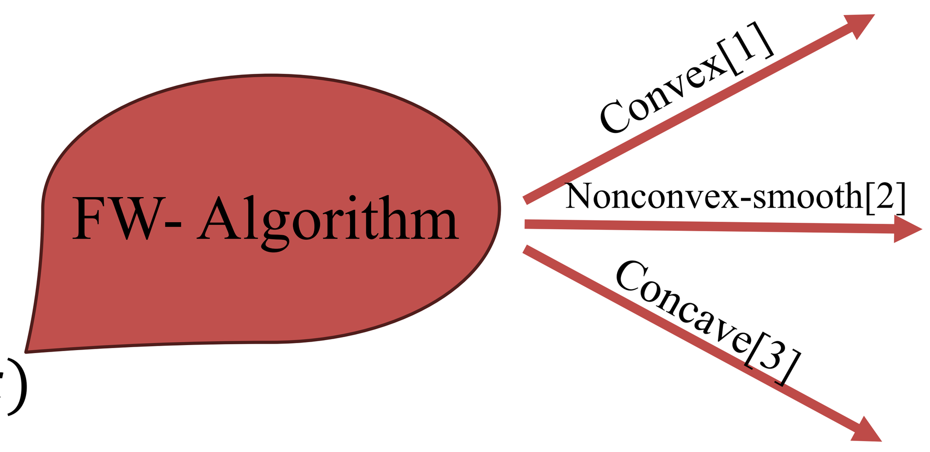
Revisiting Frank-Wolfe for Nonconvex Optimization

Hoomaan Maskan[1], Suvrit Sra[2], Alp Yurtsever[1]
[1] Department of Mathematics & Mathematical Statistics, Umeå University
[2] School of Computation, Information and Technology, Technical University of Munich

Motivation

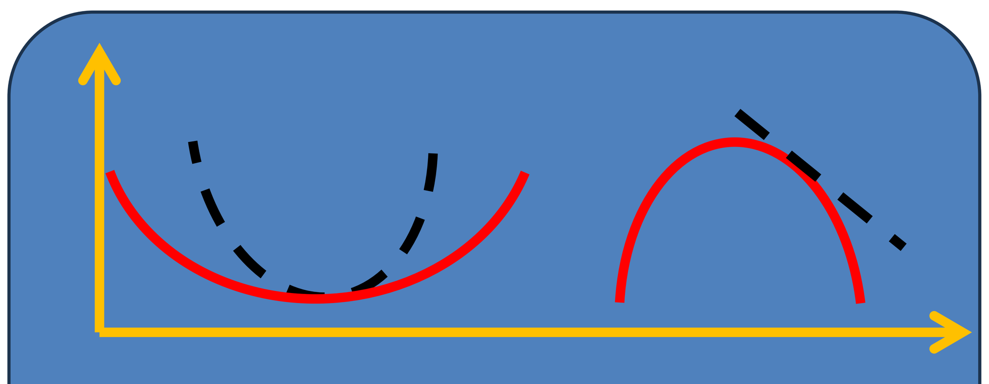


$\min_{x\in\mathcal{D}} \phi(x)$

$\phi(x_k) - \phi(x^*) \le \mathcal{O}(1/k)$ — Bounded Curvature Constant

$\min_{0\le\tau\le K} \max_{x\in\mathcal{D}} \langle\nabla\phi(x_\tau), x_\tau - x\rangle \le \mathcal{O}(\frac{1}{\sqrt{k}})$ — Smoothness

$\min_{0\le\tau\le K} \max_{x\in\mathcal{D}} \langle\nabla\phi(x_\tau), x_\tau - x\rangle \le \mathcal{O}(\frac{1}{k})$ — Concavity



- Choice of the stationarity measure

Stationarity Measures

Assume smooth $\phi(x)$

$$\min_{x\in\mathcal{D}} \frac{L\|x\|^2}{2} - \left(\frac{L\|x\|^2}{2} - \phi(x)\right)$$

Difference of Convex (DC)

DCA

$$x_{t+1} = argmin_{x\in\mathcal{D}} \frac{L\|x - x_t\|^2}{2} - \langle\nabla\phi(x_t), x_t - x\rangle$$

Modified Gap I:

$$gap_{\mathcal{D}}^{L}(y) = \max_{x\in\mathcal{D}}\left\{\langle\nabla\phi(y), y - x\rangle - \frac{L}{2}\|x - y\|^2\right\}$$

Lemma 1: Modified Gap I is a suitable measure to first-order stationarity.

$$\min_{x\in\mathcal{D}} \left(\frac{L\|x\|^2}{2} + \phi(x)\right) - \frac{L\|x\|^2}{2}$$

DCA

$$x_{t+1} = prox^{h}_{\frac{1}{L}\phi + I_{\mathcal{D}}}(x_t) = argmin_{x\in\mathcal{D}} \ \phi(x) + \frac{L\|x - x_t\|^2}{2}$$

Modified Gap II:

$$gap_{\mathcal{D}}^{L}(y) = \max_{x\in\mathcal{D}}\left\{\phi(y) - \phi(x) - \frac{L}{2}\|x - y\|^2\right\}$$

Lemma 1: Modified Gap II is a suitable measure to first-order stationarity.

DC-FW

$$\min_{x\in\mathcal{D}} \ f(x) - g(x)$$

Algorithm 1(**DC-FW**)

For $t = 1,2,\ldots$ do
  For $k = 1,2,\ldots$ do
    $s_{t,k} = argmin_{x\in\mathcal{D}} \langle\nabla f(X_{t,k}) - \nabla g(x_t), x\rangle$
    $d_{t,k} = s_{t,k} - X_{t,k}$
    (Convergence Check)
    $X_{t,k+1} = X_{t,k} - \eta_{t,k} d_{t,k}$
  End For
End For
Output: $X_{t,k}$

- For $f(x) = \frac{L\|x\|^2}{2}$ and $g(x) = \left(\frac{L\|x\|^2}{2} - \phi(x)\right)$ Conditional Gradient Sliding (CGS) is recovered [4].
- $\mathcal{O}(1/\epsilon)$ calls to the FO
- $\mathcal{O}(1/\epsilon^2)$ calls to the LMO

Corollary:
The DC-FW algorithm generates a sequence of solutions that satisfies

$$\min_{0\le\tau\le t} \max_{x\in\mathcal{D}} \{f(x_\tau) - f(x) - \langle\nabla g(x_\tau), x_\tau - x\rangle\} \le \epsilon$$

Within $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ iterations. The total number of calls to the linear minimization oracle is bounded by $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$.

References

1. Jaggi, M. **Revisiting Frank-Wolfe: Projection-free sparse convex optimization**. In International conference on machine learning, pp. 427–435. PMLR, 2013.
2. Lacoste-Julien, S. **Convergence rate of Frank-Wolfe for non-convex objectives**. arXiv preprint arXiv:1607.00345, 2016.
3. Yurtsever, A. and Sra, S. **CCCP is Frank-Wolfe in disguise**. Advances in Neural Information Processing Systems, 35:35352–35364, 2022.
4. Lan, G. and Zhou, Y. **Conditional gradient sliding for convex optimization. SIAM Journal on Optimization**, 26(2):1379–1409, 2016.

UMEÅ UNIVERSITY

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM