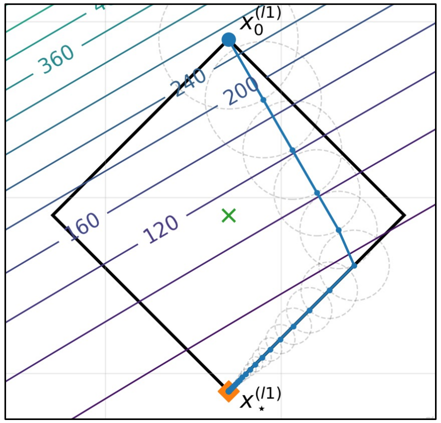
Local LMO: Constrained Gradient Optimization via a Local Linear Minimization Oracle

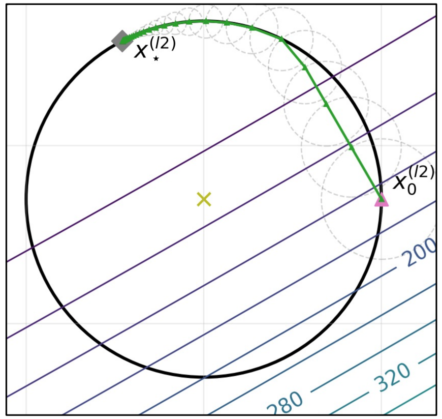
(Is it Time to Ditch Frank-Wolfe?)

Peter Richtárik
King Abdullah University of Science and Technology
Kingdom of Saudi Arabia



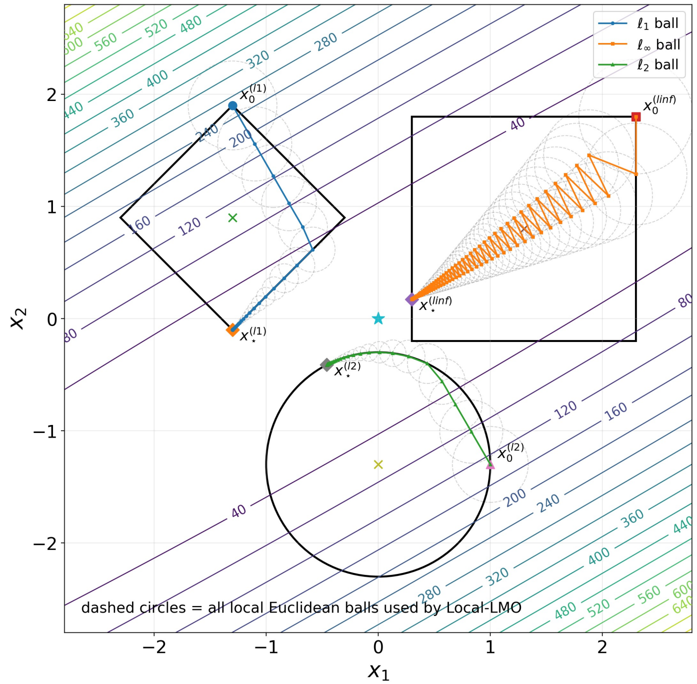
King Abdullah University of Science and Technology

