Tight Bounds for Approximate Carathéodory and Beyond: Supplementary Material

A. Overview of Convex Analysis

We give a brief overview on the theory of convex functions. For a detailed exposition we refer readers to (Nesterov, 2004).

Subgradients. A function $f : Q \subseteq \mathbb{R}^d \to \mathbb{R}$ defined on a convex domain Q is said to be convex if every point $x \in Q$ has a non-empty subgradient $\partial f(x) = \{g \in \mathbb{R}^d ; f(y) \geq f(x) + g^\top (y-x), \forall y \in Q\}$. Geometrically, this means that a function is convex iff it is the maximum of all its supporting hyperplanes, i.e. $f(x) = \max_{x_0, g \in \partial f(x_0)} f(x_0) + g^\top (x - x_0)$. When there is a unique element in $\partial f(x)$ we call it the gradient and denote it by $\nabla f(x)$. We will sometimes abuse notation and refer to $\nabla f(x)$ as an arbitrary element of $\partial f(x)$ even when it is not unique.

Strong convexity and smoothness. We say that a function $f : Q \subseteq \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex with respect to norm $\|\cdot\|$ if for all $x, y \in Q$ and all subgradients $g \in \partial f(x)$:

$$f(y) - f(x) - g^\top (y-x) \geq \frac{1}{2}\mu \|y-x\|^2$$

A function is said to be σ-smooth with respect to the $\|\cdot\|$ if for all $x, y \in Q$ and $g \in \partial f(x)$:

$$f(y) - f(x) - g^\top (y-x) \leq \frac{1}{2}\sigma \|y-x\|^2$$

Bregman divergence and the Hessian. Every continuously differentiable f induces a concept of 'distance' known as the Bregman-divergence: given $x, y \in Q$, we define $D_f(y\|x) := f(y) - f(x) - \nabla f(x)^\top (y-x)$ as the second order error when computing $f(y)$ using the linear approximation of f around x. The fact that f is convex guarantees $D_f(y\|x) \geq 0$.

If the subgradient of f is unique everywhere, we can define μ-strong convexity and σ-smoothness with respect to the Bregman divergence, as $D_f(y\|x) \geq \frac{1}{2}\sigma \|x-y\|^2$ and $D_f(y\|x) \leq \frac{1}{2}\mu \|x-y\|^2$. If f is also twice-differentiable, a simple way to compute its strong convexity and smoothness parameters is by bounding the $\|\cdot\|$-eigenvalues of the Hessian. If $\mu \cdot \|w\|^2 \leq w^\top \nabla^2 f(x) w \leq \sigma \cdot \|w\|^2$ for all $x \in Q$ and $w \neq 0$, then f is μ-strongly convex and σ-smooth. This is because:

$$\begin{aligned} D_f(y\|x) &= \int_0^1 [\nabla f(x + (y-x)t) - \nabla f(x)]^\top (y-x) dt \\ &= \int_0^1 \int_0^s (y-x)^\top \nabla^2 f(x + (y-x)s)(y-x) ds dt \end{aligned}$$

Lipschitz constant. We say that a convex function is ρ-Lipschitz with respect to norm $\|\cdot\|$ if $\|\nabla f(x)\|_* \leq \rho$. Note that ρ-Lipschitz continuity requires a bound on the *dual* norm, since

$$\begin{aligned} |f(y) - f(x)| &= \left| \int_0^1 \nabla f(x + t(y-x))^\top (y-x) dt \right| \\ &\leq \int_0^1 \left| \nabla f(x + t(y-x))^\top (y-x) \right| dt \\ &\leq \int_0^1 \|\nabla f(x + t(y-x))\|_* \cdot \|y-x\| dt \\ &\leq \rho \cdot \|y-x\| \end{aligned}$$

Fenchel duality. It is useful to write a convex function as the maximum of its supporting hyperplanes. One way to do that is using the *Fenchel transform*. When defining Fenchel transforms, it is convenient to identify a function $f : Q \to \mathbb{R}$ to its extension $\tilde{f} : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ such that $\tilde{f}(x) = f(x)$ for $x \in Q$ and $\tilde{f}(x) = \infty$ otherwise. Given that identification, we can define the Fenchel transform of a function f as the function $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ given by $f^*(z) = \sup_{x \in \mathbb{R}^d} z^\top x - f(x)$. If f is convex, the Fenchel transformation is self-invertible, i.e., $(f^*)^* = f$ or equivalently: $f(x) = \max_z z^\top x - f^*(z)$. Notice that the previous expression is a way to write any convex function as a maximum over linear functions in x parametrized by z. The *Fenchel inequality* $f(x) + f^*(z) \geq z^\top x$ follows directly from the definition of the Fenchel transform.

Envelope Theorem. When writing a convex function $f = \max_i f_i$ as a maximum of other convex function (typically linear functions), the Envelope Theorem gives a way to compute derivatives. Its statement is quite intuitive: since gradients are local objects, the gradient of f at a certain point is the gradient of the function f_i being maximized at that point. Formally, if $f(x) = \max_z g(x, z)$ where $g(x, z)$ is convex in x for every fixed z, then if $f(x_0) = g(x_0, z_0)$, then $\partial_x g(x_0, z_0) \subseteq \partial f(x_0)$. A direct application of this theorem is in computing the gradients of the Fenchel dual: $\nabla f^*(z) = \arg\min_x \{z^\top x - f(x)\}$ and $f^*(z) = z^\top \nabla f^*(z) - f(\nabla f^*(z))$.

Smoothness and strong convexity duality. Finally, we will use the following duality theorem:

Theorem A.1. *The function $f : Q \to \mathbb{R}$ is a $(1/\sigma)$-strongly convex function with respect to $\|\cdot\|$ if and only if its Fenchel dual $f^* : \mathbb{R}^d \to \mathbb{R}$ is a σ-smooth with respect to $\|\cdot\|_*$.*

Proof. Here we prove that σ-strong convexity of a function implies $(1/\sigma)$-smoothness of its dual, since this is the direction we will use. We refer to (Kakade et al., 2012; Shalev-Shwartz, 2007) for a proof of the converse.

Fix $z_1, z_2 \in \mathbb{R}^d$ and let $y_i \in \partial f^*(z_i) = \arg\max_{y \in Q} z_i^\top y - f(y)$. Since f is strongly convex, there in an unique maximum, so we can write $y_i = \nabla f^*(z_i)$. Also, $f^*(z_i) = z_i^\top y_i - f(y_i)$. Since the Fenchel transform is self-dual, $f(y_i) = \max_z y_i^\top z - f^*(z) = z_i^\top y_i - f^*(z_i)$. In particular, this means that $z_i \in \partial f(y_i)$.

Using the strong-convexity of f, we can write:

$$f(y_2) - f(y_1) - z_1^\top (y_2 - y_1) \geq \frac{1}{2\sigma} \|y_1 - y_2\|^2$$

$$f(y_1) - f(y_2) - z_2^\top (y_1 - y_2) \geq \frac{1}{2\sigma} \|y_1 - y_2\|^2$$

Summing the expressions above and applying Holder's inequality, we get:

$$\frac{1}{\sigma} \|y_1 - y_2\|^2 \leq (z_2 - z_1)^\top (y_2 - y_1) \leq \|z_1 - z_2\|^* \cdot \|y_1 - y_2\|$$

Therefore:

$$\sigma \cdot \|z_1 - z_2\|_* \geq \|y_1 - y_2\| = \|\nabla f^*(z_1) - \nabla f^*(z_2)\|$$

which implies the smoothness bound:

$$\begin{aligned} & D_{f^*}(z_2 \| z_1) \\ & = \int_0^1 [\nabla f^*(z_1 + t(z_2 - z_1)) - \nabla f^*(z_1)]^\top (z_2 - z_1) dt \\ & \leq \frac{1}{2} \sigma \|z_1 - z_2\|_*^2 \end{aligned}$$

□

B. A primer on Mirror Descent

For the sake of completeness, we will present here an elementary exposition of the Mirror Descent Framework, which is used in our proof. For a complete exposition we refer to Nemirovskii (Ben-Tal & Nemirovski, 2001) or Bubeck (Bubeck, 2014).

