On the von Neumann and Frank-Wolfe Algorithms with Away Steps

Javier Peña* Daniel Rodríguez† Negar Soheili‡

July 16, 2015

Abstract

The von Neumann algorithm is a simple coordinate-descent algorithm to determine whether the origin belongs to a polytope generated by a finite set of points. When the origin is in the *interior* of the polytope, the algorithm generates a sequence of points in the polytope that converges linearly to zero. The algorithm's rate of convergence depends on the radius of the largest ball around the origin contained in the polytope.

We show that under the weaker condition that the origin is in the polytope, possibly on its boundary, a variant of the von Neumann algorithm that includes *away steps* generates a sequence of points in the polytope that converges linearly to zero. The new algorithm's rate of convergence depends on a certain geometric parameter of the polytope that extends the above radius but is always positive. Our linear convergence result and geometric insights also extend to a variant of the Frank-Wolfe algorithm with away steps for minimizing a strongly convex function over a polytope.

*Tepper School of Business, Carnegie Mellon University, USA, jfp@andrew.cmu.edu

†Department of Mathematical Sciences, Carnegie Mellon University, USA, drod@cmu.edu

‡College of Business Administration, University of Illinois at Chicago, USA, nazad@uic.edu

1 Introduction

Assume $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ with $\|a_i\|_2 = 1,\ i = 1, \ldots, n$. The von Neumann algorithm, communicated by von Neumann to Dantzig in the late 1940s, is a simple algorithm to solve the feasibility problem:

$$\text{Is } 0 \in \operatorname{conv}(A) = \operatorname{conv}\{a_1, \ldots, a_n\}?$$

More precisely, the algorithm finds an approximate solution to the problem

$$Ax = 0,\ x \in \Delta_{n-1} = \{x \in \mathbb{R}^n_+ : \|x\|_1 = 1\}. \tag{1}$$

The algorithm starts from an arbitrary point $x_0 \in \Delta_n$. At the k-th iteration the algorithm updates the current trial solution $x_k \in \Delta_{n-1}$ as follows. First, if finds the column a_j of A that forms the widest angle with $y_k := Ax_k$. If this angle is acute, i.e., $A^{\mathrm{T}} y_k > 0$, then the algorithm halts as the vector y_k separates the origin from $\operatorname{conv}(A)$. Otherwise the algorithm chooses $x_{k+1} \in \Delta_{n-1}$ so that Ax_{k+1} is the minimum-norm convex combination of Ax_k and a_j. Let $e_j \in \Delta_{n-1}$ denote the n-dimensional vector with j-th component equal to one and all other components equal to zero. To ease notation, we shall write $\|\cdot\|$ for $\|\cdot\|_2$ throughout the paper.

Von Neumann Algorithm

1. pick $x_0 \in \Delta_{n-1}$; put $y_0 := Ax_0$; $k := 0$.
2. for $k = 0, 1, 2, \ldots$
   if $A^{\mathrm{T}} y_k > 0$ then HALT: $0 \notin \operatorname{conv}(A)$
   $j := \operatorname*{argmin}_{i=1,\ldots,n} \langle a_i, y_k \rangle$;
   $\theta_k := \operatorname*{argmin}_{\theta \in [0,1]} \{\|y_k + \theta(a_j - y_k)\|\}$;
   $x_{k+1} := (1 - \theta_k) x_k + \theta_k e_j$;
   $y_{k+1} := (1 - \theta_k) y_k + \theta_k a_j$;
   end for

The von Neumann algorithm can be seen as a kind of coordinate-descent method for finding a solution to (1): At each iteration the algorithm judiciously selects a coordinate j and *increases* the weight of the j-th component of x_k while decreasing all of the others via a line-search step. Like other currently popular coordinate-descent and first-order methods for convex optimization, the main attractive features of the von Neumann algorithm are its simplicity and low computational cost per iteration. Another attractive feature is its convergence rate. Epelman and Freund [6] showed that the speed of convergence of the von Neumann algorithm can be characterized in terms of the following *condition measure* of the matrix A:

$$\rho(A) := \max_{y \in \mathbb{R}^m : \|y\| = 1} \min_{i=1,\ldots,n} \langle a_i, y \rangle . \tag{2}$$

The condition measure $\rho(A)$ was introduced by Goffin [8] and later independently studied by Cheung and Cucker [3]. The latter set of authors showed that $|\rho(A)|$ is also a certain *distance to ill-posedness* in the spirit introduced and developed by Renegar [15, 16].

Observe that $\rho(A) > 0$ if and only if $0 \notin \text{conv}(A)$, and $\rho(A) < 0$ if and only if $0 \in \text{int}(\text{conv}(A))$. When $\rho(A) > 0$, this condition measure is closely related to the concept of margin in binary classification [19] and with the minimum enclosing ball problem in computational geometry [5]. The quantity $\rho(A)$ also has the following geometric interpretation. If $\rho(A) > 0$ then

$$\rho(A) = \min\{\|y\| : y \in \text{conv}(A)\}, \tag{3}$$

and if $\rho(A) \leq 0$ then

$$|\rho(A)| = \max\{r : \|y\| \leq r \Rightarrow y \in \text{conv}(A)\}. \tag{4}$$

In particular, $|\rho(A)| = \mathsf{dist}(0, \partial\text{conv}(A))$.

