Frank-Wolfe / Conditional Gradient algorithm

Alexandre Gramfort

alexandre.gramfort@telecom-paristech.fr

Telecom ParisTech

Master 2 Data Science, Univ. Paris Saclay
Optimisation for Data Science

Constrained optimization problem

We consider the constrained optimization problem $(\mathcal{P})$:

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

- where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is the **objective function**
- $\mathcal{D}$ is the **domain** which we assume is a **convex set**.

$\rightarrow$ Assuming f is smooth how would you solve this?
$\rightarrow$ Give me examples in machine learning of such a problem.

Constrained optimization problem

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

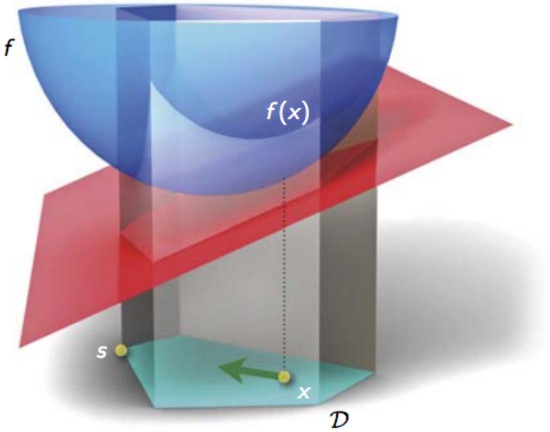


Image courtesy of Martin Jaggi (cf. [Jag13]).

Many applications

- network flows / transportation problems
- greedy selection and sparse optimization
- low-rank matrix factorizations, many other matrices
- wavelets (infinite-dimensional)
- structured sparsity and structured prediction
- total-variation-norm for image de-noising
- submodular optimization
- boosting
- training deep networks

LMO and linearization

- The Linear Minimization Oracle

$$\mathrm{LMO}_{\mathcal{D}}(d) = \operatorname*{arg\,min}_{s \in \mathcal{D}} \langle d, s \rangle$$

- Linearization

$$\min_{s \in \mathcal{D}} f(x) + \langle \nabla f(x), s - x \rangle$$

- Idea:

$$x^{k+1} \approx \operatorname*{arg\,min}_{s \in \mathcal{D}} \mathrm{LMO}_{\mathcal{D}}(\nabla f(x^k))$$

- Step depends on domain $\mathcal{D}$ and $\nabla f(x^k)$, hence the name **conditional gradient**.

Convergence

- Marguerite Frank and Philip Wolfe showed in [FW56] that:

$$f(x^k) - f(x^*) \leq \mathcal{O}(1/k)$$

- Provided that:
  - f is smooth and convex
  - $\mathcal{D}$ is bounded and convex

Rem: Same rates as projected gradient method but with simpler iterations. It is a projection free algorithm.

Frank-Wolfe / Conditional Gradient algorithm

1: $x^0 \in \mathbf{D}$
2: **for** $k = 0$ to n **do**
3: $\quad s = \mathrm{LMO}_{\mathcal{D}}(\nabla f(x^k))$
4: $\quad \gamma = \frac{2}{k+2}$
5: $\quad x^{k+1} = (1-\gamma)x^k + \gamma s$
6: **end for**
7: **return** x^{k+1}

With line search:

$$\gamma = \arg\min_{\gamma \in [0,1]} f((1-\gamma)x^k + \gamma s)$$

Convergence proof

Theorem

For each $k \geq 1$, the iterates x^k of the Frank-Wolfe algorithm satisfy

$$f(x^k) - f(x^*) \leq \frac{2C_f}{t+2} \quad .$$

Convergence proof

PROOF. Let C_f a "curvature" constant such that:

$$f(y) \leq f(x) + \gamma \underbrace{\langle s - x, \nabla f(x) \rangle}_{-g(x)} + \frac{\gamma^2}{2} C_f$$

for all $x, s \in \mathcal{D}$, $y = x + \gamma(s - x)$, $\gamma \in [0, 1]$.
Writing $h(x^k) = f(x^k) - f(x^*)$ for the error on objective, we have:

$$\begin{aligned}
h(x^{k+1}) &\leq h(x^k) - \gamma g(x^k) + \frac{\gamma^2}{2} C_f && (\text{Definition of } C_f) \\
&\leq h(x^k) - \gamma h(x^k) + \frac{\gamma^2}{2} C_f && (h \leq g \text{ by convexity}) \\
&= (1 - \gamma) h(x^k) + \frac{\gamma^2}{2} C_f .
\end{aligned}$$

From here, the decrease rate follows from a simple lemma.

