Massachusetts Institute of Technology

Submodular Functions and Machine Learning

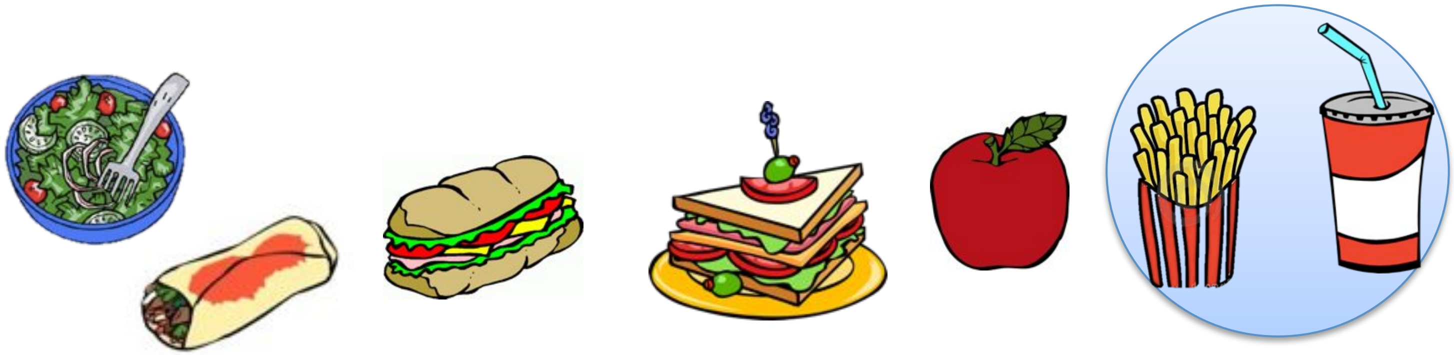
MLSS Kyoto

Stefanie Jegelka

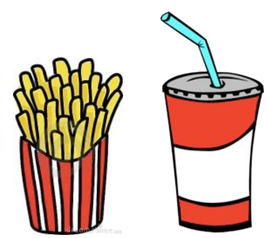
MIT

Set functions

ground set

$\mathcal{V} = \Big\{$  $\Big\}$

$$F : 2^{\mathcal{V}} \to \mathbb{R}$$

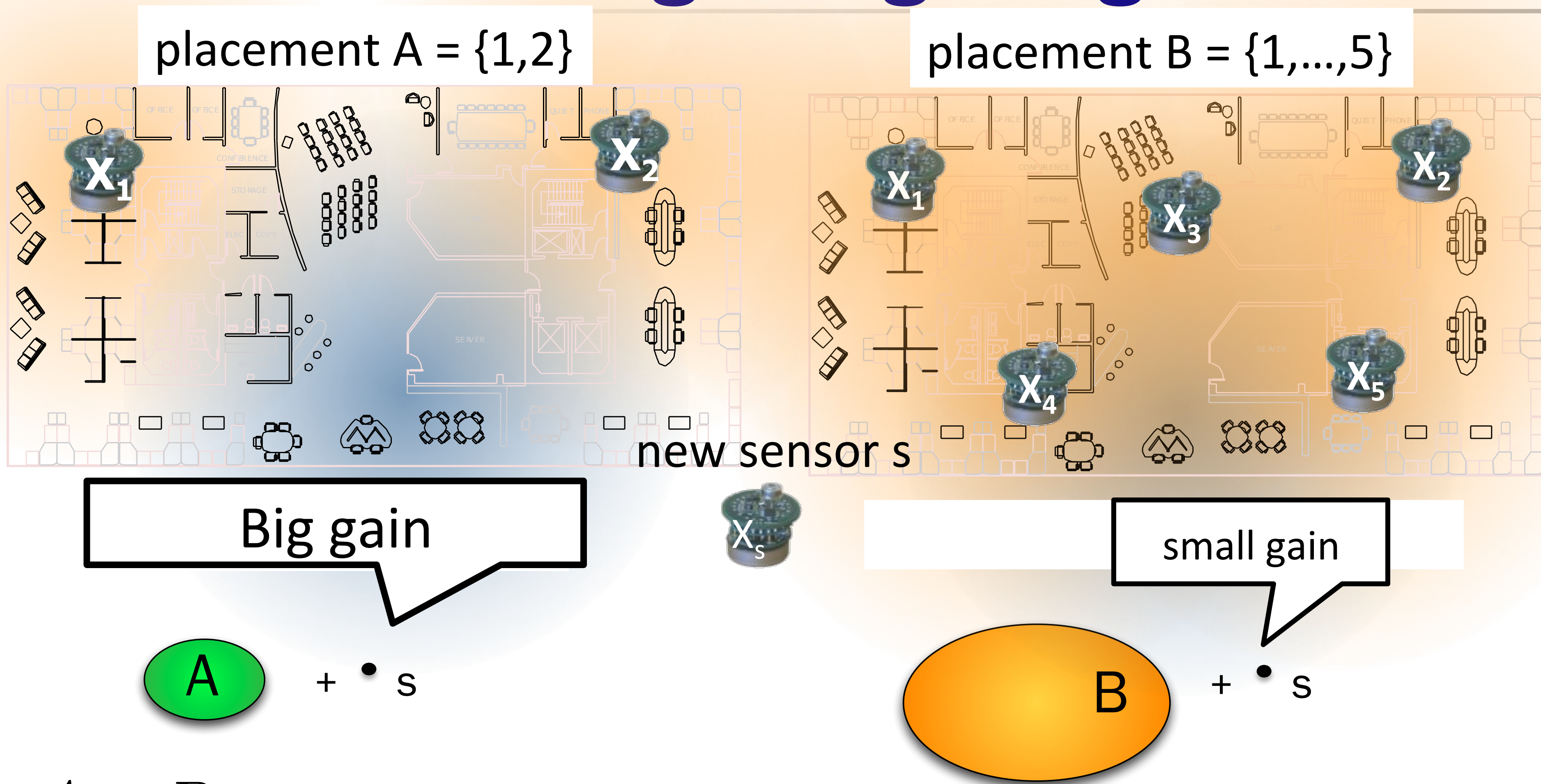