The goal of Mirror Descent is to minimize a convex function $f : Q \subseteq \mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant ρ with respect to norm $\|\cdot\|$. To motivate Mirror Descent, it is useful to think of dot products $y^\top x$ as a product of vectors in two different vector spaces, which can be thought as *vectors* vs *linear forms* or *column vectors* vs *row vectors*. In the spirit of Hölder's inequality, we can think of x as living in the $\mathbb{R}^d$ space equipped with $\|\cdot\|$ norm while y lives in $\mathbb{R}^d$ equipped with the dual norm $\|\cdot\|_*$. When we approximate $f(y) - f(x) \approx \nabla f(x)^\top (y - x)$, the second term is a dot-product of a vector in the domain $y - x$, which we call the *primal space* and measure using $\|\cdot\|$ norm and a gradient vector, which we call the *dual space* and measure with dual norm $\|\cdot\|_*$.

Keeping the discussion in the previous paragraph in mind, we can revisit the most intuitive method to minimize convex functions: *gradient descent*. The gradient descent method consists in following the directions of steepest descent, which is the direction opposite to the gradient. This leads to an iteration of the type: $y_{t+1} = y_t - \eta \cdot \nabla f(y_t)$. In the view of primal space and dual space, this iteration suddenly looks strange, because one is summing a primal vector y_t with a dual vector $\nabla f(y_t)$ which live in different spaces. In some sense, the gradient descent for Lipschitz convex functions only makes sense in the ℓ_2 norm, in which $\|\cdot\| = \|\cdot\|_*$ (see the subgradient descent method in (Nesterov, 2004)).

This motivated the idea of a map $M : \mathbb{R}^d \to Q$ connecting the primal and the dual space. The idea in the mirror descent algorithm is to keep two vectors (y_t, z_t) one in the primal space and one in the dual space. In each iteration we compute $\nabla f(y_t)$, obtaining a dual vector and update:

$$z_{t+1} = z_t - \eta \nabla f(y_t) \qquad\qquad y_{t+1} = M(z_{t+1})$$

It is convenient in the analysis to think of this map as the gradient of a convex function $M = \nabla \omega^*$. In the usual setup, we define the *mirror map*, which is a convex function $\omega : Q \to \mathbb{R}$, σ^{-1}-strongly convex with respect to $\|\cdot\|$. Let $\omega^* : \mathbb{R}^d \to \mathbb{R}$ be the Fenchel-dual $\omega^*(z) = \sup_{y \in Q} z^\top y - \omega(y)$ which is a σ-smooth convex function with respect to $\|\cdot\|_*$ by Theorem A.1.

Notice that ω^* is defined as a maximum over linear functions of z indexed by y. The result known as the envelope theorem states that $\nabla \omega^*(z_0)$ is the gradient of the linear function maximized at z_0. Therefore: $\nabla \omega^*(z_0) = y \in \arg\max_y \{z_0^\top y - \omega(y)\}$. This in particular implies that $\nabla \omega^*(z) \in Q$ since $\omega(y) = \infty$ for $y \notin Q$.

Using the definition of ω and ω^* we can define the Mirror Descent iteration as:

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

Now, we are ready to prove Theorem 2.1. We restate it:

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

Proof. The idea of the proof is to bound the growth of $\omega^*(z_t)$ using smoothness property of ω^*:

$$\begin{aligned} \omega^*(z_t) &\leq \omega^*(z_0) + \sum_{t=0}^{T-1} \nabla \omega^*(z_t)^\top (z_{t+1} - z_t) + \frac{\sigma}{2} \|z_{t+1} - z_t\|_*^2 \\ &= \omega^*(0) - \sum_{t=0}^{T-1} \eta \nabla y_t^\top \nabla f(y_t) + \frac{\sigma}{2} \eta^2 \|\nabla f(y_t)\|_*^2 \end{aligned}$$

By the Fenchel inequality $\omega^*(z_t) \geq z_t^\top y - \omega(y) = (z_0 - \sum_t \eta \nabla f(y_t))^\top y - \omega(y^*)$ for all $y \in Q$. Combining with the previous inequality and re-arranging the terms, we get:

$$\eta \sum_t \nabla f(y_t)^\top (y_t - y) \leq \omega(y) + \omega^*(z_0) - \nabla \omega(y_0)^\top y + \frac{\sigma}{2} \eta^2 \rho^2 T$$

The gradient of $\omega^*(z_0) = \sup_y z_0^\top y - \omega(y)$ corresponds by the envelope theorem to y maximizing $z_0^\top y - \omega(y)$. Therefore, since $y_0 = \nabla \omega^*(z_0)$, $\omega^*(z_0) = z_0^\top y_0 - \omega(y_0)$. Substituting $\omega^*(z_0)$ in the above expression and using the definition of Bregman divergence, we get:

$$\eta \sum_t \nabla f(y_t)^\top (y_t - y) \leq D_\omega(y \| y_0) + \frac{\sigma}{2} \eta^2 \rho^2 T$$

Rearranging the terms and using that $D_\omega(y \| y_0) \leq D$, we obtain:

$$\frac{1}{T}\sum_t \nabla f(y_t)^\top (y_t - y) \leq \frac{D}{\eta T} + \frac{\sigma \eta \rho^2}{2} = \sqrt{\frac{2D\sigma\rho^2}{T}} \text{ for } \eta = \sqrt{\frac{2D}{T\sigma\rho^2}}$$

So for $T \geq \frac{2\sigma D \rho^2}{\epsilon^2}$, $\frac{1}{T}\sum_t \nabla f(y_t)^\top (y_t - y) \leq \epsilon$. □

Corollary B.2. *In the conditions of the previous theorem, for* $\bar{y}_t = \frac{1}{T}\sum_{t=1}^T y_t$*,* $f(\bar{y}_t) - f^* \leq \epsilon$*, where* $f^* = \min_{y \in Q} f(y)$

Proof. Let $y^* = \arg\min_{y \in Q} f(y)$. Applying the previous theorem with $y = y^*$ we get:

$$f(\bar{y}_t) - f(y^*) \leq \frac{1}{T}\sum_t f(y_t) - f(y^*) \leq \frac{1}{T}\sum_t \nabla f(y_t)^\top (y_t - y^*) \leq \epsilon$$

where both inequalities follow from convexity of f. □

C. Mirror Map

Proposition C.1. *For* $1 < q \leq 2$*, the function* $\omega : \boldsymbol{B}_q(1) \to \mathbb{R}$*,* $\omega(y) = \frac{1}{2}\|y\|_q^2$ *is* $(q-1)$*-strongly convex with respect to the* ℓ_q *norm and* $\max_{y \in \boldsymbol{B}_q(1)} D_\omega(y \| 0) = \frac{1}{2}$.

While a similar statement is shown to hold in (Shalev-Shwartz, 2007), we provide a proof here for completeness.

