Frank-Wolfe and Greedy Optimization

for Learning with Big Data

Part I: Martin Jaggi
ETH Zurich

Part II: Zaid Harchaoui
Inria
Tutorial Outline

**Part I:** (Martin Jaggi)
- Frank-Wolfe basics and history
- convergence analysis and geometry independence
- lower bounds
- demo and applications for atomic domains
- faster rates under additional assumptions

**Part II:** (Zaid Harchaoui)
- Greedy Optimization and Frank-Wolfe
- Frank-Wolfe methods for:
  - composite optimization
  - non-smooth optimization
  - block-structured optimization
Greedy Pursuit Algorithms
in Statistics, Harmonic Analysis, and Signal Processing

Problem:
Approximate a given function \( b \) by a sparse convex combination \( x \) of functions in dictionary \( A = \{ a_1, a_2, \ldots, a_p, \ldots \} \)

\[
\min_{x \in \mathcal{D}} \| x - b \|_2^2
\]

where

\[
\mathcal{D} := \text{conv}(A)
\]
Greedy Pursuit in Statistics

Projection Pursuit

Projection pursuit regression model

\[ E(Y | Z = z) = \sum_{m=1}^{M} \alpha_m a_m(z) \]

where \( a_1, \ldots, a_M \) are ridge functions

\[ a_m(z) := \frac{1}{1 + \exp \left( -\frac{v_m^T z - 1/2}{2} \right)} \]

References

[ Friedman and Tukey, 1974 ] exploratory projection pursuit
[ Friedman and Stuetzle, 1981 ] projection pursuit regression
[ Huber, 1985 ] theoretical analysis
Greedy Pursuit in Signal Processing

Matching Pursuit

Consider a signal

\[ b(\tau) = \sum_{m=1}^{\infty} \alpha_m a_m(\tau) \]

where \( a_1, \ldots, a_m, \ldots \) belong to a redundant dictionary of atomic functions.

Time-Frequency Gabor Wavelets

Consider \( g(\tau) \) a real, continuously differentiable function. In addition, assume that \( \|g\| = 1, \int g \neq 0, \) and \( g(0) = 0. \) Then, a redundant dictionary of atomic functions can be generated as

\[ a_\beta(\tau) := \frac{1}{\sqrt{s}} g\left(\frac{\tau - u}{s}\right) e^{i \xi \tau} \]

where \( \beta = (s, \xi, u) \) gathers resp. scale, frequency modulation, and transition.
Greedy Pursuit in Signal Processing

Matching Pursuit

Assume an M-term atomic decomposition

\[ x(\tau) = \sum_{m=1}^{M} \alpha_m a_m(\tau) \]

where \( a_1, \ldots, a_m, \ldots \) belong to a redundant dictionary of atomic functions.

Time-Frequency Gabor Wavelets

Consider \( g(\tau) \) a real, continuously differentiable function. In addition, assume that \( \|g\| = 1, \int g \neq 0, \) and \( g(0) = 0. \) Then, a redundant dictionary of atomic functions can be generated as

\[ a_\beta(\tau) := \frac{1}{\sqrt{s}} g \left( \frac{\tau - u}{s} \right) e^{i \xi \tau} \]

where \( \beta = (s, \xi, u) \) gathers resp. scale, frequency modulation, and transition.
Greedy Pursuit in Signal Processing

Matching Pursuit

At iteration $t$ we have the decomposition

$$x(t)(\tau) = \sum_{m=1}^{t} \alpha_m a_m(\tau)$$

Residuals

$$r(t) := b - x(t)$$

Time-Frequency Gabor Wavelets

Consider $g(\tau)$ a real, continuously differentiable function. In addition, assume that $\|g\| = 1$, $\int g \neq 0$, and $g(0) = 0$. Then, a redundant dictionary of atomic functions can be generated as

$$a_\beta(\tau) := \frac{1}{\sqrt{s}} g\left(\frac{\tau - u}{s}\right) e^{i\xi \tau}$$

where $\beta = (s, \xi, u)$ gathers resp. scale, frequency modulation, and transition.
Greedy Pursuit in Signal Processing

