Linear Convergence of Modified Frank-Wolfe Methods for Ellipsoid Optimization Problems

Michael J. Todd

Cornell University
Joint work with Damla Ahipasaoglu and Peng Sun

Fields Institute, October 2009

October 27, 2009

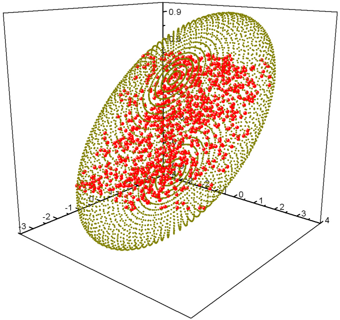
Outline

Linear Convergence of Modified Frank-Wolfe Methods for Ellipsoid Optimization Problems

- Two geometric optimization problems
- Applications
- Duality
- Optimality conditions
- Rank-one update first-order algorithms
- Global convergence
- Linear convergence
- Conclusions

1. Minimum Volume Enclosing Ellipsoids

Given m points $\mathcal{X} := \{x_1, x_2, \ldots, x_m\} \subset \mathbf{R}^n$ which span $\mathbf{R}^n$, the Minimum Volume Enclosing Ellipsoid (MVEE) problem seeks an ellipsoid $E_*(\mathcal{X})$ which is centered at the origin (wlog), covers all the points, and has minimum volume.



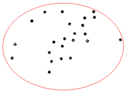
Data Analysis and Computational Geometry

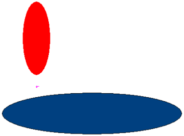
Suppose we are given a finite set S of points in $\mathbf{R}^n$.

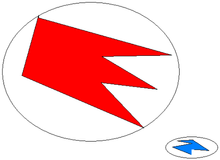
a) Detecting outliers:
choose points far from the center of a minimum volume enclosing ellipsoid.
b) Testing the worth of a cluster $T \subseteq S$:
measure by the volume of $E_*(T)$.
c) Finding a small representative subset $T \subseteq S$:
use a core set (small subset with same minimum-volume ellipsoid) of $E_*(S)$.
In all cases, desirable linear invariance properties.

Sufficient condition for moving bodies S and T in R^n not to hit:

$$E_*(S) \cap E_*(T) = \emptyset.$$







2. Geometry

The set

$$\mathcal{E}(H,\bar{x}) := \{x \in \mathbf{R}^n : (x-\bar{x})^T H (x-\bar{x}) \leq n\}$$

for $\bar{x} \in \mathbf{R}^n$ and $H \succ 0$ is an ellipsoid in $\mathbf{R}^n$ with center $\bar{x}$ and shape defined by H.

We have

$$\mathrm{vol}(\mathcal{E}(H,\bar{x})) = \mathrm{const}(n)/\sqrt{\det H},$$

and minimizing the volume of $\mathcal{E}(H,\bar{x})$ is equivalent to minimizing

$$-\ln\det H.$$

3. MVEE Formulation

The MVEE problem can be formulated as follows:

$$(P) \qquad \begin{array}{rrcl} \min_H & f(H) & := & -\ln\det H \\ & x_i^T H x_i & \leq & n,\ i = 1, \ldots, m, \\ & H & \succ & 0. \end{array}$$

Problem (P) is convex, with linear inequality constraints.

This is also an SDP problem with added centering term.
Interior-point methods can be applied to the problem with barrier function $-\ln\det$ on S^n_{++}.

The LogDet Function

Define f on symmetric $n \times n$ matrices by

$$f(H) := -\ln\det H$$

if H is positive definite, $+\infty$ otherwise.
Note: if $\hat{f}(x) := -\sum \ln x_j$, then $f = \hat{f} \circ \lambda$, with $\lambda(H)$ the vector of eigenvalues of H: this is a spectral function as studied by A. Lewis.

$$Df(H)[E] = -H^{-1} \bullet E := -\mathrm{Tr}(H^{-1}E),$$

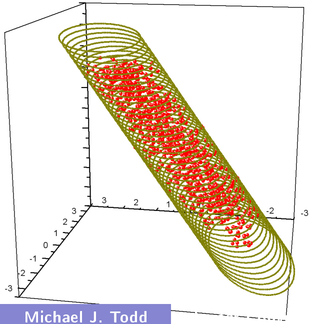
$$D^2 f(H)[E,E] = H^{-1}EH^{-1} \bullet E.$$

Hence f is convex.

