Simple steps are all you need: Frank-Wolfe and generalized self-concordant functions

Alejandro Carderera [1], Mathieu Besançon [2], Sebastian Pokutta [2, 3]

[1]Georgia Institute of Technology, [2]Zuse Institute Berlin, [3]Technische Universität Berlin

Abstract

Generalized self-concordance is a key property present in many learning problems. We establish the convergence rate of a simple Frank-Wolfe variant that uses a $\gamma_t = 2/(t+2)$ step size, obtaining a $O(1/t)$ convergence rate in primal and Frank-Wolfe gap. This avoids the use of second-order information or the need to estimate local smoothness parameters. We also show improved rates when the feasible region is uniformly convex or polyhedral.

Motivation

We consider the problem:

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}), \tag{1}$$

where f is (M,ν) generalized self-concordant (GSC) and $\mathcal{X}$ is a compact convex set. We solve Problem (1) armed with:

- **Zeroth/First Order Oracle** (Z/FOO)
- **Linear Minimization Oracle** (LMO)
- **Domain Oracle** (DO).

Focus on *Frank-Wolfe* (FW) [1], a.k.a. *Conditional Gradient* (CG) [2], algorithms. Typically, in order to solve Problem (1), existing algorithms utilize second-order oracles and obtain $O(1/t)$ rates in primal gap [3]:

*Can we match existing rates in the literature without second-order information? **Yes!***

Our contributions are:

1. `Monotonous FW (M-FW)`: A simple variant that achieves $O(1/t)$ convergence in primal and FW gap.
2. `Backtracking FW (B-FW)`: We show that FW with the line search of [4] achieves improved rates when $\mathcal{X}$ is uniformly convex.
3. `Backtracking AFW (B-AFW)`: We show that AFW [5] with the line search of [4] achieves improved rates when $\mathcal{X}$ is polyhedral.

`Monotonous/Backtracking FW`

Focusing on the first two contributions, the algorithms proposed are:

Algorithm 1 `M/B-FW`

```
1: for $t = 0$ to ... do
2:     $\mathbf{v}_t \leftarrow \operatorname{argmin}_{\mathbf{v}\in\mathcal{X}} \langle \nabla f(\mathbf{x}_t), \mathbf{v}\rangle$
       Option 1: M-FW
3:     $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + 2/(t+2)(\mathbf{v}_t - \mathbf{x}_t)$
4:     if $\mathbf{x}_{t+1} \notin \operatorname{dom}(f)$ or $f(\mathbf{x}_{t+1}) > f(\mathbf{x}_t)$ then
5:         $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t$
       Option 2: B-FW
6:     $\gamma_t, L_t \leftarrow \texttt{Backtrack}(f, \mathbf{x}_t, \mathbf{v}_t - \mathbf{x}_t, L_{t-1}, 1)$
7:     $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \gamma_t(\mathbf{v}_t - \mathbf{x}_t)$
```

where $\texttt{Backtrack}(f, \mathbf{x}_t, \mathbf{v}_t - \mathbf{x}_t, L_{t-1}, 1)$ is the backtracking line search of [4], which automatically estimates the local smoothness parameter, with the modification that we also check if the point we are moving towards is inside $\operatorname{dom}(f)$.

`Backtracking AFW`

Given a polytope $\mathcal{X}$, one can use the AFW algorithm [5] with the modified version of the backtracking line search of [4] mentioned in the previous section to obtain the following convergence in primal and Frank-Wolfe gap:

Convergence rate of `B-AFW`

Let $\mathcal{X}$ be a polytope, and f be a GSC function, then `B-AFW` requires $O(\log 1/\epsilon)$ iterations to achieve an ϵ-optimal solution in primal gap or Frank-Wolfe gap.

References

[1] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.

[2] B. T. Polyak, "Minimization methods in the presence of constraints," *Itogi Nauki i Tekhniki. Seriya" Matematicheskii Analiz*", 1974.

[3] P. Dvurechensky, K. Safin, S. Shtern, and M. Staudigl, "Generalized self-concordant analysis of Frank-Wolfe algorithms," *arXiv preprint arXiv:2010.01009*, 2020.

[4] F. Pedregosa, G. Negiar, A. Askari, and M. Jaggi, "Linearly convergent Frank–Wolfe with backtracking line-search," in *Proceedings of AISTATS'20*.

[5] J. Guélat and P. Marcotte, "Some comments on Wolfe's 'away step'," *Mathematical Programming*, vol. 35, no. 1, pp. 110–119, 1986.

[6] T. Kerdreux, A. d'Aspremont, and S. Pokutta, "Projection-free optimization on uniformly convex sets," in *Proceedings of AISTATS'21*.

