Optimization for Machine Learning
CS-439

Lecture 8: Frank-Wolfe algorithm

Nicolas Flammarion

EPFL – github.com/epfml/OptML_course

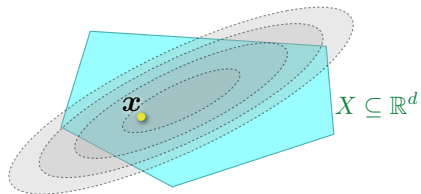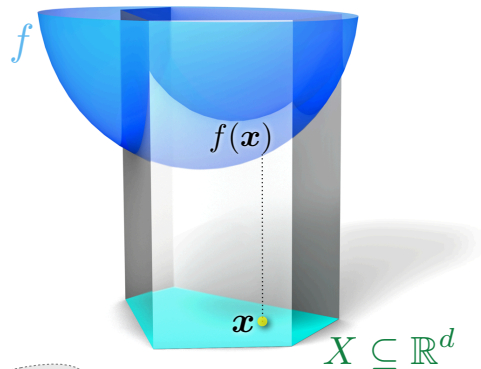April 26, 2020

Chapter 9

Frank-Wolfe

Constrained Optimization

Constrained Optimization Problem

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{x} \in X \end{aligned}$$
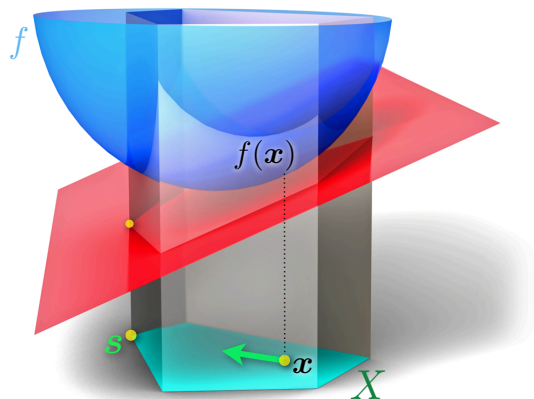
Frank-Wolfe Algorithm

Frank-Wolfe Algorithm:

$$\mathbf{s} := \mathrm{LMO}(\nabla f(\mathbf{x}_t)),$$
$$\mathbf{x}_{t+1} := (1-\gamma)\mathbf{x}_t + \gamma\mathbf{s},$$

for timesteps $t = 0, 1, \ldots,$ and
stepsize $\gamma := \frac{2}{t+2}$.

Linear Minimization Oracle:

$$\mathrm{LMO}(\mathbf{g}) := \operatorname*{argmin}_{\mathbf{s}\in X} \langle \mathbf{s}, \mathbf{g} \rangle$$

Properties

- Aways feasible: $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t \in X$.
  $\mathbf{x}_{t+1}$ is on line segment $[\mathbf{s}, \mathbf{x}_t]$, for $\gamma \in [0, 1]$.

- Reduces non-linear to linear optimization

- Projection-free

- Sparse iterates (in terms of corners $\mathbf{s}$ used)

Invented and analyzed 1956 by Marguerite Frank and Philip Wolfe.

Example

Lasso Regression

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|^2 \ \ s.t. \ \ \|\mathbf{x}\|_1 \leq 1$$

L1-ball is the convex hull of the unit basis vectors:
$X = \{\mathbf{x} \mid \|\mathbf{x}\|_1 \leq 1\} = \operatorname{conv}(\{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_n\})$.

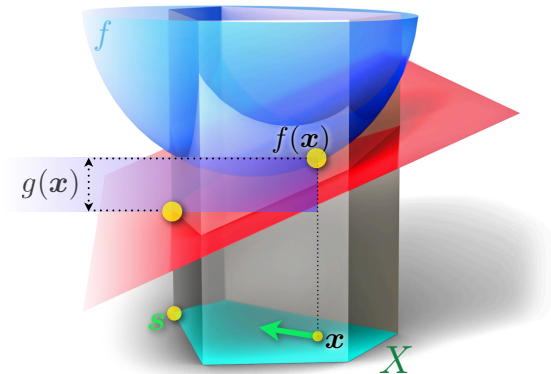- $\nabla f(\mathbf{x}) = \mathbf{g} := A^\top(A\mathbf{x} - \mathbf{b})$
- $\operatorname{LMO}(\mathbf{g}) = -\operatorname{sign}(g_i)\mathbf{e}_i$ with $i := \underset{i\in[n]}{\operatorname{argmax}} |g_i|$

simpler than projection onto L1-ball !

Duality Gap

Duality Gap

$$g(\mathbf{x}) := \langle \mathbf{x} - \mathbf{s}, \nabla f(\mathbf{x}) \rangle$$

Certificate for optimization quality:

$$
\begin{aligned}
g(\mathbf{x}) &= \max_{\mathbf{s} \in X} \langle \mathbf{x} - \mathbf{s}, \nabla f(\mathbf{x}) \rangle \\
&\geq \langle \mathbf{x} - \mathbf{x}^{\star}, \nabla f(\mathbf{x}) \rangle \\
&\geq f(\mathbf{x}) - f(\mathbf{x}^{\star})
\end{aligned}
$$

Stepsize variants

$$\gamma_t := \frac{2}{t+2},$$

$$\gamma_t := \operatorname*{argmin}_{\gamma \in [0,1]} f\big((1-\gamma)\mathbf{x}_t + \gamma\mathbf{s}\big), \qquad \text{(line-search)}$$

$$\gamma_t := \min\Big\{\frac{g(\mathbf{x}_t)}{L\,\|\mathbf{s}-\mathbf{x}_t\|^2}, 1\Big\}, \qquad \text{(gap-based)}$$

Convergence in $\mathcal{O}(1/\varepsilon)$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and smooth with parameter L, and $\mathbf{x}_0 \in X$. Then choosing any of the above stepsizes, the Frank-Wolfe algorithm yields

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{2L \operatorname{diam}(X)^2}{T + 1}$$

Where $\operatorname{diam}(X) := \max_{\mathbf{x}, \mathbf{y} \in X} \|\mathbf{x} - \mathbf{y}\|$ is the diameter of X.