ELLIIT Symposium on Optimization for Learning
Lund, Sweden, May 6-8, 2026

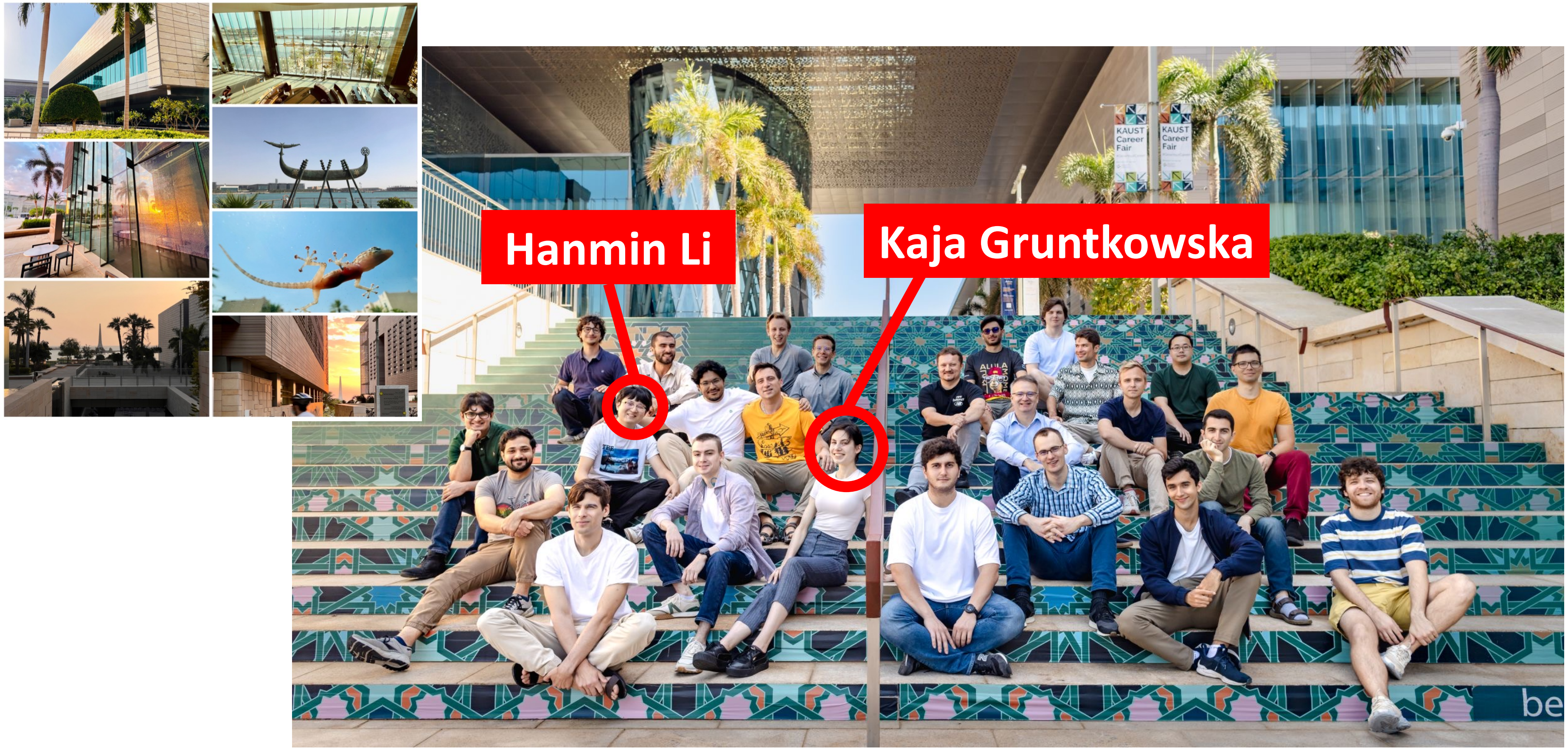


Talk Outline

- Introduction
- New Algorithm
- Convergence Theory
- Toy Experiments



Optimization & Machine Learning Lab @ KAUST



Part 1
Introduction

Fundamental Optimization Problem: Constrained Optimization

$$\min_{x \in \mathcal{X}} f(x)$$

loss / cost / objective function

$f : \mathbb{R}^d \to \mathbb{R}$

differentiable

constraint / feasible set

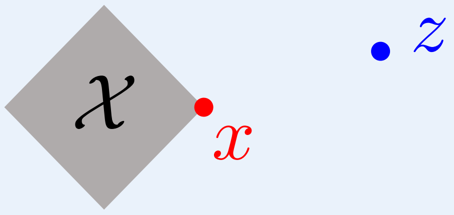
$\emptyset \neq \mathcal{X} \subseteq \mathbb{R}^d$

closed & convex

$$\mathcal{X}_\star := \arg\min_{x \in \mathcal{X}} f(x) \text{ is nonempty}$$

Three Ways of Handling Constraints

Projection Oracle

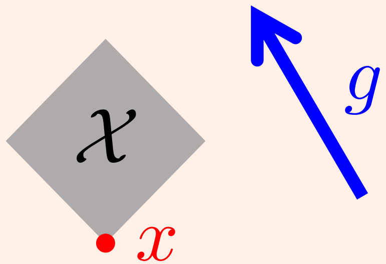


$$\arg\min_{x \in \mathcal{X}} \|x - z\|$$

Projected GD
Mirror Descent

Strong theory
Practical

(Global) Linear Minimization Oracle

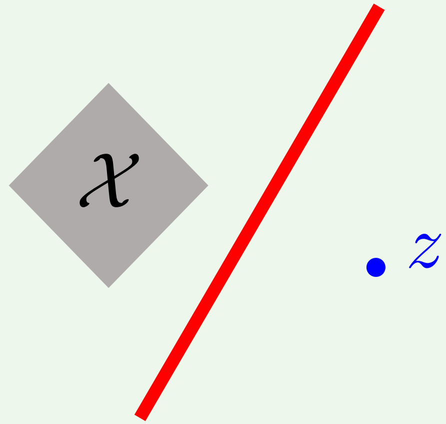


$$\arg\min_{x \in \mathcal{X}} \langle g, x \rangle$$

Frank–Wolfe
Conditional GD

Weak theory
Practical

Separation Oracle



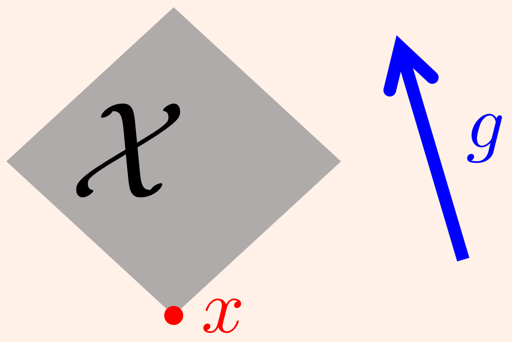
Ellipsoid Method

Strong theory
Not Practical

Can we design a projection-free gradient method with projected-GD-grade theory?

New Oracle: Local LMO

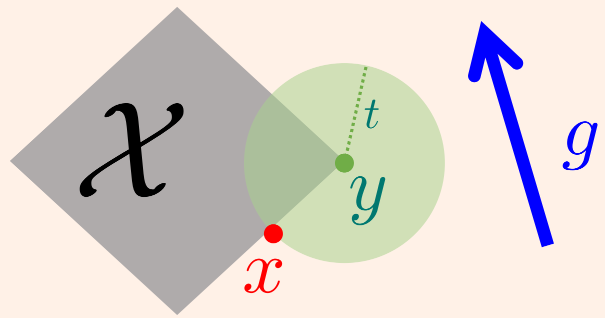
(Global) LMO



$$\arg\min_{x\in\mathcal{X}} \langle g, x\rangle$$

Frank–Wolfe
Conditional GD

(Local) LMO



$$\arg\min_{x\in\mathcal{X}\cap\mathcal{B}(y,t)} \langle g, x\rangle$$

Local LMO

Part 2
New Algorithm: Local LMO

New Method: Local LMO

$$\min_{x \in \mathcal{X}} f(x)$$

nonempty, closed & convex

differentiable &
not necessarily convex!

