Efficient Minimization of Decomposable Submodular Functions

Peter Stobbe and Andreas Krause

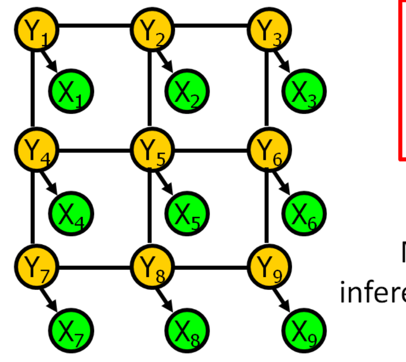
Structures in Optimization

- Convexity useful for continuous functions
  - $f(\boldsymbol{x}+\theta \boldsymbol{h})-f(\boldsymbol{x}) \le \theta(f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x}))$
  - Minimization tractable if convex.
- Similar Submodular discrete functions:
  - Domain of f: subsets of finite set E
  - $f(A\cup B\cup C)-f(A\cup B)\le f(A\cup C)-f(A)$
  - 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:
$P(x_1,\dots,x_n,y_1,\dots,y_n) = \prod_{(i,j)} \psi_{ij}(y_i,y_j)\prod_i \phi_i(x_i,y_i)$

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

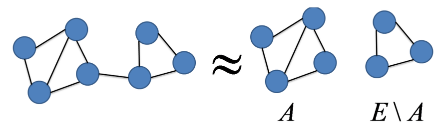
Submodular

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

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

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

- Factorizing random variables. [Narasimhan and Bilmes 2004]



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

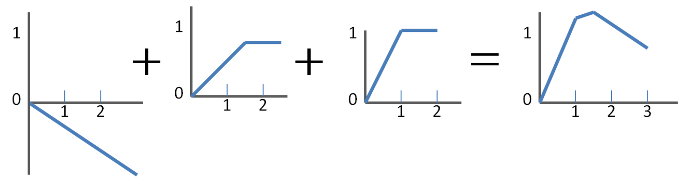
- Clustering [Narasimhan et al 2005]

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!

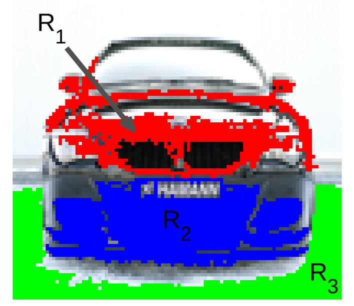
Decomposable submodular functions

- Definition:
$$f(A)=\sum_j \phi_j(\boldsymbol{w}_j\cdot \boldsymbol{e}_A)$$
ϕ_j concave, $\boldsymbol{w}_j\ge \boldsymbol{0}$, and $\boldsymbol{e}_A[k]=\begin{cases}0 \text{ if } k\notin A\\ 1 \text{ if } k\in A\end{cases}$.
- Key example - *Threshold Potentials*:
$$f(A)=\min(y,\boldsymbol{w}\cdot\boldsymbol{e}_A)$$
- All decomposable functions can be expressed as a modular part plus a sum/integral of threshold potentials.

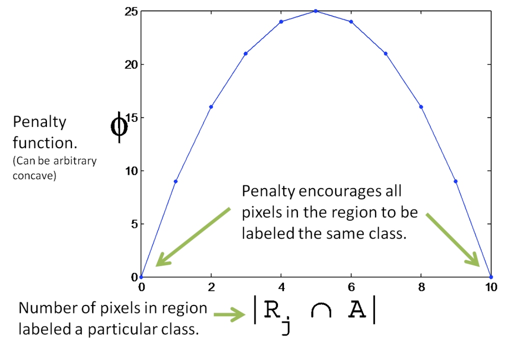


Example concave function as a sum of linear plus thresholds

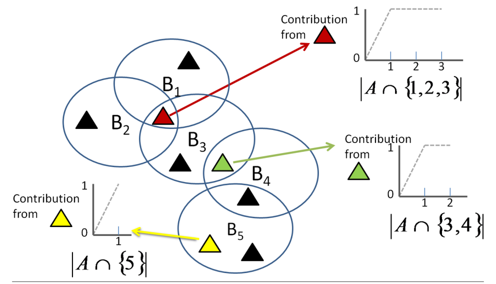
- Concave cardinality functions: A strict subclass of decomposable functions:
$$f(A)=\sum_j \phi_j(|R_j\cap A|)$$
- Higher Order Potentials for MAP inference



