Learning with Submodular Functions: A Convex Optimization Perspective

Francis Bach
Sierra project-team, INRIA - Ecole Normale Supérieure

Thanks to R. Jenatton, J. Mairal, G. Obozinski
December 2011
Convex optimization with combinatorial structure

• Supervised learning
  – Minimize regularized empirical risk from data \((x_i, y_i), i = 1, \ldots, n:\)
    \[
    \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \Omega(f)
    \]
  – \(\mathcal{F}\) is often a vector space, formulation often convex

• Introducing discrete structures within a vector space framework
  – Trees, graphs, etc.
  – Many different approaches (e.g., stochastic processes)

• Submodularity allows the incorporation of discrete structures
Outline

• **Submodular functions**
  – Links with convexity through Lovász extension
  – Optimization on submodular polyhedra

• **Structured sparsity-inducing norms**
  – Relaxation of the penalization of supports
  – Examples
  – Unified algorithms and analysis

• **Approximate submodular function minimization**
Outline

- **Submodular functions**
  - Links with convexity through Lovász extension
  - Optimization on submodular polyhedra

- **Structured sparsity-inducing norms**
  - Relaxation of the penalization of supports
  - Examples
  - Unified algorithms and analysis

- **Approximate submodular function minimization**

- (for more details, see tutorial / technical report on web page)
Submodular functions

• $F : 2^V \rightarrow \mathbb{R}$ is submodular if and only if

\[
\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cap B) + F(A \cup B)
\]

\[
\iff \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}
\]
Submodular functions

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

  $\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cap B) + F(A \cup B)$

  $\iff \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A)$ is non-increasing

- **Intuition 1**: defined like concave functions ("diminishing returns")
  - Example: $F : A \mapsto g(\text{Card}(A))$ is submodular if $g$ is concave
Submodular functions

- \( F : 2^V \to \mathbb{R} \) is submodular if and only if

\[
\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cap B) + F(A \cup B)
\]
\[
\iff \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}
\]

- **Intuition 1**: defined like concave functions (“diminishing returns”)
  - Example: \( F : A \mapsto g(\text{Card}(A)) \) is submodular if \( g \) is concave

- **Intuition 2**: behave like convex functions
  - Polynomial-time minimization, conjugacy theory
Submodular functions

- $F : 2^V \to \mathbb{R}$ is **submodular** if and only if

\[
\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cap B) + F(A \cup B)
\]

\[\iff \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}\]

- **Intuition 1**: defined like concave functions ("diminishing returns")
  - Example: $F : A \mapsto g(\text{Card}(A))$ is submodular if $g$ is concave

- **Intuition 2**: behave like convex functions
  - Polynomial-time minimization, conjugacy theory

- Used in several areas of signal processing and machine learning
  - Total variation/graph cuts (Chambolle, 2005; Boykov et al., 2001)
  - Optimal design (Krause and Guestrin, 2005)
Submodular functions - Examples

- Concave functions of the cardinality: $g(|A|)$

- Cuts

- Entropies
  - $H((X_k)_{k \in A})$ from $p$ random variables $X_1, \ldots, X_p$
  - Gaussian variables $H((X_k)_{k \in A}) \propto \log \det \Sigma_{AA}$
  - Functions of eigenvalues of sub-matrices

- Network flows
  - Efficient representation for set covers

- Rank functions of matroids
Submodular functions - Lovász extension

- Subsets may be identified with elements of \( \{0, 1\}^p \)

- Given any set-function \( F \) and \( w \) such that \( w_{j_1} \geq \cdots \geq w_{j_p} \), define:

\[
f(w) = \sum_{k=1}^p w_{j_k} [F(\{j_1, \ldots, j_k\}) - F(\{j_1, \ldots, j_{k-1}\})]
\]

- If \( w = 1_A \), \( f(w) = F(A) \) \( \Rightarrow \) extension from \( \{0, 1\}^p \) to \( \mathbb{R}^p \)
- \( f \) is piecewise affine and positively homogeneous

