The Spectral Bundle Method for Eigenvalue Optimization and Semidefinite Relaxations

Christoph Helmberg (TU Chemnitz)

Overview

Convex functions and the subdifferential

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, a vector $g \in \mathbb{R}^n$ is a subgradient of f at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall y \in \mathbb{R}^n \qquad \text{"subgradient ineq."}$$

Convex functions and the subdifferential

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, a vector $g \in \mathbb{R}^n$ is a subgradient of f at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall y \in \mathbb{R}^n \qquad \text{“subgradient ineq.”}$$

For $\gamma = f(x) - \langle g, x \rangle$ the pair (γ, g) defines a (global) linear minorant $f_{(\gamma,g)}$ of f: $\quad f_{(\gamma,g)}(y) := \gamma + \langle g, y \rangle \leq f(y)$

Convex functions and the subdifferential

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, a vector $g \in \mathbb{R}^n$ is a subgradient of f at x if

$$f(y) \geq f(x) + \langle g, y - x\rangle \quad \forall y \in \mathbb{R}^n \qquad \text{“subgradient ineq.”}$$

For $\gamma = f(x) - \langle g, x\rangle$ the pair (γ, g) defines a (global) linear minorant $f_{(\gamma,g)}$ of f: $\quad f_{(\gamma,g)}(y) := \gamma + \langle g, y\rangle \leq f(y)$

The subdifferential of f at x is the set of all subgradients of f at x,

$$\partial f(x) = \{g : f(y) \geq f(x) + \langle g, y - x\rangle \quad \forall y \in \mathbb{R}^n\}.$$

(for differentiable convex f, $\partial f(x) = \{\nabla f(x)\}$)

Convex functions and the subdifferential

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, a vector $g \in \mathbb{R}^n$ is a subgradient of f at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall y \in \mathbb{R}^n \qquad \text{“subgradient ineq.”}$$

For $\gamma = f(x) - \langle g, x \rangle$ the pair (γ, g) defines a (global) linear minorant $f_{(\gamma,g)}$ of f: $\quad f_{(\gamma,g)}(y) := \gamma + \langle g, y \rangle \leq f(y)$
The subdifferential of f at x is the set of all subgradients of f at x,

$$\partial f(x) = \{g : f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall y \in \mathbb{R}^n\}.$$

(for differentiable convex f, $\partial f(x) = \{\nabla f(x)\}$)
A closed proper convex function $f : \mathbb{R}^n \to \mathbb{R}$ is the supremum over its linear minorants $\mathcal{M}$,

$$f(y) = \sup_{(\gamma,g)\in\mathcal{M}} \gamma + \langle g, y \rangle$$

(all supporting hyperplanes of the epigraph of f)

Convex functions and the subdifferential

Given a convex function $f : \mathbb{R}^n \to \mathbb{R}$, a vector $g \in \mathbb{R}^n$ is a subgradient of f at x if

$$f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall y \in \mathbb{R}^n \qquad \text{"subgradient ineq."}$$

For $\gamma = f(x) - \langle g, x \rangle$ the pair (γ, g) defines a (global) linear minorant $f_{(\gamma,g)}$ of f: $f_{(\gamma,g)}(y) := \gamma + \langle g, y \rangle \leq f(y)$

The subdifferential of f at x is the set of all subgradients of f at x,

$$\partial f(x) = \{g : f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall y \in \mathbb{R}^n\}.$$

(for differentiable convex f, $\partial f(x) = \{\nabla f(x)\}$)

A closed proper convex function $f : \mathbb{R}^n \to \mathbb{R}$ is the supremum over its linear minorants $\mathcal{M}$,

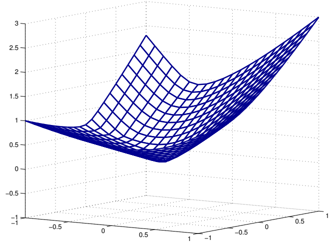
$$f(y) = \sup_{(\gamma,g)\in\mathcal{M}} \gamma + \langle g, y \rangle$$

(all supporting hyperplanes of the epigraph of f)

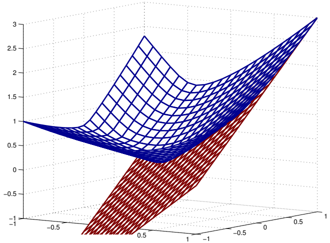
Minimize nonsmooth convex functions → subgradient and bundle methods Hiriart-Urruty and Lemaréchal 1993

Proximal Bundle Method [Lemaréchal78,Kiwiel90]

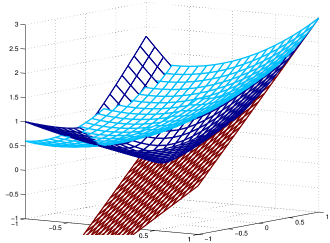
convex function



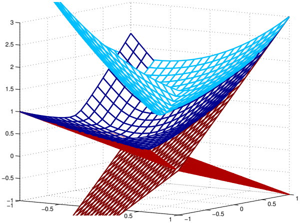
cutting plane model with $g \in \partial f(\hat{y})$



solve augmented model $\longrightarrow y^+$



improve cutting plane model in y^+



The main steps of Bundle Methods

Input: a convex function given by a first order oracle

1. Find a candidate by solving the quadratic model
2. Evaluate the function and determine a subgradient (oracle)
3. Decide on
    - null step
    - descent step
4. Update the model and iterate

A polyhedral cutting model and its quadratic model

A closed proper convex function $f : \mathbb{R}^n \to \mathbb{R}$ is the supremum over its linear minorants $\mathcal{M}$,

$$f(y) = \sup_{(\gamma,g)\in\mathcal{M}} \gamma + \langle g, y \rangle \qquad \forall y \in \mathbb{R}^n.$$

A polyhedral cutting model and its quadratic model

A closed proper convex function $f : \mathbb{R}^n \to \mathbb{R}$ is the supremum over its linear minorants $\mathcal{M}$,

$$f(y) = \sup_{(\gamma,g)\in\mathcal{M}} \gamma + \langle g, y\rangle \qquad \forall y \in \mathbb{R}^n.$$

Any subset $\widehat{\mathcal{M}} \subseteq \mathcal{M}$ yields a minorizing cutting model,

$$f_{\widehat{\mathcal{M}}}(y) := \sup_{(\gamma,g)\in\widehat{\mathcal{M}}} \gamma + \langle g, y\rangle \ \leq \ f(y) \qquad \forall y \in \mathbb{R}^n.$$

A polyhedral cutting model and its quadratic model

A closed proper convex function $f : \mathbb{R}^n \to \mathbb{R}$ is the supremum over its linear minorants $\mathcal{M}$,

$$f(y) = \sup_{(\gamma,g)\in\mathcal{M}} \gamma + \langle g, y \rangle \qquad \forall y \in \mathbb{R}^n.$$

Any subset $\widehat{\mathcal{M}} \subseteq \mathcal{M}$ yields a minorizing cutting model,

$$f_{\widehat{\mathcal{M}}}(y) := \sup_{(\gamma,g)\in\widehat{\mathcal{M}}} \gamma + \langle g, y \rangle \ \le \ f(y) \qquad \forall y \in \mathbb{R}^n.$$

Finite $\widehat{\mathcal{M}}$ yields a polyhedral model and may be written as

$$f_{\widehat{\mathcal{M}}}(y) = \max_{\xi_i \ge 0, \sum \xi_i = 1} \sum \xi_i (\gamma_i + g_i^T y).$$

A polyhedral cutting model and its quadratic model

A closed proper convex function $f : \mathbb{R}^n \to \mathbb{R}$ is the supremum over its linear minorants $\mathcal{M}$,

$$f(y) = \sup_{(\gamma,g)\in\mathcal{M}} \gamma + \langle g, y \rangle \qquad \forall y \in \mathbb{R}^n.$$

Any subset $\widehat{\mathcal{M}} \subseteq \mathcal{M}$ yields a minorizing cutting model,

$$f_{\widehat{\mathcal{M}}}(y) := \sup_{(\gamma,g)\in\widehat{\mathcal{M}}} \gamma + \langle g, y \rangle \ \leq \ f(y) \qquad \forall y \in \mathbb{R}^n.$$

Finite $\widehat{\mathcal{M}}$ yields a polyhedral model and may be written as

$$f_{\widehat{\mathcal{M}}}(y) = \max_{\xi_i \geq 0, \sum \xi_i = 1} \sum \xi_i(\gamma_i + g_i^T y).$$

The quadratic model penalizes deviations from a current center of stability $\hat{y}$ by a quadratic term with a weight $u > 0$,

$$\min_{y\in\mathbb{R}^n} \ f_{\widehat{\mathcal{M}}}(y) + \frac{u}{2}\|y - \hat{y}\|^2.$$

Its minimizer is the next candidate y^+.

Solving the augmented model $\min f_{\widehat{\mathcal{M}}}(y) + \frac{u}{2}\|y - \hat{y}\|^2$

$$\min_{y} \max_{\xi_i \geq 0, \sum \xi_i = 1} \sum \xi_i(\gamma_i + g_i^T y) + \frac{u}{2}\|y - \hat{y}\|^2$$

$$= \max_{\xi_i \geq 0, \sum \xi_i = 1} \min_{y} \sum \xi_i(\gamma_i + g_i^T y) + \frac{u}{2}\|y - \hat{y}\|^2$$

Solving the augmented model $\min f_{\widehat{\mathcal{M}}}(y) + \frac{u}{2}\|y - \hat{y}\|^2$

$$\min_{y} \max_{\xi_i \geq 0, \sum \xi_i = 1} \sum \xi_i(\gamma_i + g_i^T y) + \frac{u}{2}\|y - \hat{y}\|^2$$

$$= \max_{\xi_i \geq 0, \sum \xi_i = 1} \min_{y} \sum \xi_i(\gamma_i + g_i^T y) + \frac{u}{2}\|y - \hat{y}\|^2$$

Solve unconstrained quadratic inner optimization over y explicitly:

$$\boxed{y^+(\xi) = \hat{y} - \tfrac{1}{u} \sum \xi_i g_i}$$

[u “step size/trust region control”]

Solving the augmented model $\min f_{\widehat{\mathcal{M}}}(y) + \frac{u}{2}\|y - \hat{y}\|^2$

$$\min_{y} \max_{\xi_i \geq 0, \sum \xi_i = 1} \sum \xi_i(\gamma_i + g_i^T y) + \frac{u}{2}\|y - \hat{y}\|^2$$

$$= \max_{\xi_i \geq 0, \sum \xi_i = 1} \min_{y} \sum \xi_i(\gamma_i + g_i^T y) + \frac{u}{2}\|y - \hat{y}\|^2$$

Solve unconstrained quadratic inner optimization over y explicitly:

$$\boxed{y^+(\xi) = \hat{y} - \tfrac{1}{u} \sum \xi_i g_i} \qquad [u \text{ “step size/trust region control”}]$$

Substitute for y to obtain a (convex) quadratic problem in ξ,

$$(\mathrm{QP}) \qquad \begin{array}{ll} \max & \sum \xi_i(\gamma_i + g_i^T \hat{y}) - \frac{1}{2u}\|\sum \xi_i g_i\|^2 \\ \text{s.t.} & \sum \xi_i = 1 \\ & \xi \geq 0. \end{array}$$

small if $|\widehat{\mathcal{M}}|$ is small, finds “a best” convex combination
→ “best aggregate (minorant)” $(\gamma^+, g^+) = \sum \xi_i^+ (\gamma_i, g_i)$ $\qquad [(\gamma_k^+, g_k^+)]$
→ new candidate $y^+ = y^+(\xi^+)$. $\qquad [y_k]$

The Algorithm

Input: $y_0 = \hat{y}_1$, $\widehat{\mathcal{M}}_1$, $\kappa \in (0,1)$, $\varepsilon > 0$, $k = 1$.

1. Solve (QP) $\quad\rightarrow\quad (\gamma_k^+, g_k^+)$ and y_k.
   If $\quad f(\hat{y}_k) - f_{(\gamma_k^+, g_k^+)}(y_k) < \varepsilon(|f(\hat{y}_k)| + 1) \quad$ then **stop**.

The Algorithm

Input: $y_0 = \hat{y}_1$, $\widehat{\mathcal{M}}_1$, $\kappa \in (0,1)$, $\varepsilon > 0$, $k = 1$.

1. Solve (QP) $\quad \rightarrow \quad (\gamma_k^+, g_k^+)$ and y_k.
   If $\quad f(\hat{y}_k) - f_{(\gamma_k^+, g_k^+)}(y_k) < \varepsilon(|f(\hat{y}_k)| + 1) \quad$ then **stop**.
2. Compute $f(y_k)$ and subgradient g_k^s, yields also γ_k^s.

The Algorithm

Input: $y_0 = \hat{y}_1$, $\widehat{\mathcal{M}}_1$, $\kappa \in (0,1)$, $\varepsilon > 0$, $k = 1$.