$$x_{k+1} \in \arg \min_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z \rangle$$

current iterate
$x_k \in \mathcal{X}$

admissible radius

No Constraints:
Local LMO = Gradient Descent

$$\min_{x \in \mathcal{X}} f(x)$$

$$\mathcal{X} = \mathbb{R}^d$$

Local LMO

$$x_{k+1} = \arg \min_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z \rangle$$

$$= \arg \min_{z \,:\, \|z - x_k\| \le t_k} \langle \nabla f(x_k), z \rangle$$

$$= x_k - t_k \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$$

Effective stepsize

$t_k = \|x_{k+1} - x_k\|$

Subspace Constraints:
Local LMO = Gradient Descent in the Subspace

$$\min_{x \in \mathcal{X}} f(x)$$

$\mathcal{X}$ = linear subspace of $\mathbb{R}^d$

Local LMO

$$x_{k+1} = \begin{cases} x_k - t_k \dfrac{\mathrm{Proj}_{\mathcal{X}}(\nabla f(x_k))}{\| \mathrm{Proj}_{\mathcal{X}}(\nabla f(x_k)) \|} & \text{if} \quad \mathrm{Proj}_{\mathcal{X}}(\nabla f(x_k)) \neq 0 \\ x_k & \text{if} \quad \mathrm{Proj}_{\mathcal{X}}(\nabla f(x_k)) = 0 \end{cases}$$

Projection Oracle vs Linear Minimization Oracle

	Projected GD	Global LMO (Frank-Wolfe)	Local-LMO (new!)
Projection-free	✗	✓	✓
Works even when $\mathrm{diam}\,(\mathcal{X}) = +\infty$	✓	✗	✓
Rate independent of $\mathrm{diam}(\mathcal{X})$	✓	✗	✓
Reduces to GD when $\mathcal{X} = \mathbb{R}^d$	✓	✗	✓
Linear convergence in smooth strongly convex regime	✓	✗	✓
Converges for non-smooth functions	✓	✗	✓
Easily implementable stepsize rules	✓	✓	✗

Part 3
Convergence Theory

Master Theorem

Two Radius Admissibility Rules

Type-I Admissibility:

$$\nabla f(x_k) \neq 0, \quad 0 < t_k \leq \frac{\langle \nabla f(x_k),\, x_k - x_\star \rangle}{\|\nabla f(x_k)\|}$$

constrained minimizer

$$x_\star \in \arg\min_{x \in \mathcal{X}} f(x)$$

Type-II Admissibility:

$$\nabla f(x_k) \neq \nabla f(x_\star), \quad 0 < t_k \leq \frac{\langle \nabla f(x_k) - \nabla f(x_\star),\, x_k - x_\star \rangle}{\|\nabla f(x_k) - \nabla f(x_\star)\|}$$

Master Theorem

$$x_{k+1} \in \arg \min_{z \in \mathcal{X} \cap \mathcal{B}(x_k, t_k)} \langle \nabla f(x_k), z \rangle$$

nonempty, closed & convex

$x_k \in \mathcal{X}$

differentiable &
not necessarily convex!

Admissible radius (Type-I or Type-II)

$$\nabla f(x_k) \neq 0, \quad 0 < t_k \leq \frac{\langle \nabla f(x_k), x_k - x_\star \rangle}{\|\nabla f(x_k)\|}$$

$$\nabla f(x_k) \neq \nabla f(x_\star), \quad 0 < t_k \leq \frac{\langle \nabla f(x_k) - \nabla f(x_\star), x_k - x_\star \rangle}{\|\nabla f(x_k) - \nabla f(x_\star)\|}$$