Epelman and Freund [6] showed the following properties of the von Neumann algorithm. When $\rho(A) < 0$ the algorithm generates iterates $x_k \in \Delta_{n-1},\ k = 1, 2, \ldots$ such that

$$\|Ax_k\|^2 \leq \left(1 - \rho(A)^2\right)^k \|Ax_0\|^2. \tag{5}$$

On the other hand, the iterates $x_k \in \Delta_{n-1}$ also satisfy $\|Ax_k\|^2 \leq \frac{1}{k}$ as long as the algorithm has not halted. In particular, if $\rho(A) > 0$ then by (3) the algorithm must halt with a certificate of infeasibility $A^{\mathrm{T}} y_k > 0$ for $0 \notin \text{conv}(A)$ in at most $\frac{1}{\rho(A)^2}$ iterations. The latter bound is identical to a classical convergence bound for the perceptron algorithm [2, 14]. This is not a coincidence as there is a nice duality between the perceptron and the von Neumann algorithms [13, 17].

We show that a variant of the von Neumann algorithm with *away steps* has the following stronger convergence properties. When $0 \in \text{conv}(A)$, possibly on its boundary, the algorithm generates a sequence $x_k \in \Delta_{n-1}$ satisfying

$$\|Ax_k\|^2 \leq \left(1 - \frac{w(A)^2}{16}\right)^{k/2} \|Ax_0\|^2. \tag{6}$$

The quantity $w(A)$ is a kind of *relative width* of $\text{conv}(A)$ that is at least as large as $|\rho(A)|$. However, unlike $|\rho(A)|$ the relative width $w(A)$ is positive for any non-zero matrix $A \in \mathbb{R}^{m \times n}$ provided $0 \in \text{conv}(A)$. When $\rho(A) > 0$, or equivalently $0 \notin \text{conv}(A)$, the von Neumann algorithm with away steps finds a certificate of infeasibility $A^{\mathrm{T}} y_k > 0$ for $0 \notin \text{conv}(A)$ in at most $\frac{8}{\rho(A)^2}$ iterations.

We show that a linear convergence result similar to (6) also holds for a version of the Frank-Wolfe algorithm with away steps for minimizing a strongly convex function with a Lipschitz gradient over a polytope. These linear convergence results are in the same spirit as the results established in [9, 10, 11] as well as some linear convergence results for the randomized Kaczmarz algorithm [18] and for the methods of randomized coordinate descent and iterated projections [12]. Our main contributions are the succinct and transparent proofs of these linear convergence results that highlight the role of the relative width $w(A)$ and a closely related *restricted width* $\varrho(A)$. Our presentation unveils a deep connection between problem conditioning as encompassed by the quantities $w(A), \varrho(A)$ and the behavior of the von Neumann and Frank-Wolfe algorithms with away steps. We also provide some lower bounds on $w(A)$ and $\varrho(A)$ in terms of certain radii quantities that naturally extend $\rho(A)$. We note that the linear convergence results in [11] are stated in terms of a certain *pyramidal width* whose geometric intuition and properties appear to be less understood than those of $w(A)$ and $\varrho(A)$. We also note that during the review process of this manuscript we also became aware of the related and independent work of Beck and Shtern [1]. In contrast to our geometric approach, the approach followed by Beck and Shtern is primarily founded on convex duality.

The rest of the paper is organized as follows. In Section 2 we describe a von Neumann Algorithm with Away Steps and establish its main convergence result in terms of the relative width $w(A)$. Section 3 extends our main result to the more general problem of minimizing a quadratic function over the polytope $\operatorname{conv}(A)$. Section 4 presents the same ideas for more general strongly convex functions with Lipschitz gradient. Finally, Section 5 discusses some properties of the relative and restricted widths.

2 Von Neumann Algorithm with Away Steps

Throughout this section we assume $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ with $\|a_i\| = 1,\ i = 1, \ldots, n$. We next consider a variant of the von Neumann Algorithm that includes so-called "away" steps. To that end, at each iteration, in addition to a "regular step" the algorithm considers an alternative "away step". Each of these away steps identifies ℓ such that the ℓ-th component of x_k is positive and *decreases* the weight of the ℓ-th component of x_k. The algorithm needs to keep track of the *support*, that is, the set of positive entries of a vector. To that end,

given $x \in \mathbb{R}^n_+$, let the support of x be defined as

$$S(x) := \{i \in \{1, \ldots, n\} : x_i > 0\}.$$

Von Neumann Algorithm with Away Steps

1. pick $x_0 \in \Delta_{n-1}$; put $y_0 := Ax_0$; $k := 0$; .
2. for $k = 0, 1, 2, \ldots$
   if $A^{\mathrm{T}} y_k > 0$ then HALT: $0 \notin \mathrm{conv}(A)$
   $j := \underset{i=1,\ldots,n}{\mathrm{argmin}} \langle a_i, y_k \rangle$; $\ell := \underset{i \in S(x_k)}{\mathrm{argmax}} \langle a_i, y_k \rangle$;
   if $\|y_k\|^2 - \langle a_j, y_k \rangle > \langle a_\ell, y_k \rangle - \|y_k\|^2$ then (regular step)
   $a := a_j - y_k$; $u := e_j - x_k$; $\theta_{\max} := 1$
   else (away step)
   $a := y_k - a_\ell$; $u := x_k - e_\ell$; $\theta_{\max} := \frac{(x_k)_\ell}{1-(x_k)_\ell}$
   endif
   $\theta_k := \underset{\theta \in [0, \theta_{\max}]}{\mathrm{argmin}} \{\|y_k + \theta a\|\}$;
   $y_{k+1} = y_k + \theta_k a$; $x_{k+1} := x_k + \theta_k u$;
   end for

Define the relative width $w(A)$ of $\mathrm{conv}(A)$ as

$$w(A) := \min_{x \geq 0, Ax \neq 0} \max_{\ell, j} \left\{ \frac{\langle Ax, a_\ell - a_j \rangle}{\|Ax\|} : \ell \in S(x),\ j \in \{1, \ldots, n\} \right\}. \tag{7}$$

It is easy to show that $w(A) \geq |\rho(A)|$ when $0 \in \mathrm{conv}(A)$. In Section 5 below we discuss some properties of $w(A)$. In particular, we will formally prove the intuitively clear property that $w(A) > 0$ for any nonzero matrix $A \in \mathbb{R}^{m \times n}$ such that $0 \in \mathrm{conv}(A)$.