$F\Big($  $\Big) =$ cost of buying items together, or

utility, or

probability, ...

We will assume:
- $F(\emptyset) = 0$
- black box “oracle” to evaluate *F*

Diminishing marginal gains

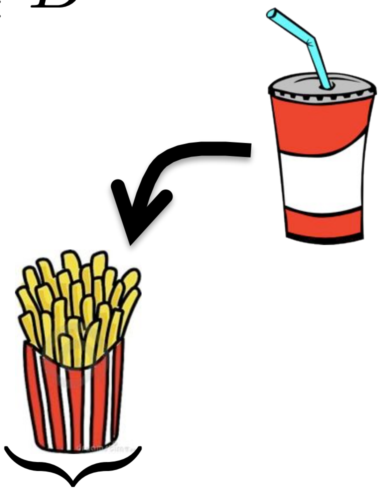


$$A \subseteq B$$

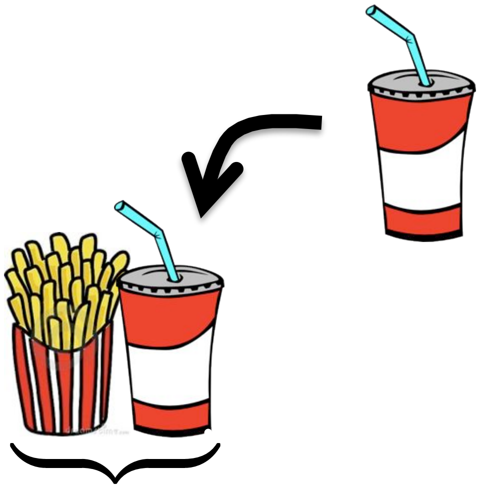
$$F(A \cup s) - F(A) \quad \geq \quad F(B \cup s) - F(B)$$

