Projection Efficient Subgradient Method and Optimal Nonsmooth Frank-Wolfe Method

Praneeth Netrapalli

Microsoft Research India

Kiran Thekumparampil

UIUC

Prateek Jain

Google India Research

Sewoong Oh

Univ. of Washington

Outline

- Constrained Nonsmooth Convex Problem
- What is currently known
- Our results
- Techniques: Moreau-Yosida smoothing and Gradient sliding

Constrained Nonsmooth Convex Minimization

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

Nonsmooth Convex function

Convex Constraint Set

$$\mathcal{C} \subset \mathbb{R}^d$$

GOAL:

Find an ε-suboptimal *feasible* solution $\tilde{x}$

$$f(\tilde{x}) - \min_{x \in \mathcal{C}} f(x) \leq \varepsilon \quad \text{and} \quad \tilde{x} \in \mathcal{C}$$

- High dimensional: $d \gg \text{poly}(\epsilon^{-1})$
- Prefer methods with convergence rate independent of d

Assumptions

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

- G-Lipschitz continuous function $|f(x) - f(y)| \leq G\|x - y\|$

- Bounded Constraint Set $D = \max_{x,y \in \mathcal{C}} \|x - y\|$

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

Assumptions

- G-Lipschitz continuous function $|f(x) - f(y)| \leq G\|x - y\|$
- Bounded Constraint $\|x - y\|$

For $f(x) = |x|$

$$\partial f(x) = \begin{cases} -1 \text{ if } x < 0 \\ [-1,1] \text{ if } x = 0 \\ 1 \text{ if } x > 0 \end{cases}$$

Black-box Orac

Oracle Complexity
=
Num. of Oracle calls

- First-order Oracle (FO) $\mathrm{FO}(x) \in \partial f(x)$ sub-gradient
- Projection Oracle (PO) $\mathrm{PO}(x) = \mathcal{P}_{\mathcal{C}}(x) = \arg\min_{y \in \mathcal{C}} \|x - y\|^2$

Black-box oracles. Cannot use methods like IPM, Augmented Lagrangian Methods

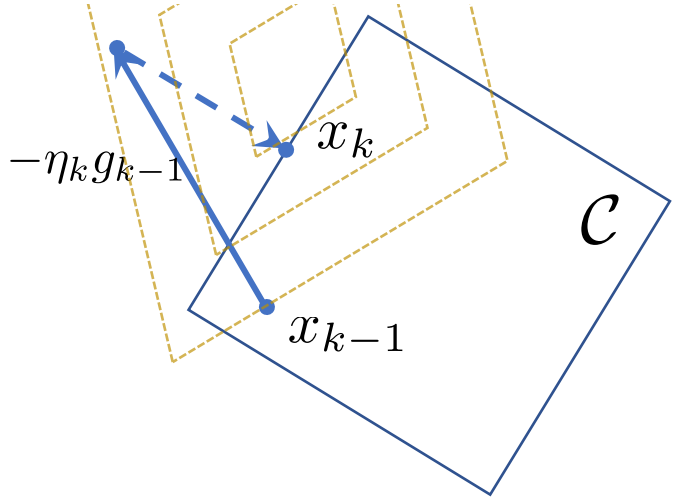
Projected subGradient Descent (PGD)

$x_0 \in \mathcal{C}$

For all $k = 1, 2, \ldots$

$$g_{k-1} = \mathrm{FO}(x_{k-1}) \in \partial f(x_{k-1})$$

$$x_k = \mathcal{P}_{\mathcal{C}}(x_{k-1} - \eta_k g_{k-1})$$



- Simple and practical. Workhorse of machine learning
- Guarantee: $f\left(\frac{1}{k}\sum_{i=1}^{k} x_i\right) - \min_{x \in \mathcal{C}} f(x) \leq \epsilon$ if $k \geq \Theta\left(\frac{G^2 D^2}{\epsilon^2}\right)$
- # FO calls = # PO calls = $\Theta\left(\frac{G^2 D^2}{\epsilon^2}\right)$

Is this the best we can do?

- Yes, finding an ε-suboptimal *feasible* solution requires

$$\text{number FO calls} \geq \Omega\left(\frac{G^2 D^2}{\varepsilon^2}\right)$$ [NemYud83]

- No, we don't know how many projections (PO) are required!

$$\text{number PO calls} \geq \ ?$$

PGD : $x_k = \mathcal{P}_{\mathcal{C}}(x_{k-1} - \eta_k g_{k-1})$

Projection and subgradient descent steps are coupled!