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



- Example from queuing systems [Itoko 2007]. If $\boldsymbol{u},\boldsymbol{v}\ge\boldsymbol{0}$, ϕ *nonincreasing* concave
$$f(A)=(\boldsymbol{u}\cdot\boldsymbol{e}_A)\phi(\boldsymbol{v}\cdot\boldsymbol{e}_A)$$

Overview of method and contributions

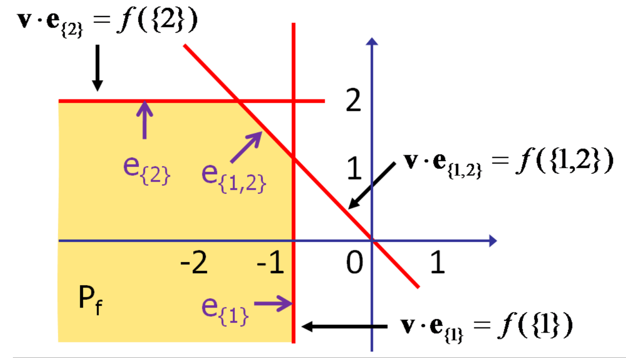
- Reformulate as (nonsmooth) convex minimization problem
- Use modern technique of smooth minimization of nonsmooth functions
- Novel stopping criterion

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: $\tilde f(\boldsymbol{e}_A)=f(A)$
  - A corner is optimal: $\{\boldsymbol{e}_A: A\subset E\}\cap \arg\min_{\boldsymbol{x}\in[0,1]^n}\tilde f(\boldsymbol{x})\ne\emptyset$.
- Definition, assuming $f(\emptyset)=0$.
$$\tilde f(\boldsymbol{x})=\sup_{\boldsymbol{v}\in P_f}\boldsymbol{v}\cdot\boldsymbol{x}.$$
$$P_f=\{\boldsymbol{v}\in\mathbb{R}^n: \boldsymbol{v}\cdot\boldsymbol{e}_A\le f(A), \text{ for all } A\in 2^E\}.$$
P_f is *Submodular polyhedron* associated with f.

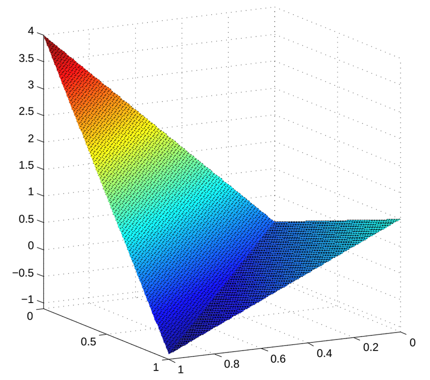
Example P_f with $E=\{1,2\}$.



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

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

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



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 $\boldsymbol{x}$ components. Choose σ: $\boldsymbol{x}[\sigma(1)]\ge\ldots\ge\boldsymbol{x}[\sigma(n)]$. Let $S_k=\{\sigma(1),\ldots,\sigma(k)\}$.
$$\tilde f(\boldsymbol{x})=\sum_{k=1}^n \boldsymbol{x}[\sigma(k)](f(S_k)-f(S_{k-1})).$$
- Can compute extension and also a subgradient in linear time:
$$\partial\tilde f(\boldsymbol{x})\ni\sum_{k=1}^n \boldsymbol{e}_{\sigma(k)}(f(S_k)-f(S_{k-1})).$$
- Can use projected subgradient descent, but slow convergence $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ iterations to achieve ε accuracy. Impractical.

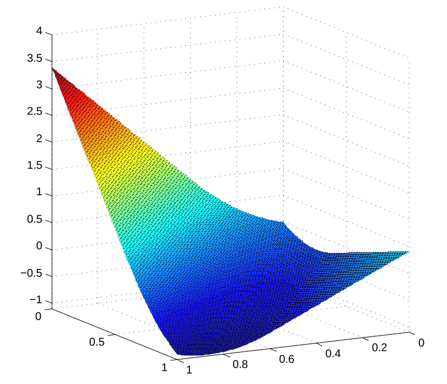
Smooth Minimization of Nonsmooth Functions

- 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: $h(\boldsymbol{x})=\sup_{\boldsymbol{y}\in C}\boldsymbol{x}\cdot\boldsymbol{y}$. If C easy to project onto, then h can be smoothed and an accelerated gradient descent scheme can be applied.