Matching Pursuit

1. Initialize: \( x^{(0)} := 0 \).
2. for \( t = 0, \ldots, T \) do
   - Compute \( \beta_t := \arg \max_{\beta} |\langle r^{(t)}, a_{\beta} \rangle| \)
   - Update \( x^{(t+1)} := \sum_{\ell=0}^{t-1} \langle r^{(\ell)}, a_{\beta_{\ell}} \rangle a_{\beta_{\ell}} + \langle r^{(t)}, a_{\beta_t} \rangle a_{\beta_t} \) \( x^{(t)} \)
3. end
Matching Pursuit - Another View

\[ f(x) := \frac{1}{2} \| x - b \|_n^2 \]

\[ r^{(t)} = -\nabla f(x^{(t)}) \]

1. Initialize: \( x^{(0)} := 0 \).
2. for \( t = 0, \ldots, T \) do
   - Compute \( \beta_t := \arg \max_{\beta} |\langle \nabla f(x^{(t)}), a_{\beta} \rangle| \)
   - Update \( x^{(t+1)} := x^{(t)} - \langle \nabla f(x^{(t)}), a_{\beta_t} \rangle a_{\beta_t} \)
3. end
Matching Pursuit and Frank-Wolfe: Two Cousins

Matching Pursuit

Initialize $\mathbf{x}^{(0)}$

for $t = 0, \ldots, T$ do

- $\beta_t := \arg \max_\beta |\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{a}_\beta \rangle|$
- $\mathbf{x}^{(t+1)} := \mathbf{x}^{(t)} - \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{a}_{\beta_t} \rangle \mathbf{a}_{\beta_t}$

end

[ Mallat, Zhang, 1993 ]

Frank-Wolfe

Initialize $\mathbf{x}^{(0)}$

for $t = 0, \ldots, T$ do

- $\mathbf{a}_{\beta_t} := \text{LMO}_D(\nabla f(\mathbf{x}^{(t)}))$
- $\mathbf{x}^{(t+1)} := (1 - \gamma)\mathbf{x}^{(t)} + \gamma \mathbf{a}_{\beta_t}$

end

[ Frank, Wolfe, 1956 ]
Matching Pursuit and Frank-Wolfe
Two Cousins

Similarities

- Optimization over Atomic Sets
- Adds one Atom at a Time
- one LMO-call per iteration

Differences

- slightly different Step-Length / Direction
- Theory Guarantees: Optimization Accuracy for FW / Recovery for (O)MP
Application: Speech Processing

In Figure 12.14(b), the two chirps whose frequencies increase and decrease linearly are decomposed in many Gabor atoms. To improve the representation of signals with frequency and scale parameters that are used to characterize the signal structure, Matching pursuits in Gabor dictionaries provide sparse representation of oscillatory signals, which have for example been carried in cognitive neurophysiology for the analysis of gamma oscillations in Electro-Encephalograms (EEG) signals which are highly non-stationary. In the time-frequency plane, Gabor functions are localized around an oriented segment in the time-frequency plane. Such atoms can represent more efficiently progressive frequency variations of the signal.

Matching pursuits in Gabor dictionaries provide sparse representation of oscillatory signals, which have for example been carried in cognitive neurophysiology for the analysis of gamma oscillations in Electro-Encephalograms (EEG) signals which are highly non-stationary. In the time-frequency plane, Gabor functions are localized around an oriented segment in the time-frequency plane. Such atoms can represent more efficiently progressive frequency variations of the signal.

The dark blobs of various sizes are the Wigner-Ville distributions of a Gabor function with frequency and scale parameters that are used to characterize the signal structure selected by the matching pursuit. Figure 12.15: Speech recording of the word “greasy” sampled at 16 kHz. The signal has 5782 samples, and the sound recovered from these atoms is noise whose time-frequency energy is spread over a high-frequency interval. Most of the energy is characterized by a few time-frequency atoms. For example, studies have shown that the period of seizure initiation by analyzing the selected atom properties [25, 347] allows for the prediction of epilepsy patterns [21] all of which are composed in many Gabor atoms. To improve the representation of signals, Matching pursuits in Gabor dictionaries provide sparse representation of oscillatory signals, which have for example been carried in cognitive neurophysiology for the analysis of gamma oscillations in Electro-Encephalograms (EEG) signals which are highly non-stationary. In the time-frequency plane, Gabor functions are localized around an oriented segment in the time-frequency plane. Such atoms can represent more efficiently progressive frequency variations of the signal.
Orthogonal Matching Pursuit: going further

