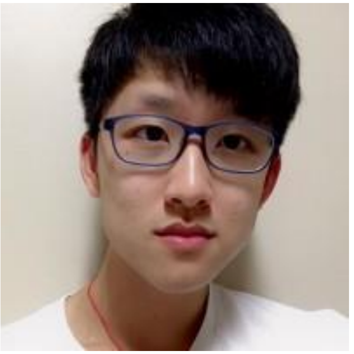
On the smoothed complexity of Frank-Wolfe methods

Luis Rademacher, UC Davis

ITA, February 2020

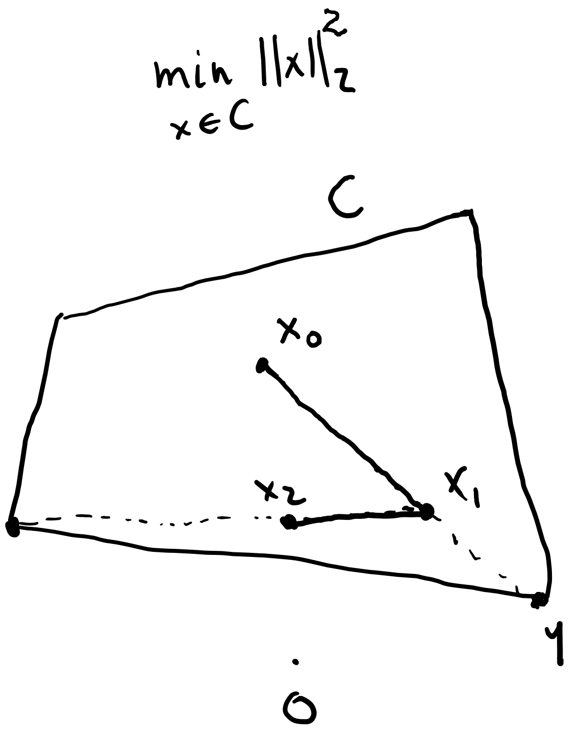
Joint work with Chang Shu



Frank-Wolfe methods

$$\min f(x)$$
$$s.t.\ x \in C$$

- $C \subseteq R^d$: a compact convex set
- $f: C \to R$: a differentiable function
- Basic iterative Frank-Wolfe method, to minimize:
    1. Start from any point $x_0 \in C$. Let $k = 0$.
    2. Repeat
        a. Find minimum, y, of $x \mapsto \left(\nabla f(x_k)\right)^T x$ over C.
        b. Let $x_{k+1} = x_k + \alpha^*(y - x_k)$, where α^* is a suitable step size.
        c. Let $k = k + 1$.



Wolfe's method

- A specialized refinement of F-W for

$$\min \|x\|_2^2$$
$$s.t.\ x \in P$$

where P is a polytope (bounded convex polyhedron).

Complexity

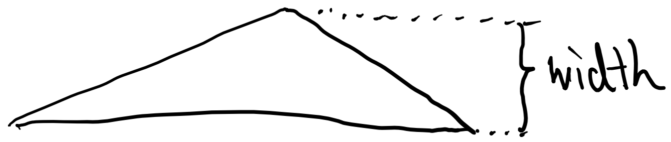
- [De Loera, Haddock, Rademacher] Exponential time lower bound for Wolfe’s method.
- Many results on linear convergence of F-W.
- [Lacoste-Julien, Jaggi ‘13 ‘15], [Beck, Shtern ‘15 ‘16], [Peña Rodriguez Soheili ’15 ‘17] [Peña Rodriguez] Global linear convergence of certain variations of F-W:
  - F-W with away steps,
  - pairwise F-W,
  - Wolfe’s method,

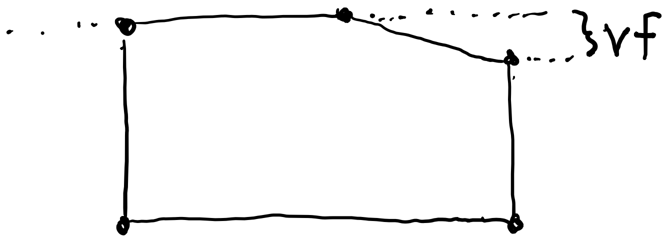
  when feasible region is a polytope $C = \mathrm{conv}(A)$.

  Speed depends on a condition number of $\boldsymbol{C}$.

