Tight Bounds for Approximate Carathéodory and Beyond

Vahab Mirrokni [*1] **Renato Paes Leme** [*1] **Adrian Vladu** [*2] **Sam Chiu-wai Wong** [*3]

Abstract

We present a deterministic nearly-linear time algorithm for approximating any point inside a convex polytope with a sparse convex combination of the polytope's vertices. Our result provides a constructive proof for the Approximate Carathéodory Problem (Barman, 2015), which states that any point inside a polytope contained in the ℓ_p ball of radius D can be approximated to within ϵ in ℓ_p norm by a convex combination of $O\left(D^2p/\epsilon^2\right)$ vertices of the polytope for $p \geq 2$. While for the particular case of $p = 2$, this can be achieved by the well-known Perceptron algorithm, we follow a more principled approach which generalizes to arbitrary $p \geq 2$; furthermore, this naturally extends to domains with more complicated geometry, as it is the case for providing an approximate Birkhoff-von Neumann decomposition. Secondly, we show that the sparsity bound is tight for ℓ_p norms, using an argument based on anti-concentration for the binomial distribution, thus resolving an open question posed by Barman. Experimentally, we verify that our deterministic optimization-based algorithms achieve in practice much better sparsity than previously known sampling-based algorithms. We also show how to apply our techniques to SVM training and rounding fractional points in matroid and flow polytopes.

1. Introduction

The (exact) Carathéodory Theorem is a fundamental result in convex geometry which states that any point u in a polytope $P \subseteq \mathbb{R}^n$ can be expressed as a convex combination of $n + 1$ vertices of P. The approximate version states that if one is willing to tolerate an error of ϵ in ℓ_p norm, $O\left(D^2p/\epsilon^2\right)$ vertices suffice to approximate u, where D is the radius of the smallest ℓ_p ball enclosing P. The key significance of the approximate Carathéodory Theorem is that the bound it provides is *dimension-free*, and consequently allows us to approximate any point inside the polytope with a *sparse* convex combination of vertices.

The Approximate Carathéodory Problem Given a polytope P contained inside the ℓ_p ball of radius D, and $u \in P$, find vertices $v_1, \ldots, v_k$ of P such that $k = O\left(D^2p/\epsilon^2\right)$ and $\left\| \frac{1}{k}\sum_{i=1}^{k} v_i - u \right\|_p \leq \epsilon$.

The ℓ_2 version of this result is quite an old observation. The earliest record is perhaps due to Novikoff (1962) who showed that the ℓ_2 version of Approximate Carathéodory can be obtained as a byproduct of the analyis of the Perceptron Algorithm (as pointed out by (Blum et al., 2016)). The fact that a sparse approximation can be obtained by a very simple and efficient algorithms found many applications in Machine Learning. Shalev-Shwartz et al. (2010) use it to minimize the loss of a linear predictor using a small number of features. Garber & Hazan (2013) use it to speed up conditional grandient methods.

The results described above focus on the ℓ_2 norm. The interest for approximate Caratheodory in higher ℓ_p-norms was sparked by a recent result of Barman (2015) who used it to improve algorithms for computing Nash equilibria in game theory and algorithms for the k-densest subgraph in combinatorial optimization. Another area where higher norms are widely applied is in functional analysis where the approximate Caratheodory Theorem is often referred as Maurey's Lemma (Pisier, 1980).

Both Barman's proof and Maurey's original proof start from a solution $u = \sum_{i=1}^{n+1} \lambda_i v_i$ of the exact Carathéodory problem, interpret the coefficients λ_i of the convex combination as a probability distribution and generate a sparse solution by sampling from the distribution induced by λ. Concentration inequalities are then applied to argue that the average sampled solution is close to u in ℓ_p-norm. The proof is clean and elegant, but is computationally expensive since it involves first computing a solution to the exact Carathéodory

[*]Equal contribution [1]Google Research, New York, NY, USA [2]MIT, Cambridge, MA, USA [3]UC Berkeley, Berkeley, CA, USA. Correspondence to: Vahab Mirrokni <mirrokni@google.com>, Renato Paes Leme <renatoppl@google.com>, Adrian Vladu <avladu@mit.edu>, Sam Chiu-wai Wong <samcwong@berkeley.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

problem, which can take $O(n^\omega)$ even if the vertices are given explicitly. The situation becomes even worse for polytopes where it is not desirable to maintain an explicit representation of all its vertices (e.g. the matching polytope) since there may be exponentially many of them.

This is contrast with simple iterative solutions like the Perceptron for ℓ_2, which runs in nearly-linear time. The first question we explore in this paper is how to obtain a deterministic nearly-linear time algorithm for higher ℓ_p norms. Our algorithm runs in $O(D^2p/\epsilon^2)$ iterations, each of which takes linear time.[1]

Secondly, we resolve an open question posed by (Barman, 2015), who observed that the bound for the ℓ_2 bound was tight and asked whether the ℓ_p bound was also tight. Barman gave a $\Omega((D/\epsilon)^{p/(p-1)})$ lower bound for $p \geq 2$. We resolve the question by showing that the $O(D^2p/\epsilon^2)$ is tight by exhibiting a polytope P in the radius-D ℓ_p ball and a point u inside for which all convex combinations of $o(D^2p/\epsilon^2)$ vertices are more than ϵ-far from u in ℓ_p-norm.

Even though the dependence on ϵ cannot be improved in general, it can be greatly improved in a special case. If u is far away from the boundary of P, i.e., if the ball of radius r around u is contained in P, then there exists a solution to the approximate Carathéodory problem with $k = O\left(\frac{D^2p}{r^2}\log\left(\frac{r}{\epsilon}\right)\right)$.

For the positive result, our technique involves writing approximate Carathéodory as a convex minimization problem and solving it by running *Mirror Descent* on a dual convex function obtained via Sion's Theorem. Our technique is inspired by the similarity with the problems of computing Nash equilibria in games and solving packing-covering LPs. When $p = \log n$, our bound has the same sparsity as Lipton and Young (Lipton & Young, 1994) and Plotkin, Shmoys and Tardos (Plotkin et al., 1991).

The view of approximate Carathéodory as solving a zero-sum game also leads to our lower-bound, adapting a method of Klein and Young (Klein & Young, 2015) for proving conditional lower bounds on the running time for solving positive LPs.