Idea:

- Before adding a new atom, fully optimize over the previous atoms
- Error is orthogonal to the subspace spanned by the previous atoms
- Similar idea to fully corrective FW (simplicial decomposition), and FW with away steps
Orthogonal Matching Pursuit: going further

1. Initialize: $\mathbf{x}^{(0)} := \mathbf{0}$.
2. for $t = 0, \ldots, T$ do
   - Compute $\beta_t = \arg \max_{\beta} \left| \langle (\mathbf{I} - \Pi^{(t)}) \left( \nabla f(\mathbf{x}^{(t)}) \right), \mathbf{a}_{\beta} \rangle \right|$
   - Update $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \langle (\mathbf{I} - \Pi^{(t)}) \left( \nabla f(\mathbf{x}^{(t)}) \right), \mathbf{a}_{\beta^t} \rangle \mathbf{a}_{\beta^t}$
3. end
Orthogonal Matching Pursuit: theoretical results

Theorem (Support recovery) Assume that $T < 1/(8\sqrt{2} \text{Coherence}(\mathcal{A})) - 1$. For any signal $\mathbf{x}$, OMP generates an $T$-term approximant $\mathbf{x}^{(T)}$ which satisfies

$$\|\mathbf{b} - \mathbf{x}^{(T)}\|_2 \leq 8\sqrt{T}\|\mathbf{b} - \mathbf{x}^{\text{Opt}}\|_2$$

where $\mathbf{x}^{\text{Opt}}$ is the optimal $T$-term approximation of $\mathbf{x}$.

[ Gilbert et al. 2003 ]
(O)MP and Frank-Wolfe
Two Cousins

<table>
<thead>
<tr>
<th>Similarities</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>✤ Optimization over Atomic Sets</td>
<td>✤ slightly different</td>
</tr>
<tr>
<td>✤ Adds one Atom at a Time</td>
<td>✤ Step-Length / Direction</td>
</tr>
<tr>
<td>✤ one LMO-call per iteration</td>
<td>✤ Theory Guarantees:</td>
</tr>
<tr>
<td></td>
<td>✤ Optimization Accuracy for FW</td>
</tr>
<tr>
<td></td>
<td>✤ / Recovery for (O)MP</td>
</tr>
</tbody>
</table>
Greedy Optimization Meets Frank-Wolfe

**Convex optimization**
methods applied to

\[
\min_{x \in \mathcal{D}} \| x - b \|^2_2
\]

Frank-Wolfe

fully corrective
Frank-Wolfe

\[i := \arg \max_i |\nabla f(x)_i|\]

\(\ell_1\)-ball

**Signal processing**
sparse/direct recovery methods

recover a sparse \(x\) from a noisy measurement \(b\)

Matching Pursuit

OMP

**Problem**
Approximate a given function \(b\) by a sparse convex combination \(x\) of functions in dictionary \(A = \{a_1, a_2, \ldots, a_p, \ldots\}\)

**min**

\[x \in \mathcal{D}\]

\(D_k = \text{conv}(A)\)

**References**
Tutorial Outline

**Part I:** (Martin Jaggi)
- Frank-Wolfe basics and history
- convergence analysis and geometry independence
- lower bounds
- demo and applications for atomic domains
- faster rates under additional assumptions

**Part II:** (Zaid Harchaoui)
- Greedy Optimization and Frank-Wolfe
  - Frank-Wolfe methods for:
    - composite optimization
    - non-smooth optimization
    - block-structured optimization
Composite Frank-Wolfe

Regularized formulation (Lagrangian / penalized)

$$\min_x f(x) + \lambda \|x\| A$$
Composite Frank-Wolfe
Atomic Norms

\[ \|x\|_A := \inf_{r \geq 0} \{ r \mid x \in r \cdot \text{conv}(A) \} =: D \]
Composite Frank-Wolfe