1. Solve (QP) $\quad\rightarrow\quad (\gamma_k^+, g_k^+)$ and y_k.
   If $\quad f(\hat{y}_k) - f_{(\gamma_k^+, g_k^+)}(y_k) < \varepsilon(|f(\hat{y}_k)| + 1)\quad$ then **stop**.
2. Compute $f(y_k)$ and subgradient g_k^s, yields also γ_k^s.
3. If $f(\hat{y}_k) - f(y_k) > \kappa[f(\hat{y}_k) - f_{(\gamma_k^+, g_k^+)}(y_k)]$
   then *descent step*: set $\hat{y}_{k+1} = y_k$,
   else *null step*: $\hat{y}_{k+1} = \hat{y}_k$ unchanged.

The Algorithm

Input: $y_0 = \hat{y}_1$, $\widehat{\mathcal{M}}_1$, $\kappa \in (0,1)$, $\varepsilon > 0$, $k = 1$.

1. Solve (QP) $\quad\rightarrow\quad (\gamma_k^+, g_k^+)$ and y_k.
   If $\quad f(\hat{y}_k) - f_{(\gamma_k^+, g_k^+)}(y_k) < \varepsilon(|f(\hat{y}_k)| + 1)$ then **stop**.
2. Compute $f(y_k)$ and subgradient g_k^s, yields also γ_k^s.
3. If $f(\hat{y}_k) - f(y_k) > \kappa[f(\hat{y}_k) - f_{(\gamma_k^+, g_k^+)}(y_k)]$
   then *descent step*: set $\hat{y}_{k+1} = y_k$,
   else *null step*: $\hat{y}_{k+1} = \hat{y}_k$ unchanged.
4. Find a new model so that $\{(\gamma_k^+, g_k^+), (\gamma_k^s, g_k^s)\} \subseteq \widehat{\mathcal{M}}_{k+1}$.
   Update the weight u, set $k \leftarrow k+1$, **goto** 1.

The Algorithm

Input: $y_0 = \hat{y}_1$, $\widehat{\mathcal{M}}_1$, $\kappa \in (0,1)$, $\varepsilon > 0$, $k = 1$.

1. Solve (QP) $\rightarrow$ (γ_k^+, g_k^+) and y_k.
   If $f(\hat{y}_k) - f_{(\gamma_k^+, g_k^+)}(y_k) < \varepsilon(|f(\hat{y}_k)| + 1)$ then **stop**.
2. Compute $f(y_k)$ and subgradient g_k^s, yields also γ_k^s.
3. If $f(\hat{y}_k) - f(y_k) > \kappa[f(\hat{y}_k) - f_{(\gamma_k^+, g_k^+)}(y_k)]$
   then *descent step*: set $\hat{y}_{k+1} = y_k$,
   else *null step*: $\hat{y}_{k+1} = \hat{y}_k$ unchanged.
4. Find a new model so that $\{(\gamma_k^+, g_k^+), (\gamma_k^s, g_k^s)\} \subseteq \widehat{\mathcal{M}}_{k+1}$.
   Update the weight u, set $k \leftarrow k+1$, **goto** 1.

Theorem. Let $\varepsilon = 0$ then the sequence of descent steps $\{\hat{y}_k\}$ satisfies $f(\hat{y}_k) \rightarrow \inf_y f$ and (plus some conditions) $g_k^+ \rightarrow 0$.

[Lemaréchal78, Kiwiel90,...]

Important step in the proof of convergence:

Lemma. For an infinite sequence of null steps y_k

$$f(y_k) - f_{(\gamma_k^+, g_k^+)}(y_k) \;\rightarrow\; 0 \quad \text{and} \quad y_k \;\rightarrow\; \operatorname*{argmin}_y \; f(y) + \frac{u}{2}\|y - \hat{y}\|^2.$$

Thus,

either descent step after finitely many iterations

or $\hat{y}$ optimal.

Important step in the proof of convergence:

Lemma. For an infinite sequence of null steps y_k

$$f(y_k) - f_{(\gamma_k^+, g_k^+)}(y_k) \rightarrow 0 \quad \text{and} \quad y_k \rightarrow \operatorname*{argmin}_y f(y) + \frac{u}{2}\|y - \hat{y}\|^2.$$

Thus,

either descent step after finitely many iterations

or $\hat{y}$ optimal.

The minimizer of $f(\cdot) + \|\cdot - \hat{y}\|$ is the "proximal point" of $\hat{y}$. [Rockafellar76]

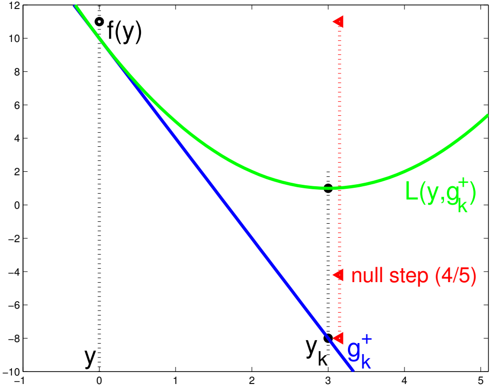
For null steps, y_k converges to the proximal point and $f_{\widehat{\mathcal{M}}_k}(y_k)$ to its value.

Idea: By $(\gamma_k^+, g_k^+) \in \widehat{\mathcal{M}}_{k+1}$ the next QP-value cannot decrease:

$$\min_y \max_{(\gamma,g)\in\widehat{\mathcal{M}}_{k+1}} L(y,(\gamma,g)) := \gamma + \langle g, y\rangle + \|y-\hat{y}\|^2$$

Idea: By $(\gamma_k^+, g_k^+) \in \widehat{\mathcal{M}}_{k+1}$ the next QP-value cannot decrease:

$$\min_y \max_{(\gamma,g)\in\widehat{\mathcal{M}}_{k+1}} L(y,(\gamma,g)) := \gamma + \langle g, y\rangle + \|y-\hat{y}\|^2$$

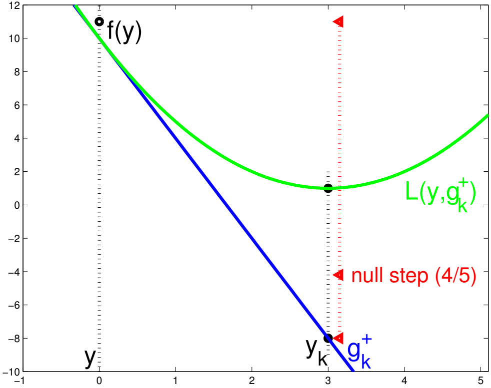


$$L(y_k,(\gamma_k^+, g_k^+)) + \|y \quad - y_k\|^2 =$$

$$= L(y \quad ,(\gamma_k^+, g_k^+))$$

Idea: By $(\gamma_k^+, g_k^+) \in \widehat{\mathcal{M}}_{k+1}$ the next QP-value cannot decrease:

$$\min_y \max_{(\gamma,g)\in\widehat{\mathcal{M}}_{k+1}} L(y,(\gamma,g)) := \gamma + \langle g, y\rangle + \|y-\hat{y}\|^2$$

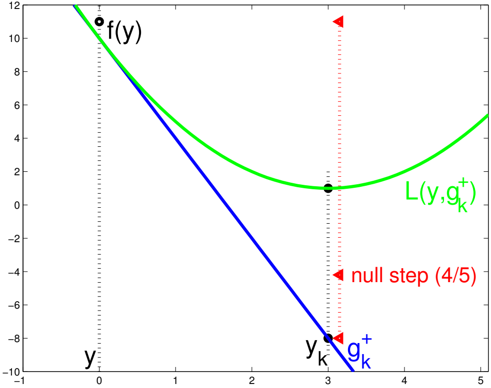


$$L(y_k,(\gamma_k^+, g_k^+)) + \|y_{k+1} - y_k\|^2 =$$

$$= L(y_{k+1},(\gamma_k^+, g_k^+))$$

Idea: By $(\gamma_k^+, g_k^+) \in \widehat{\mathcal{M}}_{k+1}$ the next QP-value cannot decrease:

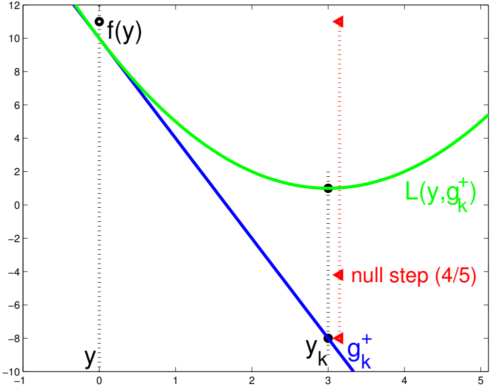
$$\min_y \max_{(\gamma,g)\in\widehat{\mathcal{M}}_{k+1}} L(y,(\gamma,g)) := \gamma + \langle g, y\rangle + \|y-\hat{y}\|^2$$



$$L(y_k,(\gamma_k^+,g_k^+)) + \|y_{k+1}-y_k\|^2 =$$
$$= L(y_{k+1},(\gamma_k^+,g_k^+))$$
$$\leq L(y_{k+1},(\gamma_{k+1}^+,g_{k+1}^+))$$
$$\leq f(\hat{y})$$

Idea: By $(\gamma_k^+, g_k^+) \in \widehat{\mathcal{M}}_{k+1}$ the next QP-value cannot decrease:

$$\min_y \max_{(\gamma,g)\in\widehat{\mathcal{M}}_{k+1}} L(y,(\gamma,g)) := \gamma + \langle g, y\rangle + \|y-\hat{y}\|^2$$



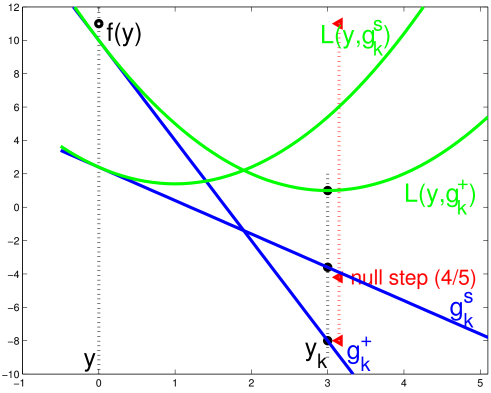
$$L(y_k,(\gamma_k^+,g_k^+)) + \|y_{k+1}-y_k\|^2 =$$
$$= L(y_{k+1},(\gamma_k^+,g_k^+))$$
$$\leq L(y_{k+1},(\gamma_{k+1}^+,g_{k+1}^+))$$
$$\leq f(\hat{y})$$

$$\Rightarrow \|y_{k+1}-y_k\|^2 \to 0$$

$$\stackrel{y \text{ bounded}}{\Rightarrow} \|y_{k+1}-y_k\| \to 0$$

Idea: By $(\gamma_k^+, g_k^+) \in \widehat{\mathcal{M}}_{k+1}$ the next QP-value cannot decrease:

$$\min_y \max_{(\gamma,g)\in\widehat{\mathcal{M}}_{k+1}} L(y,(\gamma,g)) := \gamma + \langle g, y\rangle + \|y-\hat{y}\|^2$$



$$L(y_k,(\gamma_k^+, g_k^+)) + \|y_{k+1}-y_k\|^2 =$$

$$= L(y_{k+1},(\gamma_k^+, g_k^+))$$

$$\leq L(y_{k+1},(\gamma_{k+1}^+, g_{k+1}^+))$$

$$\leq f(\hat{y})$$

$$\Rightarrow \|y_{k+1}-y_k\|^2 \to 0$$

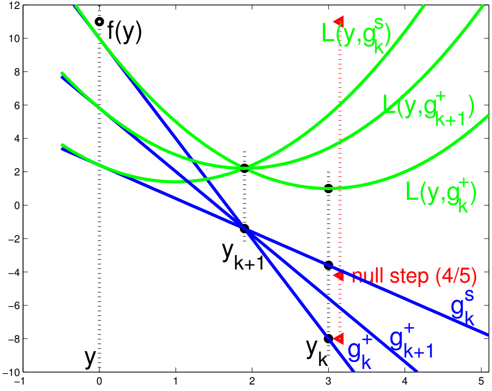
$$\overset{y \text{ bounded}}{\Rightarrow} \|y_{k+1}-y_k\| \to 0$$

In a null step, $(\gamma_k^s, g_k^s) \in \widehat{\mathcal{M}}_{k+1}$ forces y_{k+1} away from y_k:

$$f_{(\gamma_k^s, g_k^s)}(y_{k+1}) \leq f_{\widehat{\mathcal{M}}_{k+1}}(y_{k+1}) = f_{(\gamma_{k+1}^+, g_{k+1}^+)}(y_{k+1})$$

Idea: By $(\gamma_k^+, g_k^+) \in \widehat{\mathcal{M}}_{k+1}$ the next QP-value cannot decrease:

$$\min_y \max_{(\gamma,g)\in\widehat{\mathcal{M}}_{k+1}} L(y,(\gamma,g)) := \gamma + \langle g, y\rangle + \|y-\hat{y}\|^2$$



$$L(y_k,(\gamma_k^+, g_k^+)) + \|y_{k+1}-y_k\|^2 =$$
$$= L(y_{k+1},(\gamma_k^+, g_k^+))$$
$$\le L(y_{k+1},(\gamma_{k+1}^+, g_{k+1}^+))$$
$$\le f(\hat{y})$$

$$\Rightarrow \|y_{k+1}-y_k\|^2 \to 0$$

$$\overset{y \text{ bounded}}{\Rightarrow} \|y_{k+1}-y_k\| \to 0$$

In a null step, $(\gamma_k^s, g_k^s) \in \widehat{\mathcal{M}}_{k+1}$ forces y_{k+1} away from y_k:

$$f_{(\gamma_k^s, g_k^s)}(y_{k+1}) \le f_{\widehat{\mathcal{M}}_{k+1}}(y_{k+1}) = f_{(\gamma_{k+1}^+, g_{k+1}^+)}(y_{k+1})$$

