Locally Accelerated Conditional Gradients

Alejandro Carderera

Joint work with J. Diakonikolas and S. Pokutta
Georgia Institute of Technology

alejandro.carderera@gatech.edu

July 29th, 2019

H. Milton Stewart School of
Industrial and Engineering Systems

Goal is smooth convex optimization.

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

Goal is smooth convex optimization.

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

Main ingredients:

First-order (FO) oracle. Given $x \in \mathcal{X}$ and a differentiable convex function $f : \mathbb{R}^n \to \mathbb{R}$, return:

$$\nabla f(x) \in \mathbb{R}^n \text{ and } f(x) \in \mathbb{R}$$

Linear optimization (LO) oracle. Given $v \in \mathbb{R}^n$, return:

$$\operatorname*{argmin}_{x \in \mathcal{X}} \langle v, x \rangle$$

Focus of our work is on the *Conditional Gradients* algorithm (CG) [1], also known as the *Frank-Wolfe* algorithm (FW) [2].

Algorithm 1 Conditional Gradients algorithm.

Input: $x_0 \in \mathcal{X}$, stepsizes $\gamma_1 \cdots \gamma_t \in [0, 1]$.

1: **for** $t = 0$ to T **do**
2: $\quad v_t = \operatorname{argmin}_{x \in \mathcal{X}} \langle \nabla f(x_t), x \rangle$
3: $\quad x_{t+1} = x_t + \gamma_t (v_t - x_t)$
4: **end for**

Advantages of CG.

Advantages of CG.

First-order. Dimensionality of modern problems makes computing second-order information infeasible.

Advantages of CG.

First-order. Dimensionality of modern problems makes computing second-order information infeasible.

Projection-free. Projection into certain feasible regions is computationally expensive: Birkhoff polytope and flow polytope are a few examples.

Advantages of CG.

First-order. Dimensionality of modern problems makes computing second-order information infeasible.

Projection-free. Projection into certain feasible regions is computationally expensive: Birkhoff polytope and flow polytope are a few examples.

Sparse solutions. Solution is a convex combination of (a typically sparse set of) extreme points.

Disadvantages of CG.

Disadvantages of CG.

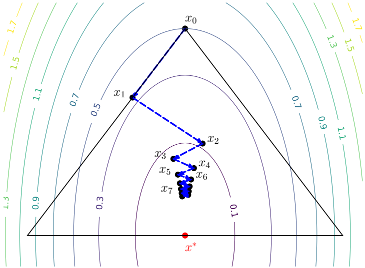
Sublinear convergence. For L-smooth and μ-strongly convex f when x^* is in a face of $\mathcal{X}$.

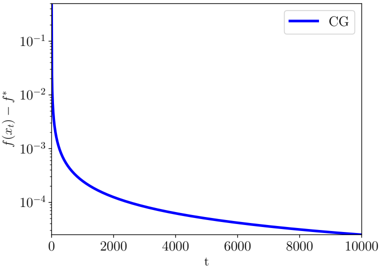
Disadvantages of CG.

Sublinear convergence. For L-smooth and μ-strongly convex f when x^* is in a face of $\mathcal{X}$.

Example (CG Convergence.)

L-smooth and μ-strongly convex f with $x \in \mathbb{R}^2$, and x^* in boundary of $\mathcal{X}$.





Linear convergence is achieved by allowing steps that decrease the weight of *bad* vertices [3]. This has led to various CG variants:

Linear convergence is achieved by allowing steps that decrease the weight of *bad* vertices [3]. This has led to various CG variants:

Away-step Conditional Gradients (AFW)

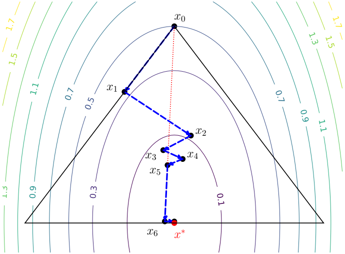


Figure: Away-step CG (AFW)

Allow steps in the direction of:

$$x - \underset{y \in \mathcal{S}}{\operatorname{argmax}} \langle \nabla f(x), y \rangle ,$$

where $\mathcal{S}$ is the active set of x.