Setup

Epigraph Form

\[
\min_{z=[x,r]} F(z) := f(x) + \lambda r
\]
Composite Frank-Wolfe Algorithm

1. Initialize: \( z^{(0)} = [0, \bar{R}] \), with \( \bar{R} > 0 \) s.t. \( \|x^*\|_A \leq \bar{R} \)
2. for \( t = 0, \ldots, T \) do
   - Compute \( \bar{z}^{(t)} = \text{LMO}(\nabla F(z^{(t)})) \)
   - Update
     \[
     z^{(t+1)} := \arg\min_{z \in C^{(t)}} F(z)
     \]
     \[
     C^{(t)} := \text{conv}\{0, z^{(t)}, \bar{R} \cdot \bar{z}^{(t)}\}
     \]
3. end
Composite Frank-Wolfe Algorithm

1. Initialize: \( z^{(0)} = [0, \overline{R}] \), with \( \overline{R} > 0 \) s.t. \( \|x^*\|_A \leq \overline{R} \)

2. for \( t = 0, \ldots, T \) do
   - Compute \( \bar{z}^{(t)} = \text{LMO}(\nabla F(z^{(t)})) \)
   - Update
     \[
     z^{(t+1)} := \arg \min_{z \in C^{(t)}} F(z)
     \]
     \[
     C^{(t)} := \text{conv}\{0, z^{(t)}, \overline{R} \cdot \bar{z}^{(t)}\}
     \]

3. end

\[
\bar{z}^{(t)} := \overline{R} \cdot [\nabla f(x^{(t)}), 1]
\]
Composite Frank-Wolfe (Conditional Gradient) Algorithm

1. Initialize: \( z^{(0)} = [0, \overline{R}] \), with \( \overline{R} > 0 \) s.t. \( \|x^*\|_A \leq \overline{R} \)
2. for \( t = 0, \ldots, T \) do
   - Compute \( \bar{z}^{(t)} = \text{LMO}(\nabla F(z^{(t)})) \)
   - Update
     \[
     z^{(t+1)} = \alpha_{t+1} \bar{z}^{(t)} + \beta_{t+1} z^{(t)}
     \]
     \[
     (\alpha_{t+1}, \beta_{t+1}) = \arg \min_{\alpha, \beta \geq 0; \alpha + \beta \leq 1} F(\alpha \bar{z}^{(t)} + \beta z^{(t)})
     \]
3. end

[Harchaoui et al., 2012; Zhang et al., 2012]
Composite Frank-Wolfe
Convergence Rate

- **Smoothness** $f$ is convex continuously differentiable with Lipschitz constant $L$.
- **Effective domain** There exists $R < 1$ such that $\|x\| \leq r$ and $r + f(x) < f(0)$ imply that $r \leq R$

**Theorem** (Convergence Rate) For each $t \geq 2$, the iterates $\{z^{(t)}\}$ of the Composite FW algorithm satisfy

$$F(z_t) - F^* \leq \frac{8LR^2}{t + 1}$$

Theoretical convergence rate is independent of $\bar{R}$

[Harchaoui et al., 2012, 2014]
Composite Frank-Wolfe
Accelerations

\[
\mathbf{z}^{(t+1)} = \arg \min_{\mathbf{z} \in \mathcal{C}_t} F(\mathbf{z})
\]

\[\mathcal{C}_t = \begin{cases}
\text{Conv}\{0; \bar{R}\bar{z}_0, \ldots, \bar{R}\bar{z}_t\}, & t \leq M,
\text{Conv}\{0; \mathbf{z}^{(t-M+1)}, \ldots, \mathbf{z}^{(t)}; \bar{R}\bar{z}^{(t-M+1)}, \ldots, \bar{R}\bar{z}^{(t)}\}, & t > M.
\end{cases}\]

Aka \textbf{Fully-corrective acceleration} [Shalev-Shwartz et al., 2012]

\[\text{Conic}\{\bar{z}_0, \ldots, \bar{z}_t\}, \quad t \leq M, \quad \text{Conic}\{\mathbf{z}^{(t-M+1)}, \ldots, \mathbf{z}_t; \bar{z}^{(t-M+1)}, \ldots, \bar{z}_t\}, \quad t > M.\]

Recovers \textbf{Atom-Descent} of [Dudik et al., 2012]
Composite Frank-Wolfe

Acceleration by local optimization

\[
\mathbf{z}^{(t+1)} = \arg \min_{u_1^{(t)}, \ldots, u_p^{(t)} \in \mathcal{C}_t} F(u_1^{(t)}, \ldots, u_p^{(t)})
\]

