Non–Euclidean High–Order Smooth Convex Optimization

Juan Pablo Contreras*, Cristóbal Guzmán*, **David Martínez-Rubio***

Universidad Diego Portales, Universidad Católica de Chile, Zuse Institute Berlin, Carlos III University of Madrid

Collaborators

Juan Pablo Contreras
(Universidad Diego Portales)

Cristóbal Guzmán
(UC Chile)

Some Problems

- Box-simplex games: $\min_{x\in[-1,1]^n} \max_{y\in\Delta^d} x^T Ay - b^T y + c^T x$.
(more general than Linear Programs, discrete optimal transport, max flow problems, etc.)

- ℓ_p-regression: $\min_{x\in\mathbb{R}^d} \|Ax - b\|_p$.

- Logistic regression (has Lipschitz gradients wrt $\|\cdot\|_\infty$):
$\min_x \sum_{i\in[n]} \log(1 + \exp(-b_i\langle a_i, x\rangle))$, for $a_i \in \mathbb{R}^d$, $b_i \in \{-1, 1\}$. .

- Etc.

Black-Box Oracle Optimization

- Under some regularity conditions, for convex $f : \mathbb{R}^d \to \mathbb{R}$, we aim to

$$\text{minimize}_x \quad f(x).$$

- We access f by querying a local oracle at some points: e.g. gradient oracle.

Optimizing f in a class $\mathcal{F}$:

- Design an algorithm $\mathcal{A}$ s.t. $\forall f \in \mathcal{F}$, finds x s.t. $f(x) - \min_y f(y) \leq \varepsilon$ with few oracle queries.
- Show that $\forall \mathcal{A}$, $\exists f \in \mathcal{F}$, s.t. $\mathcal{A}$ requires that many oracle queries.

High-order smoothness, and beyond

- For $f : \mathbb{R}^d \to \mathbb{R}$, an arbitrary norm $\|\cdot\|$, and all $x, y \in \mathbb{R}^d$:

$$\|\nabla^q f(x) - \nabla^q f(y)\|_* \leq L \|x - y\|^\nu \text{ for some } q \geq 1, \nu \in (0, 1].$$

Implies

$$\|\nabla f(y) - \nabla f_q(x)(y)\|_* \leq L \|x - y\|^{q+\nu-1}, \text{ for some } q \geq 1, \nu \in (0, 1],$$

where $f_q(y; x)$ is the q-th order Taylor expansion of f at x.

High-order smoothness, and beyond

- For $f : \mathbb{R}^d \to \mathbb{R}$, an arbitrary norm $\|\cdot\|$, and all $x, y \in \mathbb{R}^d$:

$$\|\nabla^q f(x) - \nabla^q f(y)\|_* \leq L \|x - y\|^\nu \text{ for some } q \geq 1, \nu \in (0, 1].$$

  Implies

$$\|\nabla f(y) - \nabla f_q(x)(y)\|_* \leq L \|x - y\|^{q+\nu-1}, \text{ for some } q \geq 1, \nu \in (0, 1],$$

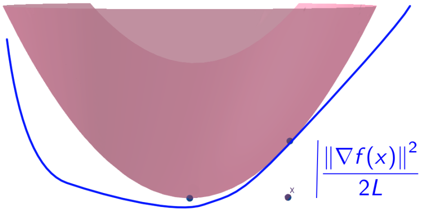
  where $f_q(y; x)$ is the q-th order Taylor expansion of f at x.

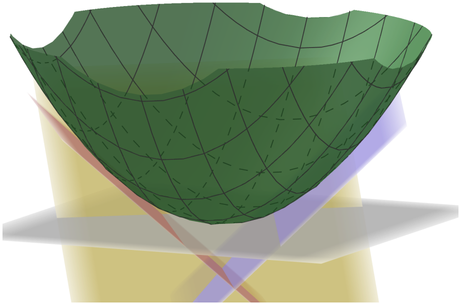
- Also in the limit when $q \to \infty$, we have a ball optimization oracle.
- That is, if we can approximately optimize f locally in unit balls, how fast can we optimize f?

Accelerated Gradient Descent (AGD) Methods

- Optimal 1st-order method for minimizing Euclidean convex, L-Lipschitz-gradient functions.

Gradient Descent	$O(\frac{LR^2}{\varepsilon})$
Accelerated Gradient Descent	$O(\sqrt{\frac{LR^2}{\varepsilon}})$





AGD is a combination of Gradient Descent and an online learning algorithm with proportional progress and instantaneous regret.
E.g. proportional to $\|\nabla f(x)\|^2$ in the unconstrained case.

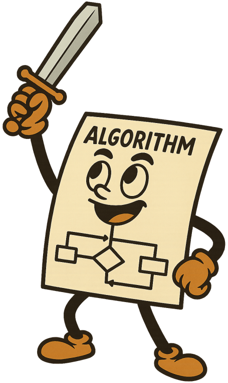
Convergence Results

Initial distance: $R_p \stackrel{\text{def}}{=} \|x_0 - x^*\|_p$; Accuracy: ε;
$m \stackrel{\text{def}}{=} \max\{2, p\}$. $\widetilde{O}_p(\cdot)$: big-O notation up to log factors and constants on p.
We use a q-th order or inexact ball oracle, $\nu \in (0, 1]$.

Algorithm	$p \in [1, \infty)$	$p = \infty$
Accelerated ($q < \infty$)	$\widetilde{O}_{q+\nu,p}\left(\left(\frac{LR_p^{q+\nu}}{\varepsilon}\right)^{\frac{m}{(m+1)(q+\nu)-m}}\right)$	–
Unaccelerated ($q < \infty$)	$\widetilde{O}_{q+\nu}\left(\left(\frac{LR_p^{q+\nu}}{\varepsilon}\right)^{\frac{1}{q}}\right)$	
ρ-Ball Oracle ($q = \infty$)	$\widetilde{O}_m\left((R_p/\rho)^{\frac{m}{m+1}}\right)$	$\widetilde{O}(R_\infty/\rho)$

Lower Bounds: Smoothing Hard Instances

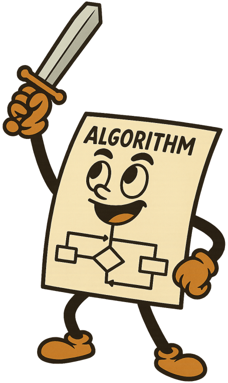
Lipschitz q-th order derivatives:



Our Algorithms	Our Lower Bounds
q-th order oracle	Local oracle
Deterministic	Possibly random algs
Single-call per round	$\mathrm{poly}(d)$ parallel queries
Any norm	Any norm
$\|\cdot\|_p$-setting: they match up to log factors	

Lower Bounds: Smoothing Hard Instances

Lipschitz q-th order derivatives:



Our Algorithms	Our Lower Bounds
q-th order oracle	Local oracle
Deterministic	Possibly random algs
Single-call per round	$\mathrm{poly}(d)$ parallel queries
Any norm	Any norm
$\|\cdot\|_p$-setting: they match up to log factors	

Inexact ball oracle: We match the lower bound in (Adil et al. 2025) that used an exact ball oracle.

Before: 1^{st}-order $\|\cdot\|_p$-LBs : $p < 2$ & $p \geq 2$ use different proofs. **Ours:** same proof. Solves an open problem on parallel 1^{st}-order convex optimization.

FTRL / Mirror Descent

- **Bregman Divergence**: $D_\psi(x,y) \stackrel{\text{def}}{=} \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle$.
- μ**-strongly convexity**: $D_\psi(x,y) \geq \frac{\mu}{2}\|x-y\|^2$.
- **FTRL algorithm:** Given 1-strongly convex ψ, initial point x_0, and vectors $g_1, \dots, g_T$ in a stream,

$$x_t \stackrel{\text{def}}{=} \operatorname{argmin}_x \left\{ \sum_{i=1}^{t-1} \langle g_i, x\rangle + \frac{D_\psi(x, x_0)}{\eta} \right\} \text{ for some } \eta > 0.$$

Then

$$\sum_{t=1}^{T} \langle g_t, x_t - u\rangle \leq \frac{D_\psi(u, x_0)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|g_t\|_*^2, \text{ for all } u.$$

FTRL / Mirror Descent

- **Bregman Divergence**: $D_\psi(x,y) \stackrel{\text{def}}{=} \psi(x) - \psi(y) - \langle \nabla\psi(y), x-y\rangle$.
- μ**-strongly convexity**: $D_\psi(x,y) \geq \frac{\mu}{2}\|x-y\|^2$.
- **FTRL algorithm:** Given 1-strongly convex ψ, initial point x_0, and vectors $g_1,\ldots,g_T$ in a stream,

$$x_t \stackrel{\text{def}}{=} \operatorname{argmin}_x \left\{ \sum_{i=1}^{t-1} \langle g_i, x\rangle + \frac{D_\psi(x,x_0)}{\eta} \right\} \text{ for some } \eta > 0.$$

Then

$$\sum_{t=1}^{T} \langle g_t, x_t - u\rangle \leq \frac{D_\psi(u,x_0)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_*^2, \text{ for all } u.$$

If $g_i = \nabla f(x_i)$ for some points x_i, then

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \nabla f(x_t), x_t - x^*\rangle.$$

- **Inexact Uniform Convexity:**

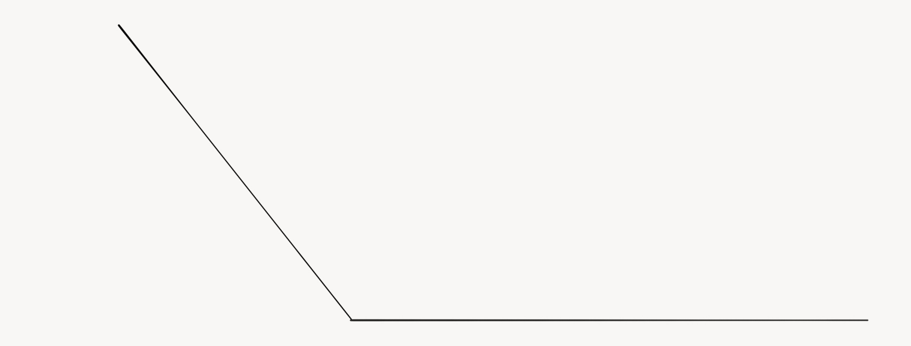
$$D_\psi(x,y) \geq \frac{\mu}{s}\|x-y\|^s - \delta, \text{ for } \delta, s \geq 0.$$

Moreau Envelope and Proximal Operator

$$M_{\lambda,f}(x) \overset{\text{def}}{=} \min_y \left\{ f(y) + \frac{1}{2\lambda} \|y - x\|_2^2 \right\}; \quad \text{prox}_{\lambda,f}(x) \overset{\text{def}}{=} \text{argmin}_y \left\{ f(y) + \frac{1}{2\lambda} \|y - x\|_2^2 \right\}.$$

By optimality $\nabla_y \left(f(y) + \frac{1}{2\lambda} \|y - x\|_2^2 \right) (\text{prox}(x)) = 0$, so

$\text{prox}(x) = x - \lambda \nabla f(\text{prox}(x))$, i.e., implicit Gradient Descent. And minimizers are preserved.