4. Minimum Area Enclosing Ellipsoidal Cylinders

Given m points $\{x_1, x_2, ..x_m\} \subset \mathbf{R}^n$ which span $\mathbf{R}^n$ and $k \leq n$, the Minimum Area Enclosing Ellipsoidal Cylinder (MAEC) problem seeks an ellipsoidal cylinder which is centered at the origin, covers all the points and has minimum area intersection with

$$\Pi := \left\{ \begin{bmatrix} y \\ 0 \end{bmatrix} \in \begin{bmatrix} \mathbf{R}^k \\ \mathbf{R}^{n-k} \end{bmatrix} \right\}.$$



5. Geometry

The set

$$\mathcal{C}(E, H_{YY}) := \{[y; z] \in \mathbf{R}^n : (y + Ez)^T H_{YY}(y + Ez) \leq k\}$$

for $E \in \mathbf{R}^{k\times(n-k)}$ and $H_{YY} \succ 0$ is a cylinder in $\mathbf{R}^n$ defined by shape matrix H_{YY} and axis direction matrix E.

Note that $\mathcal{C}(E, H_{YY}) \cap \Pi$ is an ellipsoid in $\mathbf{R}^k \times \{0\}$ with

$$\text{area}(\mathcal{C}(E, H_{YY}) \cap \Pi) = \text{const}(k)/\sqrt{\det H_{YY}},$$

and minimizing the area of $\mathcal{C}(E, H_{YY}) \cap \Pi$ is equivalent to minimizing

$$-\ln\det H_{YY}.$$

6. MAEC Formulation

The MAEC problem can be formulated as follows:

$$\begin{array}{lrcl} \min_{E,H_{YY}} & f(H_{YY}) & := & -\ln\det H_{YY} \\ & (y_i + Ez_i)^T H_{YY}(y_i + Ez_i) & \leq & k,\ i = 1,\dots,m, \end{array}$$

(nonconvex!) or equivalently

$$(\bar{P}) \qquad \begin{array}{lrcl} \min_H & \bar{f}(H) & := & -\ln\det H_{YY} \\ & x_i^T H x_i & \leq & k,\ i = 1,\dots,m, \\ & H & \succeq & 0, \end{array}$$

where $H = \begin{pmatrix} H_{YY} & H_{YZ} \\ H_{YZ}^T & H_{ZZ} \end{pmatrix}$ (we set $E = H_{YY}^{-1}H_{YZ}$).

7. Duality I: the D-optimal Design Problem

Let $X := [x_1, \ldots, x_m] \in \mathbf{R}^{n \times m}$ and $U := \mathrm{Diag}\,(u)$. Then the dual to the MVEE problem (P) can be written as

$$(D) \qquad \begin{array}{rrcl} \max_u & g(u) & := & \ln\det XUX^T \\ & e^T u & = & 1, \\ & u & \geq & 0. \end{array}$$

(D) is the statistical problem of finding a D-optimal design measure on the columns of X, which maximizes the determinant of the Fisher information matrix when estimating all parameters $\theta_1, \ldots, \theta_n$ in the linear model

$$\tilde{y} \approx x^T \theta.$$

Duality II: the D$_k$-optimal Design Problem

The dual to the MAEC problem $(\bar{P})$ can be stated as

$$
(\bar{D}) \qquad
\begin{array}{rcl}
\max_{u,K} \ \bar{g}(u,K) & := & \ln\det K \\
XUX^T - \overline{K} & := & XUX^T - \begin{pmatrix} K & 0 \\ 0 & 0 \end{pmatrix} \succeq 0 \\
e^T u & = & 1, \\
u & \geq & 0.
\end{array}
$$

