Efficient Minimization of Decomposable Submodular Functions

Peter Stobbe and Andreas Krause

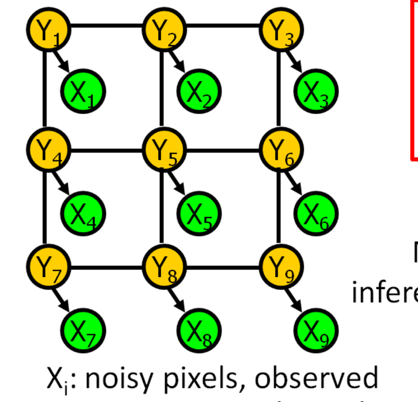
Structures in Optimization

- Convexity useful for continuous functions
  - $f(\mathbf{x}+\theta\mathbf{h}) - f(\mathbf{x}) \leq \theta(f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}))$
  - Minimization tractable if convex.
- Similar Submodular discrete functions:
  - Domain of f: subsets of finite set E
  - $f(A \cup B \cup C) - f(A \cup B) \leq f(A \cup C) - f(A)$
  - Minimization tractable if submodular.

Submodular Minimization Examples

Many important Machine Learning problems require $A^* \in \arg\min_{A \subset E} f(A)$

- Maximum A Posteri Inference of Hidden Variables for Markov Random Field [Kolmogorov et al 2004]



MRF model distribution:
$P(x_1,\ldots,x_n,y_1,\ldots,y_n) = \prod_{(i,j)} \psi_{ij}(y_i,y_j) \prod_i \phi_i(x_i,y_i)$

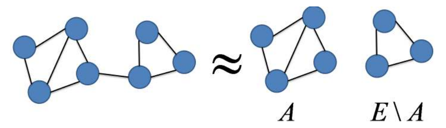
MAP inference
$\arg\max_y P(y|x) \Leftrightarrow \arg\max_y P(x,y)$
$\Leftrightarrow \arg\max_{y_i} \prod_{(i,j)} \psi_{ij}(y_i,y_j) \prod_i \phi_i(x_i,y_i)$
$\Leftrightarrow \boxed{\arg\min_A \sum_{(i,j)} f_{ij}(A) + \sum_i g_i(A)}$ Submodular

X_i: noisy pixels, observed
Y_i: "true" pixels $\in \{+1,-1\}$

Set of Pixels labeled +1 → $A = \{i : y_i = +1\}$ $\quad -\log\phi_i(x_i,y_i) \propto g_i(A) = \begin{cases} a_i & i \notin A \\ b_i & i \in A \end{cases}$

Only pairwise interactions → $-\log\psi_{ij}(y_i,y_j) \propto f_{ij}(A) = \begin{cases} 0 & \{i\} \cap A = \{j\} \cap A \\ c_{ij} & \{i\} \cap A \neq \{j\} \cap A \end{cases}$

- Factorizing random variables.



Mutual Information is submodular: $f(A) = I(X_A; X_{E \setminus A})$

[Narasimhan and Bilmes 2004]

- Clustering [Narasimhan et al 2005]

Algorithms

- General case: $O^*(n^5)$ function evaluations. [Iwata Orlin 2009]
- Min-norm algorithm. Often practical, unknown complexity. [Fujishige et al]
- More efficient special cases:
  - Pairwise potentials
    - eg. MAP for Ising model.
    - Fast mincut algorithms $O^*(n^2)$
  - Queyranne's algorithm.
    - Only symmetric functions $f(A) = f(E \setminus A)$.
    - Running time $O^*(n^3)$
  - Sum of Submodular Functions [Kolmogorov 2010]
    - Each term in sum must be relatively low-order (function of few elements).
  - Our work: Decomposable functions!