The aggregate (γ^+, g^+)

- is constructed from a dual optimal QP-solution
- is "the best" supporting hyperplane in $\operatorname{conv} \widehat{\mathcal{M}}$
- is the linear minorant holding the current solution (saddle point)
- needs to be contained in the next model to ensure convergence
- is the object "converging" to the zero subgradient

Overview

LP $\leftrightarrow$ SDP

LP	SDP
$\max \langle c,x\rangle$ s.t. $Ax=b$, $x\geq 0$	$\max \langle C,X\rangle$ s.t. $\mathcal{A}X=b$, $X\succeq 0$
$x\in\mathbb{R}^n_+$ nonneg. orthant (polyhedral)	$X\in\mathcal{S}^n_+$ pos. semidef. matrices (non-polyhedral)
$\langle c,x\rangle=\sum_i c_ix_i$	$\langle C,X\rangle=\sum_{i,j} C_{ij}X_{ij}$
$Ax=\begin{pmatrix}\langle a_1,x\rangle\\ \vdots\\ \langle a_m,x\rangle\end{pmatrix}$	$\mathcal{A}X=\begin{pmatrix}\langle A_1,X\rangle\\ \vdots\\ \langle A_m,X\rangle\end{pmatrix}$
$A^Ty=\sum_i a_iy_i$	$\mathcal{A}^Ty=\sum_i A_iy_i$
$\min \langle b,y\rangle$ s.t. $A^Ty-z=c$, $z\geq 0$	$\min \langle b,y\rangle$ s.t. $\mathcal{A}^Ty-Z=C$, $Z\succeq 0$

Example

$$\begin{array}{rl} \max & \langle C, X\rangle \\ \text{s.t.} & \langle I, X\rangle = 1 \\ & X \succeq 0 \end{array}$$

$$\begin{array}{rl} \min & y \\ \text{s.t.} & Z = yI - C \succeq 0 \end{array}$$

Example

$$\begin{array}{rl} \max & \langle C, X\rangle \\ \text{s.t.} & \langle I, X\rangle = 1 \\ & X \succeq 0 \end{array} \qquad \begin{array}{rl} \min & y \\ \text{s.t.} & Z = yI - C \succeq 0 \end{array}$$

$$\mathcal{W} := \{X \succeq 0 : \langle I, X\rangle = 1\} \quad = \quad \operatorname{conv}\left\{vv^T : \left\langle I, vv^T\right\rangle = v^T v = 1\right\}$$

and
$$\max_{\|v\|^2=1} \left\langle C, vv^T\right\rangle \quad = \quad \max_{\|v\|=1} v^T C v \quad = \quad \lambda_{\max}(C)$$

Example

$$\begin{array}{rl} \max & \langle C, X \rangle \\ \text{s.t.} & \langle I, X \rangle = 1 \\ & X \succeq 0 \end{array} \qquad \begin{array}{rl} \min & y \\ \text{s.t.} & Z = yI - C \succeq 0 \end{array}$$

$$\mathcal{W} := \{X \succeq 0 : \langle I, X \rangle = 1\} \;=\; \operatorname{conv}\left\{vv^T : \left\langle I, vv^T \right\rangle = v^T v = 1\right\}$$

$$\text{and} \qquad \max_{\|v\|^2=1} \left\langle C, vv^T \right\rangle \;=\; \max_{\|v\|=1} v^T C v \;=\; \lambda_{\max}(C)$$

---

set of primal optimal solutions:

$$\begin{aligned} & \operatorname{conv}\left\{vv^T : \left\langle I, vv^T \right\rangle = 1, v^T C v = \lambda_{\max}(C)\right\} \qquad [v = Pu] \\ = \; & \operatorname{conv}\left\{Puu^T P^T : \left\langle I, uu^T \right\rangle = 1\right\} \\ = \; & \left\{PUP^T : \langle I, U \rangle = 1, U \succeq 0\right\} \end{aligned}$$

columns of P form an orthonormal basis of the eigenspace of $\lambda_{\max}(C)$.

Example

$$\begin{array}{rl} \max & \langle C, X\rangle \\ \text{s.t.} & \langle I, X\rangle = 1 \\ & X \succeq 0 \end{array} \qquad \begin{array}{rl} \min & y \\ \text{s.t.} & Z = yI - C \succeq 0 \end{array}$$

$$\mathcal{W} := \{X \succeq 0 : \langle I, X\rangle = 1\} \;=\; \operatorname{conv}\left\{vv^T : \left\langle I, vv^T\right\rangle = v^T v = 1\right\}$$

$$\text{and} \qquad \max_{\|v\|^2=1} \left\langle C, vv^T\right\rangle \;=\; \max_{\|v\|=1} v^T C v \;=\; \lambda_{\max}(C)$$

---

set of primal optimal solutions:

$$\begin{array}{rl} & \operatorname{conv}\left\{vv^T : \left\langle I, vv^T\right\rangle = 1, v^T C v = \lambda_{\max}(C)\right\} \qquad [v = Pu] \\ = & \operatorname{conv}\left\{Puu^T P^T : \left\langle I, uu^T\right\rangle = 1\right\} \\ = & \left\{PUP^T : \langle I, U\rangle = 1, U \succeq 0\right\} \end{array}$$

columns of P form an orthonormal basis of the eigenspace of $\lambda_{\max}(C)$.

---

dual: min λ s.t. $\lambda I - C \succeq 0$ $\quad\Rightarrow\quad$ optimal $\lambda = \lambda_{\max}(C)$

SDP and Eigenvalue Optimization

For constant trace, the dual is an eigenvalue optimization problem

$$\begin{array}{rl} \max & \langle C, X\rangle \\ \text{s.t.} & \langle I, X\rangle = a \\ & \mathcal{A}X = b \\ & X \succeq 0, \end{array} \qquad \min_{y \in \mathbb{R}^m} \; a\lambda_{\max}(C - \mathcal{A}^T y) + \langle b, y\rangle$$

(E.g., many semidefinite relaxations of comb. opt. problems satisfy this.)

SDP and Eigenvalue Optimization

For constant trace, the dual is an eigenvalue optimization problem

$$\begin{array}{rl} \max & \langle C, X\rangle \\ \text{s.t.} & \langle I, X\rangle = a \\ & \mathcal{A}X = b \\ & X \succeq 0, \end{array} \qquad \min_{y \in \mathbb{R}^m} \; a\lambda_{\max}(C - \mathcal{A}^T y) + \langle b, y\rangle$$

(E.g., many semidefinite relaxations of comb. opt. problems satisfy this.)
In the following, we assume (w.l.o.g.) $a = 1$.

$$f(y) := \lambda_{\max}(C - \mathcal{A}^T y) + \langle b, y\rangle = \max_{W \in \mathcal{W}} \left\langle C - \mathcal{A}^T y, W \right\rangle + b^T y$$

is convex and nonsmooth.

SDP and Eigenvalue Optimization

For constant trace, the dual is an eigenvalue optimization problem

$$\begin{array}{rl} \max & \langle C, X \rangle \\ \text{s.t.} & \langle I, X \rangle = a \\ & \mathcal{A}X = b \\ & X \succeq 0, \end{array} \qquad \min_{y \in \mathbb{R}^m} \; a\lambda_{\max}(C - \mathcal{A}^T y) + \langle b, y \rangle$$

(E.g., many semidefinite relaxations of comb. opt. problems satisfy this.)
In the following, we assume (w.l.o.g.) $a = 1$.

$$f(y) := \lambda_{\max}(C - \mathcal{A}^T y) + \langle b, y \rangle = \max_{W \in \mathcal{W}} \left\langle C - \mathcal{A}^T y, W \right\rangle + b^T y$$

is convex and nonsmooth. By the affine chain rule,

$$\partial f(y) = \{ b - \mathcal{A}(PUP^T) : \langle I, U \rangle = 1, U \succeq 0 \}$$

$$\text{with } P^T P = I \text{ and } P^T (C - \mathcal{A}^T y) P = \lambda_{\max}(C - \mathcal{A}^T y) I.$$

Any eigenvector v to $\lambda_{\max}(C - \mathcal{A}^T y)$ yields a subgradient $b - \mathcal{A}^T(vv^T)$.

Eigenvalue Optimization in General

$$\min_{y \in \mathbb{R}^m} \lambda_{\max}(F(y))$$

with $F : \mathbb{R}^m \to \mathcal{S}^n$ a smooth matrix valued function.

Rich history in optimization,
for theory pointers see the survey by [Lewis 2003]
some algorithmic landmarks (not complete):
[Cullum Donath Wolfe 1975, Polak Wardi 1982, Fletcher 1985, Overton 1988/92, Nesterov Nemirovskii 1993, Shapiro Fan 1995, Overton Womersley 199*, Oustry 2000, Helmberg Rendl 2000, Noll Apkarian 200*, Nesterov 2007]

Here, we concentrate on affine F,

$$F(y) = C - \sum A_i y_i.$$

Overview

The Spectral Bundle Method [H.,Rendl00]

for solving large scale eigenvalue optimization problems of the form

$$f(y) := \lambda_{\max}(C - \mathcal{A}^T y) + \langle b, y \rangle .$$

Key ideas:

- The matrix $C - \sum_i A_i y_i$ inherits the structure of cost matrix and constraints $\rightarrow$ function value and subgradient can be computed efficiently by iterative methods like Lanczos methods.
- Exploit the special structure of the subdifferential in a semidefinite cutting surface model within the bundle method.

A semidefinite model for $f(y) := \lambda_{\max}(C - \mathcal{A}^T y) + b^T y$

With $\mathcal{W} = \{W \succeq 0 : \operatorname{tr} W = 1\}$

$$f(y) = \max_{W \in \mathcal{W}} \langle W, C - \mathcal{A}^T y \rangle + b^T y$$

evaluate by computing $\lambda_{\max}(C - \mathcal{A}^T y)$, [Lanczos]
any eigenvector v to $\lambda_{\max}$, $\|v\| = 1$, yields a subgradient via $vv^T \in \mathcal{W}$

A semidefinite model for $f(y) := \lambda_{\max}(C - \mathcal{A}^T y) + b^T y$

With $\mathcal{W} = \{W \succeq 0 : \operatorname{tr} W = 1\}$

$$f(y) = \max_{W \in \mathcal{W}} \langle W, C - \mathcal{A}^T y \rangle + b^T y$$

evaluate by computing $\lambda_{\max}(C - \mathcal{A}^T y)$, [Lanczos]
any eigenvector v to $\lambda_{\max}$, $\|v\| = 1$, yields a subgradient via $vv^T \in \mathcal{W}$

For any subset $\widehat{\mathcal{W}}_k \subseteq \mathcal{W}$ one obtains a cutting model

$$\boxed{f_{\widehat{\mathcal{W}}_k}(y) = \max_{W \in \widehat{\mathcal{W}}_k} \langle W, C - \mathcal{A}^T y \rangle + b^T y} \qquad \leq f(y) \quad \forall y \in \mathbb{R}^m$$

A semidefinite model for $f(y) := \lambda_{\max}(C - \mathcal{A}^T y) + b^T y$

With $\mathcal{W} = \{W \succeq 0 : \operatorname{tr} W = 1\}$

$$f(y) = \max_{W \in \mathcal{W}} \langle W, C - \mathcal{A}^T y \rangle + b^T y$$

evaluate by computing $\lambda_{\max}(C - \mathcal{A}^T y)$, [Lanczos]
any eigenvector v to $\lambda_{\max}$, $\|v\| = 1$, yields a subgradient via $vv^T \in \mathcal{W}$

For any subset $\widehat{\mathcal{W}}_k \subseteq \mathcal{W}$ one obtains a cutting model

$$\boxed{f_{\widehat{\mathcal{W}}_k}(y) = \max_{W \in \widehat{\mathcal{W}}_k} \langle W, C - \mathcal{A}^T y \rangle + b^T y} \qquad \leq f(y) \quad \forall y \in \mathbb{R}^m$$

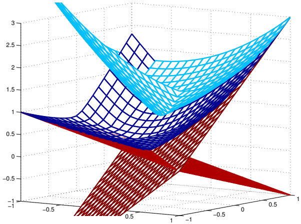
We use

$$\boxed{\widehat{\mathcal{W}}_k = \left\{ P_k U P_k^T + \alpha \overline{W}_k : \operatorname{tr} U + \alpha = 1, U \succeq 0, \alpha \geq 0 \right\}} \qquad \subseteq \mathcal{W}$$

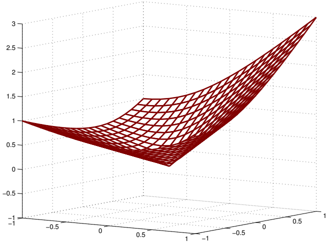
with parameters $P_k \in \mathbb{R}^{n \times r}$, $P_k^T P_k = I_r$, and a “residual” $\overline{W}_k \in \mathcal{W}$.