- \( F \) is submodular if and only if \( f \) is convex (Lovász, 1982)
  - Minimizing \( f(w) \) on \( w \in [0, 1]^p \) equivalent to minimizing \( F \) on \( 2^V \)
Submodular functions - Submodular polyhedra

- Submodular polyhedron: \( P(F) = \{ s \in \mathbb{R}^p, \forall A \subset V, s(A) \leq F(A) \} \)

- Base polyhedron: \( B(F) = P(F) \cap \{ s(V) = F(V) \} \)
Submodular functions - Submodular polyhedra

- Submodular polyhedron: \( P(F) = \{ s \in \mathbb{R}^p, \forall A \subset V, s(A) \leq F(A) \} \)

- Base polyhedron: \( B(F) = P(F') \cap \{ s(V) = F(V) \} \)

- **Link with Lovász extension** (Edmonds, 1970; Lovász, 1982):
  - if \( w \in \mathbb{R}^p_+ \), then \( \max_{s \in P(F')} w^\top s = f(w) \)
  - if \( w \in \mathbb{R}^p \), then \( \max_{s \in B(F)} w^\top s = f(w) \)

- Maximizer obtained by greedy algorithm:
  - Sort the components of \( w \), as \( w_{j_1} \geq \cdots \geq w_{j_p} \)
  - Set \( s_{j_k} = F(\{j_1, \ldots, j_k\}) - F(\{j_1, \ldots, j_{k-1}\}) \)

- Other operations on submodular polyhedra (see, e.g., Bach, 2011)
Submodular functions - Optimization

- Submodular function minimization in $O(p^6)$
  - Schrijver (2000); Iwata et al. (2001); Orlin (2009)

- Efficient active set algorithm with no complexity bound
  - Based on the efficient computability of the support function
  - Fujishige and Isotani (2011); Wolfe (1976)

- Special cases with faster algorithms: cuts, flows

- Active area of research
  - Stobbe and Krause (2010)
  - Jegelka, Lin, and Bilmes (2011)
Separable optimization on base polyhedron

- Assume each $\psi_k$ is a strictly convex function $\mathbb{R} \rightarrow \mathbb{R}$

- **Proposition**: the two following problems are dual to each other

\[
\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w) \\
\max_{s \in B(F)} \sum_{k \in V} -\psi_k(-s_k)
\]
Separable optimization on base polyhedron

• Assume each $\psi_k$ is a strictly convex function $\mathbb{R} \rightarrow \mathbb{R}$

• Proposition: the two following problems are dual to each other

$$\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$$

$$\max_{s \in B(F)} \sum_{k \in V} -\psi_k(-s_k)$$

• Proposition (Chambolle and Darbon, 2009): let $w^*$ be the solution of $\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$. Then, for $\alpha \in \mathbb{R}$,

$$\min_{A \subset V} F(A) + \sum_{j \in A} \psi'_k(\alpha)$$

has minimal minimizer $\{w^* > \alpha\}$ and maximal minimizer $\{w^* \geq \alpha\}$
From convex to combinatorial optimization

- Solving $\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$ to solve $\min_{A \subseteq V} F(A)$
  - Thresholding solutions $w$ at zero if $\forall k \in V, \psi'_k(0) = 0$
  - For quadratic functions $\psi_k(w_k) = \frac{1}{2}w_k^2$, equivalent to projecting 0 on $B(F)$ (Fujishige, 2005)
  - minimum-norm-point algorithm (Fujishige and Isotani, 2011)
From convex to combinatorial optimization and vice-versa...