We are now ready to state the main properties of the von Neumann algorithm with away steps.

Theorem 1 *Assume $x_0 \in \Delta_{n-1}$ is one of the extreme points of Δ_{n-1}.*

(a) *If $0 \in conv(A)$ then the iterates $x_k \in \Delta_{n-1}, y_k = Ax_k,\ k = 0, 1, \ldots$ generated by the von Neumann Algorithm with Away Steps satisfy*

$$\|y_k\|^2 \leq \left(1 - \frac{w(A)^2}{16}\right)^{k/2} \|y_0\|^2.$$

(b) *The iterates $x_k \in \Delta_{n-1}, y_k = Ax_k,\ k = 1, \ldots$ generated by the von Neumann Algorithm with Away Steps also satisfy*

$$\|y_k\|^2 \leq \frac{8}{k}$$

as long as the algorithm has not halted. In particular, if $0 \notin conv(A)$ *then the von Neumann Algorithm with Away Steps finds a certificate of infeasibility* $A^{\mathrm{T}} y_k > 0$ *for* $0 \notin \mathrm{conv}(A)$ *in at most* $\frac{8}{\rho(A)^2}$ *iterations.*

The crux of the proof of Theorem 1 is the following elementary lemma.

Lemma 1 *Assume* $a, y \in \mathbb{R}^m$ *satisfy* $\langle a, y \rangle < 0$. *Then*

$$\min_{\theta \geq 0} \|y + \theta a\|^2 = \|y\|^2 - \frac{\langle a, y \rangle^2}{\|a\|^2},$$

and the minimum is attained at $\theta = -\frac{\langle a, y \rangle}{\|a\|^2}$.

Proof of Theorem 1:

(a) The algorithm generates y_{k+1} by solving a problem of the form

$$\|y_{k+1}\|^2 = \min_{\theta \in [0, \theta_{\max}]} \|y_k + \theta a\|^2$$

where $a = a_j - y_k$ or $a = y_k - a_\ell$, and $-\langle a, y_k \rangle > \frac{1}{2} \langle a_\ell - a_j, y_k \rangle \geq \frac{1}{2} w(A) \|y_k\|$. If $\theta_k < \theta_{\max}$ then Lemma 1 applied to $y := y_k$ yields

$$\|y_{k+1}\|^2 = \|y_k\|^2 - \frac{\langle a, y_k \rangle^2}{\|a\|^2} \leq \|y_k\|^2 - \frac{w(A)^2}{16} \|y_k\|^2.$$

Thus each time the algorithm performs an iterate with $\theta_k < \theta_{\max}$, the value of $\|y_k\|^2$ decreases at least by the factor $1 - \frac{w(A)^2}{16}$. To conclude, it suffices to show that after N iterations the number of iterates where $\theta_k < \theta_{\max}$ is at least $N/2$. To that end, we apply the following argument from [11]: Observe that when $\theta_k = \theta_{\max}$ we have $|S(x_{k+1})| < |S(x_k)|$. On the other hand, when $\theta_k < \theta_{\max}$ we have $|S(x_{k+1})| \leq |S(x_k)| + 1$. Since $|S(x_0)| = 1$ and $|S(x)| \geq 1$ for every $x \in \Delta_{n-1}$, after any number of iterates there must have been at least as many iterates with $\theta_k < \theta_{\max}$ as there have been iterates with $\theta_k = \theta_{\max}$. Hence after N iterations, the number of iterates with $\theta_k < \theta_{\max}$ is at least $N/2$.

(b) Proceed as above but note that if the algorithm does not halt at the k-th iterate then $\langle a, y_k \rangle \leq \langle a_j - y_k, y_k \rangle \leq -\|y_k\|^2$. Thus each time the algorithm performs an iterate with $\theta_k < \theta_{\max}$, we have

$$\|y_{k+1}\|^2 \leq \|y_k\|^2 - \frac{\langle a, y_k \rangle^2}{\|a\|^2} \leq \|y_k\|^2 - \frac{\|y_k\|^4}{4}.$$

It follows by induction that if the algorithm has not halted after k iterations then we must have

$$\|y_k\|^2 \leq \frac{8}{k}.$$

If $0 \notin \text{conv}(A)$ then $\rho(A) = \min\{\|y\| : y \in \text{conv}(A)\} > 0$ and so the algorithm must halt with a certificate of infeasibility $A^{\mathrm{T}} y_k > 0$ for $0 \notin \text{conv}(A)$ after at most $\frac{8}{\rho(A)^2}$ iterations.

■

3 Frank-Wolfe Algorithm with Away Steps for Quadratic Functions

Throughout this section assume $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ is a non-zero matrix, and $f(y) = \frac{1}{2}\langle y, Qy \rangle + \langle b, y \rangle$ for a symmetric positive definite matrix $Q \in \mathbb{R}^{m \times m}$ and $b \in \mathbb{R}^m$. Consider the problem

$$\min_{y \in \text{conv}(A)} f(y) \Leftrightarrow \min_{x \in \Delta_{n-1}} f(Ax). \tag{8}$$

Problem (1) can be seen as a special case of (8) when $Q = I$ and $b = 0$. The von Neumann Algorithm can also be seen as a special case of the Frank-Wolfe Algorithm [7] for (8). This section extends the ideas and results from Section 2 to the following variant of the Frank-Wolfe algorithm with away steps. We note that this variant can be traced back to Wolfe [20] as discussed by Guélat and Marcotte [9].