Example: P holds a basis of the eigenvectors of two subgradients

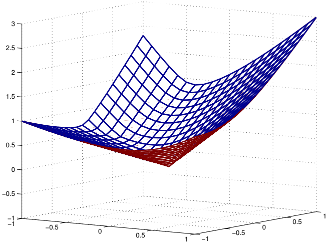
polyhedral model



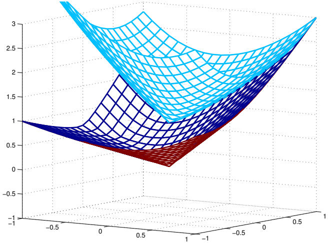
semidefinite model



model and function



quadratic semidefinite model



The Semidefinite Bundle

$$f_{\widehat{\mathcal{W}}_k}(y) = \max_{W \in \widehat{\mathcal{W}}_k} \left\langle W, C - \mathcal{A}^T y \right\rangle + b^T y \qquad \leq f(y) \quad \forall y \in \mathbb{R}^m$$

$$\boxed{\widehat{\mathcal{W}}_k = \left\{ P_k U P_k^T + \alpha \overline{W}_k : \operatorname{tr} U + \alpha = 1, U \succeq 0, \alpha \geq 0 \right\}} \qquad \subseteq \mathcal{W}$$

with parameters $P_k \in \mathbb{R}^{n \times r}$, $P_k^T P_k = I_r$, and a residual $\overline{W}_k \in \mathcal{W}$.

P should span an approximation of the eigenspace to $\lambda_{\max}$ near $\hat{y}$.

Because PUP^T spans only a face on the boundary of $\mathcal{W}$,
$\overline{W}$ is needed to span part of the interior of $\mathcal{W}$

The Semidefinite Bundle

$$f_{\widehat{\mathcal{W}}_k}(y) = \max_{W \in \widehat{\mathcal{W}}_k} \left\langle W, C - \mathcal{A}^T y \right\rangle + b^T y \qquad \leq f(y) \quad \forall y \in \mathbb{R}^m$$

$$\boxed{\widehat{\mathcal{W}}_k = \left\{ P_k U P_k^T + \alpha \overline{W}_k : \operatorname{tr} U + \alpha = 1, U \succeq 0, \alpha \geq 0 \right\}} \qquad \subseteq \mathcal{W}$$

with parameters $P_k \in \mathbb{R}^{n \times r}$, $P_k^T P_k = I_r$, and a residual $\overline{W}_k \in \mathcal{W}$.

P should span an approximation of the eigenspace to $\lambda_{\max}$ near $\hat{y}$.

Because PUP^T spans only a face on the boundary of $\mathcal{W}$,
$\overline{W}$ is needed to span part of the interior of $\mathcal{W}$

It is possible to do without $\overline{W}$ if P is “fat” enough:

Theorem (Barvinok95,Pataki98)

An SDP $\max\{\langle C, X\rangle : \mathcal{A}X = b, X \succeq 0\}$ *with finite optima also has an optimal solution of rank* r *bounded by* $\binom{r+1}{2} \leq m$.

Solving the augmented model $\min f_{\widehat{\mathcal{W}}}(y) + \frac{u}{2}\|y - \hat{y}\|^2$

$$\min_{y} \max_{W \in \widehat{\mathcal{W}}} \left\langle C - \mathcal{A}^T y, W \right\rangle + \langle b, y \rangle + \frac{u}{2}\|y - \hat{y}\|^2$$

$$= \max_{W \in \widehat{\mathcal{W}}} \min_{y} \langle C, W \rangle + \langle b - \mathcal{A}W, y \rangle + \frac{u}{2}\|y - \hat{y}\|^2$$

Solving the augmented model $\min f_{\widehat{\mathcal{W}}}(y) + \frac{u}{2}\|y - \hat{y}\|^2$

$$\min_{y} \max_{W \in \widehat{\mathcal{W}}} \left\langle C - \mathcal{A}^T y, W \right\rangle + \langle b, y \rangle + \frac{u}{2}\|y - \hat{y}\|^2$$

$$= \max_{W \in \widehat{\mathcal{W}}} \min_{y} \langle C, W \rangle + \langle b - \mathcal{A}W, y \rangle + \frac{u}{2}\|y - \hat{y}\|^2$$

Solve unconstrained quadratic inner optimization over y explicitly:

$$\boxed{y_+(W) = \hat{y} - \tfrac{1}{u}(b - \mathcal{A}W)}$$

[u "step size/trust region control"]

Solving the augmented model $\min f_{\widehat{\mathcal{W}}}(y) + \frac{u}{2}\|y - \hat{y}\|^2$

$$\min_{y} \max_{W \in \widehat{\mathcal{W}}} \left\langle C - \mathcal{A}^T y, W \right\rangle + \langle b, y \rangle + \frac{u}{2}\|y - \hat{y}\|^2$$

$$= \max_{W \in \widehat{\mathcal{W}}} \min_{y} \langle C, W \rangle + \langle b - \mathcal{A}W, y \rangle + \frac{u}{2}\|y - \hat{y}\|^2$$

Solve unconstrained quadratic inner optimization over y explicitly:

$$\boxed{y_+(W) = \hat{y} - \tfrac{1}{u}(b - \mathcal{A}W)}$$

[u "step size/trust region control"]

Substitute for y to obtain a quadratic semidefinite problem in W,

$$\text{(QSP)} \quad \begin{array}{ll} \max & \left\langle C - \mathcal{A}^T \hat{y}, W \right\rangle + \langle b, \hat{y} \rangle - \frac{1}{2u} \|b - \mathcal{A}W\|^2 \\ \text{s.t.} & W = PUP^T + \alpha \overline{W} \\ & \operatorname{tr} U + \alpha = 1 \\ & U \succeq 0, \alpha \geq 0. \end{array}$$

small if r is small ($U \in \mathcal{S}^r_+$) $\rightarrow$ interior point system matrix $\binom{r+1}{2} + 1$ [!]

$\rightarrow$ "aggregate (eps-subgradient)" $W_+ = PU_+P^T + \alpha_+ \overline{W}$ [W_k]

$\rightarrow$ new candidate $y_+ = y_+(W_+)$. [y_k]

The Algorithm

Input: $\mathcal{A}, b, C,\ y_0 = \hat{y}_1,\ \widehat{\mathcal{W}}_1,\ \kappa \in (0,1),\ \varepsilon > 0,\ k = 1.$

1. Solve (QSP) $\quad \rightarrow \quad W_k$ and y_k.
   If $\quad f(\hat{y}_k) - f_{\widehat{\mathcal{W}}_k}(y_k) < \varepsilon(|f(\hat{y}_k)| + 1) \quad$ then **stop**.
2. Compute $\lambda_{\max}(C - \mathcal{A}^T y^k)$ and eigenvector v, yields also $f(y_k)$.
3. If $f(\hat{y}_k) - f(y_k) > \kappa[f(\hat{y}_k) - f_{\widehat{\mathcal{W}}}(y_k)]$
   then *descent step*: set $\hat{y}_{k+1} = y_k$,
   else *null step*: $\hat{y}_{k+1} = \hat{y}_k$ unchanged.
4. Find new P_{k+1} and $\overline{W}_{k+1}$, so that $\boxed{\{vv^T, W_k\} \subset \widehat{\mathcal{W}}_{k+1}}$.
   Update the weight u, set $k \leftarrow k+1$, **goto** 1.

Theorem. Let $\varepsilon = 0$ then the sequence of descent steps $\{\hat{y}_k\}$ satisfies $f(\hat{y}_k) \rightarrow \inf_y f$ and (plus some conditions) $W \rightarrow X^*$.

Minimal choice in step 4 is $P_{k+1} = v$ and $\overline{W}_{k+1} = W_k$.

Overview

Eigenvalue Computation and Model Update

Important aspects in actual implementations are:

- dealing with the difficulty of clustered eigenvalues in eigenvalue computations

Eigenvalue Computation and Model Update

Important aspects in actual implementations are:

- dealing with the difficulty of clustered eigenvalues in eigenvalue computations
- exploiting the fact that iterative methods generate an increasing sequence of Ritz-values $v^T(C - \mathcal{A}^T y)v/v^T v$ converging to $\lambda_{\max}(C - \mathcal{A}^T y)$ from below by terminating early whenever the null step bound is exceeded,

Eigenvalue Computation and Model Update

Important aspects in actual implementations are:

- dealing with the difficulty of clustered eigenvalues in eigenvalue computations
- exploiting the fact that iterative methods generate an increasing sequence of Ritz-values $v^T(C - \mathcal{A}^T y)v/v^T v$ converging to $\lambda_{\max}(C - \mathcal{A}^T y)$ from below by terminating early whenever the null step bound is exceeded,
- exploiting additional Ritz-pairs from iterative methods in updating the bundle,

Eigenvalue Computation and Model Update

Important aspects in actual implementations are:

- dealing with the difficulty of clustered eigenvalues in eigenvalue computations
- exploiting the fact that iterative methods generate an increasing sequence of Ritz-values $v^T(C - \mathcal{A}^T y)v/v^T v$ converging to $\lambda_{\max}(C - \mathcal{A}^T y)$ from below by terminating early whenever the null step bound is exceeded,
- exploiting additional Ritz-pairs from iterative methods in updating the bundle,
- updating the bundle so as to keep the most important subspace in P.

Eigenvalue Computation for $A = C - \mathcal{A}^T y$ ($n > 50$)

Power method: $q_1,\ Aq_1,\ A^2q_1,\ \ldots,\ A^iq_1$

Lanczos Method: $\lambda_{\max}(A) \approx \max\limits_{v \in \operatorname{span}\{q_1, Aq_1, \ldots, A^iq_1\}} \frac{v^TAv}{v^Tv}$

Eigenvalue Computation for $A = C - \mathcal{A}^T y$ ($n > 50$)

Power method: $q_1,\ Aq_1,\ A^2 q_1,\ \ldots,\ A^i q_1$

Lanczos Method: $\lambda_{\max}(A) \approx \max\limits_{v \in \operatorname{span}\{q_1, Aq_1, \ldots, A^i q_1\}} \dfrac{v^T A v}{v^T v}$

constructs orthonormal bases Q_i of $\operatorname{span}\{q_1, Aq_1, \ldots, A^i q_1\}$ so that

$$T_i = Q_i^T A Q_i = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \ddots & \vdots \\ 0 & \beta_2 & \alpha_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_{i-1} \\ 0 & \cdots & 0 & \beta_{i-1} & \alpha_i \end{bmatrix} \in \mathcal{S}_i \rightarrow \text{eigenv. decomp. in } O(i^2).$$

Eigenvalue Computation for $A = C - \mathcal{A}^T y$ ($n > 50$)

Power method: $q_1,\ Aq_1,\ A^2 q_1,\ \ldots,\ A^i q_1$

Lanczos Method: $\lambda_{\max}(A) \approx \max\limits_{v \in \operatorname{span}\{q_1, Aq_1, \ldots, A^i q_1\}} \dfrac{v^T A v}{v^T v}$

constructs orthonormal bases Q_i of $\operatorname{span}\{q_1, Aq_1, \ldots, A^i q_1\}$ so that

$$T_i = Q_i^T A Q_i = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \ddots & \vdots \\ 0 & \beta_2 & \alpha_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_{i-1} \\ 0 & \cdots & 0 & \beta_{i-1} & \alpha_i \end{bmatrix} \in \mathcal{S}_i \rightarrow \text{eigenv. decomp. in } O(i^2).$$

$Q_i = [q_1, \ldots, q_i]$; compute q_{i+1} by orthonormalizing Aq_i to all q_j,

$$q_{i+1} = \frac{\bar{q}_{i+1}}{\|\bar{q}_{i+1}\|} \quad \text{with} \quad \underline{\bar{q}_{i+1}} = Aq_i - Q_i Q_i^T A q_i = \underline{Aq_i - \alpha_i q_i - \beta_{i-1} q_{i-1}}.$$

If $\|\bar{q}_{i+1}\| = 0 \Rightarrow$ invariant subspace found [usually $\lambda_{\max}$]

Eigenvalue Computation for $A = C - \mathcal{A}^T y$ ($n > 50$)

Power method: $q_1,\ Aq_1,\ A^2 q_1,\ \ldots,\ A^i q_1$

Lanczos Method: $\lambda_{\max}(A) \approx \max_{v \in \operatorname{span}\{q_1, Aq_1, \ldots, A^i q_1\}} \frac{v^T A v}{v^T v}$

constructs orthonormal bases Q_i of $\operatorname{span}\{q_1, Aq_1, \ldots, A^i q_1\}$ so that

$$T_i = Q_i^T A Q_i = \begin{bmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \ddots & \vdots \\ 0 & \beta_2 & \alpha_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \beta_{i-1} \\ 0 & \cdots & 0 & \beta_{i-1} & \alpha_i \end{bmatrix} \in \mathcal{S}_i \to \text{eigenv. decomp. in } O(i^2).$$

