
Linear convergence of a modified Frank-Wolfe algorithm for the minimum-volume enclosing ellipsoid problem

Damla Ahipasaoglu, Peng Sun, and Mike Todd

School of Operations Research and Industrial Engineering,

Cornell University

Fuqua School of Business, Duke University, and

SORIE, Cornell

<http://www.orie.cornell.edu/~miketodd/todd.html>

Problem

We are given m points x_1, \dots, x_m in \mathbb{R}^n , and we want to find a minimum-volume central ellipsoid containing them. This arises in data analysis, computational geometry, and optimization. If we write the ellipsoid as

$$E(0, H) := \{x \in \mathbb{R}^n : x^T H x \leq n\},$$

where $H \succ 0$, its volume is $(\det H)^{-1/2}$ times that of a ball of radius \sqrt{n} . So we can formulate the problem as:

$$(P) \quad \begin{aligned} \min_{H \succ 0} \quad & f(H) := -\ln \det H \\ & x_i^T H x_i \leq n, \quad i = 1, \dots, m. \end{aligned}$$

Primal and Dual

$$\begin{aligned} & \min_{H \succ 0} \quad f(H) := -\ln \det H \\ (P) \quad & x_i^T H x_i \leq n, \quad i = 1, \dots, m. \end{aligned}$$

After some simplification, its Lagrangian dual becomes

$$\begin{aligned} & \max_u \quad g(u) := \ln \det X U X^T \\ (D) \quad & e^T u = 1, \\ & u \geq 0, \end{aligned}$$

where $X := [x_1, \dots, x_m] \in \mathbb{R}^{n \times m}$, $U := \text{Diag}(u)$, and e denotes an m -vector of ones.

This dual problem also arises in optimal design in statistics.

Optimality conditions

Note that

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

and that for any feasible u , we have

$$u^T w(u) = n.$$

A feasible H and a feasible u are optimal iff

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

(b) $H = (XUX^T)^{-1}$.

In fact, given (b), (a) holds iff $w_i(u) = x_i^T (XUX^T)^{-1} x_i \leq n$ for all i .

Motivating Frank-Wolfe

Suppose u is updated by

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

Then rank-1 update formulae give

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

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.

FW-algorithm with Away Steps

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 .

Iteration

Stop if $\max\{w_i(u) - n, n - w_j(u)\} \leq \epsilon n$. 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 $n - w_j(u) \geq w_i(u) - n$, 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 .

Drop-iterations

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 $u_i > 0$.

Global complexity bounds

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

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).

The basis for the analyses consists of two lemmas:

Lemma 1 (Khachiyan) *If u is δ -primal feasible, $g^* - g(u) \leq n\delta$.*

Lemma 2 (Khachiyan, Todd-Yildirim) *Suppose $\delta \leq 1/2$. 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 Lemma 1, and show

Proposition 3 *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 Proposition 3 and Lemma 2 together, we obtain

Theorem 4 *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 3 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

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

If u is as in the proposition, define $z := z(u, \delta) \in \mathbb{R}^m$ 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 4 *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.

Discussion

Wolfe (1970) sketched a proof that the Frank-Wolfe method with away steps on the simplex had linear convergence, and Guélat and Marcotte (1986) gave a complete proof. Their results do not apply here, firstly because $-g$ is not boundedly convex (it goes to infinity at the vertices of the simplex if $n > 1$), but more importantly because it is **not strictly, let alone strongly, convex**.

The arguments here can also be used to show linear convergence for the situation of Wolfe and Guélat and Marcotte, under slightly weaker assumptions, again using a perturbation and Robinson's results.

The message here is that very simple methods when suitably implemented can be highly effective in solving important large-scale problems, with the effectiveness certified by a convergence proof and computational experiments.