Frank-Wolfe Algorithm with Away Steps

1. pick $x_0 \in \Delta_{n-1}$; put $y_0 := Ax_0$; $k := 0$; .
2. for $k = 0, 1, 2, \ldots$
   $j := \operatorname{argmin}_{i=1,\ldots,n} \langle a_i, \nabla f(y_k) \rangle$; $\ell := \operatorname{argmax}_{i \in S(x_k)} \langle a_i, \nabla f(y_k) \rangle$;
   if $\langle y_k - a_j, \nabla f(y_k) \rangle > \langle a_\ell - y_k, \nabla f(y_k) \rangle$ then (regular step)
   $a := a_j - y_k$; $u := e_j - x_k$; $\theta_{\max} := 1$
   else (away step)
   $a := y_k - a_\ell$; $u := x_k - e_\ell$; $\theta_{\max} := \frac{(x_k)_\ell}{1-(x_k)_\ell}$
   endif
   $\theta_k := \operatorname{argmin}_{\theta \in [0, \theta_{\max}]} f(y_k + \theta a)$
   $y_{k+1} = y_k + \theta_k a$; $x_{k+1} := x_k + \theta_k u$
   end for

Observe that the computation of θ_k in the second to last step reduces to minimizing a one-dimensional convex quadratic function over the interval $[0, \theta_{\max}]$.

We next present a general version of Theorem 1 for the above Frank-Wolfe Algorithm with Away Steps. The linear convergence result depends on a certain restricted width and diameter defined as follows. For $x \geq 0$ with $Ax \neq 0$ let

$$\varrho(A, x) := \sup\left\{\lambda > 0 : \exists u, v \in \Delta_{n-1},\ S(u) \subseteq S(x),\ Au - Av = \frac{\lambda}{\|Ax\|} Ax\right\}.$$

Define the restricted width $\varrho(A)$ and diameter $d(A)$ of $\operatorname{conv}(A)$ as follows.

$$\varrho(A) := \min_x \left\{\varrho(A, x) : x \geq 0,\ Ax \neq 0\right\}, \tag{9}$$

and

$$d(A) := \max_{u,x \in \Delta_{n-1}} \|Ax - Au\|. \tag{10}$$

It is immediate from (7) and (9) that $w(A) \geq \varrho(A)$ for all nonzero $A \in \mathbb{R}^{m\times n}$. Furthermore, the restricted width $\varrho(A)$ can be seen as an extension of the radius $\rho(A)$ defined in (2). Indeed, when $0 \in \operatorname{int}(\operatorname{conv}(A))$, we have $\operatorname{span}(A) = \mathbb{R}^m$. Hence (4) can alternatively be written as

$$|\rho(A)| := \min_{x \geq 0: Ax \neq 0} \max\left\{\lambda : \exists v \in \Delta_{n-1},\ -Av = \frac{\lambda}{\|Ax\|} Ax\right\}.$$

This implies that $\varrho(A, x) \geq |\rho(A)| + \frac{\|Ax\|}{\|x\|_1}$ for all $x \geq 0$ with $Ax \neq 0$. Hence the following inequality readily follows

$$\varrho(A) \geq |\rho(A)|.$$

Section 5 presents a stronger lower bound on $\varrho(A)$ in terms of certain variants of $\rho(A)$. In particular, we will show that $\varrho(A) > 0$, and consequently $w(A) > 0$, for any nonzero matrix $A \in \mathbb{R}^{m\times n}$ such that $0 \in \operatorname{conv}(A)$.

The linear convergence property of the von Neumann algorithm with away steps, as stated in Theorem 1(a), extends as follows.

Theorem 2 *Assume $x^* \in \Delta_{n-1}$ is a minimizer of (8). Let $y^* = Ax^*$ and $\bar{A} := Q^{1/2}\begin{bmatrix} a_1 - y^* & \cdots & a_n - y^* \end{bmatrix}$. If $x_0 \in \Delta_{n-1}$ is one of the extreme points of Δ_{n-1} then the iterates $x_k \in \Delta_{n-1}, y_k = Ax_k$ generated by the Frank-Wolfe Algorithm with Away Steps satisfy*

$$f(y_k) - f(y^*) \leq \left(1 - \frac{\varrho(\bar{A})^2}{4d(\bar{A})^2}\right)^{k/2} (f(y_0) - f(y^*)). \tag{11}$$

The proof of Theorem 2 relies on the following two lemmas. The first one is similar to Lemma 1 and also follows via a straightforward calculation.

Lemma 2 *Assume f is as above and $a, y \in \mathbb{R}^m$ satisfy $\langle a, \nabla f(y)\rangle < 0$. Then*

$$\min_{\theta \geq 0} f(y + \theta a) = f(y) - \frac{\langle a, \nabla f(y)\rangle^2}{2\langle a, Qa\rangle},$$

and the minimum is attained at $\theta = -\frac{\langle a, \nabla f(y)\rangle}{\langle a, Qa\rangle}$.

Lemma 3 *Assume $f, A, y^*, \bar{A}$ are as in Theorem 2 above. Then for all $x \in \Delta_{n-1}$*

$$\max_{\ell \in S(x), j=1,\ldots,n} \langle \nabla f(Ax), a_\ell - a_j\rangle \geq \varrho(\bar{A})\sqrt{2(f(Ax) - f(y^*))}.$$

Proof: Let $y := Ax \in \mathrm{conv}(A)$. Assume $y \neq y^*$ as otherwise there is nothing to show. For ease of notation put $\|y - y^*\|_Q^2 := \langle y - y^*, Q(y - y^*)\rangle$. It readily follows that

$$f(y) + \langle \nabla f(y), y^* - y\rangle + \frac{1}{2}\|y - y^*\|_Q^2 = f(y^*)$$

so

$$f(y) - f(y^*) = \langle \nabla f(y), y - y^*\rangle - \frac{1}{2}\|y - y^*\|_Q^2 \leq \frac{\langle \nabla f(y), y - y^*\rangle^2}{2\|y - y^*\|_Q^2}$$

where the last step follows from the inequality $a^2 + b^2 + 2ab \geq 0$.