$Q_i = [q_1, \ldots, q_i]$; compute q_{i+1} by orthonormalizing Aq_i to all q_j,

$$q_{i+1} = \frac{\bar{q}_{i+1}}{\|\bar{q}_{i+1}\|} \quad \text{with} \quad \underline{\bar{q}_{i+1}} = Aq_i - Q_i Q_i^T A q_i = \underline{Aq_i - \alpha_i q_i - \beta_{i-1} q_{i-1}}.$$

If $\|\bar{q}_{i+1}\| = 0 \Rightarrow$ invariant subspace found [usually $\lambda_{\max}$]

---

trouble: q_i loose orthogonality quickly

$\to$ complete orthogonalization, restart every n_L iterations to keep Q small

Convergence: the better the larger $\frac{\lambda_{\max} - \lambda_2}{\lambda_{\max} - \lambda_{\min}}$ [ignore multiplicities]

trouble: in eigenvalue optimization clustering around $\lambda_{\max}$ is generic

Convergence: the better the larger $\frac{\lambda_{\max} - \lambda_2}{\lambda_{\max} - \lambda_{\min}}$ [ignore multiplicities]

trouble: in eigenvalue optimization clustering around $\lambda_{\max}$ is generic

---

Spectral transformation: apply Lanczos to $p(A)$ to increase $\frac{\lambda_{\max} - \lambda_2}{\lambda_{\max} - \lambda_{\min}}$

- compute polynomial $p(\cdot)$ by matrix vector multiplications
- but: Lanczos provides best polynomial

$\longrightarrow$ tradeoff: cost of polynomial to cost of orthogonalization

Convergence: the better the larger $\frac{\lambda_{\max} - \lambda_2}{\lambda_{\max} - \lambda_{\min}}$ [ignore multiplicities]

trouble: in eigenvalue optimization clustering around $\lambda_{\max}$ is generic

---

Spectral transformation: apply Lanczos to $p(A)$ to increase $\frac{\lambda_{\max} - \lambda_2}{\lambda_{\max} - \lambda_{\min}}$

- compute polynomial $p(\cdot)$ by matrix vector multiplications
- but: Lanczos provides best polynomial

$\rightarrow$ tradeoff: cost of polynomial to cost of orthogonalization

---

Inexact evaluation for null steps

- $\lambda_{\max}(T_i) \uparrow \lambda_{\max}(A)$
- before each restart check whether $\lambda_{\max}(T_i)$ ensures null step.
  if yes $\rightarrow$ STOP

Convergence: the better the larger $\frac{\lambda_{\max}-\lambda_2}{\lambda_{\max}-\lambda_{\min}}$ [ignore multiplicities]

trouble: in eigenvalue optimization clustering around $\lambda_{\max}$ is generic

---

Spectral transformation: apply Lanczos to $p(A)$ to increase $\frac{\lambda_{\max}-\lambda_2}{\lambda_{\max}-\lambda_{\min}}$

- compute polynomial $p(\cdot)$ by matrix vector multiplications
- but: Lanczos provides best polynomial

$\to$ tradeoff: cost of polynomial to cost of orthogonalization

---

Inexact evaluation for null steps

- $\lambda_{\max}(T_i) \uparrow \lambda_{\max}(A)$
- before each restart check whether $\lambda_{\max}(T_i)$ ensures null step.
  if yes $\to$ STOP

Lanczos (Ritz-)vectors:

- $L =$ eigenvectors of $Q_i T_i Q_i^T$ [usually "Ritz vectors"]
- at exit n_L available
- often good estimates for large eigenvalues of A

$\to$ valuable for forming the bundle P

- for each eigenvalue of A, L holds at most one Ritz vector

The Bundle Update: $P, \overline{W}, L \;\rightarrow\; P_+, \overline{W}_+$

$\widehat{\mathcal{W}}_+$ must contain W_+ and vv^T for convergence.

Solving (QSP) with an interior point code yields

$$W_+ = PU_+P^T + \alpha_+\overline{W}$$

Keep "important" eigenspace of PU_+P^T in the bundle.

The Bundle Update: $P, \overline{W}, L \;\longrightarrow\; P_+, \overline{W}_+$

$\widehat{\mathcal{W}}_+$ must contain W_+ and vv^T for convergence.

Solving (QSP) with an interior point code yields

$$W_+ = PU_+P^T + \alpha_+\overline{W}$$

Keep "important" eigenspace of PU_+P^T in the bundle.

$$U_+ = [Q_1 Q_2]\begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}[Q_1 Q_2]^T$$

Q_1 holds at most n_K eigenvectors to large eigenvalues of U_+,
where "large" means $\lambda_{\min}(\Lambda_1) \geq t\lambda_{\max}(U_+)$ for some $t > 0$.

The Bundle Update: $P, \overline{W}, L \;\rightarrow\; P_+, \overline{W}_+$

$\widehat{\mathcal{W}}_+$ must contain W_+ and vv^T for convergence.

Solving (QSP) with an interior point code yields

$$W_+ = PU_+P^T + \alpha_+\overline{W}$$

Keep "important" eigenspace of PU_+P^T in the bundle.

$$U_+ = [Q_1 Q_2]\begin{bmatrix}\Lambda_1 & 0\\ 0 & \Lambda_2\end{bmatrix}[Q_1 Q_2]^T$$

Q_1 holds at most n_K eigenvectors to large eigenvalues of U_+, where "large" means $\lambda_{\min}(\Lambda_1) \geq t\lambda_{\max}(U_+)$ for some $t > 0$.

$$W_+ = PQ_1\Lambda_1 Q_1^T P^T + \underbrace{PQ_2\Lambda_2 Q_2^T P^T + \alpha_+\overline{W}}_{\rightarrow\,\overline{W}_+}$$

- keep subspace spanned by PQ_1 in the bundle
- add subspace of some n_A Lanczos vectors with largest Ritz values

---

$$P^+ = \operatorname{orth}([PQ_1, L]) \qquad \overline{W}^+ = \frac{PQ_2\Lambda_2(PQ_2)^T + \alpha^+\overline{W}}{\operatorname{tr}\Lambda_2 + \alpha^+}$$

Computer Session Thursday, 11:00-12:30

- C++ callable library ConicBundle
  (see "Software" on my home page)
- begin with explaining a given code for the max-cut relaxation
- you will then be asked to extend it to equipartition/bisection
- finally, all participants will be asked to choose some related combinatorial relaxation and to try to implement it on their own or to extract primal information for rounding.

Please participate only, if you like to implement things and to play around with optimization codes!

Overview

Box Constraints for Bundle Methods

Frequently some variables of $y \in \mathbb{R}^n$ are sign constrained (e.g., as dual variables to inequality constraints) or constrained to intervals.
For one technique to deal with this, consider the simplified scenario

$$\min_{y \in \mathbb{R}^m_+} f(y) := \sup_{(\gamma, g) \in \mathcal{M}} \gamma + \langle g, y \rangle$$

Box Constraints for Bundle Methods

Frequently some variables of $y \in \mathbb{R}^n$ are sign constrained (e.g., as dual variables to inequality constraints) or constrained to intervals.
For one technique to deal with this, consider the simplified scenario

$$\min_{y \in \mathbb{R}^m_+} f(y) := \sup_{(\gamma,g)\in\mathcal{M}} \gamma + \langle g, y\rangle$$

Extend f to $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ by setting

$$f(y) := \sup_{(\gamma,g)\in\mathcal{M},\eta\in\mathbb{R}^+} \gamma + \langle g - \eta, y\rangle \qquad (y \in \mathbb{R}^m)$$

Box Constraints for Bundle Methods

Frequently some variables of $y \in \mathbb{R}^n$ are sign constrained (e.g., as dual variables to inequality constraints) or constrained to intervals.
For one technique to deal with this, consider the simplified scenario

$$\min_{y \in \mathbb{R}^m_+} f(y) := \sup_{(\gamma,g)\in\mathcal{M}} \gamma + \langle g, y\rangle$$

Extend f to $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ by setting

$$f(y) := \sup_{(\gamma,g)\in\mathcal{M},\eta\in\mathbb{R}^+} \gamma + \langle g-\eta, y\rangle \qquad (y \in \mathbb{R}^m)$$

For a compact convex model $\widehat{\mathcal{M}} \subseteq \mathcal{M}$ the QP subproblem still satisfies

$$\inf_{y\in\mathbb{R}^m} \sup_{(\gamma,g)\in\widehat{\mathcal{M}},\eta\in\mathbb{R}^+} \gamma + \langle g-\eta, y\rangle + \frac{u}{2}\|y-\hat{y}\|^2 =$$

$$= \sup_{(\gamma,g)\in\widehat{\mathcal{M}},\eta\in\mathbb{R}^+} \inf_{y\in\mathbb{R}^m} \gamma + \langle g-\eta, y\rangle + \frac{u}{2}\|y-\hat{y}\|^2$$

Solve the inner problem for y: $\quad y^+((\gamma,g),\eta) = \hat{y} - \frac{1}{u}(g-\eta)$
but the resulting QP in (γ, g) and η might be expensive to solve.

Gauss-Seidel for Box-Constraints [H.,Kiwiel2002]

Instead of directly solving

$$\sup_{(\gamma,g)\in\widehat{\mathcal{M}},\eta\in\mathbb{R}^+} \gamma + \langle g-\eta, \hat{y}\rangle - \frac{1}{2u}\|g-\eta\|^2$$

note that for fixed (γ, g) finding optimal $\eta \geq 0$ is easy,

$$\eta_{\max}(g) := \max\{0, g - u\hat{y}\}$$

Gauss-Seidel for Box-Constraints [H.,Kiwiel2002]

Instead of directly solving

$$\sup_{(\gamma,g)\in\widehat{\mathcal{M}},\eta\in\mathbb{R}^+} \gamma + \langle g-\eta,\hat{y}\rangle - \frac{1}{2u}\|g-\eta\|^2$$

note that for fixed (γ, g) finding optimal $\eta \geq 0$ is easy,

$$\eta_{\max}(g) := \max\{0, g - u\hat{y}\}$$

yielding $\quad y^+ = y_{\min}((\gamma,g),\eta_{\max}(g)) \geq 0 \quad$ with $\quad \langle y^+, \eta_{\max}(g)\rangle = 0.$

Gauss-Seidel for Box-Constraints [H.,Kiwiel2002]

Instead of directly solving

$$\sup_{(\gamma,g)\in\widehat{\mathcal{M}},\eta\in\mathbb{R}^+} \gamma + \langle g-\eta,\hat{y}\rangle - \tfrac{1}{2u}\|g-\eta\|^2$$

note that for fixed (γ, g) finding optimal $\eta \geq 0$ is easy,

$$\eta_{\max}(g) := \max\{0, g - u\hat{y}\}$$

yielding $y^+ = y_{\min}((\gamma, g), \eta_{\max}(g)) \geq 0$ with $\langle y^+, \eta_{\max}(g)\rangle = 0$.

Starting with some $(\gamma^+, g^+) \in \widehat{\mathcal{M}}$, set $\eta^+ = \eta_{\max}(g^+)$ and iterate:

(a) For fixed η^+ find $(\gamma^+, g^+) \in \mathrm{Argmax}(QP(\eta^+))$

[as in the unconstrained case]

Gauss-Seidel for Box-Constraints [H.,Kiwiel2002]

Instead of directly solving

$$\sup_{(\gamma,g)\in\widehat{\mathcal{M}},\eta\in\mathbb{R}^+} \gamma + \langle g-\eta,\hat{y}\rangle - \frac{1}{2u}\|g-\eta\|^2$$

note that for fixed (γ, g) finding optimal $\eta \geq 0$ is easy,

$$\eta_{\max}(g) := \max\{0, g - u\hat{y}\}$$

yielding $y^+ = y_{\min}((\gamma,g),\eta_{\max}(g)) \geq 0$ with $\langle y^+, \eta_{\max}(g)\rangle = 0$.

Starting with some $(\gamma^+, g^+) \in \widehat{\mathcal{M}}$, set $\eta^+ = \eta_{\max}(g^+)$ and iterate:

(a) For fixed η^+ find $(\gamma^+, g^+) \in \text{Argmax}(QP(\eta^+))$

[as in the unconstrained case]

(b) Set $\eta^+ \leftarrow \eta_{\max}(g^+)$ and $y^+ \leftarrow y_{\min}((\gamma^+, g^+), \eta^+)$

until the error

$$f_{\widehat{\mathcal{M}}}(y^+) - f_{(\gamma^+,g^+)}(y^+) < \kappa_M[f(\hat{y}) - f_{(\gamma^+,g^+)}(y^+)]$$

is small for some $\kappa_M > 0$.

[converges, because $((\gamma^+, g^+), \eta^+)$ serves as aggregate of the model]

The Algorithm for Nonnegative Variables

Input: $y_0 = \hat{y}_1$, some $(\gamma_0^+, g_0^+) \in \widehat{\mathcal{M}}_1$, $\kappa \in (0,1)$, $\kappa_M > 0$, $\varepsilon > 0$, $k = 1$.