[NemYud83] Nemirovski, and Yudin, "Problem complexity and method efficiency in optimization", Wiley-Interscience 1983

PO calls can be very expensive!

- Example: Nuclear-norm ball constrained low-rank Matrix SVM
- n labeled training samples: sample matrix A_i and label b_i for $i = 1, \ldots, n$

$$\min_{\substack{X \in \mathbb{R}^{m \times m} \\ \|X\|_* \leq \lambda}} \quad \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - b_i \langle X, A_i \rangle)$$

Hinge loss is
Nonsmooth and Convex

$$\text{where} \quad \|X\|_* := \sum_{i=1}^{m} \sigma_i(X) \text{ , where } \sigma_i(X) \text{ is } X\text{'s } i\text{-th singular value}$$

- Projection requires full SVD $O(m^3)$

$$\mathcal{C} = \{X \mid \|X\|_* \leq \lambda\}$$

PO can be expensive.
We want PO call efficiency!

Accelerated PGD for L-smooth objectives

L-smooth function $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$

- First-order Oracle (FO) $\mathrm{FO}(x) = \nabla f(x)$ gradient

$$f(x_k) - \min_{x \in \mathcal{C}} f(x) \le \varepsilon \quad \text{if} \quad k \ge \Theta\left(\sqrt{\frac{L}{\varepsilon}}\right)$$

FO calls = # PO calls = $O\left(\sqrt{\frac{L}{\varepsilon}}\right)$

Even $1/\varepsilon$-smoothness reduces # PO calls to $O\left(\frac{1}{\varepsilon}\right)$

Accelerated Gradient Descent (AGD)

$z_0 = x_0 \in \mathcal{C}$

For $k = 1,2,\ldots$

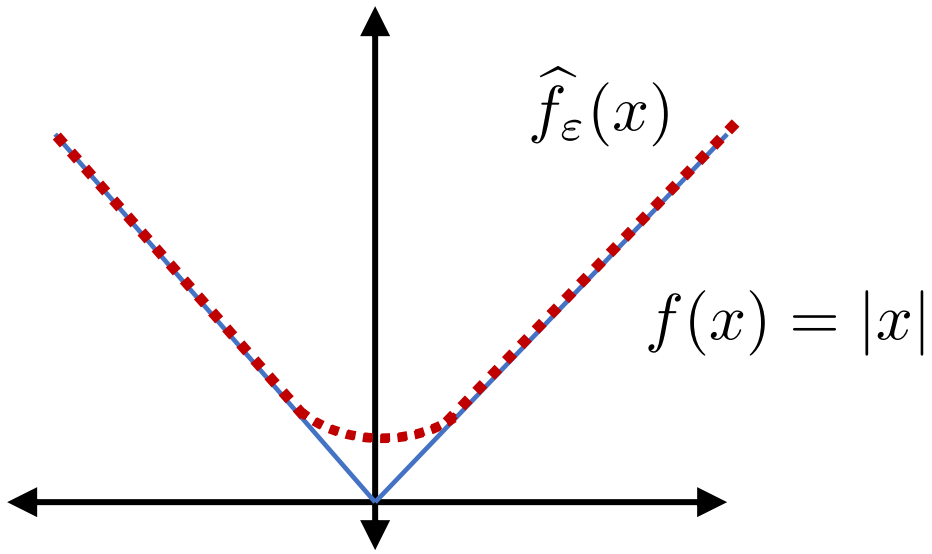
$y_k = (1 - \gamma_k)\, x_{k-1} + \gamma_k z_{k-1}$

$z_k = \mathcal{P}_{\mathcal{C}}\left(z_{k-1} - \mathrm{FO}(y_k)/\beta_k\right)$

$x_k = (1 - \gamma_k)\, x_{k-1} + \gamma_k z_k$

Randomized Smoothing of Lipschitz functions

$$\widehat{f}_\varepsilon(x) := [f * \mathcal{N}(0, \varepsilon^2\, d^{-1}\, \mathbf{I})](x) \;=\; \mathbb{E}_{\xi\sim\mathcal{N}(0,\mathbf{I})}[f(x + \varepsilon\, d^{-\frac{1}{2}}\, \xi)]$$



Close upper-bound

$$f(x) \le \widehat{f}_\varepsilon(x) \le f(x) + O(\varepsilon)$$

Convex

$L = O\left(\frac{\sqrt{d}}{\varepsilon}\right)$-smooth

[2] Duchi, Bartlett, and Wainwright, “Randomized Smoothing for stochastic optimization”, SIAM Opt 2012