$(\bar{D})$ is the statistical problem of finding a D$_k$-optimal design measure on the columns of X, which maximizes the determinant of a Schur Complement in the Fisher information matrix.This is related to estimating the first k parameters $\theta_1, \ldots, \theta_k$ in the linear model

$$\tilde{y} \approx x^T \theta.$$

8. Weak Duality

Consider the MVEE problem. Suppose H and u are feasible in (P) and (D) respectively. Then

$$\mathrm{Tr}(HXUX^T) = H \bullet XUX^T = \sum_i u_i x_i^T H x_i \leq n.$$

Hence we have

$$\begin{aligned} -\ln\det H - \ln\det XUX^T &= -\ln\det H\,XUX^T \\ = -n\ln(\Pi_{i=1}^n \lambda_i(H\,XUX^T))^{1/n} &\geq -n\ln\left(\frac{\sum_{i=1}^n \lambda_i(H\,XUX^T)}{n}\right) \\ \geq -n\ln\left(\frac{n}{n}\right) &\geq 0. \end{aligned}$$

A similar derivation holds for the MAEC problem. Suppose H, u and K are feasible in $(\bar{P})$ and $(\bar{D})$ respectively. Then

$$0 \leq H \bullet (XUX^T - \overline{K}) = \sum_i u_i x_i^T H x_i - H \bullet \overline{K} \leq k - H_{YY} \bullet K.$$

Hence we have

$$\begin{aligned} -\ln\det H_{YY} - \ln\det K &= -\ln\det H_{YY}K \\ = -k\ln(\Pi_{i=1}^k \lambda_i(H_{YY}K))^{1/k} &\geq -k\ln\left(\frac{\sum_{i=1}^k \lambda_i(H_{YY}K)}{k}\right) \\ \geq -k\ln\left(\frac{k}{k}\right) &\geq 0. \end{aligned}$$

9. Optimality Conditions

For the MVEE problem, we have strong duality for feasible solutions if

(i) $u_i > 0$ only if $x_i^T H x_i = n$; and

(ii) $H = H(u) := (XUX^T)^{-1}$.

We say u is an ϵ-approximate optimal solution if

(a) $x_i^T H(u) x_i \leq (1+\epsilon)n$, $i = 1, \dots, m$,

(b) $u_i > 0$ implies $x_i^T H(u) x_i \geq (1-\epsilon)n$.

For the MAEC problem, we have strong duality if

(i) $H \bullet (XUX^T - \overline{K}) = 0$;

(ii) $u_i > 0$ only if $x_i^T H x_i = (y_i + Ez_i)^T H_{YY}(y_i + Ez_i) = k$; and

(iii) $H_{YY} = K^{-1}$.

For optimal u, condition (i) implies $E(ZUZ^T) = -(YUZ^T)$ and $K = YUY^T - E(ZUZ^T)E^T$. Assuming ZUZ^T is invertible, $E = E(u)$ and $K = K(u)$ are uniquely defined and smooth in u.

We say u is an ϵ-approximate optimal solution if

(a) $(y_i + Ez_i)^T K^{-1}(y_i + Ez_i) \leq (1 + \epsilon)k$, $i = 1, \ldots, m$; and

(b) $u_i > 0$ implies $(y_i + Ez_i)^T K^{-1}(y_i + Ez_i) \geq (1 - \epsilon)k$.

10. A Frank-Wolfe-Type Algorithm

We will analyze a first-order method for (D). Note that

$$w(u) := \nabla g(u) = (x_i^T (XUX^T)^{-1} x_i)_{i=1}^m.$$

Suppose u is updated by

$$u_+ := (1-\tau)u + \tau e_i.$$

Then rank-1 update formulae give

$$(XU_+X^T)^{-1} = \frac{1}{(1-\tau)} \left((XUX^T)^{-1} - \frac{\tau (XUX^T)^{-1} x_i x_i^T (XUX^T)^{-1}}{1 - \tau + \tau w_i(u)} \right)$$

and

$$\det XU_+X^T = (1-\tau)^{n-1} (1 - \tau + \tau w_i(u)) \det XUX^T,$$

so that it is easy to update w after such an update, and it is easy to perform a line search on τ to maximize $g(u_+)$.