1. (Candidate finding) Set $\eta^+ = \eta^k_{\max}(g^+_{k-1})$.
   (a) For fixed η^+ find $(\gamma^+, g^+) \in \operatorname{Argmax}(QP_k(\eta^+))$.
   (b) Set $\eta^+ \leftarrow \eta^k_{\max}(g^+)$ and $y^+ \leftarrow y^k_{\min}((\gamma^+, g^+), \eta^+)$.
   (c) If $f(\hat{y}_k) - f_{(\gamma^+, g^+)}(y_k) < \varepsilon(|f(\hat{y}_k)| + 1)$ then **stop**.
   (d) If $f_{\widehat{\mathcal{M}}_k}(y^+) - f_{(\gamma^+, g^+)}(y^+) < \kappa_M[f(\hat{y}) - f_{(\gamma^+, g^+)}(y^+)]$ goto (a).
   (e) Set $y_k = y^+$, $(\gamma_k^+, g_k^+) = (\gamma^+, g^+)$, $\eta_k^+ = \eta^+$.

The Algorithm for Nonnegative Variables

Input: $y_0 = \hat{y}_1$, some $(\gamma_0^+, g_0^+) \in \widehat{\mathcal{M}}_1$, $\kappa \in (0,1)$, $\kappa_M > 0$, $\varepsilon > 0$, $k = 1$.

1. (Candidate finding) Set $\eta^+ = \eta^k_{\max}(g^+_{k-1})$.
   (a) For fixed η^+ find $(\gamma^+, g^+) \in \mathrm{Argmax}(QP_k(\eta^+))$.
   (b) Set $\eta^+ \leftarrow \eta^k_{\max}(g^+)$ and $y^+ \leftarrow y^k_{\min}((\gamma^+, g^+), \eta^+)$.
   (c) If $f(\hat{y}_k) - f_{(\gamma^+,g^+)}(y_k) < \varepsilon(|f(\hat{y}_k)| + 1)$ then **stop**.
   (d) If $f_{\widehat{\mathcal{M}}_k}(y^+) - f_{(\gamma^+,g^+)}(y^+) < \kappa_M[f(\hat{y}) - f_{(\gamma^+,g^+)}(y^+)]$ goto (a).
   (e) Set $y_k = y^+$, $(\gamma_k^+, g_k^+) = (\gamma^+, g^+)$, $\eta_k^+ = \eta^+$.
2. Compute $f(y_k)$ and subgradient g_k^s, yields also γ_k^s.
3. If $f(\hat{y}_k) - f(y_k) > \kappa[f(\hat{y}_k) - f_{(\gamma_k^+,g_k^+)}(y_k)]$
   then *descent step*: set $\hat{y}_{k+1} = y_k$,
   else *null step*: $\hat{y}_{k+1} = \hat{y}_k$ unchanged.
4. Find a new model so that $\{(\gamma_k^+, g_k^+), (\gamma_k^s, g_k^s)\} \subseteq \widehat{\mathcal{M}}_{k+1}$.
   Update the weight u, set $k \leftarrow k+1$, **goto** 1.

Theorem. Let $\varepsilon = 0$ then the sequence of descent steps $\{\hat{y}_k\}$ satisfies $f(\hat{y}_k) \to \inf_{y \geq 0} f$ and (plus some conditions) $g_k^+ - \eta_k^+ \to 0$.
[in fact, doing (a) and (b) just once suffices for convergence]

Overview

Primal Aggregation in Lagrangian Relaxation [folklore]

Bundle methods are often employed for solving Lagrangian relaxations of linear constraints,

$$\begin{array}{ll} \max & c^T x \\ \text{s.t.} & Ax \le b \\ & x \in \operatorname{conv} \Omega \end{array} \quad \Leftrightarrow \quad \max_{x \in \operatorname{conv} \Omega} c^T x + \inf_{y \ge 0} (b - Ax)^T y$$

Primal Aggregation in Lagrangian Relaxation [folklore]

Bundle methods are often employed for solving Lagrangian relaxations of linear constraints,

$$\begin{array}{ll} \max & c^T x \\ \text{s.t.} & Ax \leq b \\ & x \in \operatorname{conv} \Omega \end{array} \qquad \Leftrightarrow \qquad \max_{x \in \operatorname{conv} \Omega} c^T x + \inf_{y \geq 0} (b - Ax)^T y$$

No duality gap under a regularity assumption (e.g., $\operatorname{conv} \Omega$ compact):

$$\min_{y \geq 0} f(y) := b^T y + \max_{x \in \Omega} (c - A^T y)^T x$$

Primal Aggregation in Lagrangian Relaxation [folklore]

Bundle methods are often employed for solving Lagrangian relaxations of linear constraints,

$$\begin{array}{ll} \max & c^T x \\ \text{s.t.} & Ax \leq b \\ & x \in \operatorname{conv} \Omega \end{array} \quad \Leftrightarrow \quad \max_{x \in \operatorname{conv} \Omega} c^T x + \inf_{y \geq 0} (b - Ax)^T y$$

No duality gap under a regularity assumption (e.g., $\operatorname{conv} \Omega$ compact):

$$\min_{y \geq 0} f(y) := b^T y + \max_{x \in \Omega} (c - A^T y)^T x$$

Bundle methods generate $y_k \to y_* \in \operatorname{Argmin}_{y \geq 0} f(y) \quad (\text{if} \neq \emptyset)$,
but what about x^*?

Primal Aggregation in Lagrangian Relaxation [folklore]

Bundle methods are often employed for solving Lagrangian relaxations of linear constraints,

$$\begin{array}{ll} \max & c^T x \\ \text{s.t.} & Ax \le b \\ & x \in \operatorname{conv}\Omega \end{array} \quad \Leftrightarrow \quad \max_{x \in \operatorname{conv}\Omega} c^T x + \inf_{y \ge 0} (b - Ax)^T y$$

No duality gap under a regularity assumption (e.g., $\operatorname{conv}\Omega$ compact):

$$\min_{y \ge 0} f(y) := b^T y + \max_{x \in \Omega} (c - A^T y)^T x$$

Bundle methods generate $y_k \to y_* \in \operatorname{Argmin}_{y \ge 0} f(y) \quad (\text{if} \neq \emptyset)$,
but what about x^*?

Evaluating $f(y_k)$ requires solving $\max_{x \in \Omega} (c - A^T y)^T x$ and yields

$$x_k^s \in \operatorname*{Argmax}_{x \in \Omega} (c - A^T y)^T x$$

$$\gamma_k^s = c^T x_k^s$$

$$g_k^s = b - Ax_k^s$$

Quadratic Subproblem for $\widehat{\mathcal{M}} = \{(\gamma_1, g_1), \dots, (\gamma_{h_k}, g_{h_k})\}$

$$\min_{y \geq 0} \max_{(\gamma_i, g_i) \in \widehat{\mathcal{M}}, \eta \geq 0} \gamma_i + (g_i - \eta)^T y + \frac{1}{2}\|y - \hat{y}\|^2$$

equivalently (for fixed $\eta \geq 0$)

$$\begin{array}{ll} \max & \sum \xi_i(\gamma_i + (g_i - \eta)^T \hat{y}) - \frac{1}{2}\left\|\sum \xi_i g_i - \eta\right\|^2 \\ \text{s.t.} & \xi^T e = 1 \\ & \xi \geq 0. \end{array}$$

Need only two: $(\gamma^+, g^+) = \sum \xi_i^+ (\gamma_i, g_i)$ and the new (γ^s, g^s)

Theorem

If $\operatorname{Argmin} f \neq \emptyset$ *(and ++),*
the proximal bundle method yields $(\sum \xi_i g_i - \eta) \to 0$ *and* $\sum \xi_i \gamma_i \to f_*$.

In Lagrangian relaxation $\gamma_i = c^T x_i$, $g_i = b - Ax_i$ for $x_i \in \Omega$ (or $\operatorname{conv} \Omega$)

$$\begin{array}{rcll} \sum \xi_i g_i - \eta & = & b - A(\sum \xi_i x_i) - \eta & \to 0 \qquad [\eta \geq 0 \text{ slacks}] \\ & & c^T(\sum \xi_i x_i) & \to f_* \end{array}$$

Accumulation points of $\sum \xi_i^k x_i^k$ (++) are optimal solutions (for $\operatorname{conv} \Omega$)

Quadratic Subproblem for convex compact $\widehat{\Omega}$ (e.g., $\widehat{\mathcal{W}}$ for SDP)

$$\begin{array}{ll} \max & c^T x + (b - Ax - \eta)^T \hat{y} - \frac{1}{2} \|b - Ax - \eta\|^2 \\ \text{s.t.} & x \in \widehat{\Omega} \end{array}$$

Need only two in the next $\widehat{\Omega}_+$:

- the aggregate solution $x^+ \in \widehat{\Omega}$
- and a new $x^s \in \Omega$ supplied by the oracle

Primal Approximation in Lagrangian Relaxation:
Theorem $\Rightarrow$ for an appropriate subsequence

$$\begin{aligned} b - Ax_k^+ - \eta \quad &\to 0 \\ c^T x_k^+ \quad &\to f_* \end{aligned}$$

Accumulation points of x_k^+ $(++)$ are optimal solutions (for $\operatorname{conv} \Omega$)

Primal Aggregation for Large Scale SDPs

- $W_+ = PU_+P^T + \alpha_+\overline{W} \to X_*$

For huge X storing $\overline{W}$ in full may be too expensive, but

Primal Aggregation for Large Scale SDPs

- $W_+ = PU_+P^T + \alpha_+\overline{W} \to X_*$

For huge X storing $\overline{W}$ in full may be too expensive, but

- by the bundle update rule, α_+ is mostly small and PU_+P^T may suffice,

Primal Aggregation for Large Scale SDPs

- $W_+ = PU_+P^T + \alpha_+\overline{W} \to X_*$

For huge X storing $\overline{W}$ in full may be too expensive, but

- by the bundle update rule, α_+ is mostly small and PU_+P^T may suffice,
- aggregation of $\overline{W}$ may be restricted to the required support exclusively

Primal Aggregation for Large Scale SDPs

- $W_+ = PU_+P^T + \alpha_+\overline{W} \to X_*$

For huge X storing $\overline{W}$ in full may be too expensive, but

- by the bundle update rule, α_+ is mostly small and PU_+P^T may suffice,
- aggregation of $\overline{W}$ may be restricted to the required support exclusively
- the bundle method does not need $\overline{W}$, but only $\langle C, \overline{W} \rangle$ and $\mathcal{A}\overline{W}$

Primal Aggregation for Large Scale SDPs

- $W_+ = PU_+P^T + \alpha_+\overline{W} \to X_*$

For huge X storing $\overline{W}$ in full may be too expensive, but

- by the bundle update rule, α_+ is mostly small and PU_+P^T may suffice,
- aggregation of $\overline{W}$ may be restricted to the required support exclusively
- the bundle method does not need $\overline{W}$, but only $\langle C, \overline{W} \rangle$ and $\mathcal{A}\overline{W}$

The quadratic semidefinite subproblem

$$\begin{array}{ll} \max & -\frac{1}{2}\begin{bmatrix}\operatorname{svec} U\\ \alpha\end{bmatrix}^T \begin{bmatrix}Q_{11} & q_{12}\\ q_{12}^T & q_{22}\end{bmatrix}\begin{bmatrix}\operatorname{svec} U\\ \alpha\end{bmatrix} + \begin{bmatrix}c_1\\ c_2\end{bmatrix}^T\begin{bmatrix}\operatorname{svec} U\\ \alpha\end{bmatrix} + d\\ \text{s.t.} & \alpha + \operatorname{tr} U = 1\\ & \alpha \ge 0, U \succeq 0 \end{array}$$

where

$$Q_{11} = \frac{1}{u}\sum_{i=1}^{m} \operatorname{svec}(P^T A_i P)\operatorname{svec}(P^T A_i P)^T \qquad c_1 = \operatorname{svec}(P^T[\mathcal{A}^T(\tfrac{1}{u}b - \hat{y}) + C]P)$$

$$q_{12} = \tfrac{1}{u}\operatorname{svec}(P^T\mathcal{A}^T(\mathcal{A}\overline{W})P) \qquad c_2 = (\langle \tfrac{1}{u}b - \hat{y}, \mathcal{A}\overline{W}\rangle + \langle C, \overline{W}\rangle)$$

$$q_{22} = \tfrac{1}{u}\langle \mathcal{A}\overline{W}, \mathcal{A}\overline{W}\rangle \qquad d = \langle b, \hat{y} - \tfrac{1}{2u}b\rangle$$

Overview

Dynamic Bundle Methods [H. 2004]

If Lagrangian relaxation is applied to a primal cutting plane approach,

$$\begin{array}{ll} \max & c^T x \\ \text{s.t.} & Ax \le b \\ & x \in \operatorname{conv}\Omega \end{array} \qquad \Leftrightarrow \qquad \max_{x \in \operatorname{conv}\Omega} c^T x + \inf_{y \ge 0} (b - Ax)^T y$$

then $Ax \le b$ is constantly changing, so the dimension of the dual problem changes as well → dynamic bundles methods [BelloniSagastizabal2009]

Dynamic Bundle Methods [H. 2004]