AGD with Randomized Smoothing

$$\widehat{f}_\varepsilon(x) = \mathbb{E}[f(x + \varepsilon\, d^{-\frac{1}{2}}\, \xi)]$$

$$\nabla \widehat{f}_\varepsilon(x) \approx \frac{1}{T_\varepsilon} \sum_{i=1}^{T_\varepsilon} \mathrm{FO}(x + \varepsilon\, d^{-\frac{1}{2}}\, \xi_i) \text{ , where } T_\varepsilon = O\left(\frac{d^{-\frac{1}{4}}}{\varepsilon}\right)$$

AGD + Randomized Smoothing

$$y_k = (1-\gamma_k)\, x_{k-1} + \gamma_k z_{k-1}$$

$$\widehat{\nabla}_k = \frac{1}{T_\varepsilon} \sum_{i=1}^{T_\varepsilon} \mathrm{FO}(y_k + \varepsilon\, d^{-\frac{1}{2}} \xi_i)$$

$$z_k = \mathcal{P}_{\mathcal{C}}\left(z_{k-1} - \widehat{\nabla}_k / \beta_k\right)$$

$$x_k = (1-\gamma_k)\, x_{k-1} + \gamma_k z_k$$

$$O\left(\sqrt{\frac{L}{\varepsilon}}\right) = O\left(\frac{d^{1/4}}{\varepsilon}\right) \text{iterations} \qquad L = O\left(\frac{\sqrt{d}}{\varepsilon}\right)$$

$$x\ O\left(\frac{d^{-\frac{1}{4}}}{\varepsilon}\right) \textit{ FO/grad } = O\left(\frac{1}{\varepsilon^2}\right) \text{ FO calls}$$

$$x \quad 1 \qquad \text{PO/proj } = O\left(\frac{d^{1/4}}{\varepsilon}\right) \text{ PO calls}$$

Dimension dependent due to randomization

[2] Duchi, Bartlett, and Wainwright, “Randomized Smoothing for stochastic optimization”, SIAM Opt 2012

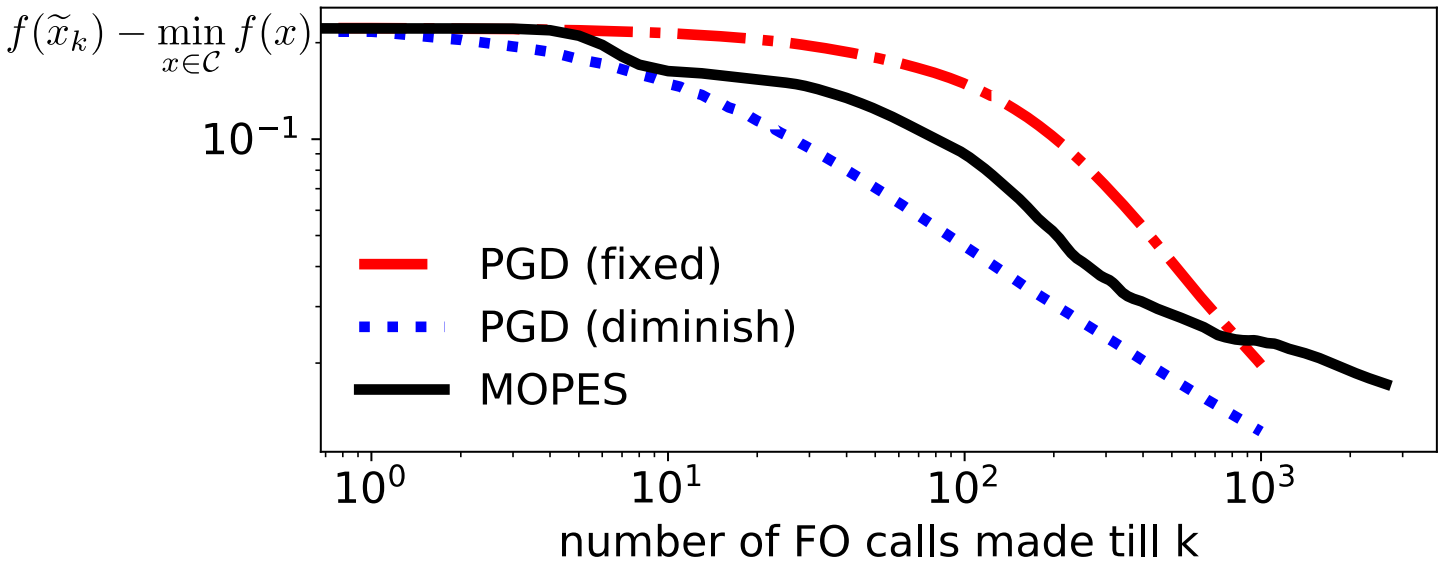
Our result: only $O\left(\frac{1}{\varepsilon}\right)$ projections

