Optimization for Data Science

Zeroth Order Optimization for Black box Adversarial Attacks

Giacomo Virginio, Mikhail Kolobov, Anna Putina

Department of Mathematics, Università degli Studi di Padova

September 15, 2023

Abstract

In this project, our primary objective is to analyze three gradient-free modifications (*SGFFW* [3], *FZCGS* [1] and *ZO-SCGS* [2]) of the original Frank-Wolfe algorithm. We aim to gain a deep understanding of the theory behind these algorithms and to evaluate their performance in a practical scenario.

The aforementioned algorithms are specifically designed for constrained stochastic non-convex optimization problems. They focus on enhancing the iteration complexity, which depends on the number of oracle queries, in comparison to existing algorithms. Furthermore, the algorithms aspire to be competitive with their first-order counterparts.

Following a theoretical summary of these methods, we conduct practical tests, subjecting the algorithms to a black-box attacks scenario reported in Section 4.3 [1].

1 Introduction

In the examined articles, the minimization constrained optimization problem takes one of the following forms:

1. *Stochastic*

$$\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x}) = \min_{x\in\mathcal{C}} \mathbb{E}_{\mathbf{y}\sim\mathcal{P}}[F(\mathbf{x};\mathbf{y})], \tag{1}$$

where $\mathcal{C} \in \mathbb{R}^d$ is a closed convex set [3], [2];

2. *Finite-sum*

$$\min_{\mathbf{x}\in\mathcal{C}a} F(\mathbf{x}) = \min_{\mathbf{x}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{2}$$

where $\mathcal{C} \subset \mathbb{R}^d$ denotes a closed convex feasible set [1].

One of the potential solutions to the problems Eq. (1), (2) is the utilization of projection-free methods, such as the Frank-Wolfe algorithm. Furthermore, in the papers, they emphasize a stochastic variant of this method that relies on a zeroth-order oracle (function queries). Derivative-free optimization finds its motivation in scenarios where the analytical form of the function is either unavailable or where evaluating the gradient is computationally prohibitive. Hence, the application of such algorithms is driven by tangible practical benefits. In the articles, innovative and more refined modifications are introduced, which demand fewer oracle queries to converge to a solution. It's worth noting that the initial assumptions about the problem vary slightly. For instance, in *SGFFW* [3] and *FZCGS* [1], they tackle non-convex smooth functions, whereas in *ZO-SCGS* [2], their focus is on convex but non-smooth functions.

1.1 Frank-Wolfe Algorithm

The Frank-Wolfe algorithm (first-order) is a versatile optimization method employed in solving constrained optimization problems. It is especially suitable for scenarios where the constraint set is defined by a large number of linear constraints or where projection onto the constraint set is computationally expensive.

The core idea of the Frank-Wolfe algorithm revolves around iteratively updating the solution by performing a linear approximation of the objective function. The algorithm then proceeds by moving towards a direction that minimizes this approximation while ensuring that the solution remains within the constraints. This direction is determined by solving a linear optimization subproblem. The algorithm converges to the optimal solution by iteratively refining the approximation and adjusting the current solution.

1. Computation of the gradient of the objective function at the current solution:

$$\nabla f(x_k)$$

2. Solving a linear optimization subproblem to find a feasible direction d_k that minimizes the linear approximation of the objective function:

$$d_k = \arg\min_{d\in\mathcal{C}} \langle \nabla f(x_k), d\rangle$$

where $\mathcal{C}$ represents the feasible set or constraint set.

3. Updating the current solution x_k using a step size γ_k:

$$x_{k+1} = x_k + \gamma_k \cdot (d_k - x_k)$$

The Frank-Wolfe algorithm is a powerful optimization technique for large-scale constrained problems and an efficient choice for finding solutions while maintaining sparsity and handling complex constraints.

1.2 Zeroth Order Optimization

The fundamental idea behind zeroth-order optimization is to efficiently explore the function space with minimal reliance on assumptions about the function's mathematical properties. This is achieved through a combination of sampling, interpolation, and search strategies that guide the optimization process.

When the gradient of a function is not available, we can utilize the difference of the function value with respect to two random points to estimate it. One well-known method for such estimation, among many others, is the coordinate-wise gradient estimator.

$$\hat{\nabla} f(\mathbf{x}) = \sum_{j=1}^{d} \frac{f(\mathbf{x} + \mu_j \mathbf{e}_j) - f(\mathbf{x} - \mu_j \mathbf{e}_j)}{2\mu_j} \mathbf{e}_j, \quad (3)$$

where $\mu_j > 0$ is the smoothing parameter, and $\mathbf{e}_j \in \mathbb{R}^d$ denotes the basis vector where only the j-th element is 1 and all the others are 0 .