Proof. We want to bound $\omega(y) - \omega(x) - g^\top (y - x)$ for all $g \in \partial\omega(x)$. For all x in the interior of the ball $\boldsymbol{B}_q(1)$ there is a unique subgradient which we represent by $\nabla\omega(x)$. In the border of $\boldsymbol{B}_q(1)$, however, there are multiple subgradients. First we claim that we need only to bound $\omega(y) - \omega(x) - \nabla\omega(x)^\top (y - x)$ where $\nabla\omega(x)$ denotes the gradient of the function $\frac{1}{2}\|y\|_q^2$. In order to see that, notice that if g is a subgradient in a point x and $y \in \boldsymbol{B}_q(1)$ then:

$$\omega(x + t(y - x)) - \omega(x) - g^\top (y - x) \geq 0$$

by the definition of subgradient. Dividing the expression by t and taking the limit when $t \to 0+$, we get: $\nabla\omega(x)^\top (y - x) \geq g^\top (y - x)$, so in particular: $\omega(y) - \omega(x) - g^\top (y - x) \geq \omega(y) - \omega(x) - \nabla\omega(x)^\top (y - x)$.

This observation allows us to bound the strong convexity parameter of ω by looking at the $\|\cdot\|_q$-eigenvalues of the Hessian of ω. In particular, we will show that for all $w \in \mathbb{R}^d$, $w^\top \nabla^2\omega(y) w \geq (q-1)\|w\|_q^2$.

To make the notation simpler, we define $\textsc{Pow} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ as $\textsc{Pow}(y, p) = (|y_i|^p \cdot \operatorname{sgn}(y_i))_i$. This allows us to represent $\nabla \|y\|_q$ in a succinct form: since

$$\partial_i \|y\|_q = \frac{1}{q}(\|y\|_q^q)^{\frac{1}{q}-1} x_i^q q \cdot \operatorname{sgn}(x_i) = \|y\|_q^{1-q} x_i^{q-1} \operatorname{sgn}(y_i)$$

so we can write $\nabla \|y\|_q = \|y\|_q^{1-q} \cdot \textsc{Pow}(y, q-1)$. Therefore:

$$\nabla\omega(y) = \nabla\left[\frac{1}{2}\|y\|_q^2\right] = \|y\|_q \cdot \nabla\|y\|_q = \|y\|_q^{2-q} \cdot \textsc{Pow}(y, q-1)$$

Now, to compute the Hessian, we have:

$$\nabla^2\omega(y) = (2-q)\|y\|_q^{2-2q} \cdot \textsc{Pow}(y, q-1)\textsc{Pow}(y, q-1)^\top + (q-1)\|y\|_q^{2-q}\,\textsc{Diag}(|y_i|^{q-2})$$

where $\textsc{Diag}(|y_i|^{q-2})$ is the diagonal matrix with x_i^{q-2} in the diagonal. Using the fact that $1 < q \leq 2$, we can write:

$$\begin{aligned} w^\top \nabla^2 \omega(y) w &= (2-q) \cdot \|y\|_q^{2-q} \left[\textsc{Pow}(y, q-1)^\top w\right]^2 + (q-1) \cdot \|y\|_q^{2-q} \sum_i |y|_i^{q-2} w_i^2 \\ &\geq (q-1) \left(\sum_i |y|_i^q\right)^{\frac{2-q}{q}} \cdot \left(\sum_i |y|_i^{q-2} w_i^2\right) \\ &= (q-1) \left[\left(\sum_i |y_i|^{\frac{q(2-q)}{2}\cdot\frac{2}{2-q}}\right)^{\frac{2-q}{2}} \cdot \left(\sum_i (|y_i|^{\frac{q(q-2)}{2}} w_i^q)^{\frac{2}{q}}\right)^{\frac{q}{2}}\right]^{\frac{2}{q}} \end{aligned}$$

The last equality is a convoluted re-writing of the previous expression, but allows us to apply Hölder's inequality. Recall that Hölder's inequality states that $\|z_1\|_a \cdot \|z_2\|_b \geq z_1^\top z_2$ whenever $\frac{1}{a} + \frac{1}{b} = 1$. Applying this inequality with $a = \frac{2}{2-q}$ and $b = \frac{2}{q}$, we get:

$$w^\top \nabla^2 \omega(y) w \geq (q-1) \cdot \left(\sum_i |y_i|^{\frac{q(2-q)}{2}} \cdot |y_i|^{\frac{q(q-2)}{2}} w_i^q\right)^{\frac{2}{q}} = (q-1) \cdot \left(\sum_i w_i^q\right)^{\frac{2}{q}} = (q-1) \cdot \|w\|_q^2$$

□

Finally, we need to show how to compute the Fenchel dual ω^* and the mirror map $\nabla\omega^*$ efficiently:

Proposition C.2. *The Fenchel dual of the function ω defined in Proposition C.1 can be computed explicitly:*

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

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: $\phi(z)_i = \operatorname{sgn}(z_i) \cdot |z_i|^{p-1} / \|z\|_p^{p-1}$.

Proof. By the definition of Fenchel duality:

$$\omega^*(z) = \max_{y \in \boldsymbol{B}_q(1)} z^\top y - \frac{1}{2}\|y\|_q^2 = \max_{0 \leq \lambda \leq 1} \left[\max_{\hat{y}; \|\hat{y}\|_q = 1} \lambda z^\top \hat{y} - \frac{1}{2}\lambda^2\right] = \max_{0 \leq \lambda \leq 1} \lambda \|z\|_p - \frac{1}{2}\lambda^2$$

where the second equality follows from writting $y = \lambda\hat{y}$ for $0 \leq \lambda \leq 1$ and $\|\hat{y}\|_p = 1$. The optimal value of $\hat{y}$ is $\phi(z)$. The expression $\lambda\|z\|_p - \frac{1}{2}\lambda^2$ is maximized at $\lambda = \|z\|_p$. Since λ is restricted to lie between 0 and 1, the optimal λ must be $\min(1, \|z\|_p)$.

If $\|z\|_p \leq 1$, $\lambda = \|z\|_p$ and $\omega^*(z) = \frac{1}{2}\|z\|_p^2$. If $\|z\|_p > 1$, then $\lambda = 1$ and $\omega^*(z) = \|z\|_p - \frac{1}{2}$.

By the envelope theorem, $\nabla\omega^*(z) = \hat{y} \cdot \lambda = \phi(z) \cdot \min(1, \|z\|_p)$. Now, it simple to check that ϕ has the desired properties:

$$\|\phi(z)\|_q^q = \sum_i |z_i|^{q(p-1)} / \|z\|_p^{q(p-1)} = \sum_i |z_i|^p / \|z\|_p^p = 1$$

$$z^\top \phi(z) = \sum_i z_i^p / \|z\|_p^{p-1} = \|z\|_p$$

□

Combining the previous results, we obtain:

Theorem C.3. *Given n points $v_1, \dots, v_n \in \boldsymbol{B}_p(1) \subseteq \mathbb{R}^d$ with $p \geq 2$ 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 $\|u' - u\|_p \leq \epsilon$.*

Proof. The number of iterations of Mirror Descent (and consequently the sparsity bound) $T = 4p/\epsilon^2$ can be obtained by substituting $D = 1/2$ and $\sigma^{-1} = (q-1)^{-1} = p-1$ from Proposition C.1 in Theorem 3.2.

For the running time, notice that the time per iteration is dominated by the computation of the subgradient of f. The most expensive step is the computation of $V^\top y$ which takes dn operations, which is the size of matrix V.

