Approximate Frank-Wolfe Algorithms over Graph-structured Support Sets

Baojian Zhou [1], Yifan Sun [2]

July 19, 2022

[1]Fudan University

[2]Stony Brook University

Graph-structured convex optimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}), \text{ subject to } \boldsymbol{x} \in \mathcal{D}(C, \mathbb{M}), \text{ where}$$

Graph-structured convex optimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}), \text{ subject to } \boldsymbol{x} \in \mathcal{D}(C, \mathbb{M}), \text{ where}$$

Graph-structured set:

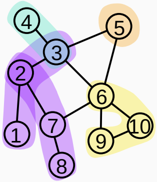
$\mathcal{D}(C, \mathbb{M}) \triangleq \text{conv}\{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq C, \text{supp}(\boldsymbol{x}) \in \mathbb{M}\}$ is a convex hull of interesting support sets described by $\mathbb{M}$, which contains a collection of allowed structures of the problem.

Graph-structured convex optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}),\ \text{subject to } \boldsymbol{x}\in\mathcal{D}(C,\mathbb{M}),\ \text{where}$$

Graph-structured set:

$\mathcal{D}(C,\mathbb{M}) \triangleq \operatorname{conv}\{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq C, \operatorname{supp}(\boldsymbol{x}) \in \mathbb{M}\}$ is a convex hull of interesting support sets described by $\mathbb{M}$, which contains a collection of allowed structures of the problem.

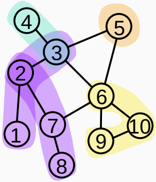


Example $\mathbb{M}$: It is defined on a 10-node graph where each colored region is a subgraph. Elements of $\mathbb{M}$ are these colored region subgraphs.

Graph-structured convex optimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}), \text{ subject to } \boldsymbol{x} \in \mathcal{D}(C, \mathbb{M}), \text{ where}$$

Graph-structured set:
$\mathcal{D}(C, \mathbb{M}) \triangleq \text{conv}\{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq C, \text{supp}(\boldsymbol{x}) \in \mathbb{M}\}$ is a convex hull of interesting support sets described by $\mathbb{M}$, which contains a collection of allowed structures of the problem.



Example $\mathbb{M}$: It is defined on a 10-node graph where each colored region is a subgraph. Elements of $\mathbb{M}$ are these colored region subgraphs.

Applications: Graph-structured compressive sensing, Gene-pathway finding, etc...

Frank-Wolfe (FW) algorithm

FW-type algorithm: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t(\rho \boldsymbol{v}_t - \boldsymbol{x}_t)$, where $\boldsymbol{v}_t$ is the minimizer of *graph-structured LMO*

LMO: $\boldsymbol{v}_t \in \underset{\boldsymbol{v} \in \mathcal{D}(C, \mathbb{M})}{\operatorname{argmin}} \left\langle a\boldsymbol{x}_t + b\nabla f(\boldsymbol{x}_t), \boldsymbol{v} \right\rangle .$

Setting $a = 0, b = 1, \eta_t = 2/(t+2)$ with $\rho = 1$ recovers the standard FW.

Frank-Wolfe (FW) algorithm

FW-type algorithm: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t(\rho \boldsymbol{v}_t - \boldsymbol{x}_t)$, where $\boldsymbol{v}_t$ is the minimizer of *graph-structured LMO*

LMO: $\boldsymbol{v}_t \in \underset{\boldsymbol{v} \in \mathcal{D}(C, \mathbb{M})}{\operatorname{argmin}} \langle a\boldsymbol{x}_t + b\nabla f(\boldsymbol{x}_t), \boldsymbol{v} \rangle$.

Setting $a = 0, b = 1, \eta_t = 2/(t+2)$ with $\rho = 1$ recovers the standard FW.

Fact: LMO is NP-hard to solve when $\mathcal{D}(C, \mathbb{M})$ is complex.

Approximate LMOs and adversarial examples

Additive approximate LMO[Dunn and Harshbarger, 1978] finds $\bar{\boldsymbol{v}}_t$:

$\langle \nabla f(\boldsymbol{x}_t), \bar{\boldsymbol{v}}_t \rangle \leq \min_{\boldsymbol{v} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} \rangle + \mathcal{O}(\frac{\epsilon}{t})$

Approximate LMOs and adversarial examples

Additive approximate LMO[Dunn and Harshbarger, 1978] finds $\bar{\boldsymbol{v}}_t$:
$\langle \nabla f(\boldsymbol{x}_t), \bar{\boldsymbol{v}}_t \rangle \leq \min_{\boldsymbol{v} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} \rangle + \mathcal{O}(\frac{\epsilon}{t})$

Multiplicative approximate LMO[Locatello et al., 2017] returns
$\bar{\boldsymbol{v}}_t$: $\underbrace{\langle \nabla f(\boldsymbol{x}_t), \bar{\boldsymbol{v}}_t - \boldsymbol{x}_t \rangle}_{\bar{g}_t(\boldsymbol{x}_t)} \leq \delta \cdot \underbrace{\min_{\boldsymbol{v} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} - \boldsymbol{x}_t \rangle}_{g_t(\boldsymbol{x}_t)}$, where $\delta \in (0, 1]$.

Approximate LMOs and adversarial examples

Additive approximate LMO[Dunn and Harshbarger, 1978] finds $\bar{\boldsymbol{v}}_t$:
$\langle \nabla f(\boldsymbol{x}_t), \bar{\boldsymbol{v}}_t \rangle \leq \min_{\boldsymbol{v} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} \rangle + \mathcal{O}(\frac{\epsilon}{t})$

Multiplicative approximate LMO[Locatello et al., 2017] returns $\bar{\boldsymbol{v}}_t$: $\underbrace{\langle \nabla f(\boldsymbol{x}_t), \bar{\boldsymbol{v}}_t - \boldsymbol{x}_t \rangle}_{\bar{g}_t(\boldsymbol{x}_t)} \leq \delta \cdot \underbrace{\min_{\boldsymbol{v} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} - \boldsymbol{x}_t \rangle}_{g_t(\boldsymbol{x}_t)}$, where $\delta \in (0, 1]$.

Adversarial examples setup: We show that these two popular approximate LMOs are generally impractical to obtain. Suppose $\tau \in (0, 1/2)$ and $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{x} - \boldsymbol{x}^\top \boldsymbol{b}$, $\boldsymbol{b} = [1, 1, 1, 1, \tau, \cdots, \tau]^\top$. Take $C = 1$ and then the solution is $\boldsymbol{x}^* = [\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, \ldots, 0]^\top$ with $\nabla f(\boldsymbol{x}^*) = -[\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \tau, \cdots, \tau]^\top$. When $\boldsymbol{x}_t = \boldsymbol{x}^*$, $\boldsymbol{v}_t = \boldsymbol{x}^*$, and the duality gap $g_t(\boldsymbol{x}^*) = 0$.

Gap-additive bound cannot decay properly: Any suboptimal $\bar{\boldsymbol{v}}_t$ gives $g_t(\boldsymbol{x}^*) - \bar{g}_t(\boldsymbol{x}^*) = 1 - \sqrt{\frac{3}{4} + \tau^2} > 0$, which is strictly positive and constant in t, hence violating additive approximate LMO.

Dual approximation oracle

Inner Product Operator (IPO): Given $\boldsymbol{z} \in \mathbb{R}^d$, $\mathcal{D} \subseteq \mathbb{R}^d$, and approximation factor $\delta \in (0, 1]$, it returns $\boldsymbol{v}$ such that

$$(\delta, \boldsymbol{z}, \mathcal{D})\text{-IPO}: \qquad \langle \boldsymbol{z}, \boldsymbol{v} \rangle \leq \delta \cdot \min_{\boldsymbol{s} \in \mathcal{D}} \langle \boldsymbol{z}, \boldsymbol{s} \rangle .$$

Dual approximation oracle

Inner Product Operator (IPO): Given $\boldsymbol{z} \in \mathbb{R}^d$, $\mathcal{D} \subseteq \mathbb{R}^d$, and approximation factor $\delta \in (0, 1]$, it returns $\boldsymbol{v}$ such that

$$(\delta, \boldsymbol{z}, \mathcal{D})\text{-IPO}: \qquad \langle \boldsymbol{z}, \boldsymbol{v} \rangle \leq \delta \cdot \min_{\boldsymbol{s} \in \mathcal{D}} \langle \boldsymbol{z}, \boldsymbol{s} \rangle .$$

Dual Maximization Oracle(DMO) finds an $S \in \mathbb{M}$ such that

$$(\delta, \boldsymbol{z}, \mathcal{D})\text{-DMO}: \qquad \|\boldsymbol{z}_S\|_2 \geq \delta \cdot \max_{S' \in \mathbb{M}} \|\boldsymbol{z}_{S'}\|_2 .$$

Dual approximation oracle

Inner Product Operator (IPO): Given $\boldsymbol{z} \in \mathbb{R}^d$, $\mathcal{D} \subseteq \mathbb{R}^d$, and approximation factor $\delta \in (0, 1]$, it returns $\boldsymbol{v}$ such that

$$(\delta, \boldsymbol{z}, \mathcal{D})\text{-IPO}: \qquad \langle \boldsymbol{z}, \boldsymbol{v} \rangle \leq \delta \cdot \min_{\boldsymbol{s} \in \mathcal{D}} \langle \boldsymbol{z}, \boldsymbol{s} \rangle .$$

Dual Maximization Oracle(DMO) finds an $S \in \mathbb{M}$ such that

$$(\delta, \boldsymbol{z}, \mathcal{D})\text{-DMO}: \qquad \|\boldsymbol{z}_S\|_2 \geq \delta \cdot \max_{S' \in \mathbb{M}} \|\boldsymbol{z}_{S'}\|_2 .$$

Theorem (IPO $\Leftrightarrow$ DMO)

Given the set $\mathcal{D}$ and suppose $S \in (\delta, \boldsymbol{z}, \mathcal{D})$-DMO. Define the approximate supporting vector $\tilde{\boldsymbol{v}}_t \triangleq -C \cdot \boldsymbol{z}_S / \|\boldsymbol{z}_S\|_2$. Then, $\tilde{\boldsymbol{v}}_t \in (\delta, \boldsymbol{z}, \mathcal{D})$-IPO.

Dual approximation oracle

Inner Product Operator (IPO): Given $\boldsymbol{z} \in \mathbb{R}^d$, $\mathcal{D} \subseteq \mathbb{R}^d$, and approximation factor $\delta \in (0, 1]$, it returns $\boldsymbol{v}$ such that

$$(\delta, \boldsymbol{z}, \mathcal{D})\text{-IPO}: \qquad \langle \boldsymbol{z}, \boldsymbol{v} \rangle \leq \delta \cdot \min_{\boldsymbol{s} \in \mathcal{D}} \langle \boldsymbol{z}, \boldsymbol{s} \rangle \,.$$

Dual Maximization Oracle(DMO) finds an $S \in \mathbb{M}$ such that

$$(\delta, \boldsymbol{z}, \mathcal{D})\text{-DMO}: \qquad \|\boldsymbol{z}_S\|_2 \geq \delta \cdot \max_{S' \in \mathbb{M}} \|\boldsymbol{z}_{S'}\|_2.$$

Theorem (IPO ⇔ DMO)

Given the set $\mathcal{D}$ and suppose $S \in (\delta, \boldsymbol{z}, \mathcal{D})$-DMO. Define the approximate supporting vector $\tilde{\boldsymbol{v}}_t \triangleq -C \cdot \boldsymbol{z}_S / \|\boldsymbol{z}_S\|_2$. Then, $\tilde{\boldsymbol{v}}_t \in (\delta, \boldsymbol{z}, \mathcal{D})$-IPO.