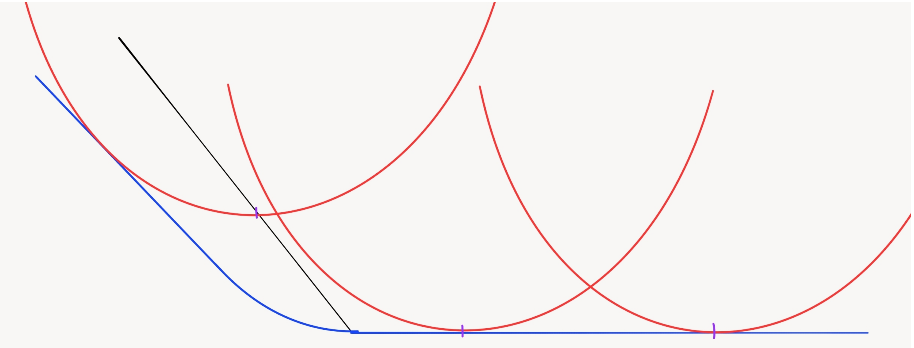


Moreau Envelope and Proximal Operator

$$M_{\lambda,f}(x) \stackrel{\text{def}}{=} \min_y \left\{ f(y) + \frac{1}{2\lambda} \|y - x\|_2^2 \right\}; \quad \operatorname{prox}_{\lambda,f}(x) \stackrel{\text{def}}{=} \operatorname{argmin}_y \left\{ f(y) + \frac{1}{2\lambda} \|y - x\|_2^2 \right\}.$$

By optimality $\nabla_y \left(f(y) + \frac{1}{2\lambda} \|y - x\|_2^2 \right) (\operatorname{prox}(x)) = 0$, so

$\operatorname{prox}(x) = x - \lambda \nabla f(\operatorname{prox}(x))$, i.e., implicit Gradient Descent. And minimizers are preserved.



Non-Euclidean Proximal Point Step

Let $\|\cdot\|$ be an arbitrary norm. Traditionally people use the non-Euclidean Moreau envelope

$$M_\psi(x) \stackrel{\text{def}}{=} \min_y \left\{ f(y) + \frac{1}{\lambda} D_\psi(y, x) \right\}, \quad \text{prox}_{\lambda}(x) \stackrel{\text{def}}{=} \text{argmin}_y \left\{ f(y) + \frac{1}{\lambda} D_\psi(y, x) \right\}.$$

for ψ being strongly convex wrt $\|\cdot\|$.

Non-Euclidean Proximal Point Step

Let $\|\cdot\|$ be an arbitrary norm. **We use**

$$M(x) \stackrel{\text{def}}{=} \min_y \left\{ f(y) + \frac{1}{(q+\nu)\lambda} \|y-x\|^{q+\nu} \right\}, \quad \text{prox}_{\lambda}(x) \stackrel{\text{def}}{=} \text{argmin}_y \left\{ f(y) + \frac{1}{(q+\nu)\lambda} \|y-x\|^{q+\nu} \right\}.$$

Non-Euclidean Proximal Point Step

Let $\|\cdot\|$ be an arbitrary norm. **We use**

$$M(x) \stackrel{\text{def}}{=} \min_y \left\{ f(y) + \frac{1}{(q+\nu)\lambda} \|y-x\|^{q+\nu} \right\}, \quad \text{prox}_{\lambda}(x) \stackrel{\text{def}}{=} \text{argmin}_y \left\{ f(y) + \frac{1}{(q+\nu)\lambda} \|y-x\|^{q+\nu} \right\}.$$

M is **not smooth** in general but satisfies a **descent condition** and **controlled subgradient norm**:

$$M_\lambda(x) - M_\lambda(\text{prox}_\lambda(x)) \geq \frac{1}{(q+\nu)\lambda} \|\text{prox}_\lambda(x) - x\|^{q+\nu},$$

and

$$\|g_x\|_* = \frac{1}{\lambda} \|\text{prox}(x) - x\|^{q+\nu-1} \text{ and } \langle g_x, \text{prox}(x) - x \rangle = \frac{1}{\lambda} \|\text{prox}(x) - x\|^{q+\nu}.$$

Non-Euclidean Proximal Point Step

Let $\|\cdot\|$ be an arbitrary norm. **We use**

$$M(x) \stackrel{\text{def}}{=} \min_y \left\{ f(y) + \frac{1}{(q+\nu)\lambda} \|y-x\|^{q+\nu} \right\}, \quad \text{prox}_{\lambda}(x) \stackrel{\text{def}}{=} \text{argmin}_y \left\{ f(y) + \frac{1}{(q+\nu)\lambda} \|y-x\|^{q+\nu} \right\}.$$