□

D. Missing Proofs from Section 3.1

Proof of Theorem 3.5. Let $\mathsf{ApproxCara}(u)$ represent the convex combination x of vertices of P returned by the algorithm from Theorem C.3, when receiving as input the vertices of P and the point u, and solving for precision $r/2$. Then let $e_0 = u$, $x_i = \mathsf{ApproxCara}(e_{i-1})$, $e_i = 2(e_{i-1} - Vx_i)$, for $i \in \{1, \dots, \beta\}$ where $\beta = \log(r/\epsilon)$. Note that $\|e_i\|_p = 2\|e_{i-1} - Vx_i\|_p \le 2 \cdot \frac{r}{2} \le r$, hence $e_i \in P$, so the input to ApproxCara is always well defined. Let $\overline{x} = \sum_{i=1}^{\beta} \frac{1}{2^{i-1}} \cdot x_i \in \left(\sum_{i=1}^{\beta} 2^{-(i-1)}\right)\Delta = \frac{2^{-\beta}-1}{2^{-1}-1} \cdot \Delta = 2(1-\epsilon/r)\Delta$.

Let us bound the error when approximating u with $V\overline{x}$:

$$\begin{aligned}
\|V\overline{x} - u\|_p &= \left\| V\left(\sum_{i=1}^{\beta} \frac{1}{2^{i-1}} \cdot x_i\right) - u \right\|_p = \left\| \sum_{i=1}^{\beta} \frac{1}{2^{i-1}} \cdot Vx_i - e_0 \right\|_p = \left\| \sum_{i=2}^{\beta} \frac{1}{2^{i-1}} Vx_i + Vx_1 - e_0 \right\|_p \\
&= \left\| \sum_{i=2}^{\beta} \frac{1}{2^{i-1}} Vx_i - \frac{1}{2} e_1 \right\|_p = \frac{1}{2} \left\| \sum_{i=2}^{\beta} \frac{1}{2^{i-2}} Vx_i - e_1 \right\|_p = \dots \\
&= \frac{1}{2^{\beta-1}} \left\| \sum_{i=\beta}^{\beta} \frac{1}{2^{i-\beta}} Vx_i - e_{\beta-1} \right\|_p = \frac{1}{2^{\beta-1}} \left\| Vx_\beta - e_{\beta-1} \right\|_p \le \frac{1}{2^{\beta-1}} \cdot \frac{r}{2} \\
&= r/2^\beta = \epsilon
\end{aligned}$$

Each of the $\beta = \log(r/\epsilon)$ iterations requires a call to ApproxCara for precision $r/2$, which produces a solution with sparsity $O(p/r^2)$. Hence $\overline{x}$ will have $O\left(\frac{p}{r^2} \log \frac{r}{\epsilon}\right)$ nonzero coordinates. □

Proof of Corollary 3.6. Let $v_i' = v_i - u$ for all i. This corresponds to translating P such that u is placed at the origin. By the triangle inequality, this at most doubles the radius of the origin-centered ℓ_p ball circumscribing the polytope. Applying Theorem 3.5 we obtain a vector $x \in 2(1-r/\epsilon)\Delta$ such that $\left\| \sum_{i \in \mathrm{supp}(x)} x_i v_i' \right\|_p \le \epsilon$. Let $x_i' = x_i / \|x\|_1 \in \Delta$. This satisfies $\left\| \sum_{i \in \mathrm{supp}(x)} x_i' v_i' \right\|_p \le \epsilon / \|x\|_1$. Hence $\left\| \sum_{i \in \mathrm{supp}(x)} x_i' v_i' \right\|_p = \left\| \sum_{i \in \mathrm{supp}(x)} x_i' (v_i - u) \right\|_p = \left\| \sum_{i \in \mathrm{supp}(x)} x_i' v_i - u \right\|_p \le \frac{\epsilon}{2(1-\epsilon/r)} \le \epsilon$. □

E. Analysis via Conditional Gradient Methods

We first show that the Frank Wolfe algorithm described in Section 3.2 provides the same guarantees as the Mirror Descent. Then we argue that the two algorithms are completely isomorphic. We start by reviewing the Frank-Wolfe algorithm. Instead of the standard choice of parameters $\eta_t = 2/(t+1)$ we will choose $\eta_t = 1/t$ since it has the nice feature that $x_t = \frac{1}{t} \sum_{s=1}^{t} y_t$. In various applications discussed by Barman (Barman, 2015) it is crucial that the convex combination is uniform over (a multi-set of) vertices. For completeness we provide a proof of convergence of the Frank-Wolfe algorithm. The proof follows the presentation in (Jaggi, 2013) and (Bubeck, 2014) with the small change that we choose η_t so that we get an uniform convex combination in the end.

Lemma E.1 (Frank-Wolfe). *If f is β-smooth with respect to norm $\|\cdot\|$ and $R = \max_{x,y \in X} \|x - y\|$ then the Frank Wolfe algorithm (described by the iteration FW in Section 3.2) with parameters $\eta_t = 1/t$ is such that $f(x_t) - \min_{x \in X} f(x) \le \epsilon$ for $t = \Omega\left(\frac{\beta R^2}{\epsilon}\right)$.*

Proof. Let x^* be the minimizer of f and define $\Delta_t = f(x_t) - f(x^*)$. Since $x_{t+1} = x_t + \eta_t(y_t - x_t)$, then by the definition of β-smoothness we have:

$$f(x_{t+1}) \leq f(x_t) + \eta_t \nabla f(x_t)^\top (y_t - x_t) + \frac{1}{2}\beta\eta_t^2 \left\|y_t - x_t\right\|^2$$

By the choice of y_t we have that $y_t^\top \nabla f(x_t) \leq (x^*)^\top \nabla f(x_t)$ and therefore:

$$(y_t - x_t)^\top \nabla f(x_t) \leq (x^* - x_t)^\top \nabla f(x_t) \leq f(x^*) - f(x_t) = -\Delta_t$$

so:

$$\Delta_{t+1} = f(x_{t+1}) - f(x^*) \leq \Delta_t - \eta_t \Delta_t + \frac{1}{2}\beta\eta_t^2 R^2$$

which can be re-written as: $\Delta_{t+1} \leq (1-\eta_t)\Delta_t + \frac{1}{2}\beta\eta_t^2 R^2$. Telescoping this expression we get:

$$\Delta_{t+1} \leq \prod_t (1-\eta_t)\Delta_0 + \sum_t \frac{1}{2}\frac{\beta R^2}{t^2}\eta_t = O\left(\frac{\beta R^2}{t}\right) \qquad \text{for } \eta_t = \frac{1}{t}$$

□

Instantiating the Frank-Wolfe algorithm for our problem gives another optimization solution for the approximate Caratheodory Theorem:

Theorem E.2. *If x_t, y_t are iterates of the Frank-Wolfe algorithm for $f(x) = \left\|x - u\right\|_p^2$ then $\left\|\frac{1}{t}\sum_{s=1}^t y_t - u\right\|_p \leq \epsilon$ for $t = \Omega\left(\frac{p}{\epsilon^2}\right)$.*

Proof. By combining Theorem A.1, Proposition C.1 and the fact that the Fenchel dual of $\frac{1}{2}\left\|x\right\|_p^2$ is $\frac{1}{2}\left\|x\right\|_q^2$ we obtain that $f(x)$ is $2(p-1)$-smooth. Observe that since u is a convex combination of vertices of P, then $f(x^*) = 0$. Also, given how η_t were chosen, $x_t = \frac{1}{t}\sum_{s=1}^t y_t$. This implies that $\left\|x_t - u\right\|_p^2 \leq \epsilon^2$ for $t = \Omega\left(\frac{pR^2}{\epsilon^2}\right)$. □