Thus

$$\frac{\langle \nabla f(y), y - y^*\rangle}{\|y - y^*\|_Q} \geq \sqrt{2(f(y) - f(y^*))}. \tag{12}$$

On the other hand, by the definition of $\varrho(A)$ there exist $u, v \in \Delta_{n-1}$ with $S(u) \subseteq S(x)$ and $\lambda \geq \varrho(\bar{A})$ such that $\bar{A}u - \bar{A}v = \frac{\lambda}{\|\bar{A}x\|}\bar{A}x$. Since $\bar{A}x = Q^{1/2}(Ax - y^*) = Q^{1/2}(y - y^*)$, the latter equation can be rewritten as

$$Au - Av = \frac{\lambda}{\|y - y^*\|_Q}(y - y^*). \tag{13}$$

Putting (12) and (13) together we get

$$\langle \nabla f(y), Au - Av\rangle = \frac{\lambda\langle \nabla f(y), y - y^*\rangle}{\|y - y^*\|_Q} \geq \varrho(\bar{A})\sqrt{2(f(y) - f(y^*))}.$$

To finish, observe that

$$\begin{aligned}\max_{\ell \in S(x), j=1,\ldots,n} \langle \nabla f(Ax), a_\ell - a_j\rangle &\geq \langle \nabla f(y), Au - Av\rangle \\ &\geq \varrho(\bar{A})\sqrt{2(f(Ax) - f(y^*))}.\end{aligned}$$

■

Proof of Theorem 2: This is a modification of the proof of Theorem 1(a). At iteration k the algorithm yields y_{k+1} such that

$$f(y_{k+1}) = \min_{\theta \in [0, \theta_{\max}]} f(y_k + \theta a)$$

where $a = a_j - y_k$ or $a = y_k - a_\ell$, and

$$-\langle \nabla f(y_k), a \rangle > \frac{1}{2} \langle \nabla f(y_k), a_\ell - a_j \rangle \geq \frac{1}{2} \varrho(\bar{A}) \sqrt{2(f(y_k) - f(y^*)}.$$

The second inequality above follows from Lemma 3. If $\theta_k < \theta_{\max}$ then Lemma 2 applied to $y := y_k$ yields

$$f(y_{k+1}) = f(y_k) - \frac{\langle a, \nabla f(y_k) \rangle^2}{2 \langle a, Qa \rangle} \leq f(y_k) - \frac{\varrho(\bar{A})^2}{4 d(\bar{A})^2} (f(y_k) - f(y^*)).$$

That is,

$$f(y_{k+1}) - f(y^*) \leq \left(1 - \frac{\varrho(\bar{A})^2}{4 d(\bar{A})^2}\right) (f(y_k) - f(y^*)).$$

Then proceeding as in the last part of the proof of Theorem 1(a) we obtain (11). ■

In the special case when $Q = I, b = 0, 0 \in \text{conv}(A)$, and all columns of A have norm one, we have $d(A) \leq 2$ and the minimizer y^* of (8) is 0. Thus Theorem 2 yields a weaker version of Theorem 1(a) with $w(A)$ replaced with $\varrho(A) \leq w(A)$. Conversely, a closer look at the proof of Theorem 2 reveals that the convergence bound (11) can be sharpened as follows: Replace $\varrho(\bar{A})$ with $w_f(A) \geq \varrho(\bar{A})$, where $w_f(A)$ is the following extension of $w(A)$:

$$w_f(A) := \min_{\substack{x \in \Delta_{n-1} \\ Ax \neq y^*}} \max_{\ell, j} \left\{ \frac{\langle \nabla f(Ax), a_\ell - a_j \rangle}{\sqrt{2(f(Ax) - f(y^*))}} : \ell \in S(x), j \in \{1, \ldots, n\} \right\}.$$

We have the following related conjecture concerning $w(A)$ and $\varrho(A)$.

Conjecture 1 *If $A \in \mathbb{R}^{m \times n}$ is non-zero and $0 \in \text{conv}(A)$ then $\varrho(A) = w(A)$.*

The next result shows that the ratio $\frac{\varrho(\bar{A})}{d(\bar{A})}$ in (11) can be bounded below in terms of a product of the ratio of the smallest to largest eigenvalue of Q and a second factor that depends only on $\text{conv}(\tilde{A})$ for $\tilde{A} := \begin{bmatrix} a_1 - y^* & \cdots & a_n - y^* \end{bmatrix}$. We omit the proof as it is a straightforward matrix algebra calculation.

Proposition 1 *Assume $x^* \in \Delta_{n-1}$ is a minimizer of (8). Let $y^* = Ax^*$, $\tilde{A} := \begin{bmatrix} a_1 - y^* & \cdots & a_n - y^* \end{bmatrix}$, and $\bar{A} := Q^{1/2}\tilde{A}$. Let μ, L be respectively the smallest and largest eigenvalues of Q. Then $\varrho(\bar{A}) \geq \sqrt{\mu}\varrho(\tilde{A})$ and $d(\bar{A}) \leq \sqrt{L}d(\tilde{A}) = \sqrt{L}d(A)$. In particular*

$$\frac{\varrho(\bar{A})}{d(\bar{A})} \geq \sqrt{\frac{\mu}{L}} \cdot \frac{\varrho(\tilde{A})}{d(\tilde{A})} = \sqrt{\frac{\mu}{L}} \cdot \frac{\varrho(\tilde{A})}{d(A)}.$$

As we discuss in the next section, this results readily extends to the more general problem when f is a strongly convex function with Lipschitz gradient. We discuss that in the next section.