M is **not smooth** in general but satisfies a **descent condition** and **controlled subgradient norm**:

$$M_\lambda(x) - M_\lambda(\text{prox}_\lambda(x)) \geq \frac{1}{(q+\nu)\lambda} \|\text{prox}_\lambda(x) - x\|^{q+\nu},$$

and

$$\|g_x\|_* = \frac{1}{\lambda} \|\text{prox}(x) - x\|^{q+\nu-1} \text{ and } \langle g_x, \text{prox}(x) - x \rangle = \frac{1}{\lambda} \|\text{prox}(x) - x\|^{q+\nu}.$$

Regularized Taylor subproblems: Find a point with low gradient norm of

$$f_q(y; x_k) + M \|y - x_k\|^{q+\nu},$$

for certain $M > 0$. We show **problems are convex** if $x \mapsto \|x\|^2$ is strongly convex wrt itself. E.g. p-norms, for $p \in (1, 2]$.

Lower Bound Techniques: Hard functions

- The simplest hard function for the Euclidean Lipschitz convex class for $x_0 = 0$:

$$x \mapsto \max_{i \in [d]} \left\{ x_i - \frac{i}{d} \right\} \text{ for } x \in B(0,1).$$

- If we have a point $x = (x_1, \ldots, x_k, 0, \ldots, 0)$ we only observe k of the linear functions!
- We need to observe them all to find the minimizer. We need many to approximate it.

Lower Bound Techniques: Randomized Smoothing

- Smoothing of an f that is G-Lipschitz wrt $\|\cdot\|$:

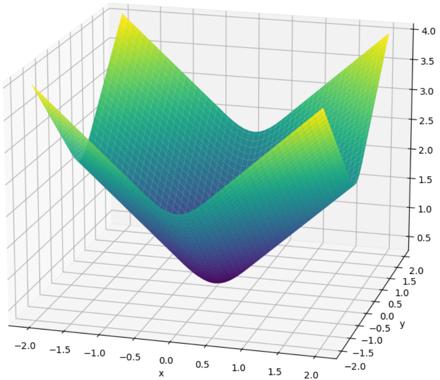
$$S_\beta[f](x) \stackrel{\text{def}}{=} \mathbb{E}_{v \sim \nu_{B_{\|\cdot\|}(\mathbf{0},\beta)}}[f(x+v)]$$

- $S_\beta[f](x)$ is also G-Lipschitz and its gradient is $\frac{dG}{\beta}$-Lipschitz wrt $\|\cdot\|$.

- $|S_\beta[f](x) - f(x)| \leq \beta G$.

- Since $S_\beta[f](x)$ depends on f locally it will preserve local hardness.

Figure: A smoothing of $x \mapsto \|x\|_1$ wrt $\|\cdot\|_1$



Lower Bound Techniques: Softmax and our Final Hard Function

- Let $A(x) = (\langle a^{(1)}, x\rangle, \ldots, \langle a^{(d)}, x\rangle)$, with $\left\|a^{(i)}\right\|_* \leq 1$.
- Define the softmax function as $\mathrm{smax}_\mu(x) \stackrel{\text{def}}{=} \mu \ln \left(\sum_{j=1}^{d} \exp(x_i/\mu)\right)$.

Lower Bound Techniques: Softmax and our Final Hard Function

- Let $A(x) = (\langle a^{(1)}, x\rangle, \ldots, \langle a^{(d)}, x\rangle)$, with $\left\|a^{(i)}\right\|_* \leq 1$.
- Define the softmax function as $\mathrm{smax}_\mu(x) \stackrel{\text{def}}{=} \mu \ln\left(\sum_{j=1}^{d} \exp(x_i/\mu)\right)$.
- $\mathrm{smax}_\mu(Ax)$ is 1-Lipschitz wrt $\|\cdot\|$.
- $\nabla^q \mathrm{smax}_\mu$ is $\widetilde{O}_q(\frac{1}{\mu^q})$-Lipschitz wrt $\|\cdot\|$.