A remarkable fact, however, is that the algorithm in Theorem E.2 and the one in Theorem 3.2 produce the same iterates. Next we prove this fact:

Proof of Theorem 3.7. Recall the the Frank-Wolfe iteration is given by:

$$y_t^{\mathtt{FW}} = \arg\min_{y \in P} \nabla f(x_{t-1}^{\mathtt{FW}})^\top y \qquad\qquad x_t^{\mathtt{FW}} = \left(1 - \frac{1}{t}\right) x_{t-1}^{\mathtt{FW}} + \frac{1}{t} \cdot y_t^{\mathtt{FW}}$$

for $f(x) = \left\|x - u\right\|_p^2$. Next we describe the Mirror Descent iteration. To avoid confusion we re-name the varibles and the function in the (MD) iteration:

$$x_{t+1}^{\mathtt{MD}} = x_t^{\mathtt{MD}} - \eta \nabla g(z_t^{\mathtt{MD}}) \qquad\qquad z_{t+1}^{\mathtt{MD}} = \nabla \omega^*(x_{t+1}^{\mathtt{MD}})$$

for $g(z) = \max_{y \in P} z^\top (u - y)$ and ω^* is as in Proposition C.2. The vertices output by the Frank Wolfe algorithm are $y_t^{\mathtt{FW}}$ and the vertices output by the Mirror Descent algorithm are $y_t^{\mathtt{MD}} = u - \nabla g(z_t^{\mathtt{MD}})$.

Let y_0 be an arbitrary point of P and let $x_0^{\mathtt{FW}} = y_0$ and $z_0^{\mathtt{MD}} = u - y_0$. We claim that for all $t = 1, 2, \dots$ we have $y_t^{\mathtt{MD}} = y_t^{\mathtt{FW}}$ and $x_t^{\mathtt{FW}} = u + \frac{1}{\eta t} x_t^{\mathtt{MD}} = \frac{1}{t}\sum_{s=1}^t y_t$.

The main observation we need to prove this fact is that $\nabla f(x) = \lambda_1(x) \cdot \phi(x - u)$ where $\lambda_1(x)$ is a non-negative real number and $\phi(\cdot)$ is the function defined in Proposition C.2. Observe that by the same proposition, we can write $\nabla \omega^*(x) = \lambda_2(x) \cdot \phi(x - u)$. The important fact about ϕ is that it is invariant by rescaling its arguments by a non-negative function, i.e., $\phi(tx) = \phi(x)$ for $t \geq 0$.

Now, we can prove the claim by induction, suppose that $y_s^{\mathtt{MD}} = y_s^{\mathtt{FW}}$ for all $s < t$ and that $x_t^{\mathtt{FW}} = u + \frac{1}{\eta t} x_t^{\mathtt{MD}} = \frac{1}{t}\sum_{s=1}^t y_t$. Then $x_{t+1}^{\mathtt{MD}} = x_t^{\mathtt{MD}} - \eta(u - y_t^{\mathtt{MD}}) = -\eta(t+1)\sum_{s=1}^t (u - y_t^{\mathtt{MD}})$ and $y_t^{\mathtt{MD}} = \arg\min_{y \in P} \nabla \omega^*(x_t^{\mathtt{MD}})^\top y = \arg\min_{y \in P} \phi(x_t^{\mathtt{MD}})^\top y$. For Frank-Wolfe, $y_{t+1}^{\mathtt{FW}} = \arg\min_{y \in P} \nabla f(x_t^{\mathtt{FW}})^\top y = \arg\min_{y \in P} \phi(x_t^{\mathtt{FW}} - u)^\top y$. The lemma follows from the fact that, by induction hypothesis, $x_t^{\mathtt{FW}} - u$ and $x_t^{\mathtt{MD}}$ are rescaled versions of the same vector. □

F. Lower bound proofs

F.1. Proof of Theorem 5.3

Proof. Let $x \in \Delta_n$ be k-sparse, i.e. $|\text{supp}\,(x)| = k$, such that $\left\|\tilde{H}x - u\right\|_p \le \epsilon$. We would like to lower bound the sparsity k in terms of ϵ and p.

We will use two main ingredients in the proof: the first is the *power mean inequality* which states that for any vector $x \in \mathbb{R}^n$, $\left(\frac{1}{n}\sum_i x_i^t\right)^{1/t}$ is non-decreasing in t. In particular, this implies that $\|x\|_t \cdot n^{-1/t}$ is non-decreasing[3]. The second fact we will use is that for every vector $\|x\|_1^2 \le \|x\|_2^2 \cdot |\text{supp}\,(x)|$. This follows from the Cauchy-Schwarz inequality: $\|x\|_1^2 = \left(\sum_{i \in \text{supp}(x)} x_i \cdot 1\right)^2 \le \left(\sum_{i \in \text{supp}(x)} x_i^2\right) \cdot \left(\sum_{i \in \text{supp}(x)} 1\right) = \|x\|_2^2 \cdot |\text{supp}\,(x)|$. Combining both results give us a bound involving the 2-norm of the error:

$$\epsilon \ge \left\|\tilde{H}x - u\right\|_p = \frac{1}{n^{1/p}} \cdot \|Hx - e_1\|_p \ge \frac{1}{n^{1/2}} \cdot \|Hx - e_1\|_2$$

where the last step follows from the power-mean inequality. Squaring both sides, we get:

$$\epsilon^2 \ge \frac{1}{n}(Hx - e_1)^\top (Hx - e_1) = \frac{1}{n}\left[x^\top H^\top Hx - 2e_1^\top Hx + 1\right] = \|x\|_2^2 - \frac{1}{n} \ge \frac{\|x\|_1^2}{|\text{supp}\,(x)|} - \frac{1}{n} = \frac{1}{k} - \frac{1}{n}$$

We used the fact that $e_1^\top Hx = \|x\|_1 = 1$ since the top row of H consists of only 1's. Hence $k \ge \left(\epsilon^2 + 1/n\right)^{-1} \ge 1/\max\left(\epsilon^2, 1/n\right) = \min\left(1/\epsilon^2, n\right)$. □

F.2. Proof of Theorem 5.1

Bounding probabilities. All the lemmas and theorems from this of this section are in the conditions of the construction described in Section 5: A and V are the random matrices previously defined, and S is a fixed subset of x-coordinates of size k.

Lemma F.1. *If the x-player plays the uniform strategy, then* $\mathbb{E}\left[\left\|V \cdot \vec{1}/n\right\|_p\right] \le \sqrt{p/n}$, *and* $\mathbb{P}\left[\left\|V \cdot \vec{1}/n\right\|_p \ge \epsilon\right] \le \sqrt{\frac{p}{n\epsilon^2}}$ *for* $n \ge p/\epsilon^2$.

Proof. The bound on the expectation follows from Khintchine's inequality, which states that for any given vectors $u_1, \ldots, u_m \in \mathbb{R}^n$ and iid uniform $\{-1, +1\}$-variables r_i

$$\mathbb{E}\left\|\sum_i r_i u_i\right\|_p \le \sqrt{p} \cdot \left(\sum_i \|u_i\|_p^2\right)^{1/2}$$

We refer to (Wolff et al.) or (Barman, 2015) for a proof. Let v_i be the columns of V. Since they are iid uniform, v_i has the same distribution of $r_i v_i$ for some uniform $\{-1, +1\}$-variable r_i. So:

$$\mathbb{E}\left\|V\left(\frac{1}{n}\vec{1}\right)\right\|_p = \mathbb{E}\left\|\sum_i r_i \frac{v_i}{n}\right\|_p \le \sqrt{p} \cdot \left(n \cdot \frac{1}{n^2}\right)^{1/2} = \sqrt{\frac{p}{n}}$$

The second part of the lemma is direct from Markov's inequality. □

For the second part, consider an x-player that is restricted to only use coordinates from S. We want to show that the y-player has a strategy that that would make all columns in S have a high value in $y^\top V$. The idea is that since n is large and k is small (in fact, a constant independent of n) there should be rows that are very skewed (i.e. have a lot more $+1$'s than -1's in the S columns). If y plays a strategy that only uses such rows, then he can force x to have high value.