The ball radius

is never too big: $$t_k \leq \|x_k - x_\star\|$$

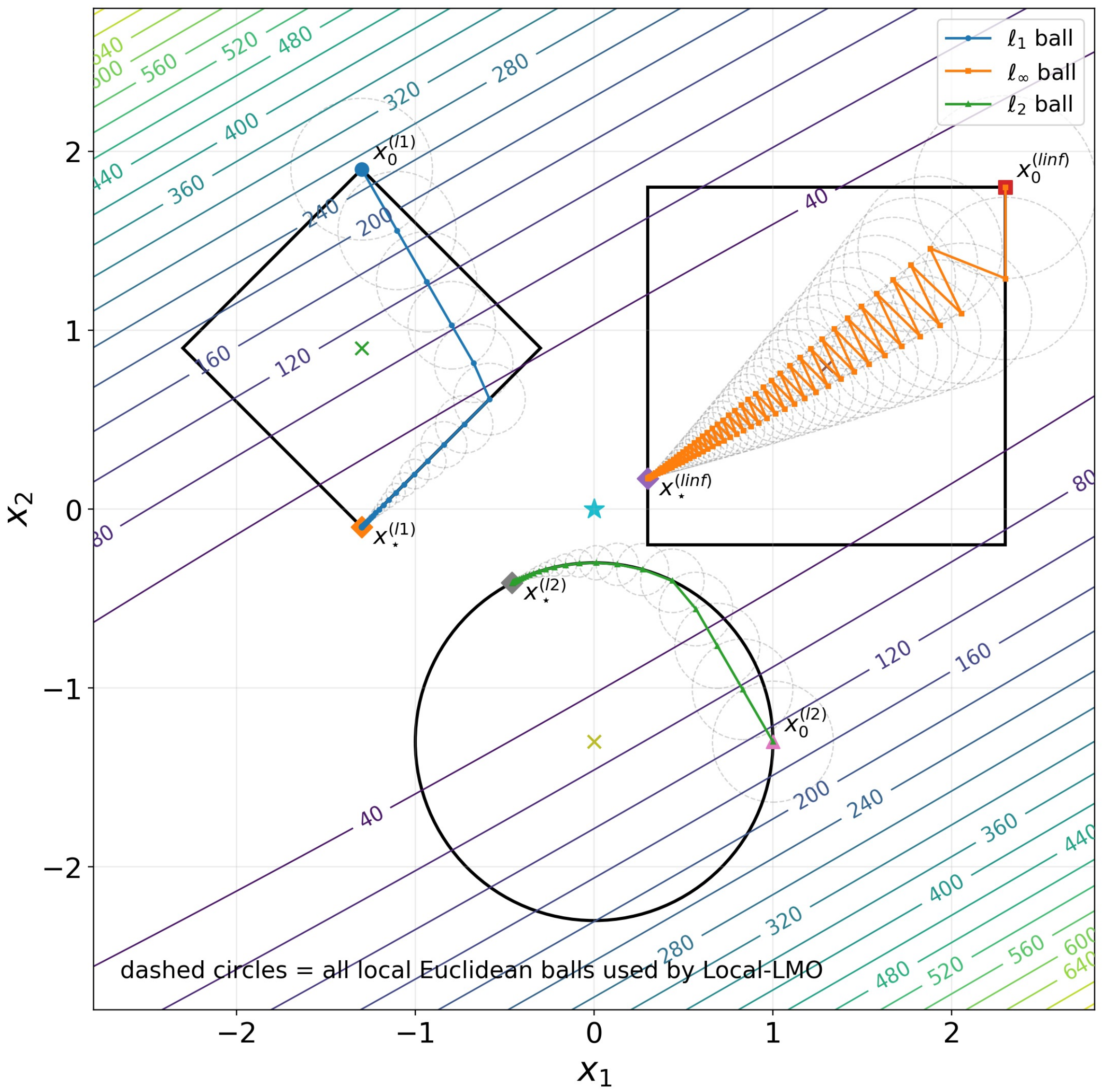
equals the effective stepsize: $$t_k = \|x_{k+1} - x_k\|$$