Smoothed Lovász Extension

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

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



2-D Example

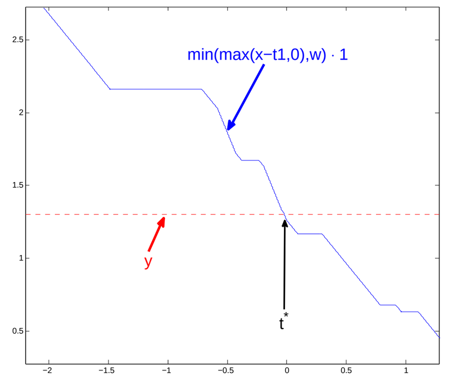
Key Insight: Smoothed Lovász Gradient for a *decomposable* function is easily computed!

Computation of Smoothed Gradient

- Since all decomposable functions will be linear combinations of threshold potentials, here assume $f(A)=\min(\boldsymbol{w}\cdot\boldsymbol{e}_A,y)$
- For $\boldsymbol{x}$ in unit cube:
$$\tilde f(\boldsymbol{x})=\max_{\boldsymbol{0}\le\boldsymbol{v}\le\boldsymbol{w},\ \boldsymbol{v}\cdot\boldsymbol{1}=y}\boldsymbol{v}\cdot\boldsymbol{x}$$
- Smoothed gradient:
$$\nabla\tilde f^\mu(\boldsymbol{x})=\arg\min_{\boldsymbol{0}\le\boldsymbol{v}\le\boldsymbol{w},\ \boldsymbol{v}\cdot\boldsymbol{1}=y}\|\boldsymbol{v}-\boldsymbol{x}/\mu\|$$
$$=\min(\max((\boldsymbol{x}-t^*\boldsymbol{1})/\mu,\boldsymbol{0}),\boldsymbol{w}).$$

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

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

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

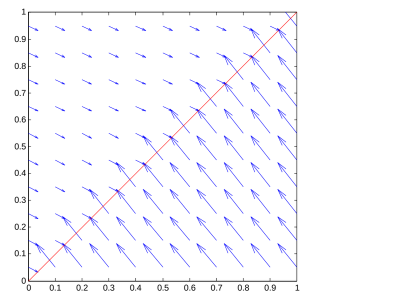
Early Stopping

- By rounding, may find optimal set before continous convergence.
- Use $\boldsymbol{g}\in\partial\tilde f(\boldsymbol{e}_A)$ to bound optimality gap of A:
$$f(A)-f^*\le\sum_{k\in A}\max(0,\boldsymbol{g}[k])+\sum_{k\in E\setminus A}\max(0,-\boldsymbol{g}[k])$$
- If $\boldsymbol{g}[k]\le 0$ for $k\in A$ and $\boldsymbol{g}[k]\ge 0$ for $k\in E\setminus A$, then A is optimal!
- Choose $\boldsymbol{g}\in\partial\tilde f(\boldsymbol{e}_A)$ from smoothed gradient

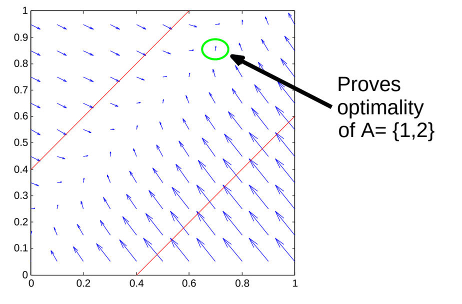
Lemma

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

Example 2-D negative gradients:

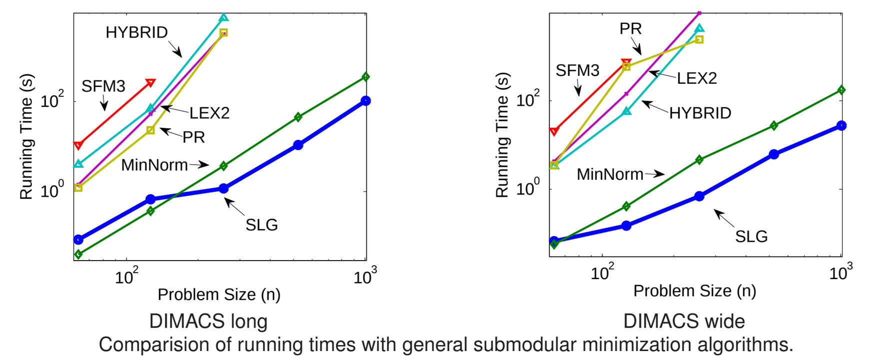


Nonsmooth



Smoothed ($\mu=.4$)