- Nuclear-norm: matrix factorization, alternating optimization
- Factorized-matrix norms: alternating optimization
- Wavelet-decomposition norm: tree-based representation
- and many others
Non-Smooth Optimization

using Frank-Wolfe

Assume $f$ is non-smooth, and that it admits the representation

$$f(x) = \max_{y \in \Delta, y \in \text{dom}\Omega} y^T x$$

See [Bertsekas, 1998; Nesterov, 2005]
Non-Smooth Optimization using Frank-Wolfe

Assume \( f \) is non-smooth. Then it can be smoothed into \( f^\mu \)

\[
 f^\mu(x) = \max_{y \in \Delta, y \in \text{dom} \Omega} y^T x - \mu \Omega(y)
\]

\( \Omega(\cdot) \) is assumed to be strongly convex over \( \Delta \)

See [Bertsekas, 1998; Nesterov, 2005]
Non-Smooth Optimization using Frank-Wolfe

1. Set accuracy $\epsilon > 0$, smoothing parameter $\mu(\epsilon)$
2. for $t = 0, \ldots, T(\epsilon)$ do
   - Compute $s^{(t)} = \text{LMO}(\nabla f^\mu(x^{(t)}))$
   - Update $x^{(t+1)} = (1 - \gamma)x^{(t)} + \gamma s^{(t)}$
3. end
4. Return $x^{(T(\epsilon)+1)}$

**Theorem** (Accuracy) Set a optimization accuracy $\epsilon$. Then, there exists a smoothing parameter $\mu(\epsilon) = O(\epsilon)$ such that after $T(\epsilon) = O(D^2/\mu\epsilon)$ we have

$$f(x^{(T)}) - f^* \leq \epsilon$$

[ Cox et al. 2013; Pierucci et al., 2013; Lan, 2013 ]
Non-Smooth Composite Optimization using Composite Frank-Wolfe

Composite Frank-Wolfe can be used for non-smooth $f$

**Theorem (Accuracy)** Set a optimization accuracy $\epsilon$. Then, there exists a smoothing parameter $\mu(\epsilon) = O(\epsilon)$ such that after $T(\epsilon) = O(R^2/\mu \epsilon)$ we have

$$F(z^{(T)}) - F^* \leq \epsilon$$
Block-Coordinate Frank-Wolfe

Cartesian Product Domain

\[ \mathcal{D} = \mathcal{D}^{(1)} \times \ldots \times \mathcal{D}^{(n)} \subseteq \mathbb{R}^m \]

\[ \mathcal{D}^{(i)} \subseteq \mathbb{R}^{m_i} \]

\[ \sum_{i=1}^{n} m_i = m \]

**Algorithm 2: Block-Coordinate Frank-Wolfe**

Let \( \mathbf{x}^{(0)} \in \mathcal{D} \)

for \( t = 0 \ldots \infty \) do

Pick \( i \in u.a.r. [n] \)

Compute \( \mathbf{s}_{(i)} := LMO_{\mathcal{D}^{(i)}}(\nabla_{(i)}f(\mathbf{x}^{(t)})) \)

Let \( \gamma := \frac{2n}{t+2n} \)

Update \( \mathbf{x}^{(t+1)}_{(i)} := (1 - \gamma)\mathbf{x}^{(t)}_{(i)} + \gamma \mathbf{s}_{(i)} \)

end

**Convergence:**

\[ \mathbb{E}[f(\mathbf{x}^{(t)})] - f^* \leq \frac{2n}{t+2n}(2C_{f}^{\text{prod}} + f(\mathbf{x}^{(0)}) - f^*) \]

[ Lacoste-Julien et al. 2013 ]
Conclusion

* Frank-Wolfe is a simple yet powerful, geometry-independent, algorithm tailored for optimization over atomic sets

* Greedy Optimization and Frank-Wolfe belong to the same family of algorithms

* New FW-type algorithms can be designed using FW as working horse
Related topics and perspectives

✦ Atomic-norm and LMO design

✦ Beyond Linear Minimization Oracles (local LMOs, approximate LMOs, etc.)

✦ Geometry-independent stochastic approximation
Tutorial Website

- Sample Code
- Bibliography
- Tutorial Slides
- Topics not covered

https://sites.google.com/site/FrankWolfeGreedyTutorial
Thanks