Decomposable submodular functions

- Definition:

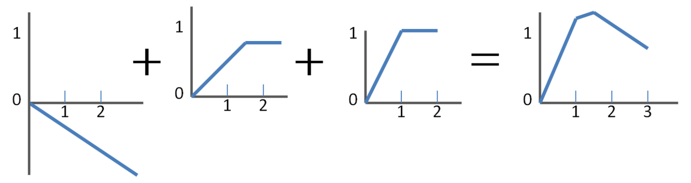
$$f(A) = \sum_j \phi_j(\mathbf{w}_j \cdot \mathbf{e}_A)$$

ϕ_j concave, $\mathbf{w}_j \geq \mathbf{0}$, and $\mathbf{e}_A[k] = \begin{cases} 0 \text{ if } k \notin A \\ 1 \text{ if } k \in A \end{cases}$.

- Key example - *Threshold Potentials*:

$$f(A) = \min(y, \mathbf{w} \cdot \mathbf{e}_A)$$

- All decomposable functions can be expressed as a modular part plus a sum/integral of threshold potentials.

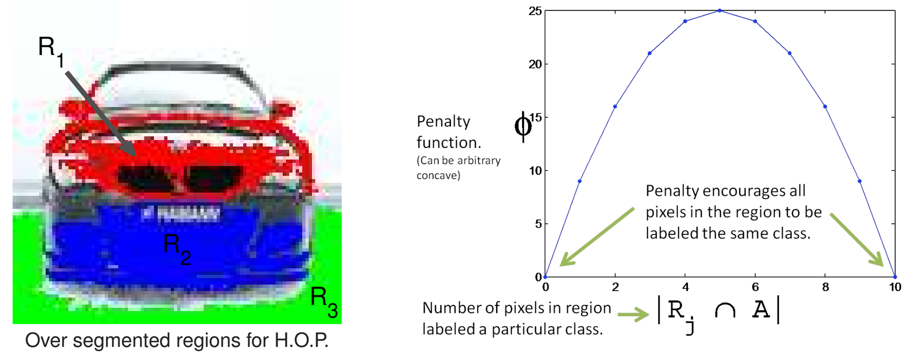


Example concave function as a sum of linear plus thresholds

- Concave cardinality functions: A strict subclass of decomposable functions:

$$f(A) = \sum_j \phi_j(|R_j \cap A|)$$

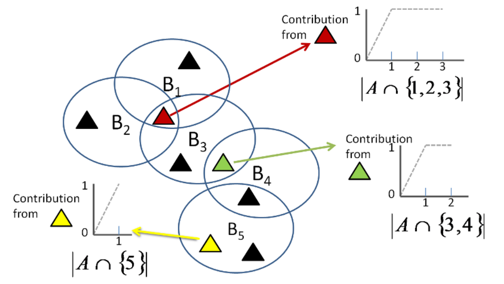
- Higher Order Potentials for MAP inference



Over segmented regions for H.O.P.

- Set cover functions: $f(A) = |\cup_{i \in A} B_i|$ where $B_i \subset F, \forall i \in E$. Let

$\mathbf{w}_k[i] = \begin{cases} 0 & k \notin B_i \\ 1 & k \in B_i \end{cases}$, then $f(A) = \sum_{k \in F} \min(1, \mathbf{w}_k \cdot \mathbf{e}_A)$.



- Example from queuing systems [Itoko 2007]. If $\mathbf{u}, \mathbf{v} \geq \mathbf{0}$, ϕ *nonincreasing* concave

$$f(A) = (\mathbf{u} \cdot \mathbf{e}_A)\phi(\mathbf{v} \cdot \mathbf{e}_A)$$

Overview of method and contributions

- Reformulate as (nonsmooth) convex minimization problem
- Use modern technique of smooth minimization of nonsmooth functions
- Novel stopping criterion

Able to solve problems with 10,000 variables in a minute.