Algorithm	Number of FO calls	Number of PO calls
PGD [1]	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$
Randomized Smoothing [2]	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{d^{1/4}}{\varepsilon}\right)$
MOPES (ours)	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon}\right)$
Lower Bound	$\Omega\left(\frac{1}{\varepsilon^2}\right)$	*Open Question!*

[1] Nemirovski, and Yudin, “Problem complexity and method efficiency in optimization”, Wiley-Interscience 1983

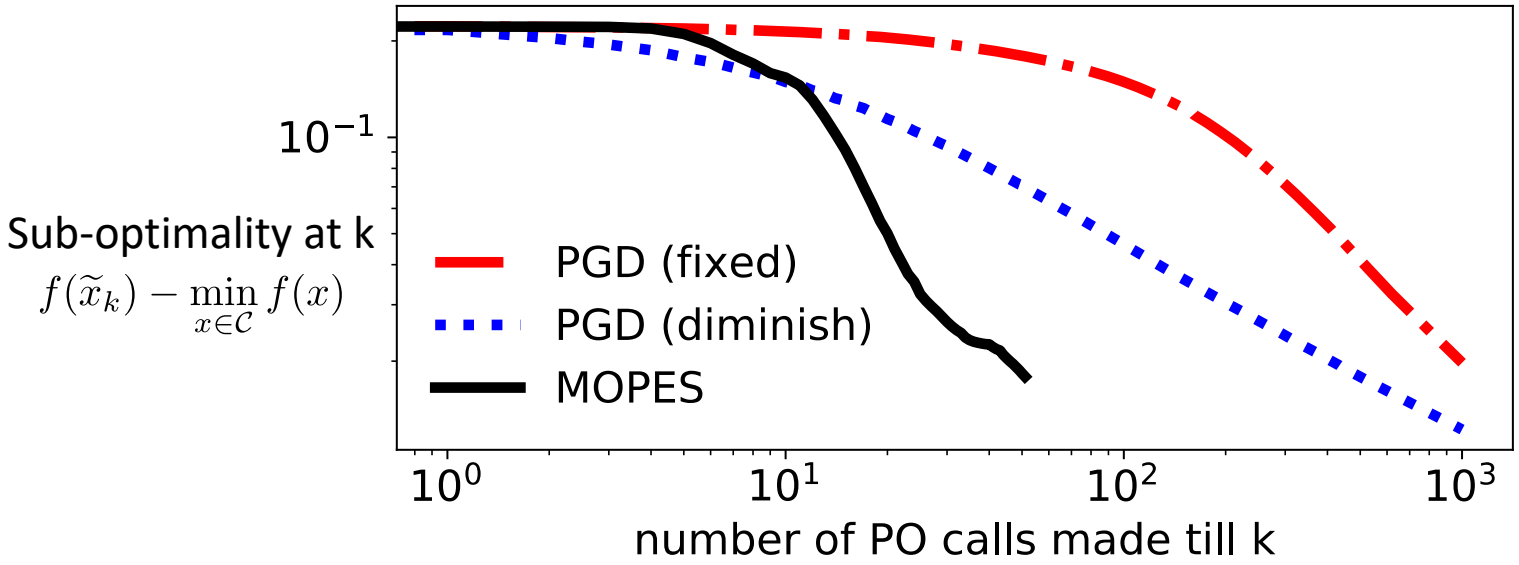
[2] Duchi, Bartlett, and Wainwright, “Randomized Smoothing for stochastic optimization”, SIAM Opt 2012

Sub-optimality at k



MOPES (our) and PGD use similar number of FO calls

MOPES (our) uses fewer PO calls than PGD!