- Solving $\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$ to solve $\min_{A \subseteq V} F(A)$
  - Thresholding solutions $w$ at zero if $\forall k \in V, \psi'_k(0) = 0$
  - For quadratic functions $\psi_k(w_k) = \frac{1}{2}w_k^2$, equivalent to projecting 0 on $B(F)$ (Fujishige, 2005)
  - minimum-norm-point algorithm (Fujishige and Isotani, 2011)

- Solving $\min_{A \subseteq V} F(A) - t(A)$ to solve $\min_{w \in \mathbb{R}^p} \sum_{k \in V} \psi_k(w_k) + f(w)$
  - General decomposition strategy (Groenevelt, 1991)
  - Efficient only when submodular minimization is efficient
Outline

• Submodular functions
  – Links with convexity through Lovász extension
  – Optimization on submodular polyhedra

• Structured sparsity-inducing norms
  – Relaxation of the penalization of supports
  – Examples
  – Unified algorithms and analysis

• Approximate submodular function minimization
Sparsity in supervised machine learning

- Observed data \((x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1, \ldots, n\)
  - Response vector \(y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n\)
  - Design matrix \(X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p}\)

- Regularized empirical risk minimization:
  \[
  \min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^\top x_i) + \lambda \Omega(w) = \min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda \Omega(w)
  \]

- Norm \(\Omega\) to promote sparsity
  - square loss + \(\ell_1\)-norm \(\Rightarrow\) basis pursuit in signal processing (Chen et al., 2001), Lasso in statistics/machine learning (Tibshirani, 1996)
  - Proxy for interpretability
  - Allow high-dimensional inference: \(\log p = O(n)\)
Sparsity in unsupervised machine learning

- **Multiple responses/signals** $y = (y^1, \ldots, y^k) \in \mathbb{R}^{n \times k}$

- **Dictionary learning**
  
  - Learn $X = (x^1, \ldots, x^p) \in \mathbb{R}^{n \times p}$ such that $\forall j$, $\|x^j\|_2 \leq 1$

  \[
  \min_{X=(x^1, \ldots, x^p)} \min_{w^1, \ldots, w^k \in \mathbb{R}^p} \sum_{j=1}^k \left\{ L(y^j, Xw^j) + \lambda \Omega(w^j) \right\}
  \]

  - Olshausen and Field (1997); Elad and Aharon (2006)

- **sparse PCA**: replace $\|x^j\|_2 \leq 1$ by $\Theta(x^j) \leq 1$
Why structured sparsity?

● Interpretability
  – Structured dictionary elements (Jenatton et al., 2009b)
  – Dictionary elements “organized” in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)
Structured sparse PCA (Jenatton et al., 2009b)

- Unstructured sparse PCA $\Rightarrow$ many zeros do not lead to better interpretability
Structured sparse PCA (Jenatton et al., 2009b)

- Unstructured sparse PCA \(\Rightarrow\) many zeros do not lead to better interpretability
Structured sparse PCA (Jenatton et al., 2009b)

- Enforce selection of convex nonzero patterns ⇒ robustness to occlusion in face identification
Structured sparse PCA (Jenatton et al., 2009b)

- Enforce selection of **convex** nonzero patterns $\Rightarrow$ robustness to occlusion in face identification
Why structured sparsity?

- Interpretability
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements “organized” in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)
Modelling of text corpora (Jenatton et al., 2010)
Why structured sparsity?

- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements “organized” in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)
Why structured sparsity?

- **Interpretability**
  - Structured dictionary elements (Jenatton et al., 2009b)
  - Dictionary elements “organized” in a tree or a grid (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)

- **Stability and identifiability**
  - Optimization problem $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda \|w\|_1$ is unstable
  - “Codes” $w^j$ often used in later processing (Mairal et al., 2009)

- **Prediction or estimation performance**
  - When prior knowledge matches data (Haupt and Nowak, 2006; Baraniuk et al., 2008; Jenatton et al., 2009a; Huang et al., 2009)

- **Numerical efficiency**
  - Non-linear variable selection with $2^p$ subsets (Bach, 2008)
$\ell_1$-norm = convex envelope of cardinality of support

- Let $w \in \mathbb{R}^p$. Let $V = \{1, \ldots, p\}$ and $\text{Supp}(w) = \{j \in V, w_j \neq 0\}$
- **Cardinality of support**: $\|w\|_0 = \text{Card}(\text{Supp}(w))$
- Convex envelope = largest convex lower bound (see, e.g., Boyd and Vandenberghe, 2004)