If Lagrangian relaxation is applied to a primal cutting plane approach,

$$\begin{array}{ll} \max & c^T x \\ \text{s.t.} & Ax \leq b \\ & x \in \operatorname{conv}\Omega \end{array} \quad \Leftrightarrow \quad \max_{x \in \operatorname{conv}\Omega} c^T x + \inf_{y \geq 0} (b - Ax)^T y$$

then $Ax \leq b$ is constantly changing, so the dimension of the dual problem changes as well $\rightarrow$ dynamic bundles methods [BelloniSagastizabal2009]

Key idea: separate with respect to the current aggregate x_k^+

Dynamic Bundle Methods [H. 2004]

If Lagrangian relaxation is applied to a primal cutting plane approach,

$$\begin{array}{ll} \max & c^T x \\ \text{s.t.} & Ax \leq b \\ & x \in \operatorname{conv} \Omega \end{array} \quad \Leftrightarrow \quad \max_{x \in \operatorname{conv} \Omega} c^T x + \inf_{y \geq 0} (b - Ax)^T y$$

then $Ax \leq b$ is constantly changing, so the dimension of the dual problem changes as well → dynamic bundles methods [BelloniSagastizabal2009]

Key idea: separate with respect to the current aggregate x_k^+

Difficulties:

- x^+ is 'never' feasible for all given constraints

→ the same inequalities may be separated again and again

→ separation routines can 'conceal' certain violated inequalities

Dynamic Bundle Methods [H. 2004]

If Lagrangian relaxation is applied to a primal cutting plane approach,

$$\begin{array}{ll} \max & c^T x \\ \text{s.t.} & Ax \le b \\ & x \in \operatorname{conv}\Omega \end{array} \quad \Leftrightarrow \quad \max_{x \in \operatorname{conv}\Omega} c^T x + \inf_{y \ge 0} (b - Ax)^T y$$

then $Ax \le b$ is constantly changing, so the dimension of the dual problem changes as well → dynamic bundles methods [BelloniSagastizabal2009]

Key idea: separate with respect to the current aggregate x_k^+

Difficulties:

- x^+ is 'never' feasible for all given constraints

→ the same inequalities may be separated again and again

→ separation routines can 'conceal' certain violated inequalities

---

What kind of separation oracle do we need?

Is it still possible to guarantee convergence to the optimal solution?

Maximum violation oracle with respect to $Ax \leq b$:

- returns inequalities from a finite inequality system

$$a_i^T x \leq b_i, \quad i \in \{1, \dots, m\}$$

- for a given x^+ the oracle either
  - asserts feasibility of x^+, or
  - returns an inequality $j \in \{1, \dots, m\}$ with

$$b_j - a_j^T x^+ \ \leq \ \min_i \ b_i - a_i^T x^+ \ < 0.$$

[many separation routines satisfy this]

Maximum violation oracle with respect to $Ax \leq b$:

- returns inequalities from a finite inequality system

$$a_i^T x \leq b_i, \quad i \in \{1, \ldots, m\}$$

- for a given x^+ the oracle either
  - asserts feasibility of x^+, or
  - returns an inequality $j \in \{1, \ldots, m\}$ with

$$b_j - a_j^T x^+ \; \leq \; \min_i \; b_i - a_i^T x^+ \; < 0.$$

[many separation routines satisfy this]

Cutting plane algorithm 1

[e.g., for max $\langle C, X \rangle$ s.t. $X \in \{X \succeq 0 : \langle I, X \rangle = a\} \cap \{X : \mathcal{A}X \leq b\}$]

1. Solve quadratic model $\longrightarrow x^+$
   If oracle(x^+) returns a new inequality, add it and go to 1
2. Evaluate function, determine subgradient
3. Decide on
   - null step
   - descent step
4. Update model and iterate

Theorem. If the primal problem (for all m constraints) has an optimal solution then the algorithm converges to an optimal solution and generates a subsequence $K \subseteq \mathbb{N}$ so that all cluster points of x_k^+, $k \in K$, are primal optimal solutions.

Theorem. If the primal problem (for all m constraints) has an optimal solution then the algorithm converges to an optimal solution and generates a subsequence $K \subseteq \mathbb{N}$ so that all cluster points of x_k^+, $k \in K$, are primal optimal solutions.

Proof idea:

1. Wait till the oracle adds no more inequalities to index set J (finite)
2. Apply convergence theorem to problem specified by subsystem J

   $\Rightarrow$ there is subsequence K with $x_k^+ \to x_J^*$ feasible and optimal for J

   $\Rightarrow$ violation $\to 0$ on inequalities J

   Maximum violation oracle $\Rightarrow$ all are satisfied for x_J^*

Theorem. If the primal problem (for all m constraints) has an optimal solution then the algorithm converges to an optimal solution and generates a subsequence $K \subseteq \mathbb{N}$ so that all cluster points of x_k^+, $k \in K$, are primal optimal solutions.

Proof idea:

1. Wait till the oracle adds no more inequalities to index set J (finite)
2. Apply convergence theorem to problem specified by subsystem J

   $\Rightarrow$ there is subsequence K with $x_k^+ \to x_J^*$ feasible and optimal for J

   $\Rightarrow$ violation $\to 0$ on inequalities J

   Maximum violation oracle $\Rightarrow$ all are satisfied for x_J^*

Is it possible to eliminate inactive inequalities during runtime?

Cutting plane algorithm 2

[e.g., for max $\langle C, X \rangle$ s.t. $X \in \{X \succeq 0 : \langle I, X \rangle = a\} \cap \{X : \mathcal{A}X \leq b\}$]

1. Solve quadratic model $\longrightarrow x^+$

   If oracle(x^+) returns a new inequality, add it and go to 1
2. Evaluate function, determine subgradient
3. Decide on
   - null step
   - descent step: delete inequalities inactive for x^+
4. Update model and iterate

Cutting plane algorithm 2

[e.g., for max $\langle C, X\rangle$ s.t. $X \in \{X \succeq 0 : \langle I, X\rangle = a\} \cap \{X : \mathcal{A}X \leq b\}$]

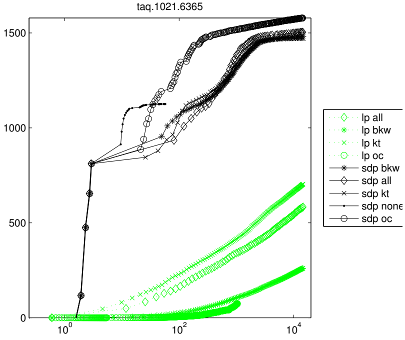
1. Solve quadratic model $\longrightarrow x^+$

   If oracle(x^+) returns a new inequality, add it and go to 1
2. Evaluate function, determine subgradient
3. Decide on
   - null step
   - descent step: delete inequalities inactive for x^+
4. Update model and iterate

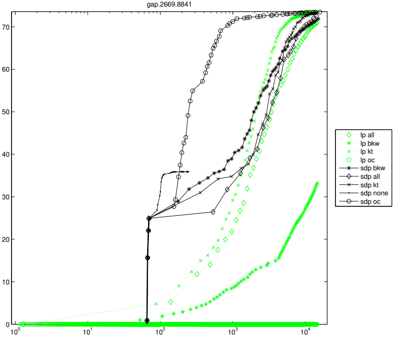
Theorem. If the primal has a strictly feasible solution then the upper bound converges to the optimal value and the algorithm generates a subsequence $K \subseteq \mathbb{N}$ so that all cluster points of x_k^+, $k \in K$, are primal optimal solutions.

The strictly feasible primal solution ensures boundedness of dual iterates

Minimum Bisection Relaxation, LP vs. SDP [AFHM2008]



1021 nodes, 6365 edges



2669 nodes, 8841 edges

Overview

Second Order Approaches

[Overton8*, OvertonWomersley95, Oustry200*]

Local quadratic convergence for correct multiplicity t in the optimum y^*,

$$C - \mathcal{A}^T y^* = [Q_1^* Q_2^*] \begin{bmatrix} \Lambda_1^* & 0 \\ 0 & \Lambda_2^* \end{bmatrix} [Q_1^* Q_2^*]^T$$

$$\lambda_1^* = \cdots = \lambda_t^* > \lambda_{t+1}^* > \cdots > \lambda_n^*$$

Second Order Approaches

[Overton8*, OvertonWomersley95, Oustry200*]
Local quadratic convergence for correct multiplicity t in the optimum y^*,

$$C - \mathcal{A}^T y^* = [Q_1^* Q_2^*] \begin{bmatrix} \Lambda_1^* & 0 \\ 0 & \Lambda_2^* \end{bmatrix} [Q_1^* Q_2^*]^T$$

$$\lambda_1^* = \cdots = \lambda_t^* > \lambda_{t+1}^* > \cdots > \lambda_n^*$$

1. Guess t_k, compute Q_1^k, Q_2^k and an interior subgradient U_k by

$$\min \|b - \mathcal{A} Q_1 U Q_1^T\|^2 \text{ s.t. } \operatorname{tr} U = 1, \; U \succeq 0$$

Second Order Approaches

[Overton8*, OvertonWomersley95, Oustry200*]

Local quadratic convergence for correct multiplicity t in the optimum y^*,

$$C - \mathcal{A}^T y^* = [Q_1^* Q_2^*] \begin{bmatrix} \Lambda_1^* & 0 \\ 0 & \Lambda_2^* \end{bmatrix} [Q_1^* Q_2^*]^T$$

$$\lambda_1^* = \cdots = \lambda_t^* > \lambda_{t+1}^* > \cdots > \lambda_n^*$$

1. Guess t_k, compute Q_1^k, Q_2^k and an interior subgradient U_k by

$$\min \|b - \mathcal{A} Q_1 U Q_1^T\|^2 \text{ s.t. } \operatorname{tr} U = 1,\ U \succeq 0$$

2. Compute the Newton candidate by solving

$$\begin{array}{ll} \min & \frac{1}{2}\|y - \hat{y}_k\|_{H_k}^2 + \langle b, y\rangle + \delta \\ \text{s.t.} & \delta I = Q_1^T (C - \mathcal{A}^T y) Q_1 \end{array}$$

where

$$H_k = 2\mathcal{A}\left((Q_1 U_k Q_1^T) \otimes (Q_2 [\lambda_1^k I - \Lambda_2^k]^{-1} Q_2^T) \right) \mathcal{A}^T \qquad \text{[regularity } \succ 0]$$

Second Order Approaches

[Overton8*, OvertonWomersley95, Oustry200*]

Local quadratic convergence for correct multiplicity t in the optimum y^*,

$$C - \mathcal{A}^T y^* = [Q_1^* Q_2^*] \begin{bmatrix} \Lambda_1^* & 0 \\ 0 & \Lambda_2^* \end{bmatrix} [Q_1^* Q_2^*]^T$$

$$\lambda_1^* = \cdots = \lambda_t^* > \lambda_{t+1}^* > \cdots > \lambda_n^*$$

1. Guess t_k, compute Q_1^k, Q_2^k and an interior subgradient U_k by

$$\min \|b - \mathcal{A} Q_1 U Q_1^T\|^2 \text{ s.t. } \operatorname{tr} U = 1,\ U \succeq 0$$

2. Compute the Newton candidate by solving

$$\begin{array}{ll} \min & \frac{1}{2}\|y - \hat{y}_k\|^2_{H_k} + \langle b, y \rangle + \delta \\ \text{s.t.} & \delta I = Q_1^T (C - \mathcal{A}^T y) Q_1 \end{array}$$

where

$$H_k = 2\mathcal{A}\left((Q_1 U_k Q_1^T) \otimes (Q_2 [\lambda_1^k I - \Lambda_2^k]^{-1} Q_2^T)\right) \mathcal{A}^T \qquad \text{[regularity } \succ 0\text{]}$$

$$[H_k]_{ij} = 2 \operatorname{tr}[(Q_1^T A_i Q_2) U_k (Q_1^T A_j Q_2)(\lambda_1^k I - \Lambda_2^k)^{-1}]$$

Adaptation of Step 2 for Spectral Bundle [H.RendlOverton]

Step 2 $\begin{array}{ll} \min & \frac{1}{2}\|y-\hat{y}\|_H^2 + \langle b, y\rangle + \delta \\ \text{s.t.} & \delta I = Q_1^T(C - \mathcal{A}^T y)Q_1 \end{array}$ is relaxed to

$$\begin{array}{ll} \min & \frac{1}{2}\|y-\hat{y}\|_H^2 + \langle b, y\rangle + \delta \\ \text{s.t.} & \delta I \succeq Q_1^T(C - \mathcal{A}^T y)Q_1, \end{array} \quad \Rightarrow \quad \delta = \lambda_{\max}(Q_1^T(C - \mathcal{A}^T y)Q_1).$$

Adaptation of Step 2 for Spectral Bundle [H.RendlOverton]

Step 2 $\begin{array}{ll} \min & \frac{1}{2}\|y-\hat{y}\|_H^2 + \langle b, y\rangle + \delta \\ \text{s.t.} & \delta I = Q_1^T(C - \mathcal{A}^T y)Q_1 \end{array}$ is relaxed to

$$\begin{array}{ll} \min & \frac{1}{2}\|y-\hat{y}\|_H^2 + \langle b, y\rangle + \delta \\ \text{s.t.} & \delta I \succeq Q_1^T(C - \mathcal{A}^T y)Q_1, \end{array} \quad \Rightarrow \quad \delta = \lambda_{\max}(Q_1^T(C - \mathcal{A}^T y)Q_1).$$