Pairwise CG

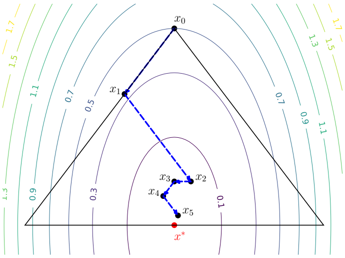


Figure: PFW

Fully-Corrective CG

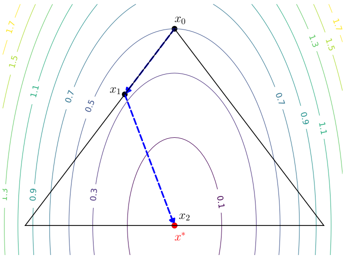


Figure: FCFW

Convergence rate for L-smooth μ-strongly convex f.

Theorem (Convergence rate of AFW, PFW and FCFW.)

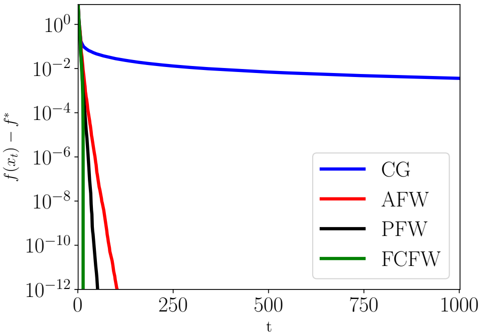
[4] Suppose that f is L-smooth μ-strongly convex over a polytope $\mathcal{X}$, the number of steps T required to reach an ϵ-optimal solution to the minimization problem verifies,

$$T = \mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right),$$

where D and δ are the diameter and pyramidal width of polytope $\mathcal{X}$

Example (CG Variant Convergence.)

L-smooth and μ-strongly convex f ($L/\mu \approx 10^8$) over the probability simplex in $\mathbb{R}^{100}$, and x^* a convex combination of 13 vertices.



CG Global Acceleration.

However, we know that optimal methods for this class of functions achieve an ϵ solution in $T = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$ first-order calls [5, 6].

Can CG achieve these convergence rates **globally**?

CG Global Acceleration.

However, we know that optimal methods for this class of functions achieve an ϵ solution in $T = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$ first-order calls [5, 6].

Can CG achieve these convergence rates **globally**?

Example ([7, 8] $f(x) = \|x\|^2$ over unit simplex in $\mathbb{R}^n$.)

We know the optimal solution is given by $x^* = \mathbb{1}/n$. CG can incorporate at most one vertex in each iteration, if we start from a vertex x_0, in iteration $t < n$ we have that:

$$f(x_t) - f(x^*) \geq \frac{1}{t} - \frac{1}{n}.$$

Considering iterations such that $t \leq \lfloor n/2 \rfloor$ and rearranging into a linear convergence contraction we have:

$$T = \Omega\left(\frac{1}{r}\log\frac{1}{\epsilon}\right),$$

where $r \leq 2\frac{\log 2t}{2t}$.

Considering iterations such that $t \leq \lfloor n/2 \rfloor$ and rearranging into a linear convergence contraction we have:

$$T = \Omega\left(\frac{1}{r}\log\frac{1}{\epsilon}\right),$$

where $r \leq 2\frac{\log 2t}{2t}$.

Convergence rate of the CG variants for this problem instance: $r = \frac{1}{4t}$.

At best a global logarithmic improvement in the convergence rate, therefore **global acceleration in Nesterov's sense is not possible**.

Conditional Gradient Sliding

Idea: Run Nesterov's Accelerated Gradient Descent, use CG to solve the projection subproblems approximately [9].

Conditional Gradient Sliding

Idea: Run Nesterov's Accelerated Gradient Descent, use CG to solve the projection subproblems approximately [9].

Results:

- Separate LO and FO oracle calls.
- Globally optimal $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$ calls to FO and $\mathcal{O}\left(\frac{LD^2}{\epsilon}+\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$ calls to LO oracles.
- Convergence rates independent of the dimension n.

Catalyst Augmented AFW.

Idea: Run Accelerated Proximal Method and solve proximal problems with a linearly convergent CG [10].

Catalyst Augmented AFW.

Idea: Run Accelerated Proximal Method and solve proximal problems with a linearly convergent CG [10].

Results:

- $\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}}\left(\frac{D}{\delta}\right)^2 \log \frac{1}{\epsilon}\right)$ Calls to FO and LO oracles.
- Convergence rates dependent of the dimension n.

Summary

Complexity for L-smooth μ-strongly convex f.