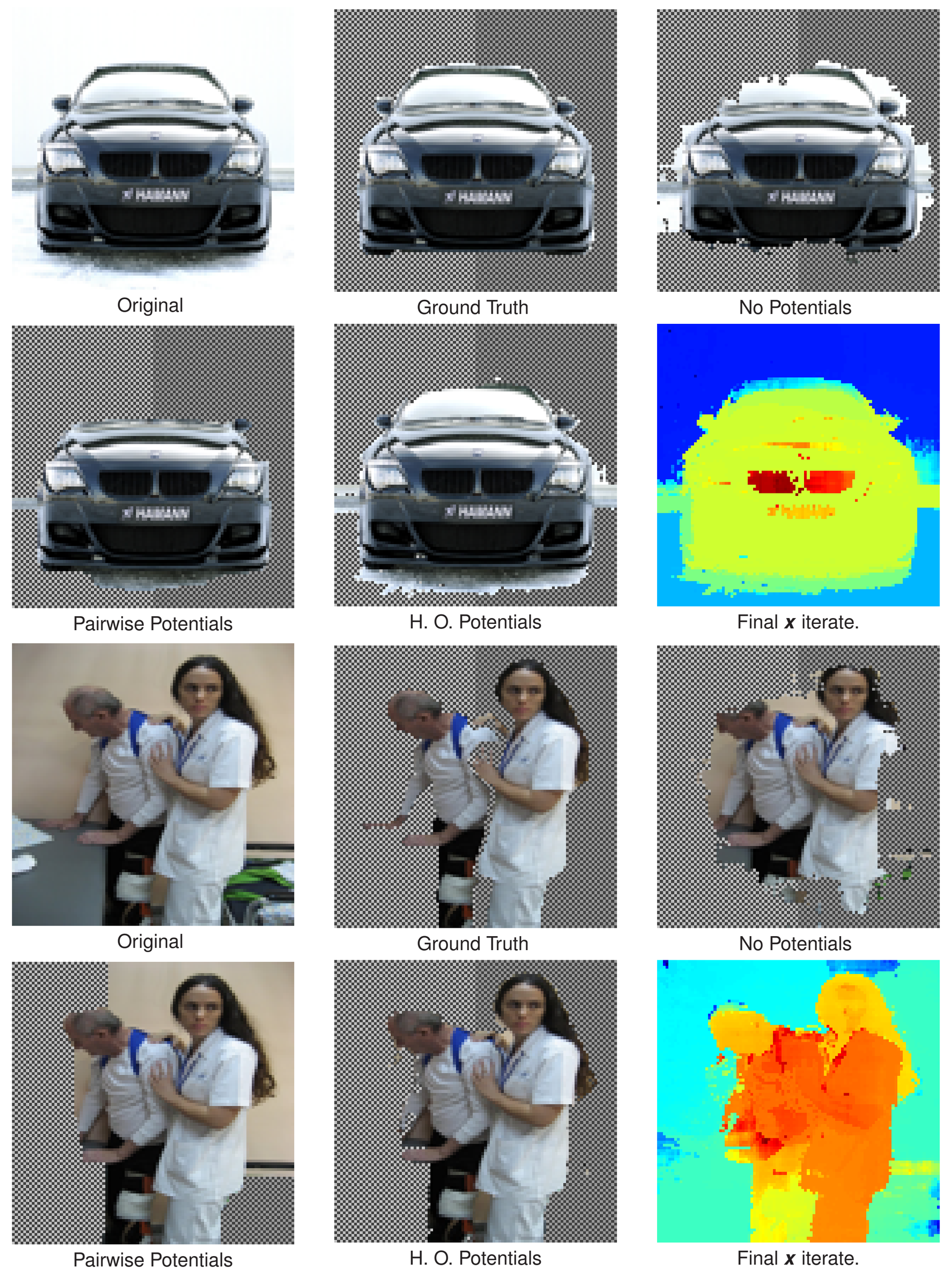
Synthetic Results Comparision



Comparison 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