drives convergence: $$\|x_{k+1} - x_\star\|^2 \leq \|x_k - x_\star\|^2 - t_k^2$$

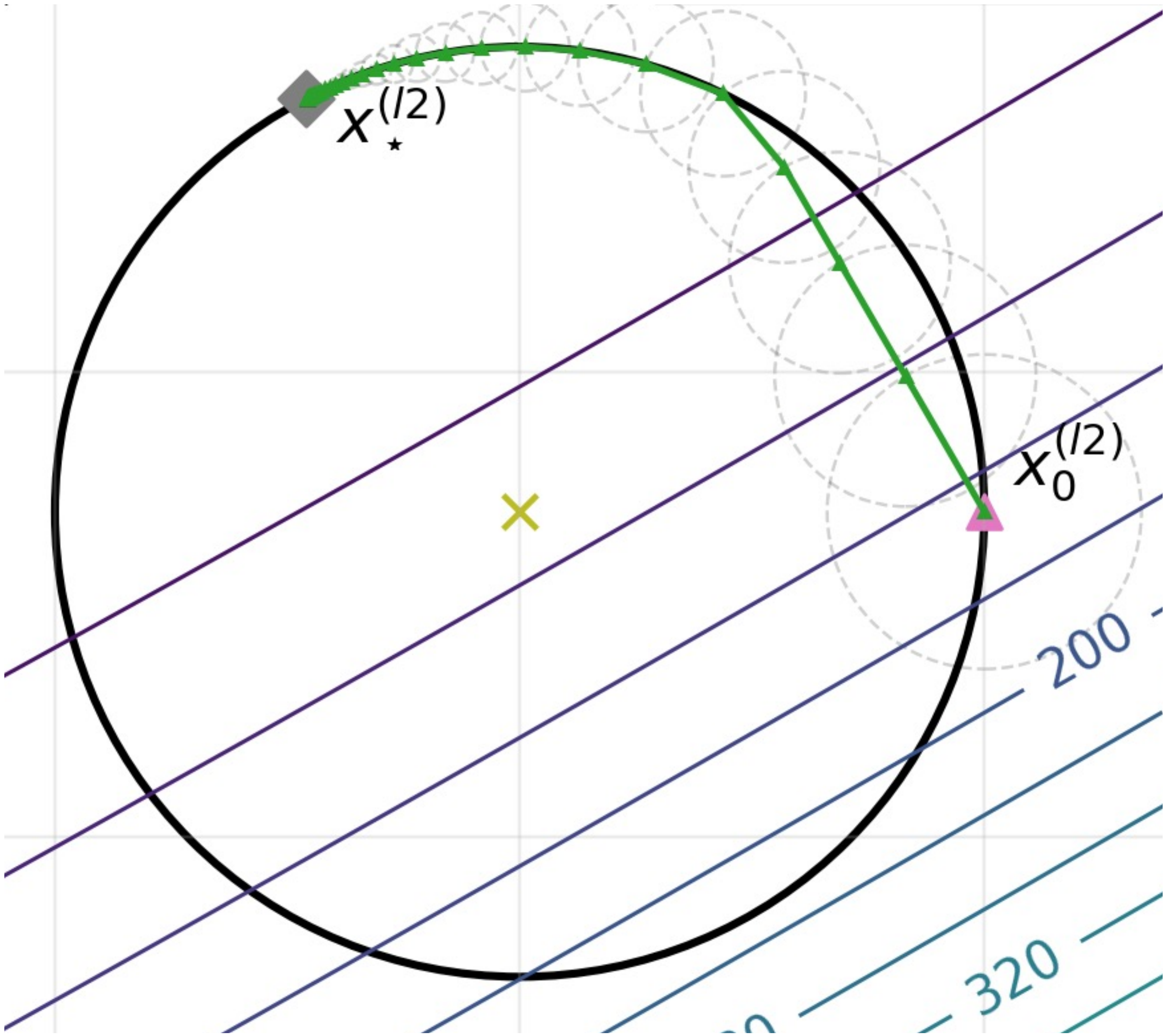
constrained minimizer

$$x_\star \in \arg \min_{x \in \mathcal{X}} f(x)$$

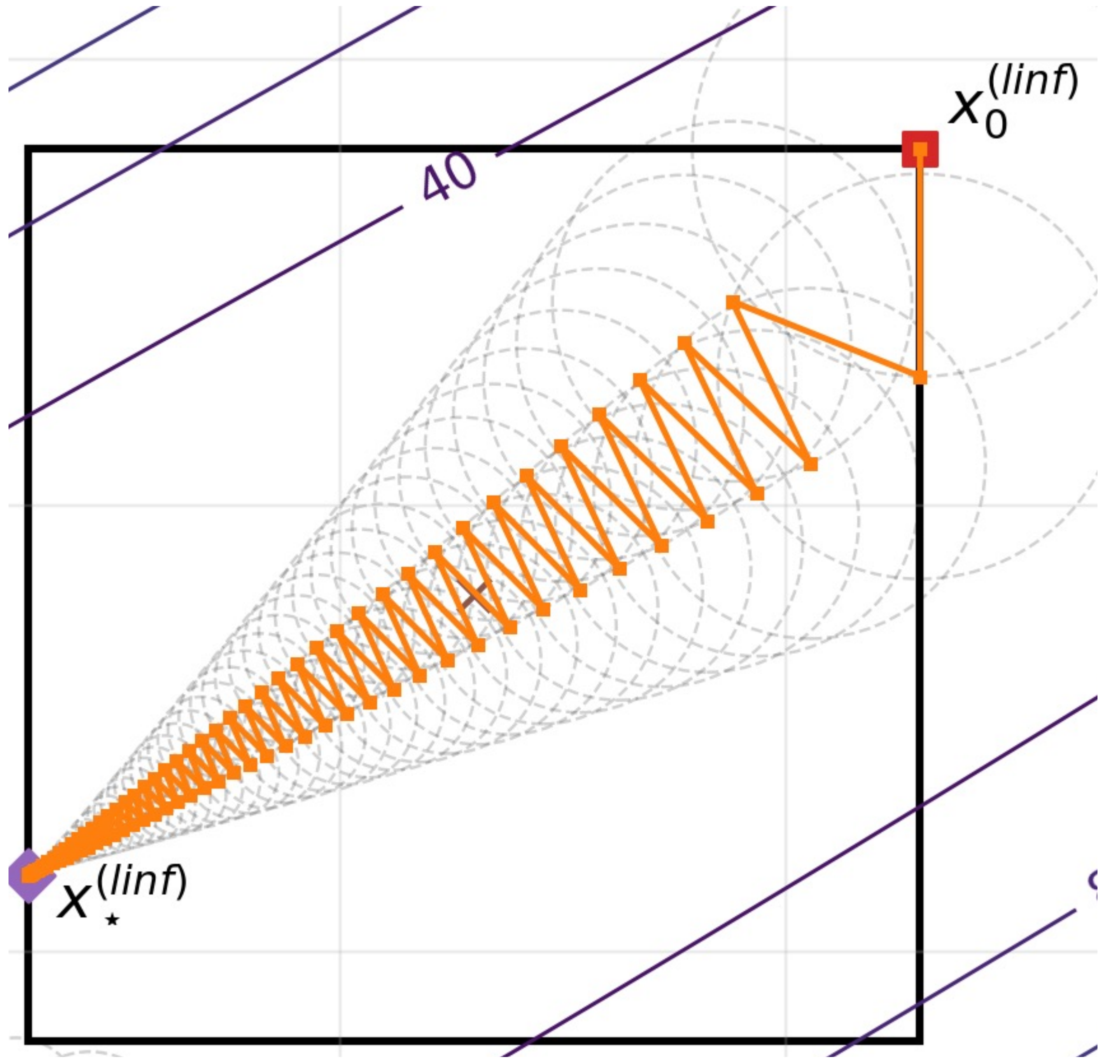
2D Examples



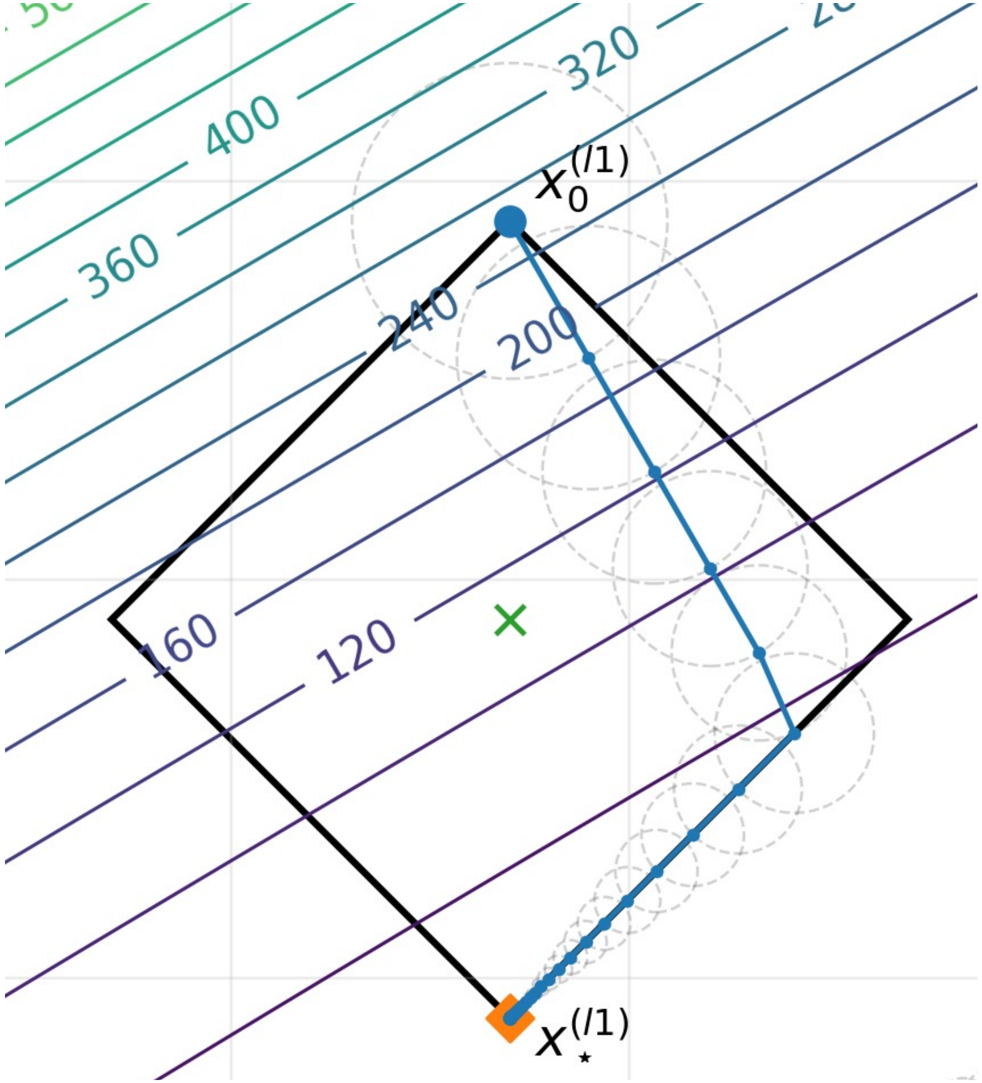
ℓ_1 ball
ℓ_∞ ball
ℓ_2 ball
$x_0^{(l1)}$
$x_*^{(l1)}$
$x_0^{(linf)}$
$x_*^{(linf)}$
$x_0^{(l2)}$
$x_*^{(l2)}$
dashed circles = all local Euclidean balls used by Local-LMO
x_1
x_2



$x_\star^{(l2)}$
$x_0^{(l2)}$
200
320



$x_0^{(linf)}$
40
$x_\star^{(linf)}$



320
400
360
240
200
160
120
$x_0^{(l1)}$
$x_\star^{(l1)}$

Theory in the
Smooth & Convex Regime

Convergence Theorem: Smooth Convex Regime

radius:

$$t_k = \frac{\|\nabla f(x_k) - \nabla f(x_\star)\|}{L}$$

constrained minimizer

$x_\star \in \arg\min_{x \in \mathcal{X}} f(x)$

Same rate as Projected GD!

In the unconstrained case:

Local LMO with this radius = GD with 1/L stepsize