Convergence proof

Lemma

Suppose a sequence of numbers $(h_k)_k$ satisfies

$$h_{k+1} \leq (1 - \gamma^k) h_k + (\gamma^k)^2 C$$

for $\gamma^k = \frac{2}{k+1}$, and $k = 0, 1, \ldots,$ and a constant C. Then

$$h_k \leq \frac{4C}{k+2}, \; k = 0, 1, \ldots$$

PROOF. Trivial by induction.

Rem: [LJJ13] shows a linear convergence if f strongly convex and use line-search.

Curvature constant vs. L-Liptschitz gradient

The curvature constant C_f is defined by:

$$C_f = \sup_{\substack{x,s\in\mathcal{D},\\ \gamma\in[0,1]\\ y=x+\gamma(s-x)}} \frac{2}{\gamma^2}(f(y)-f(x)-\langle y-x,\nabla f(x)\rangle) \ .$$

Lemma

Let f be a convex and differentiable function with its gradient ∇f being Lipschitz-continuous w.r.t. some norm $\|\cdot\|$ over the domain $\mathcal{D}$ with Lipschitz-constant $L_{\|\cdot\|} > 0$. Then:

$$C_f \leq \mathrm{diam}_{\|\cdot\|}(\mathcal{D})^2 L_{\|\cdot\|} \ .$$

PROOF. Give it a try!

Rem: For L-smooth convex function, with a bounded convex domain with have the $\mathcal{O}(1/k)$ convergence rate.

Optimality certificate (almost for free)

We solve:

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

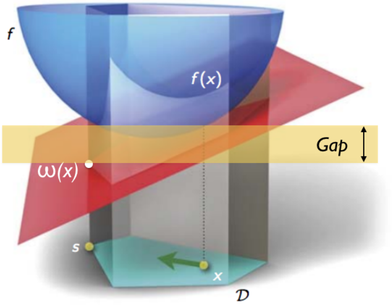
Let:

$$\omega(x) = \min_{s \in \mathcal{D}} f(x) + \langle \nabla f(x), s - x \rangle$$

Lemma (Weak duality)

$$\omega(x) \leq f(x^*) \leq f(x)$$

So if $f(x) - \omega(x) \leq \epsilon$ we have an ϵ-solution.



Special case of Atomic Sets

If

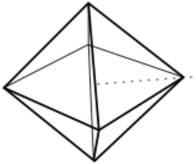
$$\mathcal{D} = \mathrm{conv}(\mathcal{A})$$

where $\mathcal{A}$ is a set (possibly infinite) of atoms/vectors.

Then we have that for every FW step $s \in \mathcal{A}$.

Example of ℓ_1 ball:

$$\mathcal{D} = \mathrm{conv}(\{e_i | i \in [n]\} \cup \{-e_i | i \in [n]\})$$



So $s = \mathrm{LMO}_{\mathcal{D}}(\nabla f(x^k)) \in \{e_i | i \in [n]\} \cup \{-e_i | i \in [n]\}$.

Let's practice

$\rightarrow$ frank_wolfe.ipynb notebook.

References

M. Frank and P. Wolfe.
An algorithm for quadratic programming.
Naval Res. Logis. Quart., 1956.

Martin Jaggi.
Revisiting frank-wolfe: Projection-free sparse convex optimization.
In *ICML*, volume 28, pages 427–435, June 2013.

S. Lacoste-Julien and M. Jaggi.
An affine invariant linear convergence analysis for frank-wolfe algorithms.
arXiv preprint arXiv:1312.7864, 2013.
https://arxiv.org/pdf/1312.7864.