To show the potential of our technique, we note that a simple extension of our method implies a new algorithm for SVM training. More specifically, we obtain $O(1/\epsilon^2)$ convergence for arbitrary kernels; each iteration only requires matrix-vector operations involving the kernel matrix, so we overcome the obstacle of having to explicitly store the kernel or compute its Cholesky factorization.

Finally, we show that our algorithm can also be obtained by an instantiation of the Frank-Wolfe algorithm. One remarkable feature of our problem is that it connects three ways in which sparsification has been done: via Mirror Descent (or more commonly, via multiplicative weight update) as in (Plotkin et al., 1991; Arora et al., 2012; Juditsky et al., 2013), via Frank-Wolfe methods (Garber & Hazan, 2013; Jaggi, 2013) and by sampling (Lipton & Young, 1994; Lipton et al., 2003).

[1] A reason to consider arbitrary ℓ_p norms for $p \geq 2$ is that they are particularly useful for inputs that are bounded in ℓ_∞ and benefit from extra structure such as k-sparsity. Then, we can run the algorithm for the $\ell_{\log k}$ norm and obtain the desired ℓ_∞ guarantee using only $O(\log k/\epsilon^2)$ points, which is an improvement over the $O(\log n/\epsilon^2)$ bound one could obtain by ignoring the sparsity.

2. Preliminaries

For $x \in \mathbb{R}^d$, we define its ℓ_p-norm as $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ for $p \geq 1$ and ℓ_∞ norm by $\|x\|_\infty = \max_i |x_i|$. We note the ℓ_p ball as $\boldsymbol{B}_p(r) = \{x \in \mathbb{R}^d; \|x\|_p \leq r\}$.

Given a norm $\|\cdot\|$, we define its dual norm $\|\cdot\|_*$ as $\|y\|_* = \max_{x:\|x\|=1} y^\top x$ so that Hölder's inequality holds with equality: $y^\top x \leq \|y\|_* \|x\|$. The dual norm of ℓ_p norm is ℓ_q norm for $\frac{1}{p} + \frac{1}{q} = 1$.

We also denote the support of x by $\text{supp}(x) = \{i|x_i \neq 0\}$.

2.1. Approximate Carathéodory problem

The (exact) Carathéodory Theorem is a fundamental result in linear algebra which bounds the number of points needed to describe a point in the convex hull of a set. More precisely, given a finite set of points $X \subseteq \mathbb{R}^d$ and $u \in \text{conv}(X) := \{\sum_{x\in X} \lambda_x \cdot x : \sum_x \lambda_x = 1, \lambda_x \geq 0\}$, there exist $d+1$ points in $x_1, \ldots, x_{d+1} \in X$ such that $u \in \text{conv}\{x_1, \ldots, x_{d+1}\}$. On the plane, in particular, every point in the interior of a convex polygon can be written as a convex combination of three of its vertices.

The approximate version of Carathéodory theorem bounds the number of points needed to describe $u \in \text{conv}(X)$ approximately. Formally, given a norm $\|\cdot\|$, an additive error parameter ϵ and a set of points $X \subseteq B_{\|\cdot\|}(1) \subseteq \mathbb{R}^d$, for given $u \in \text{conv}(X)$ we want k points $x_1, \ldots, x_k \in X$ such that there exists $u' \in \text{conv}\{x_1, \ldots, x_k\}$ and $\|u - u'\| \leq \epsilon$.

A general result of this type is given by Maurey (Pisier, 1980). For ℓ_p norm, $p \geq 2$, Barman (Barman, 2015) showed that $k \leq 4p/\epsilon^2$ points suffice. Notably, this bound is independent of the dimension of the ambient space.

Mirror Descent An overview is in the supplementary material. We simply state the guarantee here. Let ω be σ^{-1}-strongly convex w.r.t. norm $\|\cdot\|$ and ω^* be its Fenchel dual. For ρ-Lipschitz convex $f : Q \longrightarrow \mathbb{R}$ (w.r.t. norm $\|\cdot\|$), Mirror Descent computes iterates with stepsize η as follows:

$$z_{t+1} = z_t - \eta \nabla f(y_t) \qquad y_{t+1} = \nabla \omega^*(z_{t+1}) \quad \text{(MD)}$$

Let $D_\omega(y\|x) := \omega(y) - \omega(x) - \nabla\omega(x)^\top(y-x)$ be the Bregman divergence.

Theorem 2.1. *In the setup described above with $D = \max_{z\in Q} D_\omega(z\|z_0)$ and $\eta = \epsilon/\sigma\rho^2$, then in $T \geq 2D\sigma\rho^2/\epsilon^2$ iterations, it holds that $\frac{1}{T}\sum_t \nabla f(y_t)^\top(y_t - y) \leq \epsilon, \forall y \in Q$.*

3. Nearly linear time deterministic algorithm

We present a nearly linear time deterministic algorithm for the approximate Carathéodory Problem. Barman's original proof (Barman, 2015) involves solving the exact Carathéodory problem: (i) write $u = \sum_x x \cdot \lambda_x$ and interpret λ as a probability distribution over X; (iii) sample k points from X according to λ and; (iv) argue by concentration bounds (Khintchine inequality to be precise) that the expectation $\mathbb{E}\left\|u - \frac{1}{k}\sum_{i=1}^k x_i\right\|_p \leq \epsilon$. From an algorithmic point of view, this requires solving a linear program to compute λ and using randomness to sample x_i. Our main theorem shows that *neither is necessary*. There is a linear time deterministic algorithm that doesn't require a solution λ to the exact problem.

Our algorithm is based on Mirror Descent. The idea is to formulate the Carathéodory problem as an optimization problem. Inspired by early positive Linear Program solvers e.g. Plotkin-Shmoys-Tardos (Plotkin et al., 1991), we convert this to a saddle point problem and solve *its dual* using Mirror Descent. Using Mirror Descent guarantees a sparse primal certificate that would act as the desired convex combination.