$$\Rightarrow \min_{0 \le k \le K-1} \|\nabla f(x_k) - \nabla f(x_\star)\|^2 \le \frac{L^2 \|x_0 - x_\star\|^2}{K}$$

initial iterate

$x_0 \in \mathcal{X}$

f is L-smooth and strictly convex

Theory in the
Smooth & Strongly Convex Regime

Convergence Theorem: Smooth & Strongly Convex Regime

Distance radius:

$$t_k = \frac{2\sqrt{\mu L}}{L+\mu}\|x_k - x_\star\|$$

constrained minimizer

$x_\star \in \arg\min_{x \in \mathcal{X}} f(x)$

Same rate as Projected GD! Frank-Wolfe does not converge linearly!

$$\|x_K - x_\star\|^2 \leq \left(\frac{L-\mu}{L+\mu}\right)^{2K} \|x_0 - x_\star\|^2$$

f is L-smooth and μ-strongly convex

initial iterate

$x_0 \in \mathcal{X}$

Theory in the
Non-Smooth & Convex Regime

Convergence Theorem: Non-Smooth Convex Regime

Polyak radius:

$$t_k = \frac{f(x_k) - f(x_\star)}{\|\nabla f(x_k)\|}$$

constrained minimizer

$$x_\star \in \arg\min_{x \in \mathcal{X}} f(x)$$

Same rate as Projected GD!

