Self-concordant analysis of Frank-Wolfe algorithms

Pavel Dvurechensky[1] Shimrit Shtern[2] Mathias Staudigl[3] Petr Ostroukhov[4] Kamil Safin [4]

[1]WIAS [2]The Technion [3]Maastricht University [4]Moscow Institute of Physics and Technology

Maastricht University

ICML2020, July 12-July 18

Self-concordant minimization

We consider the optimization problem

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

where

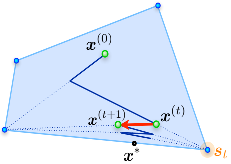
- $\mathcal{X} \subset \mathbb{R}^n$ is convex compact
- $f : \mathbb{R}^n \to (-\infty, \infty]$ is convex and thrice continuously differentiable on the open set $\operatorname{dom} f = \{x : f(x) < \infty\}$.

Given the large-scale nature of optimization problems in machine learning, first-order methods are the method of choice.

Frank-Wolfe methods

Because of great scalability and sparsity properties, *Frank-Wolfe* (FW) methods (Frank & Wolfe, 1956) received lot of attention in ML.

1. Convergence guarantees require Lipschitz continuous gradients, or finite curvature constants on f (Jaggi, 2013)
2. Even for well-conditioned (Lipschitz smooth and strongly convex) problems only sublinear convergence rates guaranteed in general.



Many canonical ML problems do not have Lipschitz gradients

- **Portfolio Optimization**

$$f(x) = -\sum_{t=1}^{T} \ln(\langle r_t, x \rangle), x \in \mathcal{X} = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^{n} x_i = 1\}.$$

- **Covariance Estimation:**

$$f(x) = -\ln(\det(X)) + \mathrm{tr}(\hat{\Sigma} X),$$
$$x \in \mathcal{X} = \{x \in \mathbb{R}^{n \times n}_{sym,+} : \|\mathrm{Vec}(X)\|_1 \leq R\}.$$

- **Poisson Inverse Problem**

$$f(x) = \sum_{i=1}^{m} \langle w_i, x \rangle - \sum_{i=1}^{m} y_i \ln(\langle w_i, x \rangle),$$
$$x \in \mathcal{X} = \{x \in \mathbb{R}^n | \|x\|_1 \leq R\}.$$

Main Results

All these function are Self-concordant (SC), and have no Lipschitz continuous gradient. Standard analysis does not apply.

Result 1: We give a unified analysis of provable convergent FW algorithms minimizing SC functions.

Result 2: Based on the theory of Local Linear Optimization Oracles (LLOO) (Lan 2013, Garber & Hazan, 2016), we construct linearly convergent variants for our base algorithms.

The analysis of FW involves

(a) a search direction

$$s(x) = \operatorname*{argmin}_{s \in \mathcal{X}} \langle \nabla f(x), s \rangle \,.$$

(b) as merit function the gap function

$$gap(x) = \langle \nabla f(x), x - s(x) \rangle$$

Standard Frank-Wolfe method:

If $gap(x^k) > \varepsilon$ then

1. Obtain $s^k = s(x^k)$;
2. Set $x^{k+1} = x^k + \alpha_k (s^k - x^k)$ for some $\alpha_k \in [0, 1]$.

Definition of SC functions

- $f : \mathbb{R}^n \to (-\infty, +\infty]$ a $\mathbf{C}^3(\mathrm{dom}\, f)$ convex function
- $\mathrm{dom}\, f$ is open set in $\mathbb{R}^n$.
- f is SC if

$$\left|\varphi'''(t)\right| \leq M\varphi''(t)^{3/2}$$

for $\varphi(t) = f(x + tv)$, $x \in \mathrm{dom}\, f$, $v \in \mathbb{R}^n$ and $x + tv \in \mathrm{dom}\, f$.