DMO is easy to obtain: The method of **Top-g + neighbor visiting** provides $\delta = \sqrt{1/\lceil s/g \rceil}$.

Approximated DMO-FW

DMO-FW-I:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t \left(\frac{C \cdot (\boldsymbol{z}_t)_{S_t}}{\|(\boldsymbol{z}_t)_{S_t}\|_2} - \boldsymbol{x}_t \right), S_t = (\delta, \nabla f(\boldsymbol{x}_t), \mathcal{D})\text{-DMO}$$

DMO-FW-II$(\boldsymbol{x}_t \in \mathcal{D}/\delta)$: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t \left(\frac{C \cdot (\boldsymbol{z}_t)_{S_t}}{\delta \cdot \|(\boldsymbol{z}_t)_{S_t}\|_2} - \boldsymbol{x}_t \right).$

Theorem (Convergence Rate of DMO-FW)

Let $s = \max_{S \in \mathbb{M}} |S|$ *be the maximal allowed sparsity. Suppose* $\nu \in (0, 1]$ *and* $\|\nabla f(\boldsymbol{x})\|_\infty \leq B$, DMO-FW-I *admits*

$$h(\boldsymbol{x}_t) \triangleq f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \begin{cases} \mathcal{O}\left(\frac{BC\sqrt{s}}{t^\nu}\right), & \|\nabla f(\boldsymbol{x}_t)\|_\infty \leq \frac{B}{t^\nu} \\ \mathcal{O}\left(\frac{BC\sqrt{s}(1-\delta)}{\delta}\right), & \textit{Otherwise.} \end{cases}$$

Furthermore,

$$\text{DMO-FW-II}(\boldsymbol{x}_t \in \mathcal{D}/\delta) : h(\boldsymbol{x}_t) \leq \frac{8LC^2}{\delta^2(t+2)}.$$

Acceleration: DMO-AccFW

DMO-AccFW (Inspired from Garber and Wolf [2021]):

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t \left(\frac{C \cdot (\boldsymbol{z}_t)_{S_t}}{\left\| (\boldsymbol{z}_t)_{S_t} \right\|_2} - \boldsymbol{x}_t \right), \boldsymbol{z}_t = \left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L \eta_t} \right).$$

Acceleration: DMO-AccFW

DMO-AccFW (Inspired from Garber and Wolf [2021]):

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t \left(\frac{C \cdot (\boldsymbol{z}_t)_{S_t}}{\left\| (\boldsymbol{z}_t)_{S_t} \right\|_2} - \boldsymbol{x}_t \right), \boldsymbol{z}_t = \left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L \eta_t} \right).$$

Quadratic Growth Condition: Exist a constant $\mu > 0$ such that

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\mu}{2} \|\boldsymbol{x} - [\boldsymbol{x}]_{\mathcal{X}^*}\|_2^2, \forall \boldsymbol{x} \in \mathcal{D}, [\boldsymbol{x}]_{\mathcal{X}^*} \triangleq \operatorname*{argmin}_{\boldsymbol{z} \in \mathcal{X}^*} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2.$$

Acceleration: DMO-AccFW

DMO-AccFW (Inspired from Garber and Wolf [2021]):

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t \left(\frac{C \cdot (\boldsymbol{z}_t)_{S_t}}{\left\| (\boldsymbol{z}_t)_{S_t} \right\|_2} - \boldsymbol{x}_t \right), \boldsymbol{z}_t = \left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right).$$

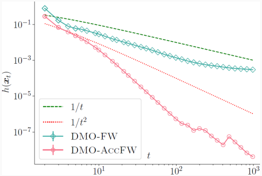
Quadratic Growth Condition: Exist a constant $\mu > 0$ such that