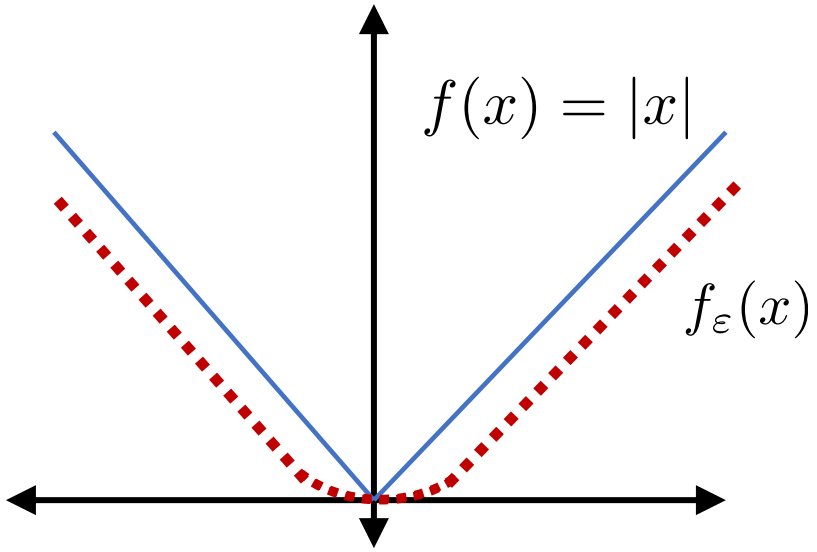
Moreau-Yosida Smoothing of Lipschitz functions

$$f_\varepsilon(x) := \min_{x'} f(x') + \frac{1}{2\varepsilon}\|x' - x\|^2$$

Infimal convolution

Approximate lower-bound

$$f(x) - O(\varepsilon) \leq f_\varepsilon(x) \leq f(x)$$



Convex

$L = O\left(\frac{1}{\varepsilon}\right)$-smooth

AGD + Moreau Smoothing

$$f_\varepsilon(x) := \min_{x'} f(x') + \frac{1}{2\varepsilon}\|x' - x\|^2$$

$$\nabla f_\varepsilon(x) = \frac{x - \mathrm{prox}_f(x)}{\varepsilon}\ , \text{where } \mathrm{prox}_f(x) = \arg\min_{x'} f(x') + \frac{1}{2\varepsilon}\|x' - x\|^2$$

subgradient method

AGD + Moreau Smoothing

$$y_k = (1-\gamma_k)\,x_{k-1} + \gamma_k z_{k-1}$$

$$y_k' \approx \mathrm{prox}_f(y_k)$$

$$z_k = \mathcal{P}_{\mathcal{C}}\left(z_{k-1} - \frac{y_k - y_k'}{\varepsilon\beta_k}\right)$$

$$x_k = (1-\gamma_k)\,x_{k-1} + \gamma_k z_k$$

$$O\left(\sqrt{\frac{L}{\varepsilon}}\right) = O\left(\frac{1}{\varepsilon}\right) \text{ iterations} \qquad L = O\left(\frac{1}{\varepsilon}\right)$$

$$\text{x } O\left(\frac{1}{\varepsilon^2}\right) \quad \textit{FO/prox} = O\left(\frac{1}{\varepsilon^3}\right) \text{ FO calls}$$

$$\text{x } 1 \quad \text{PO/proj} = O\left(\frac{1}{\varepsilon}\right) \text{ PO calls}$$

Are we wasting FO calls?

$$\min_{x \in \mathcal{X}} \left[f_\varepsilon(x) := \underbrace{\left[\min_{x'} f(x') + \frac{1}{2\varepsilon} \|x' - x\|^2 \right]}_{\text{Each inner problem requires } O\left(\frac{1}{\varepsilon^2}\right) \text{FO calls}} \right]$$

$$\min_{x \in \mathcal{X}, x'} \left[\Psi_\varepsilon(x, x') := f(x') + \frac{1}{2\varepsilon} \|x' - x\|^2 \right]$$

Decoupling the objective and the constraint

$$\min_{x \in \mathcal{C}} f(x) \quad \Longleftrightarrow \quad \min_{x \in \mathcal{C}, x'} f(x') \quad \text{subject to } x = x' \quad \overset{\lim_{\lambda \to 0^+}}{\Longleftrightarrow} \quad \min_{x \in \mathcal{C}, x'} \underbrace{f(x') + \overbrace{\frac{1}{2\lambda}\|x - x'\|^2}^{\text{penalty}}}_{=: \Psi_\lambda(x, x')}$$

- Moreau-Yosida regularization theory suggests $\lambda = \varepsilon$

$$f(\widetilde{x}) - \min_{x \in \mathcal{C}} f(x) \leq [\Psi_\varepsilon(\widetilde{x}, \widetilde{x}') - \min_{x \in \mathcal{C}, x'} \Psi_\varepsilon(x)] + O(\varepsilon)$$

- NEW GOAL: Find an ε-suboptimal solution $(\tilde{x}, \tilde{x}')$ to $\min_{x \in \mathcal{C}, x'} \Psi_\varepsilon(x, x')$