Algorithm	LO Calls	FO Calls
CG Variants	$\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$
CGS	$\mathcal{O}\left(\frac{LD^2}{\epsilon} + \sqrt{\frac{L}{\mu}} \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log\frac{1}{\epsilon}\right)$
Catalyst	$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$

Summary

Complexity for L-smooth μ-strongly convex f.

Algorithm	LO Calls	FO Calls
CG Variants	$\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$
CGS	$\mathcal{O}\left(\frac{LD^2}{\epsilon} + \sqrt{\frac{L}{\mu}} \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log\frac{1}{\epsilon}\right)$
Catalyst	$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$
What we want:	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log\frac{1}{\epsilon}\right)$

Objectives:

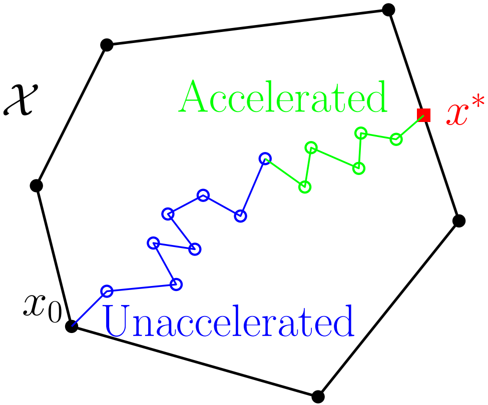
- Dimension independent global acceleration.

Objectives:

- ~~Dimension independent global acceleration.~~
- Dimension independent local acceleration.

Locally Accelerated Conditional Gradients (LaCG).

What do we mean by **local acceleration**?



After a constant number of iterations, accelerate the convergence.

Locally Accelerated Conditional Gradients (LaCG).

The key ingredients is a *Modified μAGD* algorithm [11].

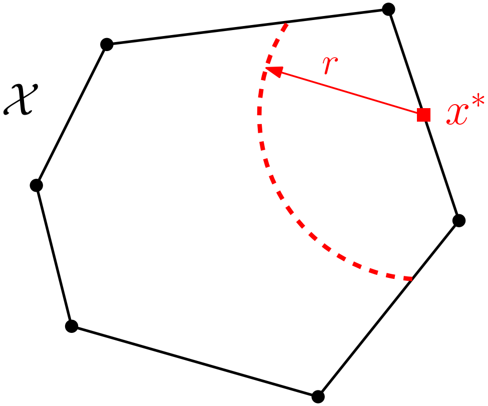
Theorem (Convergence rate of μAGD.)

Let f be L-smooth and μ-strongly convex and let $\{\mathcal{C}_i\}_{i=0}^t$ be a sequence of convex subsets of $\mathcal{X}$ such that $\mathcal{C}_i \subseteq \mathcal{C}_{i-1}$ for all i and $x^* \in \cap_{i=0}^t \mathcal{C}_i$, then the μAGD achieves an ϵ-optimal solution in:

$$T = \mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$$

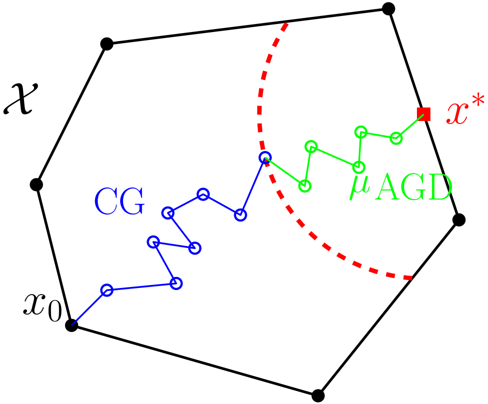
How do we build $\{\mathcal{C}_i\}_{i=0}^t$ in an efficient way?

[12] CG: $\exists r > 0$ (that depends only on f and $\mathcal{X}$) s.t. if $\|x^* - x_K\| \leq r \Rightarrow x^* \in conv\left(\mathcal{S}_t\right)$ for all $t \geq K$, where $\mathcal{S}_t$ is the active set at iteration t.

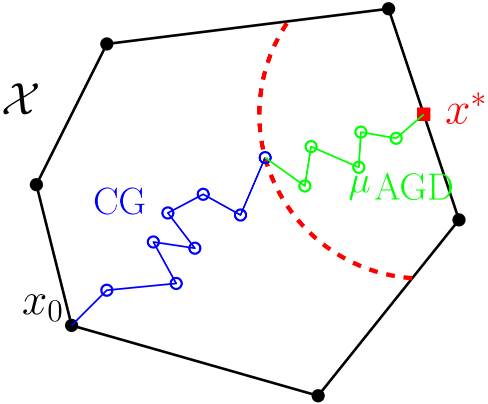


So when we are inside the red semicircle and we use $\mathcal{C}_t = \mathcal{S}_t$, acceleration is possible.