In the unconstrained case:

Local LMO with Polyak radius = GD with Polyak stepsize

$$\Rightarrow \quad f(\hat{x}_K) - f(x_\star) \leq \frac{G\|x_0 - x_\star\|}{\sqrt{K}}$$

$$\hat{x}_K := \frac{1}{K}\sum_{k=0}^{K-1} x_k$$

f has bounded gradients

$$\|\nabla f(x)\| \leq G \qquad \forall x \in \mathcal{X}$$

initial iterate

$$x_0 \in \mathcal{X}$$

Theory in the
Smooth & Non-Convex Regime

Convergence Theorem: Smooth Non-Convex Regime

Gradient mapping radius:

$$t_k = \frac{\|G(x_k)\|}{L}$$

$$G(x) = L\left[x - \mathrm{Proj}_{\mathcal{X}}\left(x - \frac{1}{L}\nabla f(x)\right)\right]$$

Same rate as Projected GD!

f is L-smooth

In the unconstrained case:
$G(x) = \nabla f(x)$
Local LMO with this radius = GD with 1/L stepsize

initial iterate
$x_0 \in \mathcal{X}$

$$\Rightarrow \min_{0 \le k \le K-1} \|G(x_k)\|^2 \le \frac{2L(f(x_0) - f_\star)}{K}$$

minimum
$f_\star = \min_{x \in \mathcal{X}} f(x)$

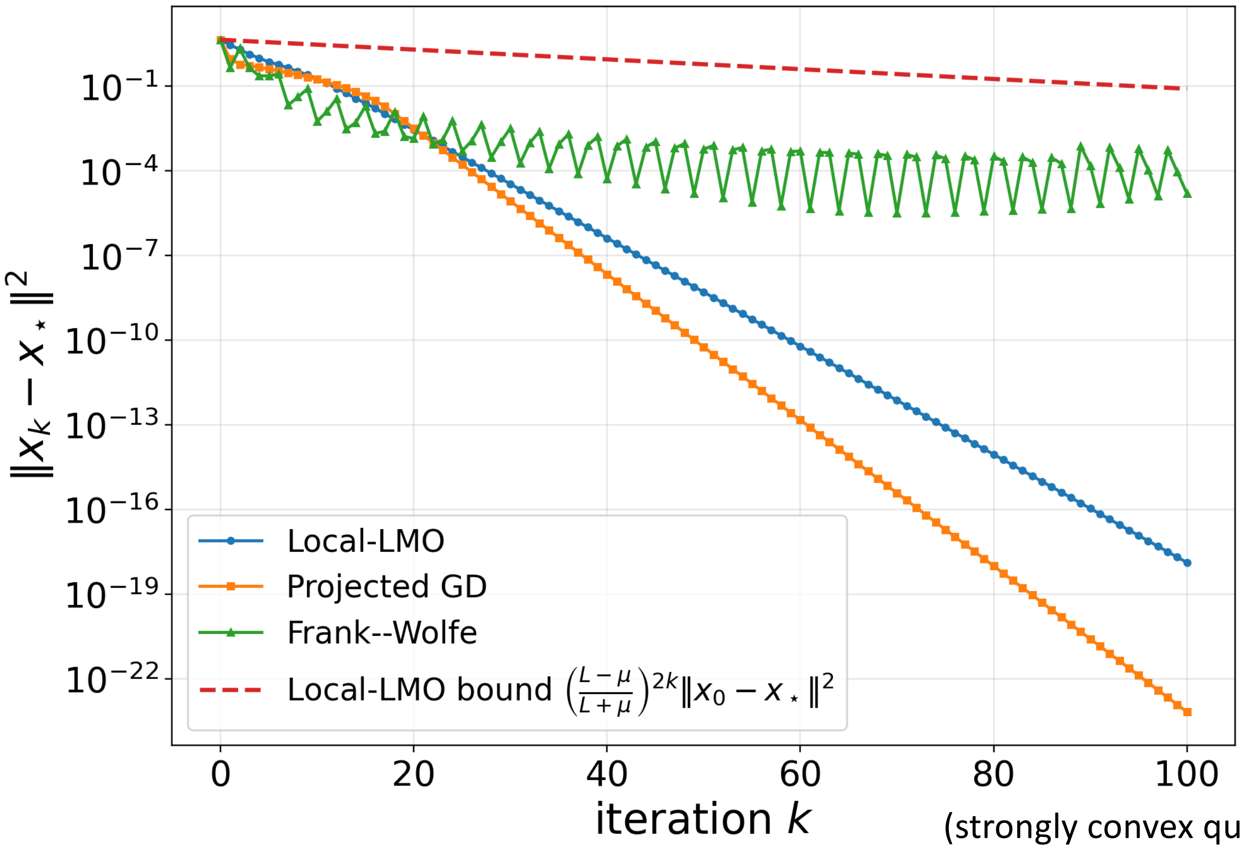
Theory in Other Regimes

Table 1: Comparison of FW and Local LMO under different regimes. Notation: $x_\star \in \mathcal{X}_\star$, $\delta_k := f(x_k) - f(x_\star)$, $R_k := \|x_k - x_\star\|$, $\Delta_k := \|\nabla f(x_k) - \nabla f(x_\star)\|$, $c_k := L_0 + L_1\|\nabla f(x_k)\|$, $c_\star := L_0 + L_1\|\nabla f(x_\star)\|$, $D := \operatorname{diam} \mathcal{X}$, and $G_\gamma(x) := \frac{1}{\gamma}\left(x - \operatorname{Proj}_{\mathcal{X}}\left(x - \gamma \nabla f(x)\right)\right)$.