4 Frank-Wolfe Algorithm with Away Steps for Strongly Convex Functions with Lipschitz Gradient

We next consider a more general version of the problem (8) where f is a μ-strongly convex and ∇f is a L-Lipschitz function.

Theorem 3 *Assume f is μ-strongly convex and ∇f is L-Lipschitz. Assume $x^* \in \Delta_{n-1}$ is a minimizer of (8). If $x_0 \in \Delta_{n-1}$ is one of the extreme points of Δ_{n-1} then the iterates $x_k \in \Delta_{n-1}, y_k = Ax_k$ generated by the Frank-Wolfe Algorithm with Away Steps satisfy*

$$f(y_k) - f(y^*) \leq \left(1 - \frac{w_f(A)^2}{4Ld(A)^2}\right)^{k/2} (f(y_0) - f(y^*)) \tag{14}$$

where

$$w_f(A) := \min_{\substack{x \in \Delta_{n-1} \\ Ax \neq y^*}} \max_{\ell, j} \left\{ \frac{\langle \nabla f(Ax), a_\ell - a_j \rangle}{\sqrt{2(f(Ax) - f(y^*))}} : \ell \in S(x), j \in \{1, \ldots, n\} \right\}.$$

Furthermore, the above parameter $w_f(A)$ satisfies

$$w_f(A) \geq \sqrt{\mu}\varrho(\tilde{A})$$

for $\tilde{A} = A - y^*$.

Proof: Since f is convex and ∇f is L-Lipschitz, we have

$$f(y) \leq f(y_k) + \langle \nabla f(y_k), y - y_k \rangle + \frac{L}{2}\|y - y_k\|^2.$$

Hence proceeding as in Theorem 2, it follows that if $\theta_k \le \theta_{\max}$ then for either $a = a_j - y_k$ or $a = y_k - a_\ell$ we have

$$\begin{aligned} f(y_{k+1}) &\le f(y_k) - \frac{\langle \nabla f(y_k), a\rangle^2}{2L\|a\|^2} \\ &\le f(y_k) - \frac{\langle \nabla f(y_k), a_\ell - a_j\rangle^2/4}{2L\|a\|^2} \\ &\le f(y_k) - \frac{w_f(A)^2}{4Ld(A)^2}(f(y_k) - f(y^*)). \end{aligned}$$

Therefore, again as in the proof of Theorem 2, it follows that

$$f(y_k) - f(y^*) \le \left(1 - \frac{w_f(A)^2}{4Ld(A)^2}\right)^{k/2} (f(y_0) - f(y^*)).$$

We next show the bound $w_f(A) \ge \sqrt{\mu}\varrho(\tilde{A})$. Since f is μ-strongly convex,

$$f(y) + \langle \nabla f(y), y^* - y\rangle + \frac{\mu}{2}\|y - y^*\|^2 \le f(y^*).$$

Thus, the inequality $a^2 + b^2 + 2ab \ge 0$ yields

$$f(y) - f(y^*) \le \frac{\langle \nabla f(y), y - y^*\rangle^2}{2\mu\|y - y^*\|^2}.$$

Hence from the construction of $\varrho(A)$ we get

$$\langle \nabla f(y), a_\ell - a_j\rangle^2 \ge \frac{\langle \nabla f(y), y - y^*\rangle^2}{\|y - y^*\|^2}\varrho(\tilde{A})^2 \ge 2\mu(f(y) - f(y^*))\varrho(\tilde{A})^2.$$

■

Observe that in a nice analogy to the bound in Proposition 1, we readily get the following lower bound on the ratio $\frac{w_f(A)}{\sqrt{L}d(A)}$ appearing in (14):

$$\frac{w_f(A)}{\sqrt{L}d(A)} \ge \sqrt{\frac{\mu}{L}} \cdot \frac{\varrho(\tilde{A})}{d(A)}.$$

5 Some properties of the restricted width

Throughout this section assume $A \in \mathbb{R}^{m\times n}$ is a nonzero matrix. As we noted in Section 3 above, when $0 \in \text{int}(\text{conv}(A))$ it follows that $\varrho(A) \ge |\rho(A)|$. Our next result establish a stronger lower bound on $\varrho(A)$ in terms of some quantities that generalize $\rho(A)$ to the case when $0 \in \partial\text{conv}(A)$. To that end, we recall some terminology and results from [4]. Assume $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m\times n}$ is a non-zero matrix.

Then there exists a unique partition $B \cup N = \{1, \ldots, n\}$ such that both $A_B x_B = 0,\ x_B > 0$ and $A_N^{\mathrm{T}} y > 0,\ A_B^{\mathrm{T}} y = 0$ are feasible. In particular, $B \neq \emptyset$ if and only if $0 \in \mathrm{conv}(A)$. Also, if $a_i = 0$ then $i \in B$.

The above canonical partition (B, N) allows us to refine the quantity $\rho(A)$ defined by (2) as follows. Let $L := \mathrm{span}(A_B)$ and $L^{\perp} := \{v \in \mathbb{R}^m : \langle v, y \rangle = 0 \ \text{ for all } \ y \in L\}$. By convention, $L = \{0\}$ and $L^{\perp} = \mathbb{R}^m$ when $B = \emptyset$. If $L \neq \{0\}$, let $\rho_B(A)$ be defined as

$$\rho_B(A) := \max_{y \in L, \|y\|=1} \min_{i \in B} \langle a_i, y \rangle .$$

Observe that if $B \neq \emptyset$, then $L = \{0\}$ only when $a_i = 0$ for all $i \in B$.