The algorithms under investigation incorporate several approaches for approximating the gradient.

2 SGFFW Algorithm

The first algorithm that we are studying in our project is *Stochastic Gradient-Free Frank-Wolfe (SGFFW)*, which combines the principles of stochastic optimization with the Frank-Wolfe framework.

SGFFW builds upon the classic Frank-Wolfe algorithm. However, instead of relying on full gradients, *SGFFW* uses stochastic gradient estimates, making it suitable for large-scale and noisy optimization problems.

In the article, *SGFFW* addresses the problem represented by eq. (1).

In the SGFFW update scheme, the linear minimization and subsequent steps differ from those in the ordinary Stochastic Frank-Wolfe method.

$$\mathbf{d}_t = (1 - \rho_t)\, \mathbf{d}_{t-1} + \rho_t \mathbf{g}(\mathbf{x}_t, \mathbf{y}_t) \quad (4)$$

$$\mathbf{v}_t = \underset{\mathbf{v} \in \mathcal{C}}{\operatorname{argmin}} \langle \mathbf{d}_t, \mathbf{v} \rangle \quad (5)$$

$$\mathbf{x}_{t+1} = (1 - \gamma_{t+1})\, \mathbf{x}_t + \gamma_{t+1} \mathbf{v}_t, \quad (6)$$

where $g(\mathbf{x}_t, \mathbf{y}_t)$ is a gradient approximation, $\mathbf{d}_0 = \mathbf{0}$ and ρ_t is a time-decaying sequence.

Key characteristics of the algorithm are highlighted as follows:

- A straightforward substitution of $\nabla f(\mathbf{x}_k)$ with its stochastic counterpart, $\nabla F(\mathbf{x}_k; \mathbf{y}_k)$, carries the potential for divergence, primarily owing to the persistent variance within gradient approximations;
- The algorithm explores three distinct gradient approximation strategies:
  1. *KWSA*
  $$\mathbf{g}(\mathbf{x}_t; \mathbf{y}) = \sum_{i=1}^{d} \frac{F(\mathbf{x}_t + c_t \mathbf{e}_i; \mathbf{y}) - F(\mathbf{x}_t; \mathbf{y})}{c_t} \mathbf{e}_i$$
  2. *RDSA*
  Sample $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I}_d)$,
  $$\mathbf{g}(\mathbf{x}_t; \mathbf{y}, \mathbf{z}_t) = \frac{F(\mathbf{x}_t + c_t \mathbf{z}_t; \mathbf{y}) - F(\mathbf{x}_t; \mathbf{y})}{c_t} \mathbf{z}_t$$
  3. *I-RDSA*
  Sample $\{\mathbf{z}_{i,t}\}_{i=1}^{m} \sim \mathcal{N}(0, \mathbf{I}_d)$,
  $$\mathbf{g}(\mathbf{x}_t; \mathbf{y}, \mathbf{z}_t) = \frac{1}{m} \sum_{i=1}^{m} \frac{F(\mathbf{x}_t + c_t \mathbf{z}_{i,t}; \mathbf{y}) - F(\mathbf{x}_t; \mathbf{y})}{c_t} \mathbf{z}_{i,t};$$
- The parameter γ_t are set as $\gamma_t = \dfrac{2}{t+8}$.

2.1 SGFFW: Convergence Analysis

It emerges that, under certain assumptions, the primal sub-optimality gap $\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)]$ in the convex case is found to be $O(\frac{d^{1/3}}{T^{1/3}})$. This matches the performance of the stochastic Frank-Wolfe algorithm, which has access to first-order information. The number of queries required by stochastic zeroth order oracle to achieve a primal gap of ϵ, i.e., $\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \leq \epsilon$, is given by $O\left(\frac{d}{\epsilon^3}\right)$.

At the same time, in a non-convex scenario, the primal sub-optimality gap and the number of queries are $O(\frac{d^{1/3}}{T^{1/4}})$ and $O\left(\frac{d^{4/3}}{\epsilon^4}\right)$, respectively.

Hence, the rate of convergence of the proposed algorithm in terms of the primal gap is showed to match its first order counterpart in terms of iterations.

3 FZCGS Algorithm

Faster Zeroth-Order Conditional Gradient Sliding (ZOCGS) method introduces a sliding technique that accelerates convergence by dynamically adjusting the step size and direction based on function evaluations. *FZOCGS* is capable of handling high-dimensional problems, non-convex functions, and noisy objective evaluations, making it suitable for a wide range of applications.

