Conditional Gradient Methods via Stochastic Path-Integrated Differential Estimator

Alp Yurtsever

alp.yurtsever@epfl.ch

joint work with

Suvrit Sra & Volkan Cevher

MIT EPFL

ICML2019 - Long Beach

EPFL

Massachusetts Institute of Technology (MIT)

Ecole Polytechnique Fédérale de Lausanne (EPFL)

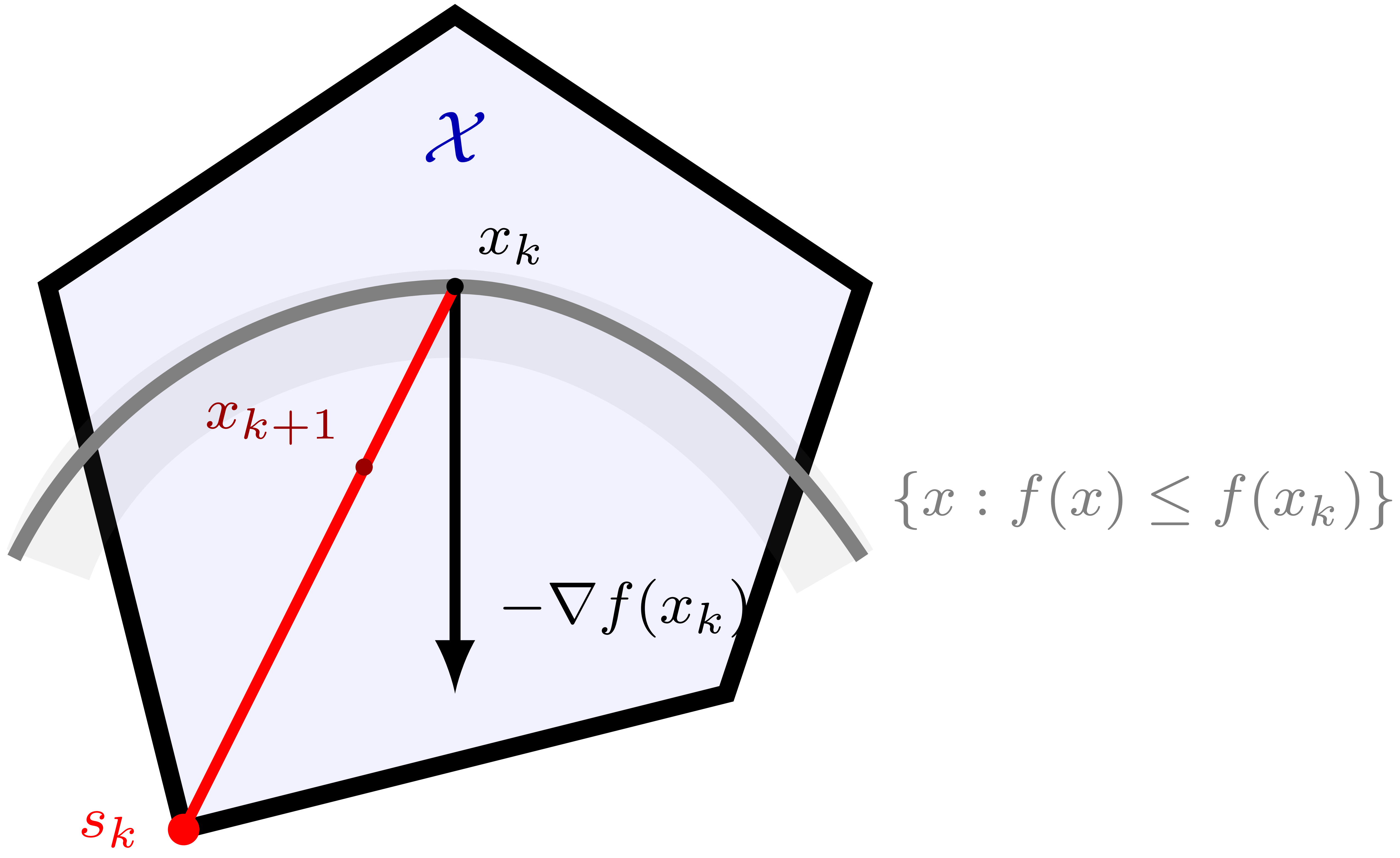
Conditional Gradient Method (CGM)

(Frank & Wolfe, 1956)
(Hazan, 2008)
(Jaggi, 2013)

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

- ▷ $\mathcal{X} \subset \mathbb{R}^d$ is a convex compact set
- ▷ $f : \mathcal{X} \to \mathbb{R}$ is a smooth function