If $N \neq \emptyset$, let $\rho_N(A)$ be defined as

$$\rho_N(A) := \max_{y \in L^{\perp}, \|y\|=1} \min_{i \in N} \langle a_i, y \rangle .$$

When $L \neq \{0\}$, it can be shown [4] that $\rho_B(A) < 0$. Likewise, when $N \neq \emptyset$ it can be shown that $\rho_N(A) > 0$. In particular, the latter implies that

$$\rho_N(A) := \max_{y \in L^{\perp}, \|y\|=1} \min_{i \in N} \langle a_i, y \rangle = \max_{y \in L^{\perp}, \|y\| \leq 1} \min_{i \in N} \langle a_i^{\perp}, y \rangle , \qquad (15)$$

where $a_i^{\perp}$ is the orthogonal projection of a_i onto $L^{\perp}$. Let $A_N^{\perp}$ denote the matrix obtained by projecting each of the columns of A_N onto $L^{\perp}$. From (15) and Lagrangian duality it follows that

$$\rho_N(A) = \min\{\|y\| : y \in \mathrm{conv}(A_N^{\perp})\}. \qquad (16)$$

Similarly, it can be shown that if $L \neq \{0\}$ then

$$|\rho_B(A)| = \max\{r : y \in L, \|y\| \leq r \Rightarrow y \in \mathrm{conv}(A_B)\}. \qquad (17)$$

Observe that (16) and (17) nicely extend (3) and (4). Indeed, (16) is identical to (3) when $B = \emptyset$. Likewise, (17) is identical to (4) when $N = \emptyset$. Furthermore, (16) and (17) imply that $\rho_N(A) = \mathsf{dist}(0, \partial \mathrm{conv}(A_N^{\perp}))$ and $|\rho_B(A)| = \mathsf{dist}_L(0, \partial \mathrm{conv}(A_B))$ thereby extending the fact that $|\rho(A)| = \mathsf{dist}(0, \partial \mathrm{conv}(A))$.

The next result shows that $\varrho(A)$ can be bounded below in terms of $\rho_B(A)$ and $\rho_N(A)$. In particular, it shows that $\varrho(A) > 0$ whenever $A \neq 0$ and $0 \in \mathrm{conv}(A)$.

Theorem 4 *Assume* $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}$ *is a nonzero matrix.*

(a) *If* $N = \emptyset$ *then* $L \neq \{0\}$ *and* $\varrho(A) \geq |\rho_B(A)|$.

(b) *If* $B = \emptyset$ *then* $\varrho(\bar{A}) \geq \rho_N(A)$ *for* $\bar{A} := \begin{bmatrix} A & 0 \end{bmatrix}$.

(c) *If* $B \neq \emptyset$ *and* $L = \{0\}$ *then* $\varrho(A) \geq \rho_N(A)$.

(d) *If* $N \neq \emptyset$ *and* $L \neq \{0\}$ *then* $\varrho(A) \geq \dfrac{|\rho_B(A)|\rho_N(A)}{\sqrt{\|A\|^2 + \rho_N(A)^2}}$, *where* $\|A\| = \max\limits_{i=1,\ldots,n} \|a_i\|$.

Proof:

(a) Assume $x \geq 0$ is such that $y := Ax \neq 0$. In this case $y \in \operatorname{span}(A_B) = L$. Hence $L \neq \{0\}$ and by (17) there exists $v \in \Delta_{n-1}$ and $r \geq |\rho_B(A)|$ such that $-Av = \frac{r}{\|Ax\|}Ax$. Thus for $u := \frac{x}{\|x\|_1}$ we have $u, v \in \Delta_{n-1}$, $S(u) \subseteq S(x)$ and $Au - Av = \left(r + \frac{\|Ax\|}{\|x\|_1}\right)\frac{1}{\|Ax\|}Ax$. It follows that $\varrho(A, x) \geq r + \frac{\|Ax\|}{\|x\|_1} > |\rho_B(A)|$.

(b) Assume $\bar{x} := \begin{bmatrix} x \\ t \end{bmatrix} \geq 0$ is such that $y := \bar{A}\bar{x} = Ax \neq 0$. From (16) it follows that $\frac{\|Ax\|}{\|x\|_1} \geq \rho_N(A)$. Thus for $u := \begin{bmatrix} \frac{x}{\|x\|_1} \\ 0 \end{bmatrix}$, $v := e_{n+1}$ we have $u, v \in \Delta_n$, $S(u) \subseteq S(\bar{x})$ and $\bar{A}u - \bar{A}v = \frac{\|Ax\|}{\|x\|_1}\frac{1}{\|Ax\|}Ax$. It follows that $\varrho(\bar{A}, \bar{x}) \geq \frac{\|Ax\|}{\|x\|_1} \geq \rho_N(A)$.

(c) Since $B \neq \emptyset$ and $L = \{0\}$, it follows that $A_B = 0$ and the columns of A_N are precisely the non-zero columns of A. Thus from part (b) we get $\varrho\left(\begin{bmatrix} A_N & 0 \end{bmatrix}\right) \geq \rho_N(A)$. To finish, observe that $\varrho(A) = \varrho(\begin{bmatrix} A_N & 0 \end{bmatrix})$ because $A_B = 0$.