This suggests that the Frank-Wolfe method (1956) might be attractive to solve (D), and this was suggested by the statisticians Fedorov (1972) and Wynn (1970). So we call this the FW-algorithm. We want to analyze the FW-algorithm with Wolfe's "away" steps (1970), which was also proposed for (D) by the statistician Atwood (1973) (hence WA-method).
At every iteration, we solve

$$\max_{\bar{u}} g(u) + w(u)^T(\bar{u} - u), \quad e^T\bar{u} = 1, \quad \bar{u} \geq 0,$$

i.e., find i that maximizes $w_i(u) - n$, and $\bar{u} = e_i$, and

$$\min_{\bar{u}} g(u) + w(u)^T(\bar{u} - u), \quad e^T\bar{u} = 1, \quad \bar{u} \geq 0, \bar{u}_k = 0 \text{ if } u_k = 0,$$

i.e., find j that maximizes $n - w_j(u)$ over j with $u_j > 0$, and $\bar{u} = e_j$.
Then we either move towards e_i or away from e_j.

Algorithms

If we only address the first half of the optimality conditions and only consider positive τ, this algorithm is due to the statisticians Fedorov and Wynn and is a specialization of the Frank-Wolfe algorithm for (D). It was analyzed by Khachiyan.

If we consider the complete optimality conditions and allow negative τ, this method was proposed by the statistician Atwood and is Frank-Wolfe's method with Wolfe's away steps. It was analyzed by Todd-Yildirim and Ahipasaoglu-Sun-Todd.

For the MAEC problem, similar but more complicated algorithms result. A change in u as above leads to rank-one updates to XUX^T, E, and K. These methods were also proposed by Fedorov and Atwood and were analyzed by Ahipasaoglu-Todd.

An Iteration of the WA Algorithm

Stop if $\max\{w_i(u) - n, n - w_j(u)\} \le \epsilon n$. Otherwise,
if $w_i(u) - n > n - w_j(u)$, replace u by

$$u_+ = (1-\tau)u + \tau e_i,$$

with $\tau > 0$ chosen optimally, i.e., move towards e_i;
if $n - w_j(u) \ge w_i(u) - n$, replace u with

$$u_+ = (1-\tau)u + \tau e_j,$$

with $\tau < 0$ chosen optimally so that u_+ remains feasible, i.e., move away from e_j.
Then update $w(u)$ and a Cholesky factorization of XUX^T.

Types of Iteration

We characterize steps as

increase-iterations: u_i increases from a positive value;
add-iterations: u_i increases from zero;
decrease-iterations: u_i decreases to a positive value; and
drop-iterations: u_i decreases to zero.

Note: #drop-iterations $\leq$ #positive components in initial u + #add-iterations.
The FW-algorithm stops when it gets an ϵ-primal feasible solution, i.e.,
u feasible and $(1+\epsilon)^{-1}(XUX^T)^{-1}$ primal feasible, or
$w_i(u) \leq (1+\epsilon)n$ for all i.
The WA-algorithm stops with u satisfying the ϵ-approximate optimality conditions, i.e., u feasible;
$w_i(u) \leq (1+\epsilon)n$ for all i; and
$w_i(u) \geq (1-\epsilon)n$ if $u_i > 0$.

11. Convergence Analysis

The FW-algorithm was analyzed by Khachiyan (1996): the number of iterations required is
$O(\frac{n}{\epsilon} + n \ln n + n \ln \ln m)$.

With a different initialization, Kumar-Yildirim (2005) achieved a bound of
$O(\frac{n}{\epsilon} + n \ln n)$.

The WA-method was analyzed by Todd-Yildirim (2005) with the KY initialization, with the same complexity bound (actually twice, because of the drop-iterations).

Each iteration requires $O(nm)$ arithmetic operations (far fewer than an interior-point method).

The basis for the analyses consists of two lemmas:

Lemma

(Khachiyan) If u is δ-primal feasible, $g^* - g(u) \le n\delta$.