Recall that we are given a finite set of points $X = \{v_1, v_2, \ldots, v_m\} \subseteq \boldsymbol{B}_p(1)$ and $u \in \text{conv}(X)$. Our goal is to produce a sparse convex combination of the points in X that is ϵ-close to u in ℓ_p-norm. Dropping the sparsity constraint for now, we can formulate this problem as:

$$\min_{x\in\Delta} \|Vx - u\|_p \qquad \text{(P-CARA)}$$

where V is a $d \times m$ matrix whose columns are the vectors $v_1, \ldots, v_m$ and $\Delta = \{x \in \mathbb{R}^d | \sum_i x_i = 1, x \geq 0\}$ is the unit simplex. We refer to P-CARA as the primal Carathéodory problem. By writing ℓ_p norm as $\|x\|_p = \max_{y:\|y\|_q=1} y^\top x$ for $\frac{1}{p} + \frac{1}{q} = 1$, P-CARA is converted to a saddle point problem:

$$\min_{x\in\Delta} \max_{y\in\boldsymbol{B}_q(1)} y^\top(Vx - u) \qquad \text{(S-CARA)}$$

Sion's Theorem (Sion, 1958) is a generalization of von Neumann's minimax theorem that allows us to swap the order of minimization and maximization for any pair of compact convex sets. This leads to dual problem: $\max_{y\in\boldsymbol{B}_q(1)} \left(\min_{x\in\Delta} y^\top(Vx - u)\right)$ which can be rewritten as:

$$-\min_{y\in\boldsymbol{B}_q(1)} \left(f(y) := \max_{x\in\Delta} y^\top(u - Vx)\right) \qquad \text{(D-CARA)}$$

Sparse solution by solving the dual. Since $u \in \text{conv}(X)$, there is a solution $x \in \Delta$ such that $u = Vx$. So P-CARA (and equivalent formulations S-CARA and D-CARA) have an optimal value of 0. Although the optimal value is known, it still helps to optimize $f(y)$ since in the process we obtain an ϵ-approximation in few iterations. If each iteration updates only one coordinate of x, then we will obtain an approximate solution with sparsity equal to the number of iterations. As we shall show, while the updates of y are not sparse, the dual certificate produced by Mirror Descent will be.

To make this statement precise, consider the gradient of f, which is obtained by applying the envelope theorem (see (Afriat, 1971)): $\nabla f(y) = u - Vx$ for $x \in \arg\max_{x\in\Delta} y^\top(u - Vx)$. This problem corresponds to maximizing a linear function over the simplex, so the optimal solution is a corner of the simplex. In other words, $\nabla f(y) = u - v_i$ where $i = \arg\max_i[-(V^\top y)_i]$. We can then use the Mirror Descent guarantee in Theorem 2.1 to bound the norm of the average gradient, as formalized in Theorem 3.2.

Remark 3.1. *In fact V does not even have to be explicitly given. All we need is to solve $i = \arg\max_i[-(V^\top y)_i]$. For explicitly given V, this can be done in dn time by picking the best vertex. Sometimes, especially in combinatorial optimization, we have a polytope (whose vertices are V) represented by its constraints. Our result states that for these alternate formulations, we can still obtain a sparse representation efficiently if we can solve the linear optimization problem over it fast.*

Theorem 3.2. *Consider a $(1/\sigma)$-strongly convex function $\omega : \boldsymbol{B}_q(1) \to \mathbb{R}$ with respect to the ℓ_q-norm, $D = \max_{y\in\boldsymbol{B}_q(1)} D_\omega(y\|0)$ and $T \geq 8D\sigma/\epsilon^2$. Let $y_1 = 0, \ldots, y_T$ be the T first iterates of the Mirror Descent algorithm (Theorem 2.1) with mirror map $\nabla\omega^*$ minimizing function f in D-CARA. If $\nabla f(y_t) = u - v_{i(t)}$, then*

$$\left\|u - \frac{1}{T}\sum_{t=1}^T v_{i(t)}\right\|_p \leq \epsilon.$$

Proof. We consider the space $y \in \boldsymbol{B}_q(1)$ equipped with the ℓ_q norm. To apply the Mirror Descent framework, we need first to show that the dual norm (the ℓ_p-norm, in this

case) of the gradient is bounded. This is easy, since in the approximate Carathéodory problem, $v_i \in \boldsymbol{B}_p(1)$, so $\|\nabla f(y)\|_p = \|u - v_i\|_p \le \|u\|_p + \|v_i\|_p \le 2$. So we can take $\rho = 2$ in Theorem 2.1.

Since $f(y) = \max_{x\in\Delta} y^\top(u - Vx)$ and $\nabla f(y) = (u - Vx)$ for $x \in \arg\max_{x\in\Delta} y^\top(u - Vx)$, then $f(y) = \nabla f(y)^\top y$. Also, since there exists x^* such that $u = Vx^*$, one has that $f(y) \ge y^\top(u - Vx^*) = 0$. Plugging those two facts in the guarantee of Theorem 2.1, we get:

$$\epsilon \ge \frac{1}{T}\sum_{t=1}^{T} \nabla f(y_t)^\top (y_t - y) = \frac{1}{T}\sum_{t=1}^{T} [f(y_t) - \nabla f(y_t)^\top y]$$
$$\ge \left[-\frac{1}{T}\sum_{t=1}^{T} \nabla f(y_t)\right]^\top y, \forall y \in \boldsymbol{B}_q(1)$$

Taking the maximum over all $y \in \boldsymbol{B}_q(1)$ we get:

$$\left\| u - \frac{1}{T}\sum_{t=1}^{T} v_{i(t)} \right\|_p = \left\| \frac{1}{T}\sum_{t=1}^{T} \nabla f(y_t) \right\|_p$$
$$= \max_{y\in\boldsymbol{B}_q(1)} \left[-\frac{1}{T}\sum_{t=1}^{T} \nabla f(y_t)\right]^\top y \le \epsilon$$

□

To complete the picture, we need to exhibit a $(1/\sigma)$-strongly convex function $\omega : \boldsymbol{B}_q(1) \to \mathbb{R}$ with a small value of $\sigma \cdot \max_{y\in\boldsymbol{B}_q(1)} D_\omega(y\|0)$ and show that the gradient of the Fenchel dual $\nabla\omega^*$ can be computed efficiently. In supplementary material we show that it suffices to use $\omega(y) = \frac{1}{2}\|y\|_q^2$. We also discuss the form of the Fenchel dual ω^* and how to compute $\nabla\omega^*$. We note that because ω is defined in the ball $\boldsymbol{B}_q(1)$ its Fenchel dual is different from that of the function $\frac{1}{2}\|y\|_q^2$ defined in $\mathbb{R}^d$.