Convergence rate of `M-FW`

Let $\mathcal{X}$ be a compact convex set and f be a GSC function, then `M-FW` satisfies:

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{4(T_\nu + 1)}{t+1} \max\{f(\mathbf{x}_0) - f(\mathbf{x}^*), C\}$$

for $t \geq T_\nu$, where C and T_ν depends on the diameter of $\mathcal{X}$, ν, M, and the largest eigenvalue of the Hessian for points $\mathbf{y} \in \mathcal{X}$ with $f(\mathbf{y}) \leq f(\mathbf{x}_0)$. Otherwise, $f(\mathbf{x}_t) \leq f(\mathbf{x}_0)$ for $t < T_\nu$.

Proof sketch: after an initial number of iterations independent of ϵ, the decreasing $2/(2+t)$ step size ensures both that the iterates remain inside $\operatorname{dom}(f)$ and that we can use a *smoothness-like* inequality from generalized self-concordance. The convergence rate follows by induction by considering two different scenarios, one in which the step size is small enough to use the aforementioned inequality to ensure primal progress, which therefore means that we do not go to Line 5 of Algorithm 1, and another case in which the step size is not small enough, but which trivially allows us to prove the desired bound.

Convergence rate of `B-FW`

Let f be a GSC function. If $\mathcal{X}$ is a compact convex set and $\mathbf{x}^* \in \operatorname{Int}(\mathcal{X} \cap \operatorname{dom}(f))$, then `B-FW` converges in primal/Frank-Wolfe gap with complexity $O(\log 1/\varepsilon)$. Otherwise, assume that $\mathcal{X}$ is a (κ, q)-uniformly convex set, then the algorithm converges with the following complexities:

Assumptions	Rate
$\min_{\mathbf{x}\in\mathcal{X}} \\|\nabla f(\mathbf{x})\\| > 0$, $q = 2$	$O(\log 1/\varepsilon)$
$\min_{\mathbf{x}\in\mathcal{X}} \\|\nabla f(\mathbf{x})\\| > 0$, $q > 2$	$O\left(\varepsilon^{-(q-2)/q}\right)$
No straight lines in $\operatorname{dom}(f)$	$O\left(\varepsilon^{-(q-1)/q}\right)$

Proof sketch: For all the cases considered, the per-iteration progress bound used stems from the backtracking line search. For the linear convergence rates, this progress bound is complemented by a *scaling condition* that relates $\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{v}_t\rangle$ to $\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^*\rangle$, which is due to either $\mathbf{x}^* \in \operatorname{Int}(\mathcal{X} \cap \operatorname{dom}(f))$ or $\min_{\mathbf{x}\in\mathcal{X}} \|\nabla f(\mathbf{x})\| > 0$ and the set being $(\kappa, 2)$-uniformly convex. For the remaining cases, we use the properties derived from the uniform convexity of the feasible region (see [6]).

Computational Results

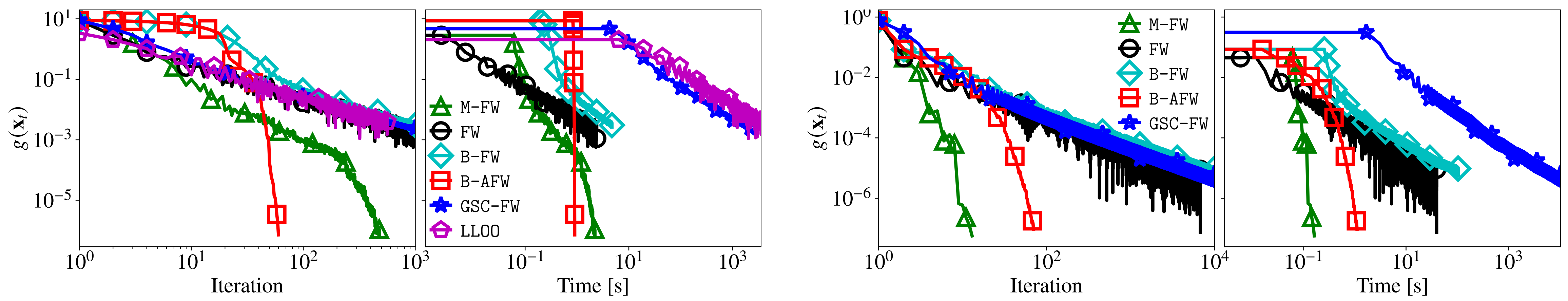

Figure 1: Performance w.r.t. iteration count/time: Algorithm comparison in terms of Frank-Wolfe gap (denoted by $g(\mathbf{x}_t)$) for a portfolio optimization problem over the probability simplex (left) and a logistic regression problem over the ℓ_1 ball (right).

Paper: https://arxiv.org/pdf/2105.13913.pdf

Code: https://github.com/ZIB-IOL/fw-generalized-selfconcordant