Lemma

(Khachiyan, Todd-Yildirim) Suppose $\delta \le 1/2$. Then

(a) If a feasible u is not δ-primal feasible, an add- or increase-iteration will improve $g(u)$ by at least $2\delta^2/7$.

(b) If a feasible u does not satisfy the δ-approximate optimality conditions, a decrease-iteration will improve $g(u)$ by at least $2\delta^2/7$.

Asymptotic linear convergence

To improve this bound, we tighten the first lemma, and show

Proposition

For some constant $M > 0$*, depending on the data, any* u *satisfying the* δ*-approximate optimality conditions for sufficiently small* δ *has* $g^* - g(u) \leq M\delta^2$.

Putting the proposition and the second lemma together, we obtain

Theorem

For some $Q > 0$*, the WA-algorithm requires at most* $Q + 56M\ln(1/\epsilon)$ *iterations to produce a feasible* u *that satisfies the* ϵ*-approximate optimality conditions.*

Proof

Proposition

For some constant $M > 0$, depending on the data, any u satisfying the δ-approximate optimality conditions for sufficiently small δ has $g^* - g(u) \leq M\delta^2$.

The proof uses the perturbed problem

$$(P(z)) \quad \begin{array}{rrcl} \min_{H \succ 0} & -\ln\det H & & \\ & x_i^T H x_i & \leq & n + z_i,\ i = 1, \ldots, m. \end{array}$$

If u is as in the proposition, define $z := z(u, \delta) \in \mathbf{R}^m$ by

$$z_i := \begin{cases} \delta n & \text{if } u_i = 0 \\ w_i(u) - n & \text{else.} \end{cases}$$

Note that $|z_i| \leq \delta n$ for each i, and $u^T z = 0$.

Analysis

Lemma

If u satisfies the δ-approximate optimality conditions, $H(u) := (XUX^T)^{-1}$ is optimal in $(P(z(u,\delta)))$, and u is a vector of Lagrange multipliers.

Let $\phi(z)$ denote the optimal value of $(P(z))$. This is convex, and any vector of Lagrange multipliers is a subgradient. So for any vector u_* of Lagrange multipliers for $(P) = (P(0))$, u as above, and $z := z(u,\delta)$,

$$g(u) = f(H(u)) = \phi(z) \geq \phi(0) + u_*^T(z-0) = g^* - (u-u_*)^T z$$

since $u^T z = 0$.
Now $\|z\| = O(\delta)$, and results of Robinson (1982) show that, for some u_*, $\|u - u_*\| = O(\delta)$, and this proves the proposition.

Convergence for the WA-Algorithm for the MAEC Problem

For the MAEC problem, assuming $\lambda_{\min}(ZUZ^T) \geq c > 0$, $\max_i\{x_i^T(XUX^T)^{-1}x_i\} < C$, we have:

- $\mathcal{O}_{c,C}(k(\ln k + k \ln\ln m + \epsilon^{-1}))$ iterations (AT).
- Each iteration takes $\mathcal{O}(nm)$ operations.
- Local linear convergence as for MVEE under a strong second-order sufficient condition (AT).
- Away steps are necessary for rapid convergence.

MVEE problems:

Table: Means of Running Times and Numbers of Iterations Required by the Algorithm to Obtain an ϵ-Approximate Solution

Dimensions			Averages	
n	m	$-\log_{10}\epsilon$	iter	time (sec.)
100	10000	10	800	2.0
200	10000	7	1894	7.5
500	10000	7	5038	142.5

MAEC problems:

Table: Means of Running Times and Numbers of Iterations Required by the Algorithm to Obtain an ϵ-Approximate Solution

Dimensions				Averages	
k	n	m	$-\log_{10}\epsilon$	iter	time (sec.)
20	100	10000	7	3366	60.2
50	100	10000	7	2897	46.4
80	100	10000	7	1328	19.6

13. Conclusions

- First-order methods can be very effective and may be necessary to handle very large instances.
- Computational complexity analysis, rate of convergence analysis, and computational experiments complement one another.
- There is much still to be understood!