Proof of Convergence in $\mathcal{O}(1/\varepsilon)$ steps

Lemma

For a step $\mathbf{x}_{t+1} := \mathbf{x}_t + \gamma(\mathbf{s} - \mathbf{x}_t)$ with arbitrary step-size $\gamma \in [0, 1]$, it holds that

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \gamma g(\mathbf{x}_t) + \tfrac{\gamma^2}{2} L \operatorname{diam}(X)^2 \ ,$$

if $\mathbf{s} = \operatorname{LMO}(\nabla f(\mathbf{x}_t))$.

Proof.

We write $\mathbf{x} := \mathbf{x}_t$, $\mathbf{y} := \mathbf{x}_{t+1} = \mathbf{x} + \gamma(\mathbf{s} - \mathbf{x})$. From the definition of smoothness of f, we have

$$\begin{aligned} f(\mathbf{y}) = \ & f(\mathbf{x} + \gamma(\mathbf{s} - \mathbf{x})) \\ \leq \ & f(\mathbf{x}) + \gamma\langle \mathbf{s} - \mathbf{x}, \nabla f(\mathbf{x})\rangle + \tfrac{\gamma^2}{2} L \operatorname{diam}(X)^2 \ . \end{aligned}$$

The lemma follows by definition of the duality gap. $\qquad\square$

Proof of Convergence in $\mathcal{O}(1/\varepsilon)$ steps

From the Lemma we know that for every step of FW, it holds that

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \gamma g(\mathbf{x}_t) + \gamma^2 C,$$

if we chose $\gamma := \frac{2}{t+2}$ and write $C := \frac{1}{2} L \operatorname{diam}(X)^2$. This bound holds also for all mentioned line-search variants *(different LHS, same RHS)*.

Writing $h(\mathbf{x}) := f(\mathbf{x}) - f(\mathbf{x}^\star)$ for the (unknown) objective error at any point $\mathbf{x}$, this implies that

$$\begin{aligned} h(\mathbf{x}_{t+1}) \leq{}& h(\mathbf{x}_t) - \gamma g(\mathbf{x}_t) + \gamma^2 C \\ \leq{}& h(\mathbf{x}_t) - \gamma h(\mathbf{x}_t) + \gamma^2 C \\ ={}& (1 - \gamma) h(\mathbf{x}_t) + \gamma^2 C \ , \end{aligned}$$

by the certificate property $h(\mathbf{x}) \leq g(\mathbf{x})$ of the duality gap.

The theorem then follows by induction (Exercise 1 of Lab 9). $\qquad\square$

Proof of Convergence in $\mathcal{O}(1/\varepsilon)$ steps

Affine Invariance

Curvature Constant

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

Algorithm is invariant to scaling (affine transformations) of the input problem.

So is C_f.

(same as Newton, but here for constrained problems)

$$C_f \le L\operatorname{diam}(X)^2 \qquad \text{for any norm!}$$

All proofs hold for C_f, instead of picking a particular $L\operatorname{diam}(X)^2$.

Convergence in Duality Gap

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and smooth with parameter L, and $\mathbf{x}_0 \in X$, $T \geq 2$. Then choosing any of the above stepsizes, the Frank-Wolfe algorithm yields a $t, 1 \leq t \leq T$ s.t.

$$g(\mathbf{x}_t) \leq \frac{27/4 \, C_f}{T+1}$$

Proof.

Idea: not all gaps can be small (use Lemma again). □

Proof (I)

Proof (II)

Proof (III)

Extensions and Use Cases

Extensions:

- ▶ Approximate LMO (of additive of multiplicative accuracy)
- ▶ Randomized LMO
- ▶ unconstrained problems (Matching Pursuit variants)

Use cases:

Whenever projection is more costly than solving a linear problem

- ▶ Lasso and other L1-constrained problems
- ▶ Matrix Completion: scalable algorithm
- ▶ Relaxation of combinatorial problems
  (e.g. matchings, network flows etc)

Applications

$$X := conv(\mathcal{A})$$

Examples	$\mathcal{A}$	$|\mathcal{A}|$	d	LMO $(\mathbf{g})$
L1-ball	$\{\pm\mathbf{e}_i\}$	$2d$	d	$\pm\mathbf{e}_i$ with $\mathrm{argmax}_i |g_i|$
Simplex	$\{\mathbf{e}_i\}$	d	d	$\mathbf{e}_i$ with $\mathrm{argmin}_i g_i$
Norms	$\{\mathbf{x}, \|\mathbf{x}\| \le 1\}$	∞	d	$\underset{\mathbf{s}, \|\mathbf{s}\| \le 1}{\mathrm{argmin}} \langle \mathbf{s}, \mathbf{g} \rangle$
Nuclear norm	$\{Y, \|Y\|_* \le 1\}$	∞	d^2	..
Wavelets	..	∞	∞	..