZCWS}(\mathbf{x}^k)$ and set $k \leftarrow k+1$.
2. let $\sigma \in \tilde{\Sigma}(-p(-\nabla f(\mathbf{x}^k)))$, and for any $i \in I_1(\mathbf{x}^k)$, $j \in I_0(\mathbf{x}^k)$ let $\ell^{i,j}$ be such that

$$\left| \left(S^{\sigma}_{[\ell^{i,j},n]} \cup I_1(\mathbf{x}^k) \cup \{j\} \right) \setminus \{i\} \right| = s$$

3. define

$$T^{i,j}_k = \left(S^{\sigma}_{[\ell^{i,j},n]} \cup I_1(\mathbf{x}^k) \cup \{j\} \right) \setminus \{i\}$$

Full-CW search method

Algorithm 7 Full-CW search method

4. take $\mathbf{z}^{i,j} \in \operatorname{argmin}\left\{f(\mathbf{y}) : y \in B, I_1(y) \subseteq T_k^{i,j}\right\}$.
5. set $(i_0, j_0) \in \operatorname{argmin}\left\{f(\mathbf{z}^{i,j}) : i \in I_1(\mathbf{x}), j \in I_0(\mathbf{x})\right\}$
6. define $\mathbf{x}^{k+1} = \mathrm{BFS}(\mathbf{z}^{i_0,j_0})$
7. if $f(\mathbf{x}^k) \leq f(\mathbf{x}^{k+1})$, then STOP and the output is $\mathbf{u} = \mathbf{x}^k$. Otherwise, $k \leftarrow k+1$ and go back to step 1.

The full-CW search method obviously find a full-CW point in a finite number of steps.

Hierarchy of Algorithms (Best to Worst)

- Full-CW search.
- ZCWS
- Gradient projection/IHT
- BAFS

Numerical Experiments

Compared Method - TGA

Greedily build the super support

Algorithm 8 TGA

Initialization: $\mathbf{x} = \mathbf{0}_n$, $S = \emptyset$.

Output: $\mathbf{x} \in C_s \cap B$

1. while $|S| < s$ do:
   1. $(j, \mathbf{x}) \in \underset{(\ell \in S^c, \mathbf{z} \in B)}{\operatorname{argmin}} \{f(\mathbf{z}) : I_1(\mathbf{z}) \subseteq S \cup \{\ell\}\}$
   2. set $S \leftarrow S \cup \{j\}$
2. Return $\mathbf{x}$

Numerical Experiment - Sparse index tracking

Objective: track an index with a small number of assets $\mathbf{x}^*$, which has a return matrix $\mathbf{A}$.

Prob. $\min \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \quad \text{s.t. } \mathbf{x} \in C_s \cap \Delta_n$

A: $\mathbf{A} \in \mathbb{R}^{72 \times 54}$, daily returns matrix

x: $\mathbf{x}$ weights vector

b: $\mathbf{b} \in \mathbb{R}^{72}$ S&P500 daily returns vector

Data 180 random sets of stocks from NYSE, 60 for each sparsity level

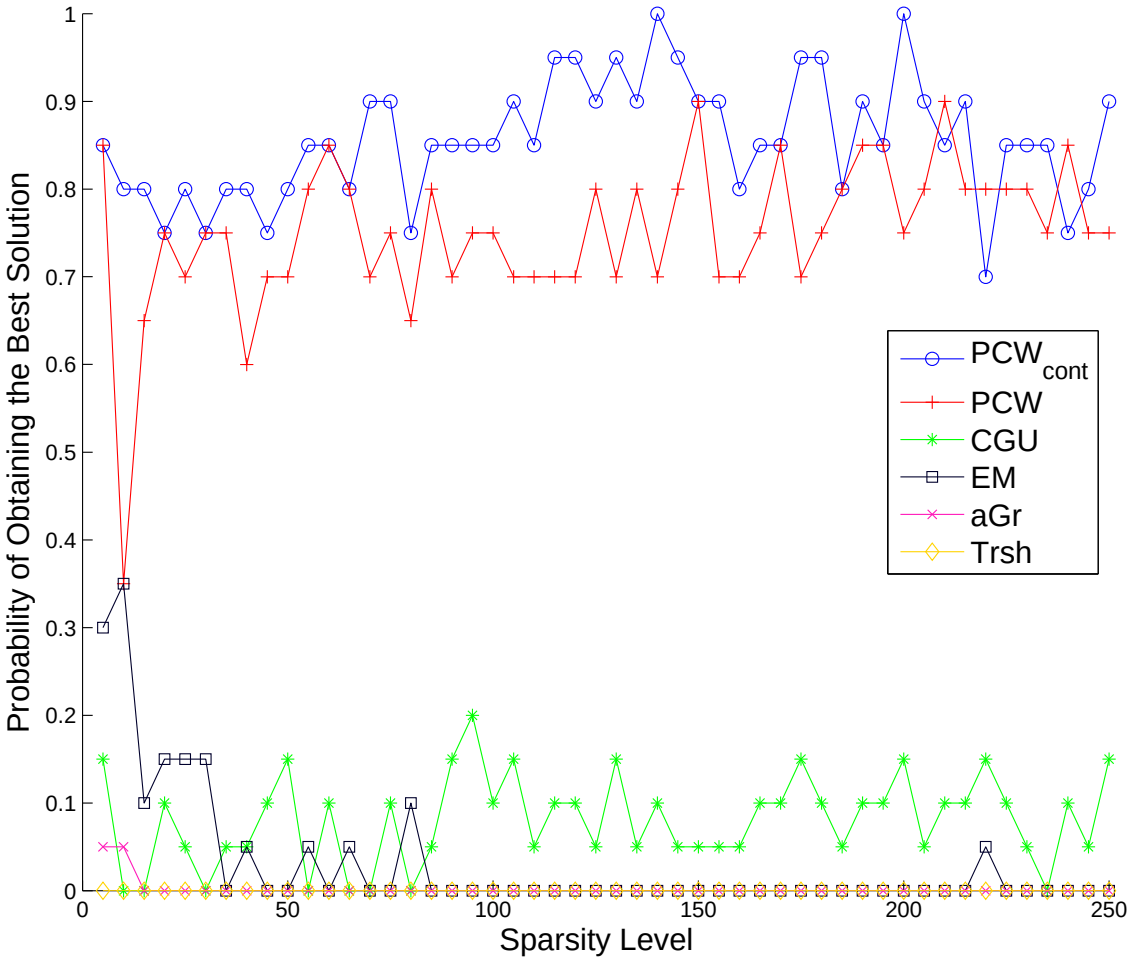
Numerical Experiment - Sparse index tracking

improver	**improved**	$s = 9$	$s = 18$	$s = 27$	**total**
ZCWS	FCWS	0	0	0	0
	IHT	60	60	60	180
	TGA	9	56	50	115
FCWS	ZCWS	33	11	17	61
	IHT	60	60	60	180
	TGA	15	56	51	122
IHT	ZCWS	0	0	0	0
	FCWS	0	0	0	0
	TGA	3	56	50	109

1. The IHT never reached a zero-cw or full-cw point.
2. The ZCWS reached a full-cw point in 66%.
3. The TGA was improved in most of the instances when $s > 9$.

sparse PCA - gene expression data (GeneChip oncology)

20 data sets. Number of variables 7129-54675.



THANK YOU FOR YOUR ATTENTION

- Beck, Hallak "On the minimization over sparse symmetric sets".
- Beck, Vaisbourd "Optimization methods for solving the sparse PCA problem".