Submodularity

$$A \subseteq B$$



A



B

$$F(A \cup s) - F(A) \quad \geq \quad F(B \cup s) - F(B)$$

extra cost:
one drink

extra cost:
free refill ☺

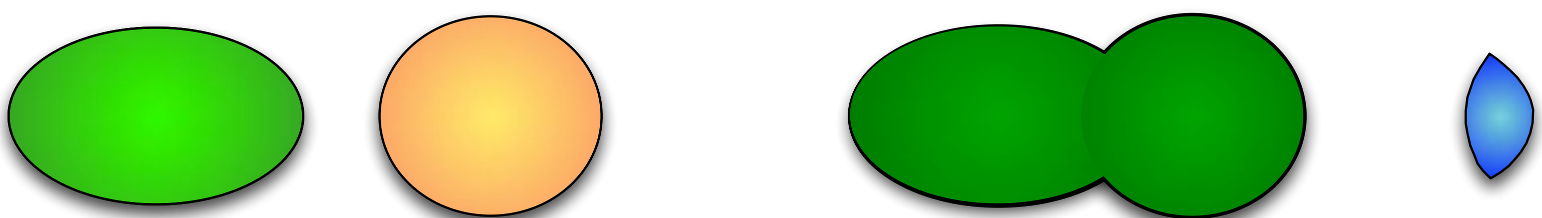
diminishing marginal costs

Submodular set functions

- Diminishing gains: for all $A \subseteq B$

$$F(\underline{A \cup e}) - F(A) \;\geq\; F(B \cup e) - \underline{F(B)}$$

- Union-Intersection: for all $S, T \subseteq \mathcal{V}$

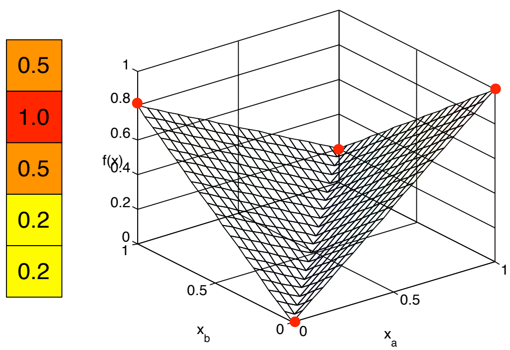


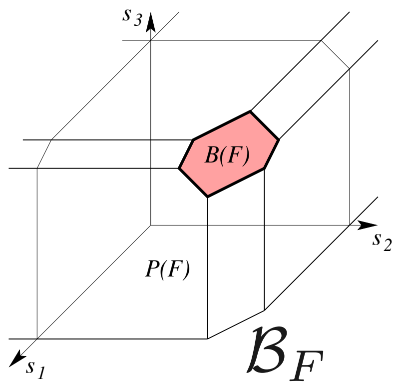
$$\underline{F(S)} + \underline{F(T)} \;\geq\; F(S \cup T) + F(S \cap T)$$

Roadmap

- Submodular set functions
  - definition & basic properties
  - links to convexity
  - special polyhedra
- Minimizing submodular functions
  - general and special cases
  - constraints
- Maximizing submodular functions

Convex relaxation



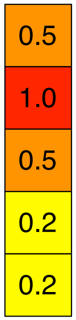


$$\min_{S \subseteq \mathcal{V}} F(S) = \min_{x \in \{0,1\}^n} F(x) = \min_{x \in [0,1]^n} f(x)$$

1. relaxation: convex optimization (non-smooth)

2. relaxation is exact!
   ➔ submodular minimization in polynomial time!
   (Grötschel, Lovász, Schrijver 1981)