[3] This can be seen by computing the derivative of $M_t(x) = \left(\frac{1}{n}\sum_i x_i^t\right)^{1/t}$ with respect to t and showing it is non-positive.

We call a row of A is *good* for the y-player if it has more than $(1/2+\epsilon)k$ 1's in the S-coordinates. Next we show that with high probability there is a large enough number of good rows available for the y-player. For this result, we need to lower bound the probability in the tail of the binomial distribution, which requires a tight anti-concentration inequality.

Anti-concentration can be derived by carefully plugging the moment generating function into the Paley-Zygmund (Paley & Zygmund, 1932) inequality, or by sharply estimating a sum of terms involving binomial coefficients. For further information we refer the reader to Tao's book (Tao).

Lemma F.2 (Chernoff bound)**.** *If $0<\rho\leq 1/2$, X_i are iid $\{0,1\}$-random variables with $\mathbb{P}[X_i=1]=\rho$ and $\sigma^2=k\rho(1-\rho)$ is the variance of $X=\sum_{i=1}^k X_i$, then there exist constants $C,c>0$ such that for all $\lambda\leq c\sigma$,*

$$\mathbb{P}\left[\left|\sum_{i=1}^k X_i-\rho\cdot k\right|\leq\lambda\sigma\right]\geq c\exp\left(-C\lambda^2\right)$$

Lemma F.3 (Anti-concentration for the binomial distribution)**.** *In the same conditions as in the previous lemma, there exist constants $\tilde{C},\tilde{c}>0$ such that for all $\lambda\leq c\sigma$,*

$$\mathbb{P}\left[\left|\sum_{i=1}^k X_i-\rho\cdot k\right|\geq\lambda\sigma\right]\geq\tilde{c}\exp\left(-\tilde{C}\lambda^2\right)$$

Lemma F.4. *There are $r=\Omega(n\exp(-O(k\epsilon^2))$ good rows with probability at least $1-\exp(-\Omega(n\exp(-O(k\epsilon^2))))$.*

Proof. Applying Lemma F.3 with $\rho=1/2$ and $\lambda=4\epsilon\sigma$ we obtain that the probability that a row is at least $k\left(\frac{1}{2}+\epsilon\right)k$ $+1$'s (or -1's, due to symmetry) is at least $\exp(-O(k\epsilon^2))$. So in expectation, there are $n\exp(-O(k\epsilon^2))$ good rows, so the result follows by applying the Chernoff bounds with $\rho=\exp(-O(k\epsilon^2))$ and $\lambda=c\sigma$. □

With high probability there will be at least $r=\Omega(n\exp(-O(k\epsilon^2))$ good rows for the y-player to play. We want to argue that if the y-player can find r good rows, then he can play $y_i=r^{-1/q}$ for each good row i, and 0 otherwise, and he will leave an x-player restricted to choosing only columns from a subset S without any good option to play.

Lemma F.5. *Let S be a fixed subset of columns of A, and let A_S be the matrix whose columns are the columns of A that belong to S. Conditioning on A_S having r good rows, with probability at least $1-k\exp\left(-\Omega(r\epsilon^2)\right)$, every column in S contains at least $r(1/2+\epsilon/2)$ $+1$'s in the r good rows.*

Proof. Sample matrix A according to the following procedure: in the first phase, sample each entry of A uniformly and independently from $\{-1,+1\}$. In the second phase, for each row, shuffle the entries in S (i.e. for each row, sample a random permutation of S and apply to the entries corresponding to those columns). In the first phase we can decide which rows are good, call those R. Conditioning on the first phase, and fixing a column $j\in S$, the entries A_{ij} for $i\in R$ are independent and uniform from $\{-1,+1\}$; the probability of A_{ij} being 1 is at least $\frac{1}{2}+\epsilon$, since this entry is a random entry from a good row sampled in the first phase. The result follows by applying the Chernoff bound with $\rho=\frac{1}{2}+\epsilon$ and $\lambda=(\sigma\epsilon)/(\rho\cdot(1-\rho))$. □

Now, we combine all the events discussed so far using the union bound:

Lemma F.6. *Fix ϵ and k. For sufficiently large n, there is a matrix A such that $V=n^{-1/p}\cdot A$ satisfies $\left\|V\cdot\vec{1}/n\right\|_p\leq\epsilon$, and for every subset S of k rows, there is a subset R of $r=\Omega(n\exp(-O(k\epsilon^2)))$ rows such that for all $i\in S$, $\sum_{j\in R}A_{ij}\geq\epsilon r$.*

Proof. The proof follows from the probabilistic method. For each subset S, with probability at least $1-\exp(-\Omega(n\exp(-O(k\epsilon^2))))$ there are r good rows (Lemma F.4) and with probability $1-k\exp\left(-\Omega(r\epsilon^2)\right)=1-k\exp\left(-\Omega(n\epsilon^2\exp(-O(k\epsilon^2)))\right)$ there are at least $(\frac{1}{2}+\epsilon)r$ many $+1$s in each column corresponding to the r rows (Lemma F.5), causing $\sum_{j\in R}A_{ij}\geq\epsilon r$. The probability that both events occur can be bounded by $1-O\left(k\exp\left(-\Omega(n\epsilon^2\exp(-O(k\epsilon^2)))\right)\right)$. Applying the union bound over all $\binom{n}{k}$ subsets S, we get:

$$1-\binom{n}{k}O\left(k\exp\left(-\Omega(n\epsilon^2\exp(-O(k\epsilon^2)))\right)\right)\geq 1-\exp\left(k\log n-O(\epsilon^2 n\exp(-O(k\epsilon^2)))\right)$$

which goes to one as $n\to\infty$ for any fixed k and ϵ. Also, as $n\to\infty$ the probability that $\left\|V(\frac{1}{n})\vec{1}\right\|_p\leq\epsilon$ also goes to 1. □

Theorem F.7 (Carathéodory lower bound)**.** *There is a matrix V whose columns have unit ℓ_p norm such that $\left\|V \cdot \vec{1}/n\right\|_p \leq \epsilon$, and for every $x \in \Delta$, $|\mathrm{supp}\,(x)| \leq k = O(p/\epsilon^2)$, $\|Vx\|_p \geq 2\epsilon$.*

Proof. Let V be the matrix obtained in Lemma F.6. From there, we have that $\left\|V \cdot \vec{1}/n\right\|_p \leq \epsilon$. Now, fix any $x \in \Delta$ with $|\mathrm{supp}\,(x)| \leq k$, and let S be the set of columns corresponding to the support of x. Let also R be the set of rows for which $\sum_{j \in R} A_{ij} \geq \epsilon r$ for all $i \in S$. Now, define $y \in \boldsymbol{B}_q(1)$ such that $y_i = r^{1/q}$ uf $i \in R$, and $y_i = 0$ otherwise:

$$\|Vx\|_p \geq y^\top Vx = n^{-1/p} \cdot (y^\top A)x \geq \frac{r\epsilon}{r^{1/q} \cdot n^{1/p}} = \epsilon \left(\frac{r}{n}\right)^{1/p}$$

We want to choose the parameters such that $\left(\frac{r}{n}\right)^{1/p} \geq 2$. Substituting $r = \Omega(n \exp(-O(k\epsilon^2)))$:

$$\left(\frac{r}{n}\right)^{1/p} = \exp\left(-O\left(\frac{k\epsilon^2}{p}\right)\right)$$

If $k \leq C \cdot \frac{p}{\epsilon^2}$ for a suitable constant C, we get $\|Vx\|_p \geq 2\epsilon$. □

G. Applications

The approach presented in the previous sections can be easily generalized or directly applied to a series of applications. Here we identify three representative applications to illustrate the usefulness of our approach. We note that there are many other possible applications in combinatorial optimization, game theory and machine learning, where a convex combination is often maintained as a subroutine of the algorithm.