$\ell_1$-norm = convex envelope of $\ell_0$-quasi-norm on the $\ell_\infty$-ball $[-1, 1]^p$
Convex envelopes of general functions of the support (Bach, 2010)

- Let $F : 2^V \to \mathbb{R}$ be a set-function
  - Assume $F$ is non-decreasing (i.e., $A \subseteq B \Rightarrow F(A) \leq F(B)$)
  - Explicit prior knowledge on supports (Haupt and Nowak, 2006; Baraniuk et al., 2008; Huang et al., 2009)

- Define $\Theta(w) = F(Supp(w))$: How to get its convex envelope?
  1. Possible if $F$ is also submodular
  2. Allows unified theory and algorithm
  3. Provides new regularizers
Submodular functions and structured sparsity

- Let $F : 2^V \rightarrow \mathbb{R}$ be a non-decreasing submodular set-function

- **Proposition**: the convex envelope of $\Theta : w \mapsto F(\text{Supp}(w))$ on the $\ell_\infty$-ball is $\Omega : w \mapsto f(|w|)$ where $f$ is the Lovász extension of $F$
Submodular functions and structured sparsity

- Let $F : 2^V \rightarrow \mathbb{R}$ be a **non-decreasing submodular set-function**

- **Proposition:** the convex envelope of $\Theta : w \mapsto F(Supp(w))$ on the $\ell_\infty$-ball is $\Omega : w \mapsto f(|w|)$ where $f$ is the Lovász extension of $F$

- **Sparsity-inducing properties:** $\Omega$ is a **polyhedral norm**

$F$ at

- $A$ if stable if for all $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$
- With probability one, stable sets are the only allowed patterns
Polyhedral unit balls

$F(A) = |A|$  
$\Omega(w) = \|w\|_1$

$F(A) = \min\{|A|, 1\}$  
$\Omega(w) = \|w\|_\infty$

$F(A) = |A|^{1/2}$  
all possible extreme points

$F(A) = 1_{\{A \cap \{1\} \neq \emptyset\}} + 1_{\{A \cap \{2,3\} \neq \emptyset\}}$  
$\Omega(w) = |w_1| + \|w_{\{2,3\}}\|_\infty$

$F(A) = 1_{\{A \cap \{1,2,3\} \neq \emptyset\}} + 1_{\{A \cap \{2,3\} \neq \emptyset\}} + 1_{\{A \cap \{3\} \neq \emptyset\}}$  
$\Omega(w) = \|w\|_\infty + \|w_{\{2,3\}}\|_\infty + |w_3|$
Submodular functions and structured sparsity

Examples

- From $\Omega(w)$ to $F(A)$: provides new insights into existing norms
  - Grouped norms with overlapping groups (Jenatton et al., 2009a)
    \[
    \Omega(w) = \sum_{G \in G} \|w_G\|_{\infty}
    \]
  - $\ell_1-\ell_\infty$ norm $\Rightarrow$ sparsity at the group level
  - Some $w_G$'s are set to zero for some groups $G$
    \[
    (\text{Supp}(w))^c = \bigcup_{G \in \mathcal{H}} G \text{ for some } \mathcal{H} \subseteq G
    \]
Submodular functions and structured sparsity

Examples

- From $\Omega(w)$ to $F(A)$: provides new insights into existing norms
  - Grouped norms with overlapping groups (Jenatton et al., 2009a)
    $$\Omega(w) = \sum_{G \in \mathcal{G}} \|w_G\|_\infty \implies F(A) = \text{Card}\left(\{G \in \mathcal{G}, G \cap A \neq \emptyset\}\right)$$
  - $\ell_1-\ell_\infty$ norm $\Rightarrow$ sparsity at the group level
  - Some $w_G$'s are set to zero for some groups $G$
    $$(\text{Supp}(w))^c = \bigcup_{G \in \mathcal{H}} G \text{ for some } \mathcal{H} \subseteq \mathcal{G}$$
  - Justification not only limited to allowed sparsity patterns
Selection of contiguous patterns in a sequence