Naively, what we would like:



Naively, what we would like:



But since the value of r is not known, we don't know when to switch from CG to μAGD.

Main ideas of LaCG:

Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \operatorname{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.

Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.
- Every H iterations restart: use $\mathcal{S}_t$ to update C_t if a vertex was added to $\mathcal{S}_t$ since the last update.

Main ideas of LaCG:

- At each iteration perform a CG variant step and a μAGD step over $\mathcal{C}_{t+1}$ and select $x_{t+1} = \text{argmin}\{x_{t+1}^{CG}, x_{t+1}^{\mu AGD}\}$.
- Every H iterations restart: use $\mathcal{S}_t$ to update C_t if a vertex was added to $\mathcal{S}_t$ since the last update.
- After a constant **burn-in phase**, acceleration will be achieved.

Convergence rate of LaCG.

Theorem (Convergence rate of LaCG.)

Let f be L-smooth and μ-strongly convex and let r be the critical radius, for:

$$t = min\left\{\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right), K + \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)\right\}$$

and $K = \frac{8L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\left(\frac{2(f(x_0)-f^*)}{\mu r^2}\right)$, *then* $f(x_t) - f(x^*) \leq \epsilon$

Convergence rate of LaCG.

Theorem (Convergence rate of LaCG.)

Let f be L-smooth and μ-strongly convex and let r be the critical radius, for:

$$t = min\left\{\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right), K + \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)\right\}$$

and $K = \frac{8L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\left(\frac{2(f(x_0)-f^*)}{\mu r^2}\right)$, *then* $f(x_t) - f(x^*) \leq \epsilon$

In fact, we often observe faster convergence even for $\|x_t - x^*\| \geq r$

Recap

If $\|x_T - x^*\| \geq r$

Algorithm	LO Calls	FO Calls
CG Variants	$\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$
CGS	$\mathcal{O}\left(\frac{LD^2}{\epsilon} + \sqrt{\frac{L}{\mu}} \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log\frac{1}{\epsilon}\right)$
Catalyst	$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$
LaCG	$\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$

Table: Complexity for L-smooth μ-strongly convex f.

Recap

If $\|x_T - x^*\| \leq r$

Algorithm	LO Calls	FO Calls
CG Variants	$\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{L}{\mu}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$
CGS	$\mathcal{O}\left(\frac{LD^2}{\epsilon} + \sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$
Catalyst	$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L-\mu}{\mu}}\left(\frac{D}{\delta}\right)^2 \log\frac{1}{\epsilon}\right)$
LaCG	$K + \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$K + \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$

Table: Complexity for L-smooth μ-strongly convex f.

K is independent of ϵ, so asymptotically optimal.

Computational Results.

Despite the faster convergence rate after the burn-in phase, how does LaCG perform with respect to other projection-free algorithms?

Simplex in $\mathbb{R}^{2000}$ with $L/\mu = 1000$.

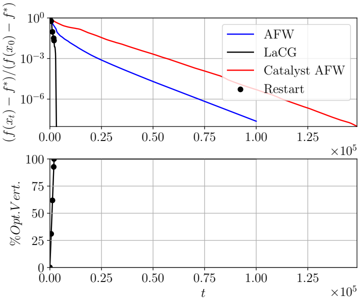


Figure: Primal gap vs. iteration

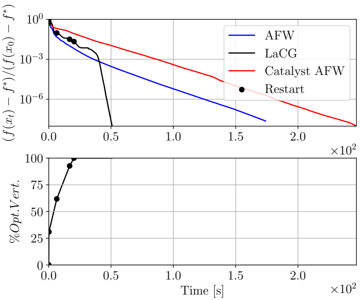


Figure: Primal gap vs. time

When close enough to $x*$ (after burn-in phase), there is a significant speedup in the convergence rate.

ℓ_1 unit ball in $\mathbb{R}^{2000}$ with $L/\mu = 100$.