G.1. 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 fast linear optimization oracle. An natural such instance is given by the matroid polytope. We denote a n-element matroid by $\mathcal{M}$ and its rank by r.

Proposition G.1. *There is an algorithm which, given a fractional point x^* contained inside the base polytope of a matroid $\mathcal{M}$, and a norm parameter $p \geq 2$, produces a distribution $\mathcal{D}$ over matroid bases supported on $O\left(\frac{p \cdot r^{2/p}}{\epsilon^2}\right)$ points, such that $\|\mathbb{E}_{x\sim\mathcal{D}}\,[x] - x^*\|_p \leq \epsilon$. Furthermore the algorithm requires $O\left(nr^{2/p}p/\epsilon^2\right)$ calls to $\mathcal{M}$'s independence oracle.*

Proof. The result follows from applying Theorem 3.2 for x^* in the convex hull of the characteristic vectors for matroid bases. Note that each of these vectors has sparsity r so their p norm is precisely $r^{1/p}$. Hence we have the desired sparsity for the support of $\mathcal{D}$. Each iteration requires maximizing a linear function over the bases of the polytope, which can be done using the standard greedy algorithm, and requires $O(n)$ calls to the independence oracle. □

Of course, there are other nice polytopes where the existence of an efficient linear optimization oracle offers advantages. To this aspect, we mention the s-t-flow polytope (i.e. the convex hull of all s-t paths), whose oracle is implemented with a single shortest path computation. This enables us to speed up the path stripping subroutine in the Raghavan-Thompson randomized rounding algorithm for approximating minimum congestion integral multicommodity flows (Raghavan & Thompson, 1991). As described in (Raghavan & Thompson, 1991) the algorithm takes $O(m^2)$, which can be improved to near linear time by carefully using link-cut trees (Kang & Payor, 2015). By contrast, approximate Carathéodory provides a lightweight algorithm for producing an approximate decomposition into integral paths, without the need of complicated data structures.

Proposition G.2. *There is an algorithm which, given a fractional s-t-flow f^* routing one unit of demand in G, and a norm parameter $p \geq 2$, produces a distribution $\mathcal{D}$ over $s-t$-paths supported on $O\left(\frac{p \cdot n^{2/p}}{\epsilon^2}\right)$ points, such that $\|\mathbb{E}_{f\sim\mathcal{D}}\,[f] - f^*\|_p \leq \epsilon$. Furthermore the algorithm requires $O\left(\frac{p \cdot n^{2/p}}{\epsilon^2}\right)$shortest path computations.*

In the setting of Raghavan-Thompson, fixing $p = \Theta(\log n)$ yields an approximate path stripping routine that runs in time $\tilde{O}(m/\epsilon^2)$.

G.2. SVM training

Support vector machines (SVM) are an extremely popular classification method, and have found ample usage in machine learning, with applications ranging from finance to neuroscience. In the era of big data it is crucial for any such method to be able to train on huge datasets. While a number of implementations (LIBLINEAR (Fan et al., 2008), P-packSVM (Zhu et al., 2009), Pegasos (Shalev-Shwartz et al., 2011)) achieve excellent convergence rates in the case of linear SVM's, handling arbitrary kernels raises a significantly harder problem. LIBLINEAR and Pegasos achieve $O(\log(1/\epsilon))$, respectively $O(1/\epsilon)$ convergence rate, but cannot be extended beyond linear kernels. The ϵ dependence for P-packSVM scales as $O(1/\epsilon)$, but it requires knowing the Cholesky factorization of the kernel matrix in advance. In our case, a simple extension of the method described in Section 3 gives $O(1/\epsilon^2)$ convergence, while only requiring matrix-vector multiplications involving the kernel matrix. So Cholesky factorization is no longer required, and the matrix does not need to be stored explicitly. In the case of linear SVM's, our method runs in nearly linear time.

Our approach is inspired from a reformulation of the training problem of Kitamura, Takeda, and Iwata (Kitamura et al., 2014), who present a method for SVM training based on Wolfe's algorithm. Their algorithm relies on a dual formulation introduced by Schölkopf et al. (Schölkopf et al., 2000) which can be easily reformulated as a convex problem over a product of two convex sets. More specifically, we are given empirical data $(x_i, y_i) \in \mathcal{X} \times \{\pm 1\}$, $1 \le i \le n$, along with a function that maps features to a Hilbert space $\Phi : \mathcal{X} \to \mathcal{H}$, which determines a kernel function $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$. Let $K \in \mathbb{R}_{n \times n}$, where $K_{ij} = k(x_i, x_j)$, $E_+ = \{e_i : y_i = +1\}$, $E_- = \{e_i : y_i = -1\}$.

In (Kitamura et al., 2014), the ν-SVM problem is reformulated as:

$$\begin{aligned} \min \quad & (\lambda_+ - \lambda_-)^\top K (\lambda_+ - \lambda_-) \\ \text{subject to} \quad & \lambda_+ \in \mathrm{RCH}_\eta(E_+) \\ & \lambda_- \in \mathrm{RCH}_\eta(E_-) \end{aligned}$$

where $\eta = \frac{2}{\nu n}$ and $\mathrm{RCH}_\eta(A) := \left\{ \sum_{a \in A} \lambda_a a | 0 \le \lambda_a \le \eta, \sum_{a \in A} \lambda_a = 1 \right\}$ is the *restricted convex hull* of set A.

Our approach to solve this problem will be to rephrase it as a saddle point problem (similar to what was done for the approximate Carathéodory problem) and apply Mirror Descent, with a suitable Mirror Map, to solve the dual. Before doing that, we introduce a few useful definitions and facts:

Definition G.3. *Let K be a symmetric positive definite matrix. Then $\|x\|_K := \sqrt{x^\top K x}$.*

Proposition G.4. *The dual norm of $\|x\|_K$ is $\|x\|_{K^{-1}}$. In other words $\|x\|_K = \max_{y : \|y\|_{K^{-1}} \le 1} \langle y, x \rangle$.*

Proof. This can be verified using Lagrange multipliers: over the unit $\|\cdot\|_{K^{-1}}$-ball the term $y^\top x$ attains its maximum at $y = Kx / \|Kx\|_{K^{-1}} = Kx/\sqrt{x^\top K x}$. We can verify that for this choice of y, $y^\top x = \frac{x^\top K x}{\sqrt{x^\top K x}} = \|x\|_K$. □

Definition G.5. *Let $\mathcal{S}_\eta = \{\lambda_+ - \lambda_- | \lambda_+ \in \mathrm{RCH}_\eta(E_+), \lambda_- \in \mathrm{RCH}_\eta(E_-)\}$.*

Proposition G.6 (Linear optimization over S_η)**.** *Linear optimization over S_η can be implemented in $\tilde{O}(n)$ time.*

Proof. The implementation of the linear optimization routine is done in near-linear time via a simple greedy algorithm. The first thing to notice is that the objective is separable, so it is sufficient to optimize separately on E_+ and E_-. This can be done easily, since we need to distribute one unit of mass over the coordinates that span E_+ and one unit of mass over the coordinates that span E_-, such that no coordinate receives more than η. Therefore adding mass to the coordinates spanning E_+ in increasing order of the weights y, and vice-versa to those spanning E_- yields the optimal solution. □

With these facts on hand, we can now proceed to describing our equivalent formulation as a saddle-point problem, which will then be solved using a similar method to the one we employed for the previous applications.