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\mu}{2} \|\boldsymbol{x} - [\boldsymbol{x}]_{\mathcal{X}^*}\|_2^2, \forall \boldsymbol{x} \in \mathcal{D}, [\boldsymbol{x}]_{\mathcal{X}^*} \triangleq \underset{\boldsymbol{z} \in \mathcal{X}^*}{\operatorname{argmin}} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2.$$

Theorem (Main result)

Assume further that $\boldsymbol{x}^*$ is on the boundary and $\delta = 1$, for all $t \geq 1$, the primal error of DMO-AccFW satisfies

$$h(\boldsymbol{x}_t) \leq \frac{4e^{4L/\mu} h(\boldsymbol{x}_0)}{(t+2)^2}.$$



Experimental evaluation

Graph-structured linear sensing: Recover a graph-sparse model $\tilde{\boldsymbol{x}}^*$ using measurements generated as

$$\boldsymbol{y} = \langle \boldsymbol{A}, \tilde{\boldsymbol{x}}^* \rangle + \boldsymbol{e}, \qquad \boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d).$$

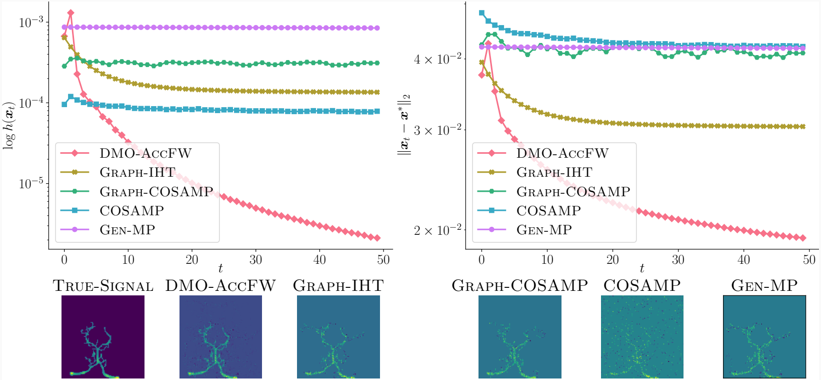
$\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is Gaussian sensing matrix with $a_{ij} \sim \mathcal{N}(0, 1/\sqrt{n})$.

Experimental evaluation

Graph-structured linear sensing: Recover a graph-sparse model $\tilde{\boldsymbol{x}}^*$ using measurements generated as

$$\boldsymbol{y} = \langle \boldsymbol{A}, \tilde{\boldsymbol{x}}^* \rangle + \boldsymbol{e}, \qquad \boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d).$$

$\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is Gaussian sensing matrix with $a_{ij} \sim \mathcal{N}(0, 1/\sqrt{n})$.



Contributions

- We demonstrate that there exist adversarial examples such that two popular inexact LMOs are impossible to obtain and propose two types of approximate FW methods.

Contributions

- We demonstrate that there exist adversarial examples such that two popular inexact LMOs are impossible to obtain and propose two types of approximate FW methods.
- We prove that the DMO is equivalent to the IPO. The standard FW admits $\mathcal{O}((1-\delta)\sqrt{s}/\delta)$ in worst case and our accelerated version has rate $\mathcal{O}(1/t^2)$.

Contributions

- We demonstrate that there exist adversarial examples such that two popular inexact LMOs are impossible to obtain and propose two types of approximate FW methods.
- We prove that the DMO is equivalent to the IPO. The standard FW admits $\mathcal{O}((1-\delta)\sqrt{s}/\delta)$ in worst case and our accelerated version has rate $\mathcal{O}(1/t^2)$.
- Initial experiments indicate that inexact FW-type methods are attractive for this type of problems and one is encouraged to find faster methods based on fully-corrective or other methods.

Code and datasets: https://github.com/baojian/dmo-fw