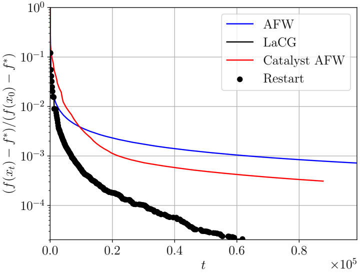


Figure: Primal gap vs. iteration

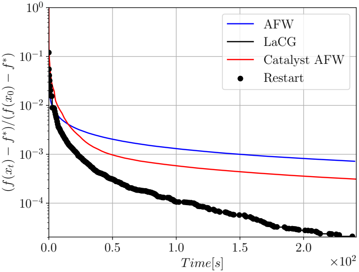


Figure: Primal gap vs. time

Birkhoff polytope in $\mathbb{R}^{40x40}$ with $L/\mu = 100$.

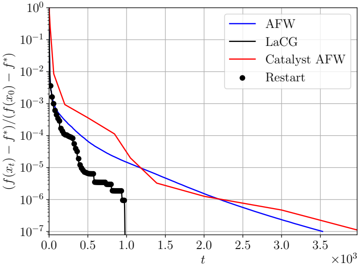


Figure: Primal gap vs. iteration

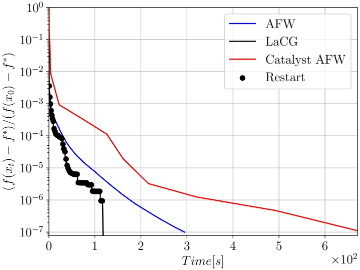


Figure: Primal gap vs. time

Video co-localization problem over flow polytope [13].

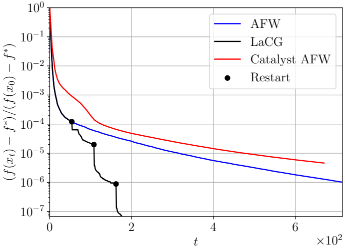


Figure: Primal gap vs. iteration

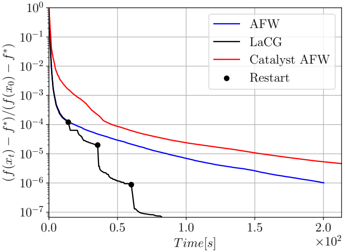


Figure: Primal gap vs. time

Thank you
for your attention.

References I

[1] Boris Teodorovich Polyak. "Minimization methods in the presence of constraints". In: *Itogi Nauki i Tekhniki. Seriya" Matematicheskii Analiz"* 12 (1974), pp. 147–197.

[2] Marguerite Frank and Philip Wolfe. "An algorithm for quadratic programming". In: *Naval research logistics quarterly* 3.1-2 (1956), pp. 95–110.

[3] Dan Garber and Elad Hazan. "Faster rates for the frank-wolfe method over strongly-convex sets". In: *32nd International Conference on Machine Learning, ICML 2015*. 2015.

[4] Simon Lacoste-Julien and Martin Jaggi. "On the Global Linear Convergence of Frank-Wolfe Optimization Variants". In: *Advances in Neural Information Processing Systems 28*. 2015, pp. 496–504.

[5] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. "Problem complexity and method efficiency in optimization". In: *Wiley-Interscience Series in Discrete Mathematics* 15 (1983).

References II

[6] Y Nesterov. “A method of solving a convex programming problem with convergence rate $O(\frac{1}{k^2})$”. In: *Soviet Math. Dokl.* Vol. 27. 1983.

[7] Guanghui Lan. “The complexity of large-scale convex programming under a linear optimization oracle”. In: *arXiv preprint arXiv:1309.5550* (2013).

[8] Martin Jaggi. “Revisiting Frank-Wolfe: Projection-free sparse convex optimization.”. In: *ICML (1)*. 2013, pp. 427–435.

[9] Guanghui Lan and Yi Zhou. “Conditional gradient sliding for convex optimization”. In: *SIAM Journal on Optimization* 26.2 (2016), pp. 1379–1409.

[10] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. “A universal catalyst for first-order optimization”. In: *Advances in neural information processing systems*. 2015, pp. 3384–3392.

References III

[11] Alejandro C., Jelena Diakonikolas, and Sebastian Pokutta. "Locally Accelerated Conditional Gradients". In: *arXiv preprint arXiv:1906.07867* (2019).

[12] Jacques Guélat and Patrice Marcotte. "Some comments on Wolfe's 'away step'". In: *Mathematical Programming* 35.1 (1986), pp. 110–119.

[13] Armand Joulin, Kevin Tang, and Li Fei-Fei. "Efficient image and video co-localization with frank-wolfe algorithm". In: *European Conference on Computer Vision*. Springer. 2014, pp. 253–268.