- Selection of contiguous patterns in a sequence

- $\mathcal{G}$ is the set of blue groups: any union of blue groups set to zero leads to the selection of a contiguous pattern
Selection of contiguous patterns in a sequence

- Selection of contiguous patterns in a sequence

- $\mathcal{G}$ is the set of blue groups: any union of blue groups set to zero leads to the selection of a contiguous pattern

- $\sum_{G \in \mathcal{G}} \|w_G\|_\infty \Rightarrow F(A) = p - 2 + \text{Range}(A)$ if $A \neq \emptyset$
Extensions of norms with overlapping groups

• Selection of rectangles (at any position) in a 2-D grids

• Hierarchies
Submodular functions and structured sparsity

Examples

- **From** $\Omega(w)$ **to** $F(A)$: provides new insights into existing norms
  - Grouped norms with **overlapping** groups (Jenatton et al., 2009a)
    
    $\Omega(w) = \sum_{G \in \mathcal{G}} \|w_G\|_{\infty} \Rightarrow F(A) = \text{Card}(\{G \in \mathcal{G}, G \cap A \neq \emptyset\})$
  
  - Justification not only limited to allowed sparsity patterns
Submodular functions and structured sparsity

Examples

- **From $\Omega(w)$ to $F(A)$**: provides new insights into existing norms
  - Grouped norms with overlapping groups (Jenatton et al., 2009a)
    \[
    \Omega(w) = \sum_{G \in \mathcal{G}} \|w_G\|_\infty \quad \Rightarrow \quad F(A) = \text{Card}(\{G \in \mathcal{G}, G \cap A \neq \emptyset\})
    \]
  - Justification not only limited to allowed sparsity patterns

- **From $F(A)$ to $\Omega(w)$**: provides new sparsity-inducing norms
  - $F(A) = g(\text{Card}(A)) \quad \Rightarrow \quad \Omega$ is a combination of order statistics
  - Non-factorial priors for supervised learning: $\Omega$ depends on the eigenvalues of $X_A^\top X_A$ and not simply on the cardinality of $A$
Unified optimization algorithms

- **Polyhedral norm** with exponentially many faces and extreme points
  - Not suitable for linear programming toolboxes

- **Subgradient** ($w \mapsto \Omega(w)$ non-differentiable)
  - Subgradient may be obtained in polynomial time $\Rightarrow$ too slow
Unified optimization algorithms

- **Polyhedral norm** with exponentially many faces and extreme points
  - Not suitable for linear programming toolboxes

- **Subgradient** \((w \mapsto \Omega(w))\) non-differentiable
  - Subgradient may be obtained in polynomial time \(\Rightarrow\) too slow

- **Proximal methods** (see, e.g., Beck and Teboulle, 2009; Bach, Jenatton, Mairal, and Obozinski, 2011)
  - \(\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda \Omega(w)\): differentiable + non-differentiable
  - Efficient when \((P):\ min_{w \in \mathbb{R}^p} \frac{1}{2}\|w - v\|_2^2 + \lambda \Omega(w)\) is “easy”

- **The proximal problem** \((P)\) is equivalent to a sequence of submodular function minimizations
  - Decomposition strategy (Groenevelt, 1991) or min-norm-point
Comparison of optimization algorithms

- Synthetic example with $p = 1000$ and $F(A) = |A|^{1/2}$

- ISTA: proximal method

- FISTA: accelerated variant (Beck and Teboulle, 2009)
Extensions

• **Unified statistical analysis** (Bach, 2010)
  – support recovery and estimation consistency

• **Extension to symmetric submodular functions**
  – Shaping level sets (Bach, 2011)

• **Avoiding artefacts linked with $\ell_\infty$-norms**
  – See poster at this workshop (Obozinski and Bach, 2011)