Input: $x_1 \in \mathcal{X}$
for $k = 1, 2, \ldots,$ **do**
  $\eta_k = 2/(k+1)$
  $s_k = \arg\min_{x \in \mathcal{X}} \langle \nabla f(x_k), x \rangle$
  $x_{k+1} = x_k + \eta_k (s_k - x_k)$
end for



Stochastic Templates

$$\underset{x \in \mathcal{X}}{\text{minimize}} \quad F(x) \qquad\qquad F(x) := \begin{cases} \mathbb{E}_\xi f(x, \xi) & \text{(expectation)} \\ \frac{1}{n} \sum_{i=1}^{n} f_i(x) & \text{(finite-sum)} \end{cases}$$

- ▷ $\mathcal{X} \subset \mathbb{R}^d$ is a convex compact set
- ▷ f and f_i are differentiable and possibly non-convex
- ▷ $\xi \sim \mathcal{P}$ is a random variable

Stochastic Templates

$$\underset{x \in \mathcal{X}}{\text{minimize}} \quad F(x)$$

$$F(x) := \begin{cases} \mathbb{E}_\xi f(x, \xi) & \text{(expectation)} \\ \frac{1}{n} \sum_{i=1}^n f_i(x) & \text{(finite-sum)} \end{cases}$$

- ▷ $\mathcal{X} \subset \mathbb{R}^d$ is a convex compact set
- ▷ f and f_i are differentiable and possibly non-convex
- ▷ $\xi \sim \mathcal{P}$ is a random variable

Assumptions

$$\mathbb{E}\nabla f(x, \xi) = \nabla F(x)$$ unbiased estimates

$$\mathbb{E}\|\nabla f(x, \xi) - \nabla F(x)\|^2 \leq \sigma^2 < +\infty, \qquad \forall x \in \mathcal{X}$$ bounded variance

$$\mathbb{E}\|\nabla f(x, \xi) - \nabla f(y, \xi)\|^2 \leq \text{Ł}\|x - y\|^2, \qquad \forall (x, y) \in \mathcal{X}^2$$ averaged smoothness

Stochastic Templates

$$\underset{x \in \mathcal{X}}{\text{minimize}} \quad F(x) \qquad F(x) := \begin{cases} \mathbb{E}_\xi f(x, \xi) & \text{(expectation)} \\ \frac{1}{n} \sum_{i=1}^n f_i(x) & \text{(finite-sum)} \end{cases}$$

- ▷ $\mathcal{X} \subset \mathbb{R}^d$ is a convex compact set
- ▷ f and f_i are differentiable and possibly non-convex
- ▷ $\xi \sim \mathcal{P}$ is a random variable

Assumptions

$$\mathbb{E} \nabla f(x, \xi) = \nabla F(x)$$ unbiased estimates

$$\mathbb{E} \|\nabla f(x, \xi) - \nabla F(x)\|^2 \leq \sigma^2 < +\infty, \qquad \forall x \in \mathcal{X}$$ bounded variance

$$\mathbb{E} \|\nabla f(x, \xi) - \nabla f(y, \xi)\|^2 \leq \text{Ł} \|x - y\|^2, \qquad \forall (x, y) \in \mathcal{X}^2$$ averaged smoothness

we study the theoretical complexity of
stochastic and finite-sum Frank-Wolfe variants

Oracle Models

- Stochastic first-order oracle (*sfo*)

  for stochastic function $\mathbb{E}_{\xi} f(x,\xi)$ with $\xi \sim \mathcal{P}$

  (*sfo*) returns $(f(x,\xi'), \nabla f(x,\xi'))$ where ξ' is an iid sample from $\mathcal{P}$

- Incremental first-order oracle (*ifo*)

  for finite-sum, (*ifo*) draws an index i from $\{1, 2, \ldots, n\}$ uniformly random

  and returns $(f_i(x), \nabla f_i(x))$

- Linear minimization oracle (*lmo*)

  given a gradient estimate $v \in \mathbb{R}^d$

  (*lmo*) returns $s \in \mathbb{R}^d$ such that $s \in \operatorname{argmin}_{x \in \mathcal{X}} \langle v, x \rangle$

State of the Art

Deterministic variants

✔ Frank-Wolfe Algorithm (FW)

(Frank & Wolfe, 1956) (Jaggi, 2013)

$\mathcal{O}(\epsilon^{-1})$ (*lmo*) and gradient complexity

in the convex setting

(Lacoste-Julien, 2016)