Convex reformulation

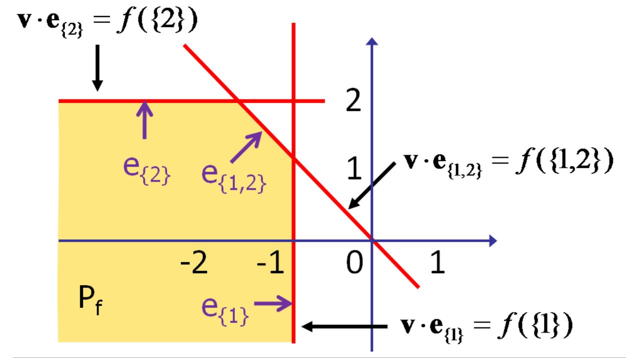
- Key Properties of *Lovász extension*: $\tilde{f} : [0,1]^n \to \mathbb{R}$
  - Convex
  - Agrees with f at corners: $\tilde{f}(\mathbf{e}_A) = f(A)$
  - A corner is optimal: $\{\mathbf{e}_A : A \subset E\} \cap \arg\min_{\mathbf{x} \in [0,1]^n} \tilde{f}(\mathbf{x}) \neq \emptyset$.
- Definition, assuming $f(\emptyset) = 0$.

$$\tilde{f}(\mathbf{x}) = \sup_{\mathbf{v} \in P_f} \mathbf{v} \cdot \mathbf{x}.$$

$$P_f = \{\mathbf{v} \in \mathbb{R}^n : \mathbf{v} \cdot \mathbf{e}_A \leq f(A), \text{ for all } A \in 2^E\}.$$

P_f is *Submodular polyhedron* associated with f.

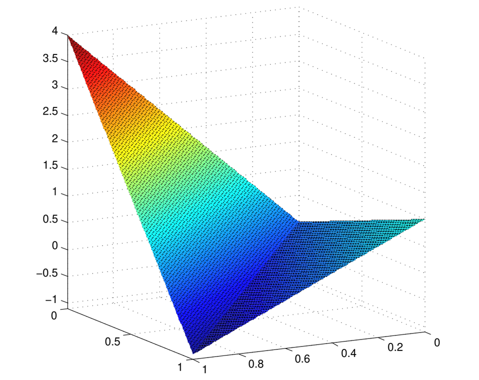
Example P_f with $E = \{1, 2\}$.



A	f(A)	e_A
$\emptyset$	0	(0,0)
{1}	-1	(1,0)
{2}	2	(0,1)
{1,2}	0	(1,1)

$\tilde{f}(\mathbf{e}_{\{1,2\}}) = \sup_{\mathbf{v} \in P_f} \mathbf{v} \cdot \mathbf{e}_{\{1,2\}} = 0 = f(\{1,2\})$

General 2-D formula: $\tilde{f}(x_1,x_2) = a|x_1 - x_2| + bx_1 + cx_2$ with $a \geq 0$.



General Lovász Extension Properties

- Piecewise linear.
- Nonsmooth at points with equal components. (eg. all corners)
- Defined by LP with exponentially many constraints.
- Computable in $O^*(n)$ time: Sort $\mathbf{x}$ components. Choose σ: $\mathbf{x}[\sigma(1)] \geq \ldots \geq \mathbf{x}[\sigma(n)]$. Let $S_k = \{\sigma(1),\ldots,\sigma(k)\}$.

$$\tilde{f}(\mathbf{x}) = \sum_{k=1}^n \mathbf{x}[\sigma(k)](f(S_k) - f(S_{k-1})).$$

- Can compute extension and also a subgradient in linear time:

$$\partial\tilde{f}(\mathbf{x}) \ni \sum_{k=1}^n \mathbf{e}_{\sigma(k)}(f(S_k) - f(S_{k-1})).$$

- Can use projected subgradient descent, but slow convergence $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ iterations to achieve ε accuracy. Impractical.

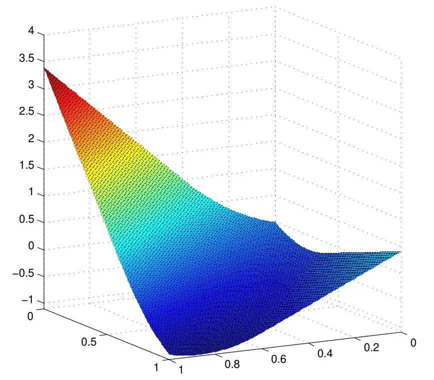
Smooth Minimization of Nonsmooth Functions

- Groundbreaking work by Nesterov [2004]
- Solves nonsmooth problems in $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ iterations
- Each iteration neglibly more work than gradient descent
- Not black-box solver; requires problem to have exploitable structure (often true)
- Example: $h(\mathbf{x}) = \sup_{\mathbf{y} \in C} \mathbf{x} \cdot \mathbf{y}$. If C easy to project onto, then h can be smoothed and an accelerated gradient descent scheme can be applied.