• **Generalization to other set-functions**
  – See same poster at this workshop (Obozinski and Bach, 2011)
Polyhedral unit balls

\[ F(A) = |A| \]
\[ \Omega(w) = \|w\|_1 \]

\[ F(A) = \min\{ |A|, 1 \} \]
\[ \Omega(w) = \|w\|_\infty \]

\[ F(A) = |A|^{1/2} \]
all possible extreme points

\[ F(A) = 1 \{ A \cap \{1\} \neq \emptyset \} + 1 \{ A \cap \{2,3\} \neq \emptyset \} \]
\[ \Omega(w) = |w_1| + \|w_{\{2,3\}}\|_\infty \]

\[ F(A) = 1 \{ A \cap \{1,2,3\} \neq \emptyset \} \]
\[ + 1 \{ A \cap \{2,3\} \neq \emptyset \} + 1 \{ A \cap \{3\} \neq \emptyset \} \]
\[ \Omega(w) = \|w\|_\infty + \|w_{\{2,3\}}\|_\infty + |w_3| \]
Outline

• **Submodular functions**
  – Links with convexity through Lovász extension
  – Optimization on submodular polyhedra

• **Structured sparsity-inducing norms**
  – Relaxation of the penalization of supports
  – Examples
  – Unified algorithms and analysis

• **Approximate submodular function minimization**
Approximate submodular function minimization

- For most machine learning applications, no need to obtain exact minimum
Approximate submodular function minimization

• For most machine learning applications, no need to obtain exact minimum

• Assume (wlog.) that $\forall k \in V, F(\{k\}) \geq 0$ and $F(V \backslash \{k\}) \geq F(V)$

• Denote $D^2 = \sum_{k \in V} \{ F(\{k\}) + F(V \backslash \{k\}) - F(V) \}$

• Proposition: $t$ iterations of subgradient descent outputs a set $A_t$ (and a certificate of optimality $s_t$) such that

$$F(A_t) - \min_{B \subset V} F(B) \leq F(A_t) - (s_t)_-(V) \leq \frac{Dp^{1/2}}{\sqrt{t}}$$
Approximate quadratic optimization on $B(F)$

- **Goal:**
  \[
  \min_{w \in \mathbb{R}^p} \frac{1}{2} \|w\|_2^2 + f(w) = \max_{s \in B(F)} -\frac{1}{2} \|s\|_2^2
  \]

- Can only maximize linear functions on $B(F)$

- **Two types of “Frank-wolfe” algorithms**

- **1. Active set algorithm ($\leftrightarrow$ min-norm-point)**
  - Sequence of maximizations of linear functions over $B(F)$
    + overheads (affine projections)
  - Finite convergence, but no complexity bounds
Approximate quadratic optimization on $B(F)$

- **Goal:** \( \min_{w \in \mathbb{R}^p} \frac{1}{2} \| w \|_2^2 + f(w) = \max_{s \in B(F)} -\frac{1}{2} \| s \|_2^2 \)

- Can only maximize linear functions on $B(F)$

- **Two types of “Frank-wolfe” algorithms**
  
  1. **Active set algorithm** (\( \Leftrightarrow \) min-norm-point)
     - Sequence of maximizations of linear functions over $B(F)$
     - + overheads (affine projections)
     - Finite convergence, but no complexity bounds
  
  2. **Conditional gradient**
     - Sequence of maximizations of linear functions over $B(F)$
     - Approximate optimality bound
Conditional gradient with line search

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)
Approximate quadratic optimization on $B(F)$

- **Proposition:** $t$ steps of conditional gradient (with line search) outputs $s_t \in B(F)$ and $w_t = -s_t$, such that

\[
\begin{align*}
f(w_t) + \frac{1}{2}\|w_t\|_2^2 - \text{OPT} &\leq f(w_t) + \frac{1}{2}\|w_t\|_2^2 + \frac{1}{2}\|s_t\|_2^2 \leq \frac{2D^2}{t}
\end{align*}
\]
Approximate quadratic optimization on $B(F)$

- **Proposition:** $t$ steps of conditional gradient (with line search) outputs $s_t \in B(F)$ and $w_t = -s_t$, such that

$$f(w_t) + \frac{1}{2} \|w_t\|^2_2 - \text{OPT} \leq f(w_t) + \frac{1}{2} \|w_t\|^2_2 + \frac{1}{2} \|s_t\|^2_2 \leq \frac{2D^2}{t}$$