$\mathcal{O}(\epsilon^{-2})$ ………

in the non-convex setting

State of the Art

Deterministic variants

✔ Frank-Wolfe Algorithm (FW)

(Frank & Wolfe, 1956) (Jaggi, 2013)

$\mathcal{O}(\epsilon^{-1})$ (*lmo*) and gradient complexity

in the convex setting

(Lacoste-Julien, 2016)

$\mathcal{O}(\epsilon^{-2})$ ………

in the non-convex setting

✔ Conditional Gradient Sliding (CGS)

(Lan & Zhou, 2016)

use accelerated gradient method

approximately solve projection step using FW

$\mathcal{O}(\epsilon^{-1})$ (*lmo*)

$\mathcal{O}(\epsilon^{-1/2})$ (*gradient*)

in the convex setting

we provide new results

in the non-convex setting

State of the Art

Stochastic variants

- ✔ Online FW (Hazan & Kale, 2012)
- ✔ Stochastic FW (Hazan & Luo, 2016) (Reddi et al., 2016)
- ✔ Stochastic FW with constant batch size (Mokhtari et al., 2018)
- ✔ Stochastic CGS (Lan & Zhou, 2016)

Variance reduced based on SVRG (Johnson & Zhang, 2013):

- ✔ SVRF / SVFW (Hazan & Luo, 2016) (Reddi et al., 2016)
- ✔ STORC (Hazan & Luo, 2016)

CGM with SPIDER

SPIDER: Stochastic Path-Integrated Differential Estimator (Fang et al., 2018)

$$v^k = \nabla_{\mathcal{S}_k}(x^k) - \nabla_{\mathcal{S}_k}(x^{k-1}) + v^{k-1}$$

Lemma (Variance bound):

$$\mathbb{E}\|\nabla F(x^k) - v^k\|^2 \leq \frac{L^2}{S_k}\|x^k - x^{k-1}\|^2 + \|\nabla F(x^{k-1}) - v^{k-1}\|^2$$
$$\leq \frac{(LD\eta_k)^2}{S_k} + \|\nabla F(x^{k-1}) - v^{k-1}\|^2$$

we introduce **SPIDER-FW**

best known rates in the non-convex setting

$\mathcal{O}(\epsilon^{-3})$ *(sfo)*	$\mathcal{O}(\sqrt{n}\epsilon^{-2})$ *(ifo)*
$\mathcal{O}(\epsilon^{-2})$ *(lmo)*	$\mathcal{O}(\epsilon^{-2})$ *(lmo)*
(expectation)	(finite-sum)

Comparison

Poster today: **Pacific Ballroom #85**

	convex				non-convex			
	finite-sum		expectation		finite-sum		expectation	
	(*ifo*)	(*lmo*)	(*sfo*)	(*lmo*)	(*ifo*)	(*lmo*)	(*sfo*)	(*lmo*)
FW	$\mathcal{O}(n\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-1})$	-	-	$\mathcal{O}(n\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	-	-
CGS	$\mathcal{O}(n\epsilon^{-1/2})$	$\mathcal{O}(\epsilon^{-1})$	-	-	$\mathcal{O}(n\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	-	-
SFW	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$
SFW-1	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	-	-	-	-
Online-FW	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$	-	-	-	-
SCGS	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$
SVRF / SVFW	$\mathcal{O}(n\ln(\epsilon^{-1})+\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	-	-	$\mathcal{O}(n+n^{2/3}\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-10/3})$	$\mathcal{O}(\epsilon^{-2})$
STORC†	$\mathcal{O}(n\ln(\epsilon^{-1})+\epsilon^{-3/2})$	$\mathcal{O}(\epsilon^{-1})$	-	-	-	-	-	-
SPIDER-FW	$\mathcal{O}(n\ln(\epsilon^{-1})+\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(n^{1/2}\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$
SPIDER-CGS	$\mathcal{O}(n\ln(\epsilon^{-1})+\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(n^{1/2}\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$

Table 1: Comparison of conditional gradient methods for stochastic optimization. Contribution of *this work* is highlighted with blue font. See Section 6 for more details.

FW (Frank & Wolfe, 1956; Jaggi, 2013) , CGS (Lan & Zhou, 2016) , SFW (Hazan & Luo, 2016; Reddi et al., 2016) , SFW-1 (Mokhtari et al., 2018) , Online-FW (Hazan & Kale, 2012) , SCGS (Lan & Zhou, 2016) , SVRF / SVFW (Hazan & Luo, 2016; Reddi et al., 2016) , STORC (Hazan & Luo, 2016)