Note that instead of directly using the kernel matrix K in the formulation, we replace it with $\tilde{K} = K + \frac{\epsilon}{2} I$. This only changes the value of the objective by at most $\epsilon/2$ and it has the advantage of making $\tilde{K}$ positive semidefinite, since it is now guaranteed to be non-degenerate. This allows us to write the objective function $(\lambda_+ - \lambda_-)^\top \tilde{K} (\lambda_+ - \lambda_-)$ as $\|\lambda_+ - \lambda_-\|_{\tilde{K}}^2$. This formulation can be easily converted to a saddle point problem:

$$\min_{\lambda\in\mathcal{S}} \|\lambda\|_{\tilde K} = \min_{\lambda\in\mathcal{S}_\eta} \max_{y:\|y\|_{\tilde K^{-1}}\le 1} y^\top\lambda = -\min_{y:\|y\|_{\tilde K^{-1}}\le 1}\left(-\min_{\lambda\in\mathcal{S}_\eta} y^\top\lambda\right) = -\min_{y:\|y\|_{\tilde K^{-1}}\le 1} f(y)$$

for $f(y) := -\min_{\lambda\in\mathcal{S}_\eta} y^\top\lambda$ defined over the $\|\cdot\|_{\tilde K^{-1}}$-ball.

The subgradients of f are easy to compute, since they require a simple linear optimization over $\mathcal{S}$:

$$\partial f(y) = -\arg\min_{\lambda\in\mathcal{S}} y^\top\lambda$$

which can be done in time $\tilde O(n)$ using the greedy algorithm described in Proposition G.6. The mirror map of choice for the domain $\{y : \|y\|_{\tilde K^{-1}} \le 1\}$will be $\omega : \{y : \|y\|_{\tilde K^{-1}} \le 1\} \to \mathbb{R}$, $\omega(y) = \frac{1}{2}\|y\|_{\tilde K^{-1}}^2$, with

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

Also, similarly to Proposition C.2, we have $\nabla\omega^*(z) = \tilde K z\cdot\min(1, 1/\|z\|_{\tilde K})$, hence $\|\nabla\omega^*(Z)\|_{\tilde K^{-1}} \le 1$.

The only thing left to do is to analyze the algorithm's iteration count by bounding the strong convexity of ω and the Lipschitz constant of f. We will do this with respect to $\|\cdot\|_2$.

Proposition G.7. *ω is $\min\left(\frac{\epsilon}{2}, \left(\|K\| + \frac{\epsilon}{2}\right)^{-1}\right)$-strongly convex with respect to $\|\cdot\|_2$, where $\|K\|$ is the spectral norm of K.*

Proof. Writing down the Hessian of the mirror map, we obtain $\nabla^2\omega(y) = \tilde K^{-1} = \left(K + \frac{\epsilon}{2}I\right)^{-1} \succeq \min\left(\frac{\epsilon}{2}, \left(\|K\| + \frac{\epsilon}{2}\right)^{-1}\right) I$. The reason for using $\tilde K$ instead of K in the formulation is now evident: if K is not full rank, then ω is not strongly convex. Adding a small multiple of the identity forces all the eigenvalues of $\tilde K$ to be at least $\epsilon/2$, and avoids the degeneracy where some of them may be zero. □

Proposition G.8. *f is $2\sqrt{\eta}$-Lipschitz with respect to $\|\cdot\|_2$.*

Proof. We simply need to bound the 2-norm of the subgradient. By the construction presented in Proposition G.6 the subgradient contains $2\cdot\lceil 1/\eta\rceil$ nonzero coordinates, $2\cdot\lfloor 1/\eta\rfloor$ of which are precisely η. This enables us to obtain a better upper bound than one would usually expect on the 2-norm of the subgradient, namely $\sqrt{2\cdot\left(\lfloor 1/\eta\rfloor\cdot\eta^2 + (1-\eta\cdot\lfloor 1/\eta\rfloor)^2\right)} \le \sqrt{2\cdot(2/\eta)\cdot\eta^2} = 2\sqrt{\eta}$. □

Proposition G.9. $\max_{y:\|y\|_{\tilde K}\le 1} \frac{1}{2}\|y\|_{\tilde K}^2 \le \frac{1}{2}$

Finally we can put everything together:

Theorem G.10. *An ϵ-approximate solution to ν-SVM can be found in $O\left(\eta\cdot\max\left(\frac{2}{\epsilon}\|K\| + \frac{\epsilon}{2}\right)/\epsilon^2\right) = O\left(\max\left(\frac{1}{\epsilon}\|K\|\right)/\left(\nu n\epsilon^2\right)\right)$ iterations.*

Proof. Follows from plugging in the parameters $\sigma = \min\left(\epsilon/2, \left(\|K\| + \epsilon/2\right)^{-1}\right)$, $L = 2\sqrt{\eta}$, $R = O(1)$ into the mirror descent algorithm. □

At this point, it makes sense to analyze the performance of our algorithm for the most common choices of SVM kernels, which only requires bounding the spectral norm of the kernel matrix; for this purpose we will simply use the trace bound. The results are summarized in the table below. The last column of the table contains the number of iterations required to find a solution down to a precision of ϵ, given that all the vectors x_i belong to the unit ℓ_2 ball.

Kernel type	Upper bound on $\|K\|$	Iteration count
Polynomial (homogeneous): $K_{ij} = \langle x_i, x_j \rangle^d$	$n \cdot \max_i \|x_i\|_2^{2d}$	$O\left(\max\left(\frac{1}{n\nu\epsilon^3}, \frac{1}{\nu\epsilon^2}\right)\right)$
Polynomial (inhomogeneous): $K_{ij} = (1 + \langle x_i, x_j \rangle)^d$	$n \cdot \left(1 + \max_i \|x_i\|_2^2\right)^d$	$O\left(\max\left(\frac{1}{n\nu\epsilon^3}, \frac{2^d}{\nu\epsilon^2}\right)\right)$
RBF: $K_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$	n	$O\left(\max\left(\frac{1}{n\nu\epsilon^3}, \frac{1}{\nu\epsilon^2}\right)\right)$
Sigmoid: $K_{ij} = \tanh(\alpha \cdot \langle x_i, x_j \rangle + c)$	n	$O\left(\max\left(\frac{1}{n\nu\epsilon^3}, \frac{1}{\nu\epsilon^2}\right)\right)$

It is worth mentioning that each iteration requires $\tilde{O}(n)$ time for computing the subgradient, and a multiplication of the kernel matrix with a vector; one advantage is that the kernel matrix does not need to be explicitly stored, as its entries can be computed on the fly, whenever needed. In the case of linear kernels, this computation is implemented in linear time since $\tilde{K}z = [x_1|\dots|x_n]^\top [x_1|\dots|x_n]\, z + \frac{\epsilon}{2}z$, which requires computing a linear combination $h = \sum_i x_i \cdot z_i$ of the vectors x, and n dot products between vectors from the training set and h.