- **Improved primal candidate through isotonic regression**
  - $f(w)$ is linear on any set of $w$ with fixed ordering
  - May be optimized using isotonic regression (“pool-adjacent-violator”) in $O(n)$ (see, e.g. Best and Chakravarti, 1990)
  - Given $w_t = -s_t$, keep the ordering and reoptimize
Approximate quadratic optimization on $B(F)$

- **Proposition**: $t$ steps of conditional gradient (with line search) outputs $s_t \in B(F)$ and $w_t = -s_t$, such that

$$f(w_t) + \frac{1}{2} \|w_t\|_2^2 - \text{OPT} \leq f(w_t) + \frac{1}{2} \|w_t\|_2^2 + \frac{1}{2} \|s_t\|_2^2 \leq \frac{2D^2}{t}$$

- **Improved primal candidate through isotonic regression**
  - $f(w)$ is linear on any set of $w$ with fixed ordering
  - May be optimized using isotonic regression ("pool-adjacent-violator") in $O(n)$ (see, e.g. Best and Chakravarti, 1990)
  - Given $w_t = -s_t$, keep the ordering and reoptimize

- **Better bound for submodular function minimization?**
From quadratic optimization on $B(F)$ to submodular function minimization

- **Proposition:** If $w$ is $\varepsilon$-optimal for $\min_{w \in \mathbb{R}^p} \frac{1}{2} \|w\|_2^2 + f(w)$, then at least a level set $A$ of $w$ is $(\frac{\sqrt{\varepsilon p}}{2})$-optimal for submodular function minimization.

- If $\varepsilon = \frac{2D^2}{t}$, $\sqrt{\varepsilon p} = \frac{Dp^{1/2}}{\sqrt{2t}}$ $\Rightarrow$ **no provable gains**, but:
  - Bound on the iterates $A_t$ (with additional assumptions)
  - Possible thresholding for acceleration
From quadratic optimization on $B(F)$ to submodular function minimization

- **Proposition**: If $w$ is $\varepsilon$-optimal for $\min_{w \in \mathbb{R}^p} \frac{1}{2} \|w\|_2^2 + f(w)$, then at least a level set $A$ of $w$ is $(\frac{\sqrt{\varepsilon p}}{2})$-optimal for submodular function minimization.

- If $\varepsilon = \frac{2D^2}{t}$, $\frac{\sqrt{\varepsilon p}}{2} = \frac{Dp^{1/2}}{\sqrt{2t}} \Rightarrow$ **no provable gains**, but:
  - Bound on the iterates $A_t$ (with additional assumptions)
  - Possible thresholding for acceleration

- **Lower complexity bound for SFM**
  - **Proposition**: no algorithm that is based only on a sequence of greedy algorithms obtained from linear combinations of bases can improve on the subgradient bound (after $p/2$ iterations).
Simulations on standard benchmark “DIMACS Genrmf-wide”, p = 575

- Submodular function minimization
  - (Left) optimal value minus dual function values \((s_t) - (V)\)
    (in dashed, certified duality gap)
  - (Right) Primal function values \(F(A_t)\) minus optimal value

![Graphs showing the relationship between the number of iterations and the function values](image)

\[
\log_{10}(\min(f) - s_t(V))
\]

\[
\log_{10}(F(A) - \min(F))
\]
Simulations on standard benchmark

- **Separable quadratic optimization**
  - (Left) optimal value minus dual function values $-\frac{1}{2}\|s_t\|^2$ (in dashed, certified duality gap)
  - (Right) Primal function values $f(w_t) + \frac{1}{2}\|w_t\|^2$ minus optimal value (in dashed, before the pool-adjacent-violator correction)
Conclusion

• Submodular functions to encode discrete structures
  – Structured sparsity-inducing norms

• Convex optimization for submodular function optimization
  – Approximate optimization using classical iterative algorithms

• Future work
  – Primal-dual optimization
  – Going beyond linear programming
References