With $\widehat{\mathcal{W}} := \{Q_1 U Q_1^T : \operatorname{tr} U = 1, U \succeq 0\}$ the problem reads

$$\min_{y} \max_{W \in \widehat{\mathcal{W}}} \left\langle W, C - \mathcal{A}^T y \right\rangle + b^T y + \frac{1}{2}\|y - \hat{y}\|_H^2$$

Adaptation of Step 2 for Spectral Bundle [H.RendlOverton]

Step 2 $\begin{array}{ll} \min & \frac{1}{2}\|y-\hat{y}\|_H^2 + \langle b, y\rangle + \delta \\ \text{s.t.} & \delta I = Q_1^T(C - \mathcal{A}^T y)Q_1 \end{array}$ is relaxed to

$$\begin{array}{ll} \min & \frac{1}{2}\|y-\hat{y}\|_H^2 + \langle b, y\rangle + \delta \\ \text{s.t.} & \delta I \succeq Q_1^T(C - \mathcal{A}^T y)Q_1, \end{array} \quad \Rightarrow \quad \delta = \lambda_{\max}(Q_1^T(C - \mathcal{A}^T y)Q_1).$$

With $\widehat{\mathcal{W}} := \{Q_1 U Q_1^T : \operatorname{tr} U = 1, U \succeq 0\}$ the problem reads

$$\min_y \max_{W \in \widehat{\mathcal{W}}} \left\langle W, C - \mathcal{A}^T y \right\rangle + b^T y + \frac{1}{2}\|y - \hat{y}\|_H^2$$

Dualize, then $y_+(W) = \hat{y} - H^{-1}(b - \mathcal{A}W)$

$$\text{(QSP)} \qquad \begin{array}{ll} \min & \frac{1}{2}\|b - \mathcal{A}W\|_{H^{-1}}^2 - \left\langle W, C - \mathcal{A}^T\hat{y} \right\rangle - \langle b, \hat{y}\rangle \\ \text{s.t.} & W = Q_1 U Q_1^T \\ & \operatorname{tr} U = 1 \\ & U \succeq 0. \end{array}$$

Scope of a second order bundle method

If QSP is solved by an interior point method with t columns, each iteration of QSP requires the factorization of a $\binom{t+1}{2}$ matrix.

For m constraints we can expect $t \approx \sqrt{m}$.
$\rightarrow$ Several $O(m^3)$ operations for each solution of QSP.

Scope of a second order bundle method

If QSP is solved by an interior point method with t columns, each iteration of QSP requires the factorization of a $\binom{t+1}{2}$ matrix.

For m constraints we can expect $t \approx \sqrt{m}$.
$\rightarrow$ Several $O(m^3)$ operations for each solution of QSP.

Typically, a full interior point code requires several $O(n^3)$ and one $O(m^3)$ operation per iteration.

$\rightarrow$ Second order SB is unlikely to be attractive for $m \geq n$, but might be relevant for small $m \leq n$ or if t is small.

$\rightarrow$ Emphasis on large n and rather small m.

Scaling Variants

Scaling Variants

- **No scaling, bounded bundle**

Scaling Variants

- **No scaling, bounded bundle**
- **No scaling, fat bundle**

Scaling Variants

- **No scaling, bounded bundle**
- **No scaling, fat bundle**
- **Modified Newton:** use explicit eigenvalue decomposition, min. norm subgradient, compute full Newton H $(+\rho I)$

Scaling Variants

- **No scaling, bounded bundle**
- **No scaling, fat bundle**
- **Modified Newton:** use explicit eigenvalue decomposition, min. norm subgradient, compute full Newton H $(+\rho I)$
- **Diagonal Newton:** use explicit eigenvalue decomposition, min. norm subgradient, compute diagonal of Newton H $(+\rho I)$

Scaling Variants

- **No scaling, bounded bundle**
- **No scaling, fat bundle**
- **Modified Newton:** use explicit eigenvalue decomposition, min. norm subgradient, compute full Newton H $(+\rho I)$
- **Diagonal Newton:** use explicit eigenvalue decomposition, min. norm subgradient, compute diagonal of Newton H $(+\rho I)$
- **Low-Rank Newton:** collect approximate subspace to large eigenvalues, use subgradient W_+ of (QSP), approximate Newton matrix with available information $(+\rho I)$

Scaling Variants

- **No scaling, bounded bundle**
- **No scaling, fat bundle**
- **Modified Newton:** use explicit eigenvalue decomposition, min. norm subgradient, compute full Newton H $(+\rho I)$
- **Diagonal Newton:** use explicit eigenvalue decomposition, min. norm subgradient, compute diagonal of Newton H $(+\rho I)$
- **Low-Rank Newton:** collect approximate subspace to large eigenvalues, use subgradient W_+ of (QSP), approximate Newton matrix with available information $(+\rho I)$
- **Diagonal Low-Rank:** Collect approximate subspace to large eigenvalues, use subgradient W_+ of (QSP) and the diagonal of approximate Newton matrix $(+\rho I)$

Low Rank Structure

$$H = 2\mathcal{A}\left((Q_1 U Q_1^T) \otimes (Q_2[\lambda_1 I - \Lambda_2]^{-1} Q_2^T)\right)\mathcal{A}^T$$

decompose $U = Q_u \Lambda_u Q_u^T$, set $\bar{Q}_1 = Q_1 Q_u$ and rewrite H as

$$H = 2\mathcal{A}\left((\bar{Q}_1 \otimes Q_2)(\Lambda_u \otimes [\lambda_1 I - \Lambda_2]^{-1})(\bar{Q}_1 \otimes Q_2^T)\right)\mathcal{A}^T$$

Truncate $[\lambda_1 I - \Lambda_2]_{1,\dots,h}$ and $Q_2 \rightarrow Q_h$,

Low Rank Structure

$$H = 2\mathcal{A}\left((Q_1 U Q_1^T) \otimes (Q_2[\lambda_1 I - \Lambda_2]^{-1} Q_2^T)\right)\mathcal{A}^T$$

decompose $U = Q_u \Lambda_u Q_u^T$, set $\bar{Q}_1 = Q_1 Q_u$ and rewrite H as

$$H = 2\mathcal{A}\left((\bar{Q}_1 \otimes Q_2)(\Lambda_u \otimes [\lambda_1 I - \Lambda_2]^{-1})(\bar{Q}_1 \otimes Q_2^T)\right)\mathcal{A}^T$$

Truncate $[\lambda_1 I - \Lambda_2]_{1,\dots,h}$ and $Q_2 \to Q_h$,
compute a QR-decomposition of $\mathcal{A}(\bar{Q}_1 \otimes Q_h) \;\to\; Q_\mathcal{A} R$

$$H_h = 2Q_\mathcal{A} \underbrace{R(\Lambda_u \otimes [\lambda_1 I - \Lambda_2]^{-1}_{1,\dots,h})R^T}_{\to\;\tilde{Q}\Lambda_H\tilde{Q}^T,\; Q_H := Q_\mathcal{A}\tilde{Q}} Q_\mathcal{A}^T$$

truncate $\Lambda_H \to \hat{\Lambda}_H, \hat{Q}_H$

$$\to \quad \hat{H} = \rho I + 2\hat{Q}_H\hat{\Lambda}_H\hat{Q}_H^T$$

for some regularization parameter $\rho > 0$.

Implementation Details

Multiplicity Detection.

Starting with the eigenvalue/vector pair following the maximum eigenvalue of (QSP)-solution $\bar{U}$ we check iteratively

- whether it is smaller than barrier parameter times $\lambda_{\max}(\bar{U})$
- whether the Ritz gap to $\lambda_{\max}(C - \mathcal{A}^T y)$ is big enough
- whether the Ritz gap is reasonable and the value is small compared to its dual value

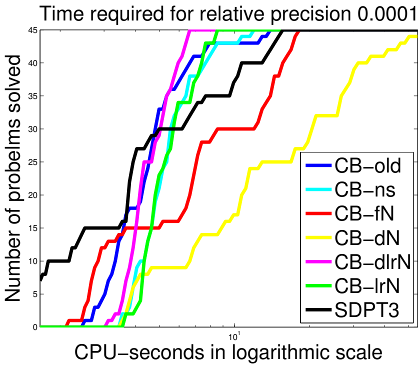
If one of the three criteria holds, this fixes the multiplicity guess t.

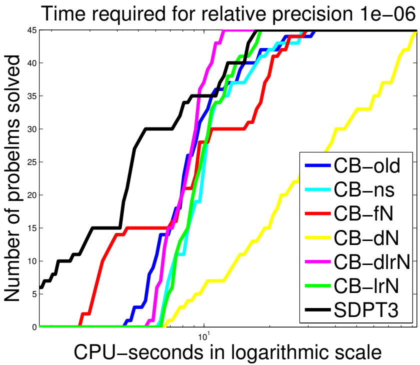
Bundle Update.

After *null steps* we include the new eigenvector, the t top most of U plus some number of the best Ritz vectors orthogonal to this subspace (taken from a collected set of vectors). We use the aggregate.

After *descent steps* we take the t best Ritz vectors into the bundle and enlarge it a bit further if this subspace differs from the old t top most bundle vectors. The aggregate is deleted if H changes.

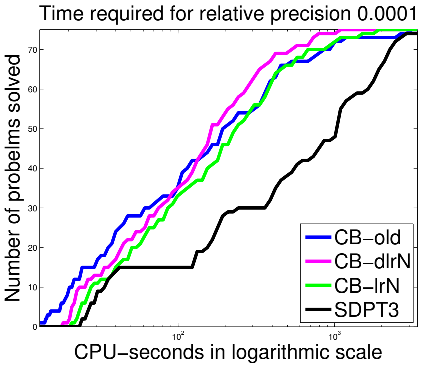
Small Instances: $n \in \{100, 300, 500\}$ and $m = 500$

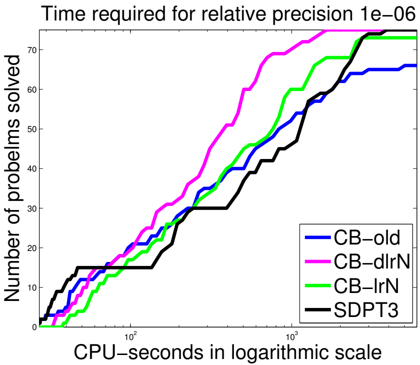




Five instances per choice of n and constraint support order $\in \{3, 5, 7\}$

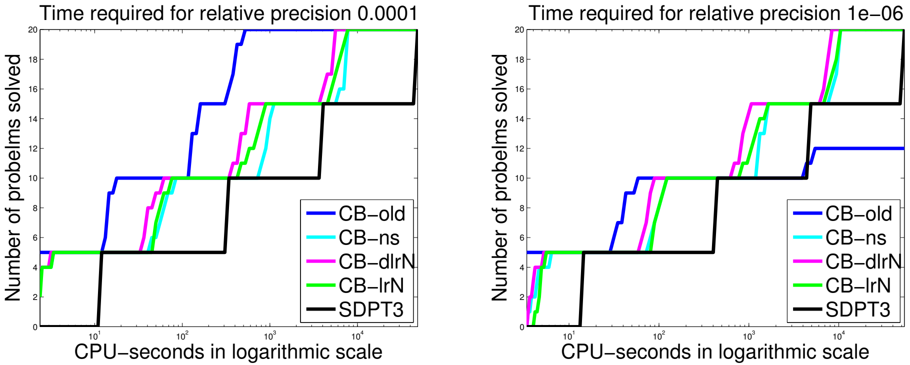
Larger Instances: $n \in \{1, \ldots, 6\} \cdot 1000$ and $m = 1000$





Five instances per choice of n and constraint support order $\in \{3, 4, 5\}$

Max-Cut 3D-Grids: n^3, $n \in \{10, 15, 20, 25\}$



Five instances with random ± 1 edge weights per choice of n

Scaling works well and behaves as expected:

- The number of oracle calls is reduced significantly
  Newton $<$ Low Rank $<$ fat Bundle
- Newton is attractive for small matrices and many constraints, but interior point methods seem preferable.
  [In the end the QSP system is of size $O(m)$.]
- Diagonal low rank scaling is attractive for large matrices and few constraints.
- Scaling allows a relative precision of 10^{-6} routinely with fast initial convergence.
- The cost of solving QSP might be reducible by Toh's approach.

Scaling works well and behaves as expected:

- The number of oracle calls is reduced significantly
  Newton $<$ Low Rank $<$ fat Bundle
- Newton is attractive for small matrices and many constraints, but interior point methods seem preferable.
  [In the end the QSP system is of size $O(m)$.]
- Diagonal low rank scaling is attractive for large matrices and few constraints.
- Scaling allows a relative precision of 10^{-6} routinely with fast initial convergence.
- The cost of solving QSP might be reducible by Toh's approach.

$\rightarrow$ Scaled SB should be a good choice for fast low precision results, cutting plane approaches, or high precision results with large matrices and few constraints.

Thank you for your attention!