Proposition 3.3. *The Fenchel dual of $\omega : \boldsymbol{B}_q(1) \to \mathbb{R}$, $\omega(y) = \frac{1}{2}\|y\|_q^2$ can be computed explicitly:*

$$\omega^*(z) = \begin{cases} \frac{1}{2}\|z\|_p^2 & \text{if } \|z\|_p \le 1 \\ \|z\|_p - \frac{1}{2} & \text{if } \|z\|_p > 1 \end{cases}$$

Also, $\nabla\omega^*(z) = \phi(z)\cdot\min(1, \|z\|_p)$ where $\phi(z)$ is a vector with ℓ_q-norm 1 such that $z^\top\phi(z) = \|z\|_p$. This function can be explicitly computed as: $\phi(z)_i = \operatorname{sgn}(z_i)\cdot |z_i|^{p-1} / \|z\|_p^{p-1}$.

Theorem 3.4. *Given n points $v_1, \dots, v_n \in \boldsymbol{B}_p(1) \subseteq \mathbb{R}^d$ with $p \ge 2$ and $u \in \operatorname{conv}\{v_1, \dots, v_n\}$, there is a deterministic algorithm of running time $O(nd\cdot p/\epsilon^2)$ that a outputs a multiset $v_{i(1)}, \dots, v_{i(k)}$ for $k = 4(p-1)/\epsilon^2$ such that $u' = \frac{1}{k}\sum_{t=1}^{k} v_{i(t)}$ and $\|u' - u\|_p \le \epsilon$.*

3.1. Improved bound when u is far from the boundary

If the point u that we are approximating is sufficiently far from the boundary of the polytope P, it is possible to make recursive calls to the algorithm described in the previous section, doubling the precision in each iteration. This allows us to obtain a significantly better sparsity guarantee.

Theorem 3.5. *Let P be a polytope contained inside the unit ℓ_p ball, and a point $u \in P$. If $\boldsymbol{B}_p(r) \subseteq P$, then there exists $x \in 2(1-\epsilon/r)\cdot\Delta$ supported at $k = O\left(\frac{p}{r^2}\cdot\log\frac{r}{\epsilon}\right)$ coordinates such that $\left\|\sum_{i\in\text{supp}(x)} x_i v_i - u\right\|_p \le \epsilon$.*

Corollary 3.6. *If $u \in P$ satisfies $\boldsymbol{B}_p(u, r) \subseteq P \subseteq \boldsymbol{B}_p(u, 1)$, $r \ge 2\epsilon$, then there exists $x \in \Delta$ supported on $k = O\left(\frac{p}{r^2}\cdot\log\frac{r}{\epsilon}\right)$ coordinates such that $\left\|\sum_{i\in\text{supp}(x)} x_i v_i - u\right\|_p \le \epsilon$.*

This highlights an interesting feature, namely that we can achieve linear convergence via an ad-hoc method, even though the dual formulation we are optimizing does not immediately exhibit strong convexity. This is achieved via iteratively rescaling the problem after solving to some fixed accuracy depending on the parameter r. The description of the improved algorithm can be found in Section D of the supplementary material. .

Even more interestingly, the primal version of this problem, which can be solved via the conditional gradient method (see Section 3.2), does exhibit strong convexity; this can then be used to provide a comparable guarantee, using a purely primal method described in (Garber & Hazan, 2015). We also mention (Lacoste-Julien & Jaggi, 2015; Shtern & Beck, 2016; Peña et al., 2016), which describe a similar phenomenon occurring under various specific assumptions involving the domain. Such methods show up under the name "accelerated Frank-Wolfe". The regimes in which they work are however different from the one we are considering here.

3.2. Sparse solution via conditional gradient methods

The algorithm described in the previous section admits a completely different analysis via conditional gradient methods, more precisely, via the Frank-Wolfe algorithm (Jaggi, 2013; Bubeck, 2014). The Frank-Wolfe method solves a problem of the type $\min_{x\in X} f(x)$ for a β-smooth convex function f over a compact convex set X via successive calls to a linear optimization oracle, that given a vector $w \in \mathbb{R}^d$ returns $x_w \in \arg\max_{x\in X} w^\top x$. Formally, start with any point $x_0 \in X$ and define the following iteration:

$$y_t = \arg\min_{y\in X} \nabla f(x_{t-1})^\top y \qquad x_t = (1-\eta_t)x_{t-1} + \eta_t\cdot y_t \tag{FW}$$

Frank-Wolfe guarantees that if η_t are suitably chosen ($\eta_t = \frac{2}{t+1}$ being a popular choice), then $f(x_t) - f(x^*) \le$

$2\beta R^2/(t+1)$ for β-smooth f w.r.t. some norm $\|\cdot\|$ and R the radius of X w.r.t. the same norm.

A remarkable fact is that the algorithm that we obtain from instantiating the Frank-Wolfe framework for our problem is completely isomorphic to the mirror descent version, in the sense that they produce the same set of vertices.

Theorem 3.7. *For $f(x) = \|x-u\|_p^2$, $X = P$ and $\eta_t = 1/t$ then for each t, the vertex y_t output by the Frank-Wolfe algorithm is the same vertex output by Mirror Descent described in Theorem 3.2.*

3.3. Connection between Frank Wolfe and Mirror Descent

Theorem 3.7 is an example of a setting where the same algorithm can be obtained from a completely primal view, through Frank-Wolfe and through a saddle point formulation, via Mirror Descent. The Frank-Wolfe approach is more standard in optimization while the Mirror Descent approach is standard in game theory and in first-order methods in Linear Programming. One might suspect that there is a deeper connection between the two algorithms.

In what follows we point out a simple but somewhat surprising observation: Frank-Wolfe methods to minimize f over a compact set X can be obtained by instantiating the Mirror Descent framework for minimizing a dualized version of f when the mirror map is the Fenchel dual of the objective itself.