Property	**Frank–Wolfe**	**Local LMO (new)**
Oracle	LMO for $\mathcal{X}$	LMO for $\mathcal{X} \cap \mathcal{B}(x_k, t_k)$
Projection-free	✓	✓
Works even if $\operatorname{diam} \mathcal{X} = \infty$	✗	✓
Reduces to GD if $\mathcal{X} = \mathbb{R}^d$	✗	✓
Rate independent of D	✗	✓
Rate for G-gradient bounded & convex f	✗[a]	$\min_{0 \le k \le K-1} \delta_k \overset{4.6}{\le} \frac{GR_0}{\sqrt{K}}$ [h] (with radii $t_k = \delta_k / \|\nabla f(x_k)\|$)
Rate for L–smooth & convex[b] f	$\delta_K = \Theta\left(\frac{LD^2}{K}\right)$	$\min_{0 \le k \le K-1} \Delta_k^2 \overset{4.2}{\le} \frac{L^2 R_0^2}{K}$ [h] (with radii $t_k = \Delta_k / L$)
Rate for (L_0, L_1)–smooth & convex f	$\delta_K = \mathcal{O}\left(\frac{(L_0 + L_1 G) D^2}{K}\right)$ [c]	$\min_{0 \le k \le K-1} \Delta_k^2 \overset{G.1}{\le} \frac{4 c_\star^2 R_0^2}{K}$ [h, d] (with radii $t_k = \Delta_k (c_k + c_\star) / 2 c_k c_\star$)
Rate for L–smooth & μ-strongly convex f	$\delta_K = \Theta\left(\frac{LD^2}{K}\right)$	$\|x_K - x_\star\|^2 \overset{4.5}{\le} \left(\frac{L-\mu}{L+\mu}\right)^{2K} R_0^2$ [h] (with radii $t_k = \frac{2\sqrt{\mu L}}{L+\mu} R_k$)
Rate for L–smooth & non-convex f	$\bar{g}_K = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ [e]	$\min_{0 \le k \le K-1} \left\|G_{1/L}(x_k)\right\|^2 \overset{H.2}{\le} \frac{2L\delta_0}{K}$ [h] (with radii $t_k = \left\|G_{1/L}(x_k)\right\| / L$)
Rate for L–smooth & projected-PŁ[f] f	✗[g]	$\delta_K \overset{H.1}{\le} \left(1 - \frac{\mu}{L}\right)^K \delta_0$ (with radii $t_k = \|G_{1/L}(x_k)\| / L$)

Part 4
Plush Toy Experiments

Local LMO vs Projected GD vs Frank-Wolfe



(strongly convex quadratic in 2 dimensions)

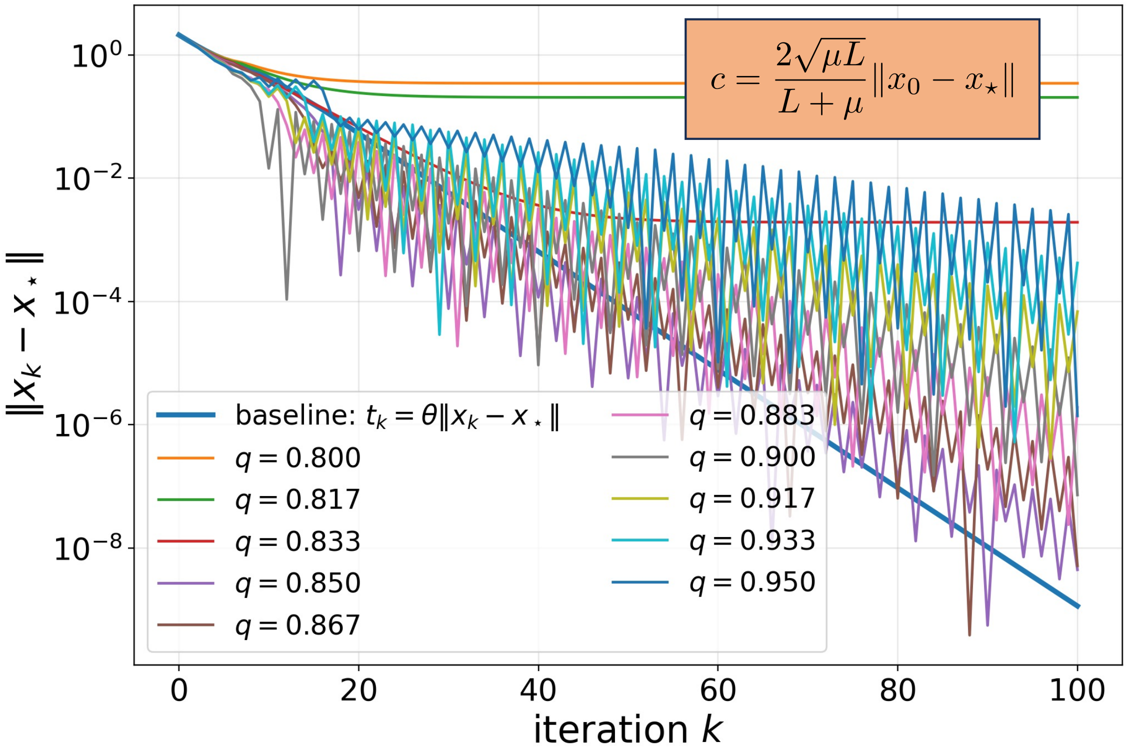
Practical Radius Rule Inspired by Theory

Theoretical radius:

$$t_k = \frac{2\sqrt{\mu L}}{L+\mu}\|x_k - x_\star\|$$

Practical radius uses two tunable parameters:

$$t_k = c \cdot q^k, \qquad q \in (0,1)$$



(strongly convex quadratic in 2 dimensions)

The End