Global linear convergence and polytope conditioning

- Linear convergence results depend on a "condition number" κ of polytope $C = \mathrm{conv}(A)$ (sketch):
$$\text{"}\ f(u_t) - f^* \le (1-\kappa)^t (f(u_0) - f^*)\ \text{"}$$
- If κ is small, convergence is slow.
- $\kappa = \frac{\text{"}something\text{"}}{\mathrm{diam}(C)}$, where "something" can be
  - [L-J J] $\mathrm{minwidth}(A) = \min_{S \subseteq A} \mathrm{width}(S)$
  - [L-J J] pyramidal width $\mathrm{PWidth}(A)$
  - [B S] vertex-facet distance $\mathrm{vf}(C) = \min_{F \in \mathrm{facets}(C)} d(\mathrm{aff}\, F, \mathrm{vertices}(C) \setminus F)$
  - [P R] facial distance $\Phi(C) = \min_{\substack{F \in \mathrm{faces}(C) \\ \emptyset \subsetneq F \subsetneq C}} d(F, \mathrm{conv}(\mathrm{vertices}(C) \setminus F))$





- Relationships:
  - [P R] $\mathrm{PWidth}(A) = \Phi(C)$
  - [L-J J] $\mathrm{minwidth}(A) \le \mathrm{PWidth}(A)$
  - **[our work]** $\Phi(C) \le \mathrm{vf}(C)$

  $\Rightarrow$ all are sandwiched between $\mathrm{minwidth}(A)$ and $\mathrm{vf}(C)$.
- All of them can be exponentially small as a function of bit-length of A [De Loera, Haddock, Rademacher].
- **[our work]** There is a 0-1 simplex in R^d where all of them are exponentially small in d (follows from observation of [L-J J], based on [Alon Vu '97]).

Smoothed analysis [Spielman Teng]

- Complexity of small random perturbations of any given input $x \in R^n$:

$$T(x+g)$$

where

1. g is N(0, $\sigma^2 I_{n\times n}$), and
2. T is "complexity" (e.g. time of an algorithm).

- (Probabilistic) polynomial smoothed complexity:

$$\max_{x \in \mathrm{R}^n, \|x\| \le 1} P_g \left[T(x+g) \ge \mathrm{poly}\left(n, \frac{1}{\sigma}, \delta\right) \right] \le \delta$$

Our results: simplex case

- Our results:
  - There is a 0-1 simplex in R^d where all condition numbers κ are exponentially small in d (follows from observation of [L-J J], based on [Alon Vu '97]).
  - "minwidth" has good smoothed complexity, implies polynomial smoothed complexity of several F-W methods for minimum norm on any simplex: Let A be matrix of vertices, then

$$P_g\left(\text{minwidth}(A+g) \geq \frac{1}{\text{poly}\left(d, \frac{1}{\sigma}\right)}\right) \geq 1 - o(1)$$

- Contrast with:
  - [De Loera, Haddock, Rademacher] Linear programming reduces to the minimum norm point on a simplex.

  ⇒No known "simple" worst case polynomial time algorithm to find the minimum norm point in a simplex.

Our results: general polytopes

- V-polytope $\mathrm{conv}(A)$.
  - **Smoothed vertex-facet distance is exponentially small.** For $g =$ “$2d$ standard Gaussian random points in R^d” and with constant probability :

$$\mathrm{vf}(g) \leq \frac{1}{c^d}$$

Proof idea for…

- … smoothed vertex-facet distance is exponentially small:
  - Want: for the convex hull of $2d$ random Gaussian points in R^d, with constant probability some vertex is exponentially close to aff(some facet) (not containing the vertex).
  - Warm-up case: given $2d$ random Gaussian points in R^d, with constant probability one point is exponentially close to span of $d - 1$ others.
  - Warm-up case relates to conditioning of random matrices and RIP in compressed sensing:
    - Warm up same as: given a $d \times 2d$ random Gaussian matrix, with constant probability there is a $d \times d$ submatrix with exponentially small σ_d.

Our results

- Bad smoothed conditioning of random matrices:
  Let A be a $d \times 2d$ random matrix with iid standard Gaussian entries. Then there exists $c > 1$ such that with constant probability
  $$\min_{S \subseteq [2d], |S|=d} \sigma_d(A_S) \leq 1/c^d$$

Proof idea:

- Warm-up case is enough: given $2d$ random Gaussian points in R^d, with constant probability one point is exponentially close to span of $d-1$ others
- Let F be the family of sets of $d-1$ columns of A. For $S \in F$, let B_S be the set of points within distance ϵ of $\mathrm{span}(S)$.
- Let $D_\epsilon = \cup_{S \in F} B_S$.
- Show that for $\epsilon = 1/c^d$ the Gaussian volume $\mathrm{G}(D_\epsilon)$ is at least a constant by lower bounding it using the **first two terms of inclusion-exclusion**:
  $$\mathrm{G}(D_\epsilon) \geq \sum_S \mathrm{G}(B_S) - \sum_{S,T} G(B_S \cap B_T)$$
- $B_S \cap B_T$ can be large if S, T share many columns. Restrict the definition of F above to a subfamily of submatrices of A having few columns in common: **packing bound = Gilbert-Varshamov bound**.

Conclusion

- We show
  - polynomial time smoothed complexity for several F-W methods for minimum norm point in a simplex.
  - Known notions of polytope conditioning do not have polynomial smoothed complexity.
  - New results about conditioning of random matrices and random polytopes.
- No smoothed exponential time lower bound for F-W, only smoothed exponential bound for known condition numbers.
  Q: Polynomial smoothed complexity for F-W via analysis of better condition number?