A similar connection between Mirror Descent and Frank-Wolfe methods for a different class of problems was shown by Bach (Bach, 2015). We believe both observations are facets of the same phenomenon.

In what follows we present a very short and clean argument for why, in our specific instance, the Frank-Wolfe and Mirror Descent yield the same results.

By writing f as the dual of its Fenchel dual and applying Sion's min-max theorem, we obtain:

$$\begin{aligned}\min_{x\in X} f(x) &= \min_{x\in X}\left[\max_{z\in X^*} z^\top x - f^*(z)\right] \\ &= \max_{z\in X^*}\left[\min_{x\in X} z^\top x - f^*(z)\right]\end{aligned}$$

Define $g(z) = \min_{x\in X} z^\top x - f^*(z)$ which is a concave function over X^*. By the envelope theorem:

$$\nabla g(z) = y - \nabla f^*(z) \text{ where } y \in \arg\min_{x\in X} z^\top x$$

The mirror descent iteration for maximizing g can be written as: $x_{t+1} = x_t + \eta_t \nabla g(z_t)$ and $z_{t+1} = \nabla\omega^*(x_{t+1})$. By choosing $\omega^*(x) = f(x)$, we exactly recover Frank-Wolfe since: $\nabla g(z_t) = y_t - \nabla f^*(\nabla f(x_t)) = y_t - x_t$, so $x_{t+1} = x_t + \eta_t\nabla g(z_t) = (1-\eta_t)x_t + \eta_t y_t$ where $y_t = \arg\min_{y\in X} z_t^\top y = \arg\min_{y\in X} \nabla f(x_t)^\top y$.

While the proof of Bach is similar in spirit to ours, it assumes a different setup. First of all, both his and our proofs work on the dual objective obtained via Sion's min-max theorem. However, instead of directly using f^* as a mirror map, Bach adds an extra strongly-convex regularizer to his objective, which he carries through as a proximal term. This guarantees that the dual problem he solves is smooth, thus achieving $1/t$ convergence rate. Distinctively, the proof we have shown above is applied directly on the dualized objective with f^* as mirror map, and only achieves $1/\sqrt{t}$ convergence rate; however, this rate is tight for the specific problem we are studying.

4. Experiments

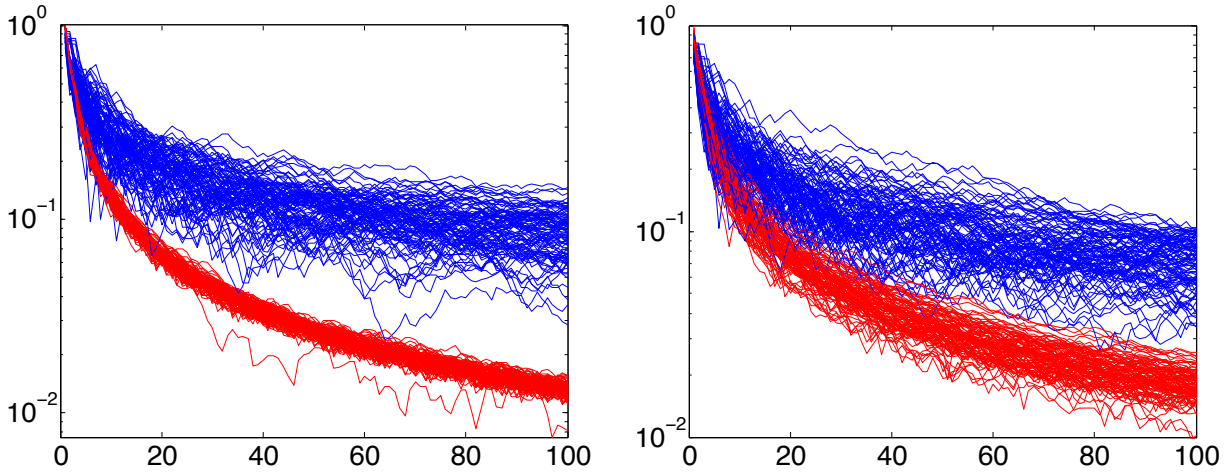
We illustrate the performance of our algorithm in two numerical experiments, presented in the figure below. We ran both the original sampling algorithm (Barman, 2015; Pisier, 1980) (where vertices are sampled from an exact convex combination) and our deterministic mirror-descent based algorithm on 100 instances. Each of these instances consisted of 1000 vectors in $\mathbb{R}^{1000}$, obtained by sampling a 1000×1000 Gaussian matrix, then scaling each column by the maximum ℓ_2 (respectively ℓ_8) column norm. For each instance we choose a convex combination of u and plot the error $\left\|u - \frac{1}{t}\sum_{s=1}^t v_s\right\|_p$ when we sample v_t at random proportionally to the exact convex combination (blue plot), and when we use the point v_t output by the t-th iteration of Mirror Descent (red plot). We do it for both the ℓ_2 and ℓ_8 norms, where in each case the input vectors are re-scaled to have unit ℓ_p-norm and the errors are measured with respect to the ℓ_p norm.

The plots from the 100 instances are overlapped, in order to highlight how mirror descent performs systematically better than random sampling.

One interesting observation is that mirror descent still performs better in practice despite the fact that rescaled Gaussian matrices are the worst-case instances for the problem, as we show in the next section, in the sense that for those families of instances, both the sampling and Mirror Descent algorithm are guaranteed to be optimal up to constant factors.

5. Lower bound

We showed that if V is a $d\times n$ matrix whose columns are contained in the unit ℓ_p ball $\boldsymbol{B}_p(1)$, then for any $x\in\Delta_n$ there is $\tilde{x}\in\Delta_n$ with $|\text{supp}\,(\tilde{x})| \le O(p/\epsilon^2)$ such that $\|Vx - Vx'\|_p \le \epsilon$, where $\text{supp}(x) = \{i | x_i \neq 0\}$.



(a) ℓ_2 error objective (b) ℓ_8 error objective

Figure 1. Quality of the solution, i.e. norm of the error (ℓ_2, respectively ℓ_8) as a function of sparsity. Blue curves correspond to sampling, red curves correspond to mirror descent. We overlapped the plots from 100 instances, in order to highlight that mirror descent performs systematically better than random sampling. This is apparent in both cases, where one can see that the red curves approach zero faster than the blue ones.