Reducing Number of Projections

$$\min_{x \in \mathcal{C}, x' \in \mathbb{R}^d} \quad \Psi_\varepsilon(x, x') \quad = \quad f(x') \quad + \quad \frac{1}{2\varepsilon}\|x - x'\|^2$$

	Composite Objective		**Unconstrained Convex**		**Constrained $\mathbf{L = 2/\varepsilon}$-smooth Convex**
	Gradient Sliding [Lan16]		subGradient Descent (no projection)		Accelerated Projected Gradient Descent (f not in objective)
# FO =	$O\left(\frac{1}{\varepsilon^2}\right)$	=	$O\left(\frac{1}{\varepsilon^2}\right)$	+	0
# PO =	$O\left(\frac{1}{\varepsilon}\right)$	=	0	+	$O\left(\sqrt{\frac{L}{\varepsilon}}\right) = O\left(\frac{1}{\varepsilon}\right)$

[Lan16] Lan, “Gradient sliding method for composite optimization”, SIAM Opt 2016

Projection Efficient Subgradient method (MOPES)

$$\min_{x\in\mathcal{C},x'}[\Psi_\varepsilon(x,x') = f(x') + \underbrace{\frac{1}{2\varepsilon}\|x-x'\|^2}_{(L=1/\varepsilon)\text{-Smooth } \phi_\varepsilon(x,x')}]$$

$$\overline{x} := (x, x')$$

Accelerated Proximal Gradient Descent $O\left(\sqrt{\frac{L}{\varepsilon}}\right) = O\left(\frac{1}{\varepsilon}\right)$ iterations

$\bar{y}_k = (1-\gamma_k)\bar{x}_{k-1} + \gamma_k \bar{z}_{k-1}$

projection: $z_k = \mathcal{P}_{\mathcal{C}}\left(z_{k-1} - \nabla_x \phi_\varepsilon(\bar{y}_k)/\beta_k\right)$

x 1 PO/proj = $O\left(\frac{1}{\varepsilon}\right)$ PO calls

$\hat{z}'_k = (z'_{k-1} - \nabla_{x'}\phi_\lambda(\bar{y}_k)/\beta_k)$

Inexact proximal *[Lan16]*: $z'_k \approx \text{prox}_{f/\beta_k}(\hat{z}'_k) := \arg\min_{z'} f(z') + \frac{\beta_k}{2}\|z' - \hat{z}'_k\|^2$

x $O\left(\frac{1}{\varepsilon}\right)$ *FO/prox* = $O\left(\frac{1}{\varepsilon^2}\right)$ FO calls

$\bar{x}_k = (1-\gamma_k)\bar{x}_{k-1} + \gamma_k \bar{z}_k$

[Lan16] Lan, “Gradient sliding method for composite optimization”, SIAM Opt 2016

Frank-Wolfe (FW) method for Smooth Convex

- Alternative to PO is Linear Minimization Oracle (LMO) $\mathrm{LMO}(g) \in \arg\min_{x \in \mathcal{C}} \langle g, x \rangle$
- LMO is often cheaper than PO, e.g. nuclear-norm ball

$x_0 \in \mathcal{C}$

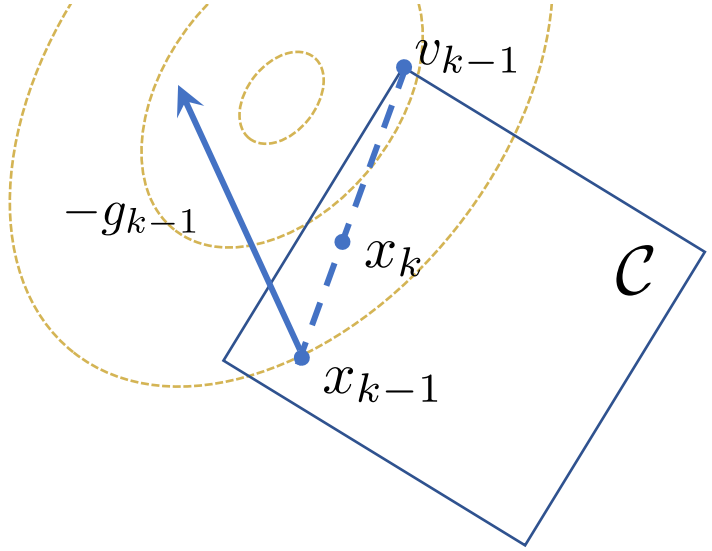
For all $k = 1, 2, \ldots$

$g_{k-1} = \mathrm{FO}(x_{k-1}) \in \partial f(x_{k-1})$

$v_{k-1} = \mathrm{LMO}(g_{k-1}) \in \arg\min_{x \in \mathcal{C}} \langle g_{k-1}, x \rangle$

$x_k = (1 - \alpha_k) \cdot x_{k-1} + \alpha_k \cdot v_{k-1}$

$O\left(\frac{L}{\varepsilon}\right)$ iterations for L-smooth function



FAILS for Nonsmooth objectives!