(d) Assume $x \geq 0$ is such that $y := Ax \neq 0$. Let $L := \operatorname{span}(A_B)$ and decompose $y = y_L + y_\perp$ where $y_\perp = A_N^\perp x_N \in L^\perp$ and $y_L = A_B x_B + (A_N - A_N^\perp)x_N \in L$. Put $r := \frac{\|y_\perp\|}{\|y\|} \in [0, 1]$. Assume $r > 0$ as otherwise $y = y_L \in \operatorname{span}(A_B)$ and the statement holds with the better bound $\varrho(A) \geq |\rho_B(A)|$ by proceeding exactly as in part (a). Since $r > 0$, we have $x_N \neq 0$. Put $r_N := \frac{\|y_\perp\|}{\|x_N\|_1}$. From (16) it follows that $r_N \geq \rho_N(A)$. Next, put $v := \frac{1}{\|x_N\|_1}\left((A_N - A_N^\perp)x_N - y_L\right)$. Observe that $\|v\| \leq \max\limits_{i \in N} \|a_i - a_i^\perp\| + \dfrac{\|y_L\|}{\|x_N\|_1} \leq \|A\| + \dfrac{r_N\sqrt{1-r^2}}{r}$ and $v \in L$. Hence by (17) there exists $\tilde{x}_B \geq 0, \|\tilde{x}_B\|_1 = 1$ such that $A_B\tilde{x}_B = cv$, where

$$c := \frac{|\rho_B(A)|r}{r\|A\| + r_N\sqrt{1-r^2}} \in (0, 1).$$

Taking $\tilde{x}_N := \frac{c}{\|x_N\|_1}x_N$ we get

$$A_N\tilde{x}_N - A_B\tilde{x}_B = \frac{c}{\|x_N\|_1}(y_\perp + y_L) = \frac{|\rho_B(A)|r_N}{r\|A\| + r_N\sqrt{1-r^2}}\frac{y}{\|y\|}.$$

Thus letting $u := (1-c)x + (0, \tilde{x}_N)$, $v = (\tilde{x}_B, 0)$ we get $u, v \in$

$\Delta_{n-1},\ S(u) \subseteq S(x)$ and

$$Au - Av = \left((1-c)\|Ax\| + \frac{|\rho_B(A)| r_N}{r\|A\| + r_N\sqrt{1-r^2}} \right) \frac{Ax}{\|Ax\|}. \quad (18)$$

Next, observe that

$$\begin{aligned} (1-c)\|Ax\| + \frac{|\rho_B(A)| r_N}{r\|A\| + r_N\sqrt{1-r^2}} &\geq \frac{|\rho_B(A)| r_N}{r\|A\| + r_N\sqrt{1-r^2}} \\ &\geq \frac{|\rho_B(A)| r_N}{\sqrt{\|A\|^2 + r_N^2}} \\ &\geq \frac{|\rho_B(A)| \rho_N(A)}{\sqrt{\|A\|^2 + \rho_N(A)^2}}. \end{aligned} \quad (19)$$

The first inequality above follows because $c \in (0,1)$, the second one follows from

$$\max_{r\in[0,1]} \left(r\|A\| + r_N\sqrt{1-r^2} \right) = \sqrt{\|A\|^2 + r_N^2},$$

and the third one follows from $r_N \geq \rho_N(A)$. Putting (18) and (19) together we get $\varrho(A,x) \geq \dfrac{|\rho_B(A)| \rho_N(A)}{\sqrt{\|A\|^2 + \rho_N(A)^2}}$. ■

References

[1] A. Beck and S. Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. Technical report, Technical Report, Faculty of Industrial Engineering and Management, Technion, 2015.

[2] H. D. Block. The perceptron: A model for brain functioning. *Reviews of Modern Physics*, 34:123–135, 1962.

[3] D. Cheung and F. Cucker. A new condition number for linear programming. *Math. Prog.*, 91(2):163–174, 2001.

[4] D. Cheung, F. Cucker, and J. Peña. On strata of degenerate polyhedral cones I: Condition and distance to strata. *Eur. J. Oper. Res.*, 19(198):23–28, 2009.

[5] K. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):63, 2010.

[6] M. Epelman and R. M. Freund. Condition number complexity of an elementary algorithm for computing a reliable solution of a conic linear system. *Math. Program.*, 88(3):451–485, 2000.

[7] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Quarterly*, 3:95–110, 1956.

[8] J. Goffin. The relaxation method for solving systems of linear inequalities. *Math. Oper. Res.*, 5:388–414, 1980.

[9] J. Guélat and P. Marcotte. Some comments on Wolfe's away step. *Math. Program.*, 35:110–119, 1986.

[10] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, volume 28 of *JMLR Proceedings*, pages 427–435, 2013.

[11] S. Lacoste-Julien and M. Jaggi. An affine invariant linear convergence analysis for Frank-Wolfe algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[12] D. Leventhal and A. Lewis. Randomized methods for linear constraints: Convergence rates and conditioning. *Math. Oper. Res.*, 35:641–654, 2010.

[13] D. Li and T. Terlaky. The duality between the perceptron algorithm and the von Neumann algorithm. In *Modeling and Optimization: Theory and Applications (MOPTA) Conference*, 2013.

[14] A. B. J. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

[15] J. Renegar. Incorporating condition measures into the complexity theory of linear programming. *SIAM J. on Optim.*, 5:506–524, 1995.

[16] J. Renegar. Linear programming, complexity theory and elementary functional analysis. *Math. Program.*, 70:279–351, 1995.

[17] N. Soheili and J. Peña. A primal–dual smooth perceptron–von Neumann algorithm. In *Discrete Geometry and Optimization*, pages 303–320. Springer, 2013.

[18] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15:262–252, 2009.

[19] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[20] P. Wolfe. Convergence theory in nonlinear programming. In *Integer and Nonlinear Programming*. North-Holland, Amsterdam, 1970.