In this section we argue that no dimension-independent bound better then $O(p/\epsilon^2)$ is possible. This shows that the sparsity bound in the approximate Carathéodory theorem is tight and improves Barman's $\Omega(1/\epsilon^{p(p-1)})$ lower bound (Barman, 2015).[2] Formally, we show that:

Theorem 5.1. *There exists a constant K such that for every $p \geq 2$ and $n \geq n_0(p)$, there exists $n \times n$ matrix V with columns of unit ℓ_p norm, and a point $u = Vx$, $x \in \Delta_n$, such that for all $\tilde{x} \in \Delta$ with sparsity $|\text{supp}\,(\tilde{x})| \leq Kp/\epsilon^2$, one has that $\|V\tilde{x} - u\|_p \geq 2\epsilon > \epsilon$.*

In other words, even though u is a convex combination of columns of V, every (Kp/ϵ^2)-sparse convex combination of columns of V has distance at least ϵ from u in ℓ_p-norm. The full proof is in supplementary material.

Our lower bound incidentally implies that the optimal rate of conditional gradient applied to a p-smooth function is $O(pR^2/\epsilon)$; this can be seen by considering the function exhibited in Theorem 3.7 and noticing that minimizing it via conditional gradient to accuracy ϵ^2 requires $\Omega(pR^2/\epsilon^2)$ iterations, since each iteration increases the number of nonzero coordinates of the solution by at most 1, but $\Omega(pR^2/\epsilon^2)$ nonzeros are required, as shown by our lower bound for approximate Carathéodory.

Here we present a simple, constructive instance from which we easily prove a $\Omega(1/\epsilon^2)$ lower bound, and sketch a tight $\Omega(p/\epsilon^2)$ bound based on the probabilistic method.

[2] In addition to this, lower bounds for the case $p = 2$ were folklore; some proofs can be found in (Jaggi, 2011; Bubeck, 2014). We point out that in this case, the simple proof from Section 5.1 follows a very different approach from the classical ones.

5.1. A simple lower bound $\Omega(1/\epsilon^2)$

This relies on Sylvester's construction of Hadamard matrices, which are defined for n's that are powers of 2. The construction is recursive as follows: $H_1 = [1]$, and for every n that is a power of 2:

$$H_{2n} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}$$

Proposition 5.2. *The Sylvester matrix H_n defined as above is Hadamard. In other words, $H_{ij} = \pm 1$ for all i, j and $H^T H = nI$ (i.e. its columns are mutually orthogonal).*

Now we consider the polytope P formed by the convex hull of the normalized columns of H. One can easily check that for the construction above the uniform combination of columns is $H \cdot \vec{1}/n = e_1$ where $\vec{1}$ is the vector of all 1's and e_i is the unit basis vector for the i-th component. We show that e_1 is at distance greater than ϵ from the convex hull of any $o(1/\epsilon^2)$ columns of H.

Theorem 5.3. *Let H_n be as above and P be the convex hull of the columns of $\tilde{H} := H/n^{1/p}$. Let $u = \tilde{H} \cdot \vec{1}/n = e_1/n^{1/p} \in P$. Then any $x \in \Delta_n$ satisfying $\left\|\tilde{H}x - u\right\|_p \leq \epsilon$ has sparsity $|\text{supp}\,(x)| \geq \min(1/\epsilon^2, n)$.*

5.2. Tight lower bound $\Omega(p/\epsilon^2)$

We now establish a tight lower bound via a probabilistic existence argument inspired by the construction of Klein

and Young (Klein & Young, 2015). The example used to exhibit the lower bound is very simple. The proof of its validity, however, is quite involved and requires a careful probability analysis. We give an overview and provide details in the supplementary material.

Overview. Recall the formulation S-CARA of the Carathéodory problem as a saddle point problem described in Section 3. If we translate all points such that $u = 0$, then we can write the problem as:

$$\min_{x \in \Delta} \max_{y \in \boldsymbol{B}_q(1)} y^\top V x$$

which can be seen as a game between a player controlling x and y. The approximate Carathéodory theorem states that if the value of the game is 0, then the x-player has a k-sparse strategy that guarantees a value of the game at most ϵ for $k = O(p/\epsilon^2)$.

For the lower bound, our goal is to design an instance of this game with value v such that for all k-sparse strategies of the x-player with $k < Cp/\epsilon^2$, the y-player can force the game to have a value strictly larger than $v + \epsilon$.

Probabilistic Construction: We define the matrix $V = n^{-1/p} \cdot A$ where A is an $n \times n$ matrix with random ± 1 entries, i.e. each entry of A is chosen at random from $\{-1, +1\}$ independently with probability $1/2$. Note the the ℓ_p norm of the columns of A is equal to $D = 1$ (as in Approximate Carathéodory). We will show that the following events happen with high probability:

1. The center of the polytope defined by the columns of V is ϵ-close to $\vec{0}$, i.e., $\left\| V \cdot \vec{1}/n \right\|_p \leq \epsilon$.

2. For each set S of k coordinates, if x is restricted to only S, the y-player can force the value of the game to be at least 2ϵ. We prove so by exhibiting a strategy for the y-player such that $y^\top V$ is at least 2ϵ for all coordinates in S.

After bounding the probabilities of the events above, the result follows by taking the union bound over all $\binom{n}{k}$ possible subsets S of cardinality k. This implies that with nonzero probability, for the matrix constructed the y-player will always be able to force $y^T V x \geq 2\epsilon$, regardless of what $o(p/\epsilon^2)$-sparse strategy the x-player chooses.

6. Applications

In the following, we discuss a number of applications of our results and techniques. We briefly describe each of them here and refer the reader to the supplementary material for complete exposition.

Approximate Birkhoff-von Neumann Decomposition The classical Birkhoff-von Neumann Theorem states that any $n \times n$ doubly stochastic matrix can be decomposed into a convex combination of at most $(n-1)^2 + 1$ permutation matrices.