Lower Bound Techniques: Softmax and our Final Hard Function

- Let $A(x) = (\langle a^{(1)}, x\rangle, \dots, \langle a^{(d)}, x\rangle)$, with $\left\|a^{(i)}\right\|_* \leq 1$.
- Define the softmax function as $\mathrm{smax}_\mu(x) \stackrel{\mathrm{def}}{=} \mu \ln\left(\sum_{j=1}^{d} \exp(x_i/\mu)\right)$.
- $\mathrm{smax}_\mu(Ax)$ is 1-Lipschitz wrt $\|\cdot\|$.
- $\nabla^q \mathrm{smax}_\mu$ is $\widetilde{O}_q(\frac{1}{\mu^q})$-Lipschitz wrt $\|\cdot\|$.
- $f_i \equiv$ softmax of $(Ax)_1 - \gamma, \dots, (Ax)_i - i\gamma$, for some $\gamma > 0$, up to some shifts.
- $h(x) \stackrel{\mathrm{def}}{=} \max_{i\in[T]} f_i(x)$.
- Hard function $g(x) = (S_{\beta/2^q} \circ S_{\beta/2^{q-1}} \circ \cdots \circ S_{\beta/2})(h)$.

A Problem Example: ℓ_∞-regression

- **Goal:** $\min_x \|Ax - b\|_\infty$ up to ε.

A Problem Example: ℓ_∞-regression

- **Goal:** $\min_x \|Ax - b\|_\infty$ up to ε.
- Approximate it by $c \log\left(\sum_{i\in[d]} \exp(\frac{1}{c}(Ax)_i)\right)$ for $c = \frac{\varepsilon}{2\log(d)}$.

A Problem Example: ℓ_∞-regression

- **Goal:** $\min_x \|Ax - b\|_\infty$ up to ε.
- Approximate it by $c \log\left(\sum_{i\in[d]} \exp(\frac{1}{c}(Ax)_i)\right)$ for $c = \frac{\varepsilon}{2\log(d)}$.
- It is $\frac{\varepsilon}{2}$ away. Solve for $\frac{\varepsilon}{2}$ accuracy.

A Problem Example: ℓ_∞-regression

- **Goal:** $\min_x \|Ax - b\|_\infty$ up to ε.
- Approximate it by $c \log \left(\sum_{i \in [d]} \exp(\frac{1}{c}(Ax)_i) \right)$ for $c = \frac{\varepsilon}{2 \log(d)}$.
- It is $\frac{\varepsilon}{2}$ away. Solve for $\frac{\varepsilon}{2}$ accuracy.
- The Hessian is locally multiplicative stable: $\gamma^{-1}\nabla^2 f(y) \preccurlyeq \nabla^2 f(x) \preccurlyeq \gamma \nabla^2 f(y)$, for $\gamma = O(1)$, $\forall y \in B_{\|\cdot\|_\infty}(x, \widetilde{O}(\varepsilon))$.

A Problem Example: ℓ_∞-regression

- **Goal:** $\min_x \|Ax - b\|_\infty$ up to ε.

- Approximate it by $c \log \left(\sum_{i \in [d]} \exp(\frac{1}{c}(Ax)_i) \right)$ for $c = \frac{\varepsilon}{2\log(d)}$.

- It is $\frac{\varepsilon}{2}$ away. Solve for $\frac{\varepsilon}{2}$ accuracy.

- The Hessian is locally multiplicative stable: $\gamma^{-1}\nabla^2 f(y) \preccurlyeq \nabla^2 f(x) \preccurlyeq \gamma \nabla^2 f(y)$, for $\gamma = O(1)$, $\forall y \in B_{\|\cdot\|_\infty}(x, \widetilde{O}(\varepsilon))$.

- One Hessian and $\widetilde{O}(1)$ gradients are enough to implement an ℓ_∞-ball optimization oracle of radius $\widetilde{O}(\varepsilon)$.

Thanks!

Questions?