Self-concordant functions

- Self-concordant (SC) function have been developed within the field of interior-point method (Nesterov & Nemirovski, 1994)
- Starting with Bach (2010), they gained a lot of interest in Machine learning and Statistics (see e.g. Tran-Dinh, Kyrillidis & Cevher; Sun & Tran-Dinh 2018; Ostrovskii & Bach 2018)
- MATLAB toolbox SCOPT

Basic estimates of SC functions

- For all $x, \tilde{x} \in \operatorname{dom} f$ we have the following bounds on function values

$$f(\tilde{x}) \geq f(x) + \langle \nabla f(x), \tilde{x} - x \rangle + \frac{4}{M^2} \omega\left(\mathrm{d}(x, \tilde{x})\right)$$

$$f(\tilde{x}) \leq f(x) + \langle \nabla f(x), \tilde{x} - x \rangle + \frac{4}{M^2} \omega_*\left(\mathrm{d}(x, \tilde{x})\right)$$

where

$$\omega(t) := t - \ln(1+t), \text{ and } \omega_*(t) := -t - \ln(1-t)$$

$$\mathrm{d}(x, y) := \frac{M}{2} \|y - x\|_x = \frac{M}{2} \left(D^2 f(x)[y - x, y - x]\right)^{1/2}.$$

Algorithm 1

Let $x_t^+ = x + t(s(x) - x), t > 0$
Obtain the non-Euclidean descent inequality:

$$f(x_t^+) \leq f(x) + \langle \nabla f(x), x_t^+ - x \rangle + \frac{4}{M^2}\omega_*(t\mathtt{e}(x))$$
$$\leq f(x) - \eta_x(t)$$

for $t \in (0, 1/\mathtt{e}(x)), \mathtt{e}(x) = \frac{M}{2}\|s(x) - x\|_x^2$.
Optimizing the per-iteration decrease w.r.t t leads to

$$\alpha(x) = \min\{1, \mathtt{t}(x)\}, \mathtt{t}(x) = \frac{gap(x)}{\mathtt{e}(x)(gap(x) + \frac{4}{M^2}\mathtt{e}(x))}.$$

Iteration Complexity

Define the approximation error : $h_k = f(x^k) - f^*$.
Let

$$S(x^0) = \{x \in \mathcal{X} | f(x) \leq f(x^0)\}, \text{ and}$$
$$L_{\nabla f} = \max_{x \in S(x^0)} \lambda_{\max}(\nabla^2 f(x)).$$

Theorem

For given $\varepsilon > 0$*, define* $N_\varepsilon(x^0) = \min\{k \geq 0 | h_k \leq \varepsilon\}$*. Then,*

$$N_\varepsilon(x^0) \leq \frac{\ln\left(\frac{h_0 \mathrm{b}}{\mathrm{a}}\right)}{\mathrm{a}} + \frac{L_{\nabla f} \operatorname{diam}(\mathcal{X})^2}{(1 + \ln(2))\varepsilon}.$$

where $\mathrm{a} = \min\left\{\frac{1}{2}, \frac{2(1 - \ln(2))}{M\sqrt{L_{\nabla f}} \operatorname{diam}(\mathcal{X})}\right\}$ *and* $\mathrm{b} = \frac{1 - \ln(2)}{L_{\nabla f} \operatorname{diam}(\mathcal{X})^2}$.

Algorithm 2: Backtracking Variant of FW

Let

$$Q(x^k, t, \mu) := f(x^k) - t \cdot gap(x^k) + \frac{t^2\mu}{2} \left\| s(x^k) - x^k \right\|_2^2 .$$

On $S(x^0) := \{x \in \mathcal{X} | f(x) \leq f(x^0)\}$, we have

$$f(x^k + t(s^k - x^k)) \leq Q(x^k, t, L_{\nabla f}).$$

Problem: $L_{\nabla f}$ is hard to estimate and numerically large.

Solution: A backtracking procedure allows us to find a local estimate for the unknown $L_{\nabla f}$ (see also Pedregosa et al. 2020)

Backtracking procedure to find the local Lipschitz constant

Algorithm 1 Function $\texttt{step}(f, v, x, g, \mathcal{L})$

Choose $\gamma_u > 1, \gamma_d < 1$
Choose $\mu \in [\gamma_d \mathcal{L}, \mathcal{L}]$
$\alpha = \min\{\frac{g}{\mu\|v\|_2^2}, 1\}$
if $f(x + \alpha v) > Q(x, \alpha, \mu)$ **then**
  $\mu \leftarrow \gamma_u \mu$
  $\alpha \leftarrow \min\{\frac{g}{\mu\|v\|_2^2}, 1\}$
end if
Return α, μ

We have for all $t \in [0, 1]$

$$f(x^{k+1}) \leq f(x^k) - t \cdot gap(x^k) + \frac{t^2 \mathcal{L}_k}{2} \left\| s^k - x^k \right\|^2$$

where $\mathcal{L}_k$ is obtained from Algorithm 1.