In (Farias et al., 2012), it was observed that such a decomposition can be used to recover a model for a probability distribution described by first order marginal information; furthermore, they showed that an *approximate* such decomposition can be recovered using a number of elements that is only linear in n rather than quadratic. More precisely, given a doubly stochastic matrix A, one can produce a convex combination of $O(n/\epsilon^2)$ permutation matrices $M_1, \ldots, M_T$ which approximates A within ϵ in Frobenius norm, i.e. $\|A - \sum_{i=1^T} p_i M_i\|_F \leq \epsilon$. A similar result can be rederived using (Garber & Hazan, 2016).

Within our framework, this is an immediate corollary. Indeed, in order to recover the result we can consider the domain to be the ℓ_2 ball of radius $\sqrt{n}$, and the doubly-stochastic input be a convex combination of permutation matrices (each of them being represented as a vector of norm $\sqrt{n}$). Then, our algorithm recovers an approximate decomposition with the same guarantees, having sparsity $O(n/\epsilon^2)$. Each call to the linear optimization oracle requires computing a minimum-cost perfect bipartite matching, which can be done in time $\tilde{O}(m \cdot \min(\sqrt{n}, m^{3/7}))$, where m is the number of nonzeros in the input (Lee & Sidford, 2014; Cohen et al., 2017).

Furthermore, the bounds easily generalize to higher norms: if instead we want to obtain a guarantee involving the element-wise ℓ_p norm of the error ($p \geq 2$), our sparsity becomes $O(n^{1/p}/\epsilon^2)$.

Fast rounding in polytopes with linear optimization oracles. The most direct application of our approach is to efficiently round a point in a polytope whenever it admits a good linear optimization oracle. An obvious such instance is the matroid polytope. Given an n-element matroid $\mathcal{M}$ of rank r and a fractional point x^* inside its base polytope, our algorithm produces a sparse distribution $\mathcal{D}$ over matroid bases such that marginals are approximately preserved in expectation. Specifically, for $p \geq 2$, $\mathcal{D}$ has a support of size $\frac{p \cdot r^{2/p}}{\epsilon^2}$, and $\|\mathbb{E}_{x \sim \mathcal{D}}[x] - x^*\|_p \leq \epsilon$; furthermore, computing $\mathcal{D}$ requires only $O\left(nr^{2/p}p/\epsilon^2\right)$ calls to $\mathcal{M}$'s independence oracle. Another example is the flow polytope.

Support vector machines (SVM). Training SVM can also be formulated as minimizing a convex function. We show that our technique of converting a problem to a saddle point formulation and solving the dual via Mirror Descent can be applied to the problem of training ν-SVMs. This is based on a formulation introduced by Schölkopf,

et al. (Schölkopf et al., 2000). Kitamura et al. (Kitamura et al., 2014) show how SVMs can be trained using Wolfe's algorithm. Replacing Wolfe's algorithm by Mirror Descent we obtain an ϵ-approximate solution in time $O\left(\max\left(\frac{1}{\epsilon}, \|K\|\right) / \left(\nu n \epsilon^2\right)\right)$, where K is the kernel matrix. This yields a constant number of iterations for polynomial and RBF kernels whenever the empirical data belong to the unit ℓ_2 ball. Our method does not need to explicitly store the kernel matrix, since every iteration only requires a matrix-vector multiplication, and the entries of the matrix can be computed on-the-fly as they are needed. In the special case of linear kernels, each iteration can be implemented in time linear in input size, yielding a nearly-linear time algorithm for linear SVM training.

Acknowledgements

AV was partially supported by NSF grants CCF-1111109 and CCF-1553428, and an internship at Google Research NYC.

References

Afriat, SN. Theory of maxima and the method of lagrange. *SIAM Journal on Applied Mathematics*, 20(3):343–357, 1971.

Arora, Sanjeev, Hazan, Elad, and Kale, Satyen. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012. doi: 10.4086/toc.2012.v008a006. URL http://dx.doi.org/10.4086/toc.2012.v008a006.

Bach, Francis R. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015. doi: 10.1137/130941961. URL http://dx.doi.org/10.1137/130941961.

Barman, Siddharth. Approximating nash equilibria and dense bipartite subgraphs via an approximate version of caratheodory's theorem. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pp. 361–369, 2015. doi: 10.1145/2746539.2746566. URL http://doi.acm.org/10.1145/2746539.2746566.

Ben-Tal, A. and Nemirovski, A. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. MPS-SIAM Series on Optimization. 2001. ISBN 9780898714913. URL https://books.google.com/books?id=kCksvznHS6oC.

Blum, Avrim, Har-Peled, Sariel, and Raichel, Benjamin. Sparse approximation via generating point sets. In *Proceedings of the Twenty-Seven Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2016.

Bubeck, Sébastien. Theory of convex optimization for machine learning. *CoRR*, abs/1405.4980, 2014. URL http://arxiv.org/abs/1405.4980.

Cohen, Michael B, Mądry, Aleksander, Sankowski, Piotr, and Vladu, Adrian. Negative-weight shortest paths and unit capacity minimum cost flow in $\tilde{O}(m^{10/7} \log W)$ time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 752–771. SIAM, 2017.

Fan, Rong-En, Chang, Kai-Wei, Hsieh, Cho-Jui, Wang, Xiang-Rui, and Lin, Chih-Jen. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008. doi: 10.1145/1390681.1442794. URL http://doi.acm.org/10.1145/1390681.1442794.

Farias, Vivek F, Jagabathula, Srikanth, and Shah, Devavrat. Sparse choice models. In *Information Sciences and Systems (CISS), 2012 46th Annual Conference on*, pp. 1–28. IEEE, 2012.

Garber, Dan and Hazan, Elad. Playing non-linear games with linear oracles. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pp. 420–428, 2013. doi: 10.1109/FOCS.2013.52. URL http://dx.doi.org/10.1109/FOCS.2013.52.

Garber, Dan and Hazan, Elad. Faster rates for the frank-wolfe method over strongly-convex sets. In *ICML*, pp. 541–549, 2015.

Garber, Dan and Hazan, Elad. A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26(3):1493–1528, 2016. doi: 10.1137/140985366. URL http://dx.doi.org/10.1137/140985366.

Jaggi, Martin. Convex optimization without projection steps. *CoRR*, abs/1108.1170, 2011. URL http://arxiv.org/abs/1108.1170.

Jaggi, Martin. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 427–435, 2013. URL http://jmlr.org/proceedings/papers/v28/jaggi13.html.

Juditsky, Anatoli, Kilinç-Karzan, Fatma, and Nemirovski, Arkadi. Randomized first order algorithms with applications to ℓ_1-minimization. *Math. Program.*, 142(1-2):269–310, 2013. doi: 10.1007/s10107-012-0575-2. URL http://dx.doi.org/10.1007/s10107-012-0575-2.

Kakade, Sham M., Shalev-Shwartz, Shai, and Tewari, Ambuj. Regularization techniques for learning with matrices. *J. Mach. Learn. Res.*, 13:1865–1890, June 2012. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2188385.2343703.

Kang, Donggu and Payor, James. Flow rounding. *CoRR*, abs/1507.08139, 2015. URL http://arxiv.org/abs/1507.08139.

Kitamura, Masashi, Takeda, Akiko, and Iwata, Satoru. Exact SVM training by wolfe's minimum norm point algorithm. In *IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2014, Reims, France, September 21-24, 2014*, pp. 1–6, 2014. doi: 10.1109/MLSP.2014.6958914. URL http://dx.doi.org/10.1109/MLSP.2014.6958914.

Klein, Philip N. and Young, Neal E. On the number of iterations for dantzig-wolfe optimization and packing-covering approximation algorithms. *SIAM J. Comput.*, 44(4):1154–1172, 2015. doi: 10.1137/12087222X. URL http://dx.doi.org/10.1137/12087222X.

Lacoste-Julien, Simon and Jaggi, Martin. On the global linear convergence of Frank-Wolfe optimization variants. In Cortes, Corinna, Lawrence, Neil D., Lee, Daniel D., Sugiyama, Masashi, and Garnett, Roman (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 496–504, 2015. URL http://papers.nips.cc/paper/5925-on-the-global-linear-convergence-of-frank-wolfe-optimization-variants.

Lee, Yin Tat and Sidford, Aaron. Path finding methods for linear programming: Solving linear programs in $O(\sqrt{rank})$ iterations and faster algorithms for maximum flow. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 424–433. IEEE, 2014.

Lipton, Richard J. and Young, Neal E. Simple strategies for large zero-sum games with applications to complexity theory. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, 23-25 May 1994, Montréal, Québec, Canada*, pp. 734–740, 1994. doi: 10.1145/195058.195447. URL http://doi.acm.org/10.1145/195058.195447.

Lipton, Richard J., Markakis, Evangelos, and Mehta, Aranyak. Playing large games using simple strategies. In *Proceedings 4th ACM Conference on Electronic Commerce (EC-2003), San Diego, California, USA, June 9-12, 2003*, pp. 36–41, 2003. doi: 10.1145/779928.779933. URL http://doi.acm.org/10.1145/779928.779933.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004. ISBN 9781402075537. URL https://books.google.com/books?id=VyYLem-l3CgC.

Novikoff, A.B.J. On convergence proofs on perceptrons. 12: 615–622, 1962.

Paley, R and Zygmund, A. A note on analytic functions in the unit circle. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pp. 266–272. Cambridge University Press, 1932.

Peña, Javier, Rodríguez, Daniel, and Soheili, Negar. On the von neumann and frank-wolfe algorithms with away steps. *SIAM Journal on Optimization*, 26(1):499–512, 2016. doi: 10.1137/15M1009937. URL https://doi.org/10.1137/15M1009937.

Pisier, Gilles. Remarques sur un résultat non publié de B. Maurey. *Séminaire Analyse fonctionnelle (dit*, pp. 1–12, 1980.

Plotkin, Serge A., Shmoys, David B., and Tardos, Éva. Fast approximation algorithms for fractional packing and covering problems. In *32nd Annual Symposium on Foundations of Computer Science, San Juan, Puerto Rico, 1-4 October 1991*, pp. 495–504, 1991. doi: 10.1109/SFCS.1991.185411. URL http://dx.doi.org/10.1109/SFCS.1991.185411.

Raghavan, Prabhakar and Thompson, Clark D. Multiterminal global routing: A deterministic approximation scheme. *Algorithmica*, 6(1):73–82, 1991. doi: 10.1007/BF01759035. URL http://dx.doi.org/10.1007/BF01759035.

Schölkopf, Bernhard, Smola, Alexander J., Williamson, Robert C., and Bartlett, Peter L. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000. doi: 10.1162/089976600300015565. URL http://dx.doi.org/10.1162/089976600300015565.

Shalev-Shwartz, Shai. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, July 2007.

Shalev-Shwartz, Shai, Srebro, Nathan, and Zhang, Tong. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM J. on Optimization*, 20(6):2807–2832, August 2010. ISSN 1052-6234. doi: 10.1137/090759574. URL http://dx.doi.org/10.1137/090759574.

Shalev-Shwartz, Shai, Singer, Yoram, Srebro, Nathan, and Cotter, Andrew. Pegasos: primal estimated sub-gradient solver for SVM. *Math. Program.*, 127(1):3–30, 2011. doi: 10.1007/s10107-010-0420-4. URL http://dx.doi.org/10.1007/s10107-010-0420-4.

Shtern, Shimrit and Beck, Amir. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, pp. 1–27, 2016.

Sion, Maurice. On general minimax theorems. *Pac. J. Math.*, 8: 171–176, 1958. ISSN 0030-8730. doi: 10.2140/pjm.1958.8.171.

Tao, T. *Topics in Random Matrix Theory*. Graduate studies in mathematics. American Mathematical Soc. ISBN 9780821885079. URL https://books.google.com/books?id=Hjq_JHLNPT0C.

Wolff, T.H., Łaba, I., and Shubin, C. *Lectures on Harmonic Analysis*. Universi Series. AMS. ISBN 9780821882863. URL https://books.google.com/books?id=i56jcHMvXuUC.

Zhu, Zeyuan Allen, Chen, Weizhu, Wang, Gang, Zhu, Chenguang, and Chen, Zheng. P-packsvm: Parallel primal gradient descent kernel SVM. In *ICDM 2009, The Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, 6-9 December 2009*, pp. 677–686, 2009. doi: 10.1109/ICDM.2009.29. URL http://dx.doi.org/10.1109/ICDM.2009.29.