Smoothed Lovász Extension

$$\tilde{f}^\mu(\mathbf{x}) = \max_{\mathbf{v} \in P_f} \mathbf{v} \cdot \mathbf{x} - \frac{\mu}{2}\|\mathbf{v}\|^2.$$

Computing General Smoothed Lovász Gradient $\leftrightarrow$ Submodular Minimization Problem. [Bach 2010] (Just as hard as orginal problem.)



2-D Example

Key Insight: Smoothed Lovász Gradient for a *decomposable* function is easily computed!

Computation of Smoothed Gradient

- Since all decomposable functions will be linear combinations of threshold potentials, here assume $f(A) = \min(\mathbf{w} \cdot \mathbf{e}_A, y)$
- For $\mathbf{x}$ in unit cube:

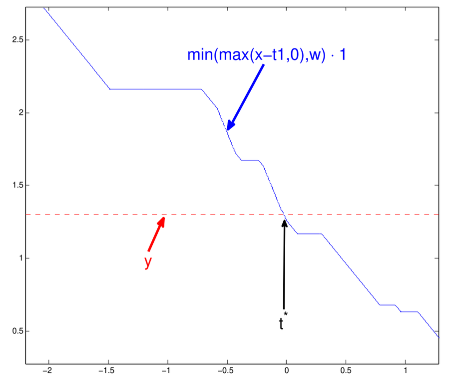
$$\tilde{f}(\mathbf{x}) = \max_{\mathbf{0} \leq \mathbf{v} \leq \mathbf{w},\ \mathbf{v} \cdot \mathbf{1} = y} \mathbf{v} \cdot \mathbf{x}$$

- Smoothed gradient:

$$\nabla\tilde{f}^\mu(\mathbf{x}) = \arg\min_{\mathbf{0} \leq \mathbf{v} \leq \mathbf{w},\ \mathbf{v} \cdot \mathbf{1} = y} \|\mathbf{v} - \mathbf{x}/\mu\| = \min(\max((\mathbf{x} - t^*\mathbf{1})/\mu, \mathbf{0}), \mathbf{w}).$$

With t^* chosen so $\nabla\tilde{f}^\mu(\mathbf{x}) \cdot \mathbf{1} = y$

Find root of monotonic continuous piecewise linear function. No explicit closed form expression, but simple to compute.



Smoothed Lovász Gradient (SLG) Algorithm

- Input: Accuracy ε; decomposable function $f(A) = \mathbf{c} \cdot \mathbf{e}_A + \sum_j d_j \min(\mathbf{w}_j \cdot \mathbf{e}_A, y_j)$.
- $D = \sum_j d_j \|\mathbf{w}_j\|_\infty$, $\mu = \frac{\varepsilon}{2D}$, $L = \frac{D}{\mu}$, $\mathbf{x}_{-1} = \mathbf{z}_{-1} = \frac{1}{2}\mathbf{1}$
- For $t = 0, 1, 2, \ldots$
  - $\mathbf{g}_t = \nabla\tilde{f}^\mu(\mathbf{x}_{t-1})/L$
  - $\mathbf{z}_t = P_{[0,1]^n}\left(\mathbf{z}_{-1} - \sum_{s=0}^t \left(\frac{s+1}{2}\right)\mathbf{g}_s\right)$
  - $\mathbf{y}_t = P_{[0,1]^n}(\mathbf{x}_t - \mathbf{g}_t)$
  - If gap$_t \leq \varepsilon/2$ stop.
  - $\mathbf{x}_t = (2\mathbf{z}_t + (t+1)\mathbf{y}_t)/(t+3)$
- $\mathbf{x}_\varepsilon = \mathbf{y}_t$
- Output: ε-optimal $\mathbf{x}_\varepsilon$ to $\min_{\mathbf{x} \in [0,1]^n} \tilde{f}(\mathbf{x})$

Theorem

SLG is guaranteed to provide an ε-optimal solution after running for $\mathcal{O}\left(\frac{D}{\varepsilon}\right)$ iterations.

Given ε-optimal $x_\varepsilon \in [0,1]^n$, can round to find corresponding ε-optimal set:

- Input: $\mathbf{x} \in [0,1]^n$; submodular function $f(A)$.
- Choose σ: $\mathbf{x}[\sigma(1)] \geq \ldots \geq \mathbf{x}[\sigma(n)]$.
- $S_k = \{\sigma(1),\ldots,\sigma(k)\}$.
- $k^* = \arg\min_k f(S_k)$, $A = S_{k^*}$
- Output: Set A satisfying $f(A) \leq \tilde{f}(\mathbf{x})$

Early Stopping

- By rounding, may find optimal set before continous convergence.
- Use $\mathbf{g} \in \partial\tilde{f}(\mathbf{e}_A)$ to bound optimality gap of A:

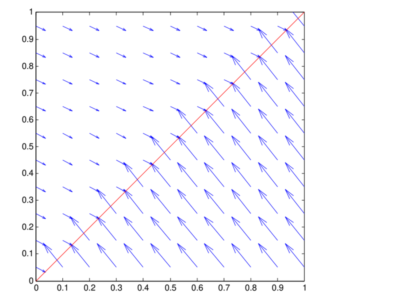
$$f(A) - f^* \leq \sum_{k \in A} \max(0, \mathbf{g}[k]) + \sum_{k \in E \setminus A} \max(0, -\mathbf{g}[k])$$

- If $\mathbf{g}[k] \leq 0$ for $k \in A$ and $\mathbf{g}[k] \geq 0$ for $k \in E \setminus A$, then A is optimal!
- Choose $\mathbf{g} \in \partial\tilde{f}(\mathbf{e}_A)$ from smoothed gradient

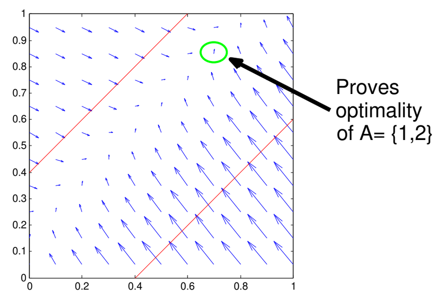
Lemma

$$\min_{k \in A, l \in E \setminus A} \mathbf{x}[k] - \mathbf{x}[l] \geq 2\mu \Rightarrow \nabla\tilde{f}^\mu(\mathbf{x}) \in \partial\tilde{f}(\mathbf{e}_A)$$

Example 2-D negative gradients:

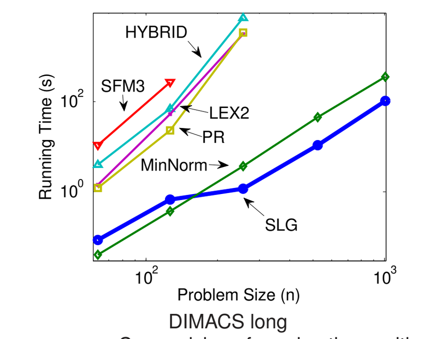


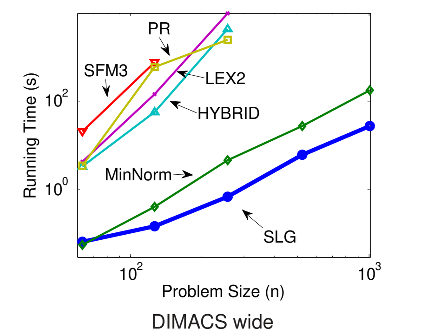
Nonsmooth



Smoothed ($\mu = .4$)

Synthetic Results Comparision

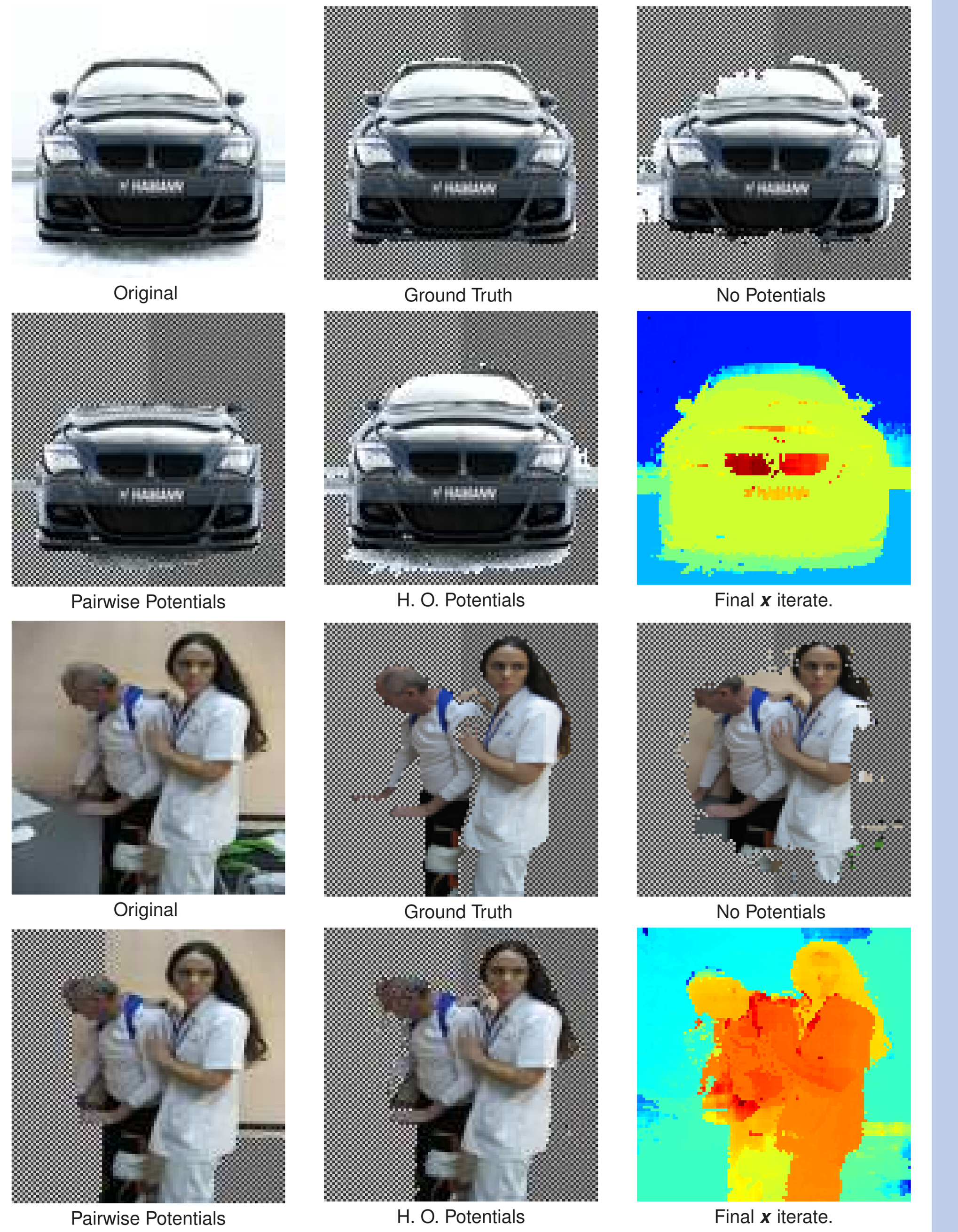




Comparision of running times with general submodular minimization algorithms.

Segmentation Results

Textonboost classification, regularized with submodular potential functions:



Minimization of H. O. Potentials takes *70 sec.* with SLG algorithm vs. *2 hrs.* for min-norm.

Conclusions

- A new class of submodular functions that can be efficiently minimized
- Apply Nesterov smoothing technique to Lovász extension
- Novel way to stop early
- Can solve larger-scale problems than previously possible

California Institute of Technology, Pasadena, CA, USA

Mail: stobbe@caltech.edu