Main Result

Theorem

Let $(x^k)_k$ be the backtracking variant of FW using Algorithm 1 as subroutine. Then

$$h_k \leq \frac{2gap(x^0)}{(k+1)(k+2)} + \frac{k\,\mathrm{diam}(\mathcal{X})^2}{(k+1)(k+2)}\bar{\mathcal{L}}_k$$

where $\bar{\mathcal{L}}_k \triangleq \frac{1}{k}\sum_{i=0}^{k-1}\mathcal{L}_i$.

Linearly Convergent FW variant

Definition (Garber & Hazan (2016))

A procedure $\mathcal{A}(x, r, c)$, where $x \in \mathcal{X}, r > 0, c \in \mathbb{R}^n$, is a LLOO with parameter $\rho \geq 1$ for the polytope $\mathcal{X}$ if $\mathcal{A}(x, r, c)$ returns a point $s \in \mathcal{X}$ such that for all $x \in B_r(x) \cap \mathcal{X}$

$$\langle c, x \rangle \geq \langle c, s \rangle \text{ and } \|x - s\|_2 \leq \rho r.$$

- Such oracles exist for any compact polyhedral domain.
- Particular simple implementation for Simplex-like domains.

Call

$$\sigma_f = \min_{x \in S(x^0)} \lambda_{\min}(\nabla^2 f(x)).$$

Theorem (Simplified version)

Given a polytope $\mathcal{X}$ with LLOO $\mathcal{A}(x, r, c)$ for each $x \in \mathcal{X}, , r \in (0, \infty), c \in \mathbb{R}^n$. Let

$$\bar{\alpha} \triangleq \min\{\frac{\sigma_f}{6L_{\nabla f}\rho^2}, 1\} \frac{1}{1 + \sqrt{L_{\nabla f}} \frac{M \operatorname{diam}(\mathcal{X})}{2}}.$$

Then,

$$h_k \leq gap(x^0) \exp(-k\bar{\alpha}/2).$$

In the paper we present a version of this Theorem without knowledge of $L_{\nabla f}$.

Numerical Performance

- Portfolio Optimization

$$f(x) = \sum_{t=1}^{T} \ln(\langle r_t, x \rangle)$$

$$\mathcal{X} = \{x \in \mathbb{R}_+^n : \sum_{i=1}^{n} x_i = 1\}.$$

- Poisson Inverse problem

$$f(x) = \sum_{i=1}^{m} \langle w_i, x \rangle - \sum_{i=1}^{m} y_i \ln(\langle w_i, x \rangle),$$

$$x \in \mathcal{X} = \{x \in \mathbb{R}^n | \, \|x\|_1 \leq R\}.$$

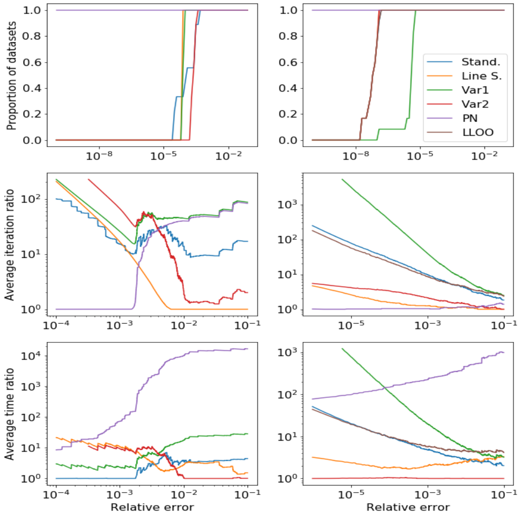


Figure: Portfolio Optimization (Right), Poisson Inverse Problem (Left)

Conclusion

- We derived various novel FW schemes with provable convergence guarantees for self-concordant minimization.
- Future directions of research include the following
  - Generalized self-concordant minimization (Sun & Tran-Dinh 2018)
  - Stochastic oracles
  - Inertial effects in algorithm design (Conditional gradient sliding (Lan & Zhou, 2016))