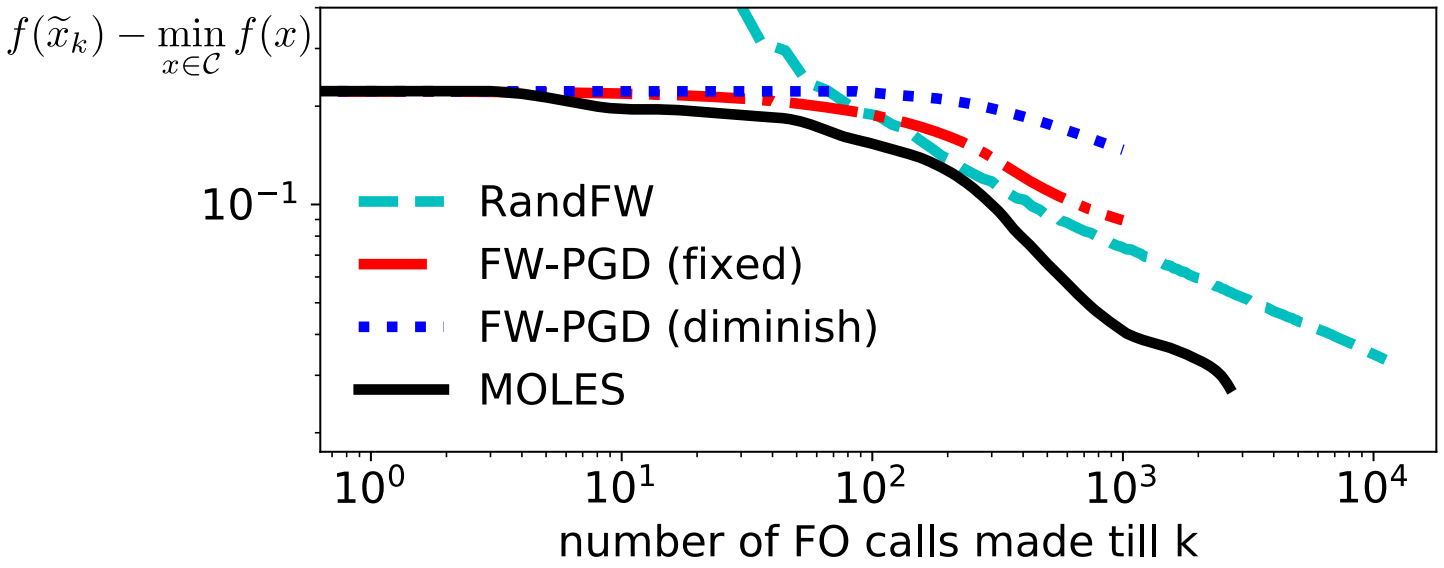
Optimal Nonsmooth FW method uses $O\left(\frac{1}{\varepsilon^2}\right)$ LMO and FO calls

Algorithm	Number of FO calls	Number of LMO calls
FW-PGD [folklore]	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^4}\right)$
Randomized Smoothing [3]	$O\left(\frac{\sqrt{d}}{\varepsilon^4}\right)$	$O\left(\frac{\sqrt{d}}{\varepsilon^2}\right)$
MOLES (ours)	$O\left(\frac{1}{\varepsilon^2}\right)$	$O\left(\frac{1}{\varepsilon^2}\right)$
Lower Bound [3]	$\Omega\left(\frac{1}{\varepsilon^2}\right)$	$\Omega\left(\frac{1}{\varepsilon^2}\right)$

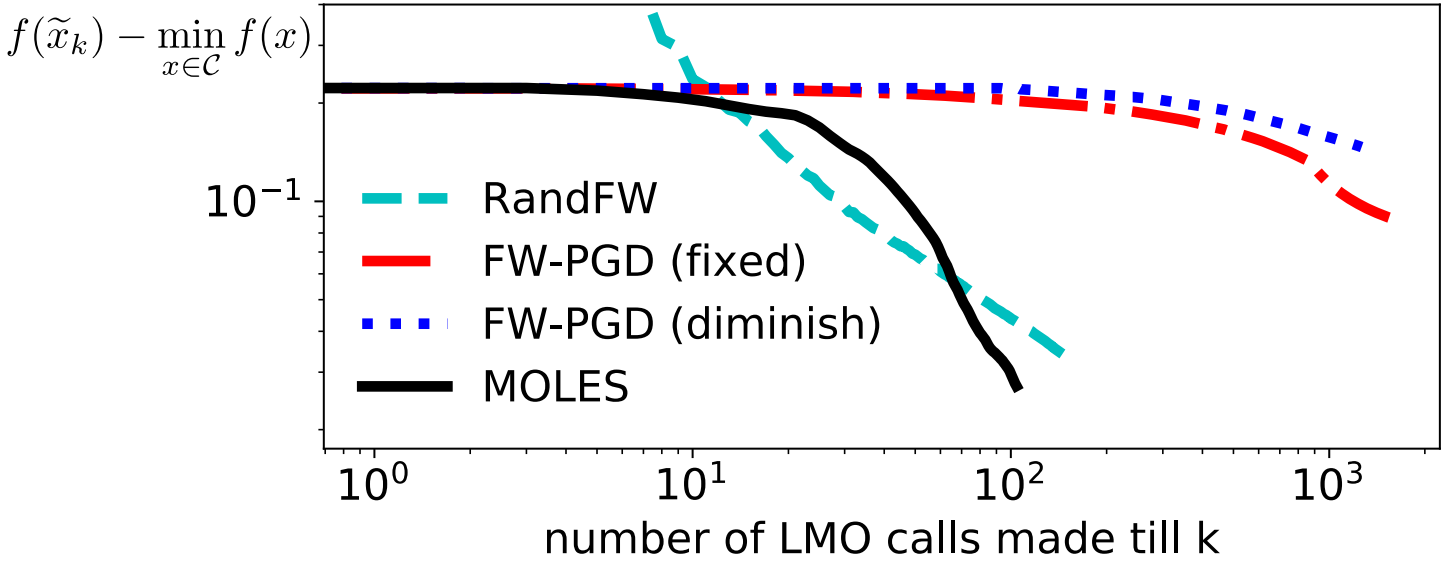
MOLES = MOPES + Inexact Projections using FW method

[3] Lan, "The Complexity of Large-scale Convex Programming under a Linear Optimization Oracle", arXiv:1309.5550

Sub-optimality at k



Sub-optimality at k



MOLES (our) uses the least number of FO and LMO calls.

Optimal Nonsmooth Frank-Wolfe method (MOLES)

$$\min_{x\in\mathcal{C},x'}[\Psi_\varepsilon(x,x') = f(x') + \frac{1}{2\varepsilon}\|x-x'\|^2]$$

$\overline{x} := (x,x')$ $\quad$ $(L = 1/\varepsilon)$-Smooth $\phi_\varepsilon(x,x')$

Projection is a $\frac{1}{\varepsilon}$-smooth problem

$$\mathcal{P}_\mathcal{X}(x) = \arg\min_{x_\mathcal{P}\in\mathcal{X}} \frac{L}{2}\|x_\mathcal{P} - x\|^2$$

Inexact MOPES

Inexact projection *[LZ16]*

$$z_k \cong \mathcal{P}_\mathcal{C}\left(z_{k-1} - \nabla_x \phi_\varepsilon(\overline{y}_k)/\beta_k\right)$$

Inexact proximal

$$z'_k \cong \mathrm{prox}_{f/\beta_k}\left(\hat{z}'_k\right)$$

$$O\left(\sqrt{\frac{L}{\varepsilon}}\right) = O\left(\frac{1}{\varepsilon}\right) \text{ iterations}$$

x $O\left(\frac{1}{\varepsilon}\right)$ *LMO/proj* = $O\left(\frac{1}{\varepsilon^2}\right)$ LMO calls

x $O\left(\frac{1}{\varepsilon}\right)$ *FO/prox* = $O\left(\frac{1}{\varepsilon^2}\right)$ FO calls

[Lan16] Lan, Zhou, "Conditional Gradient Sliding method", SIAM Opt 2016

Summary

- Constrained convex minimization method using $O(1/\varepsilon)$ projections
- Optimal nonsmooth FW method using $O(1/\varepsilon^2)$ FO and LMO calls
- We obtain similar complexities for Stochastic FO

Open questions

- Lower bound for PO calls complexity
- Extension to non-Euclidean geometry

Thank you!