In the article, *ZOCGS* addresses the problem represented by eq. (2).

3.1 FZFW Algorithm

Faster Zeroth-Order Frank-Wolfe (FZFW) method predates *FZCGS* and does not incorporate the conditional gradient sliding algorithm.

The update scheme of *FZCGS* consists of the following steps:

$$\mathbf{u}_k = \arg\max_{\mathbf{u}\in\Omega} \langle \mathbf{u}, -\hat{\mathbf{v}}_k \rangle \tag{7}$$

$$\mathbf{d}_k = \mathbf{u}_k - \mathbf{x}_k \tag{8}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma_k \mathbf{d}_k, \tag{9}$$

where $\hat{\mathbf{v}}_k$ is a gradient approximation.

Key highlights of this algorithm encompass the following:

- The usage of the coordinate-wise gradient estimator (eq. (3));
- The gradient estimation occurs at every q iterations are defined as follows:

  $$\hat{\nabla} f_{S_1}(\mathbf{x}_k) = \sum_{j=1}^{d} \frac{f_{S_1}(\mathbf{x}_k + \mu_j \mathbf{e}_j) - f_{S_1}(\mathbf{x}_k - \mu_j \mathbf{e}_j)}{2\mu_j} \mathbf{e}_j, \tag{10}$$

  and at other iterations as follows:

  $$\hat{\mathbf{v}}_k = \frac{1}{|S_2|} \sum_{i\in S_2} \left[\hat{\nabla} f_i(\mathbf{x}_k) - \hat{\nabla} f_i(\mathbf{x}_{k-1}) + \hat{\mathbf{v}}_{k-1} \right], \tag{11}$$

  where S_1 and S_2 denote the randomly selected samples;
- The estimated number of oracle queries is $O\left(\frac{n^{1/2}d}{\epsilon^2}\right)$.

3.2 FZCGS Algorithm

Although *Faster Zeroth-Order Frank-Wolfe (FZFW)* employs the same technique for gradient approximation, it upgrades $\mathbf{x}_k$ in a different way via *the conditional gradient sliding algorithm*:

- Defining $\phi(\mathbf{y}; \mathbf{x}, \nabla F(\mathbf{x}), \gamma) = \min_{\mathbf{y}\in\Omega} \langle \nabla F(\mathbf{x}), \mathbf{y} \rangle + \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{x}\|^2$;
- Optimizing $\max_{\mathbf{x}\in\Omega} \langle \phi'(\mathbf{u}_t; \mathbf{u}, \mathbf{g}, \gamma), \mathbf{u}_t - \mathbf{x} \rangle$, which is the Wolfe gap;
- Terminating when the Wolfe gap is smaller than the predefined tolerance η.

3.3 FZCGS: Convergence Analysis

Firstly, the utilization of the coordinate-wise gradient estimator significantly augments convergence performance in comparison to alternative gradient estimators. This estimator also reduces the variance introduced by the randomly selected component functions.

Furthermore, it's worth emphasizing that the incorporation of *conditional gradient sliding* has yielded a substantial reduction in the algorithm's iteration complexity, transitioning from $O\left(\frac{n^{1/2}d}{\epsilon^2}\right)$ to $O\left(\frac{n^{1/2}d}{\epsilon}\right)$.

In conclusion, it has been substantiated, through theoretical analysis, that *FZCGS* demonstrates superior convergence rates when compared to previously developed methods designed for non-convex optimization. Remarkably, its iteration complexity even outperforms that of its first-order counterparts.

4 ZO-SCGS Algorithm

4.1 ZO-SCGS: Convergence Analysis

References

[1] Hongchang Gao and Heng Huang. "Can Stochastic Zeroth-Order Frank-Wolfe Method Converge Faster for Non-Convex Problems?" In: *Proceedings of the 37th International Conference on Machine Learning.* Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 3377–

3386. URL: https://proceedings.mlr.press/v119/gao20b.html.

[2] Aleksandr Lobanov et al. *Zero-Order Stochastic Conditional Gradient Sliding Method for Non-smooth Convex Optimization.* 2023. arXiv: 2303.02778 [math.OC].

[3] Anit Kumar Sahu, Manzil Zaheer, and Soummya Kar. "Towards Gradient Free and Projection Free Stochastic Optimization". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics.* Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 16–18 Apr 2019, pp. 3468–3477. URL: https://proceedings.mlr.press/v89/sahu19a.html.