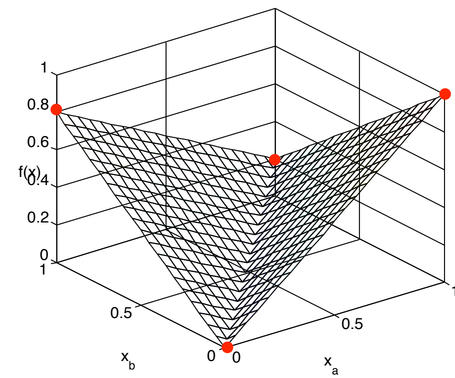
Lovász extension



$$f(x) = \sum_{i=1}^{k} \alpha_i \, F(S_i)$$

if *F* is submodular, this is equivalent to:

$$\boxed{f(x) = \max_{y \in \mathcal{B}_F} \; y^\top x}$$



Theorem *(Lovász, 1983)*
Lovász extension is convex $\Leftrightarrow$ F is submodular.

Theorem *(Lovász, 1983)*
Lovasz extension is convex $\Leftrightarrow$ F is submodular.

$$f(x) = \sum_{i=1}^{k} \alpha_i \, F(S_i)$$

$\Leftarrow$ done.

$\Rightarrow$ Let $A, B \subseteq \mathcal{V}$.

$$\underset{1_A}{\begin{bmatrix}1\\1\\0\\0\end{bmatrix}} + \underset{1_B}{\begin{bmatrix}0\\1\\1\\0\end{bmatrix}} = \begin{bmatrix}1\\2\\1\\0\end{bmatrix} = \underset{1_{A\cup B}}{\begin{bmatrix}1\\1\\1\\0\end{bmatrix}} + \underset{1_{A\cap B}}{\begin{bmatrix}0\\1\\0\\0\end{bmatrix}}$$

$$f(1_A + 1_B) \quad \leq \quad f(1_A) + f(1_B)$$

$$= f(1_{A\cup B} + 1_{A\cap B}) \qquad\qquad = F(A) + F(B)$$

$$= F(A \cup B) + F(A \cap B)$$

Submodular minimization

convex optimization

- ellipsoid method
  (Grötschel-Lovasz-Schrijver 81)
- subgradient method
- minimum-norm point / Fujishige-Wolfe algorithm
- ...

combinatorial methods

- first polynomial-time:
  (Schrijver 00, Iwata-Fleischer-Fujishige 01)
- currently fastest:

  $O(n^4 T + n^5 \log M)$ *(Iwata 03)*

  $O(n^6 + n^5 T)$ *(Orlin 09)*

Latest result: $O(n^2 T \log nM + n^3 \log^c nM)$

$O(n^3 T \log^2 n + n^4 \log^c n)$ *(Lee-Sidford-Wong 15)*

Recall: convex relaxation

Lovász extension

$$\min_{x \in [0,1]^n} f(x)$$

$$= \max_{y \in \mathcal{B}_F} \left(\sum_{i=1}^{n} \min\{y_i, 0\} \right)$$

optimal discrete solution:

$$S^* = \{\, i \mid x_i^* > 0 \,\}$$

$$= \{\, i \mid y_i^* < 0 \,\}$$

$$\begin{bmatrix} + \\ + \\ 0 \\ 0 \end{bmatrix}$$

Proximal problem

proximal problem

Lovász extension

$$\min_{x \in [0,1]^n} f(x) + \frac{1}{2}\|x\|^2$$

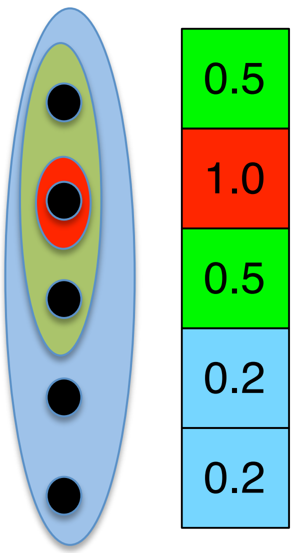
why? solves

$$\min_{S \subseteq \mathcal{V}} F(S) + \alpha|S| \quad \text{for all } \alpha$$

- Let S^α be the largest minimizer of $F(S) + \alpha|S|$
- can show: if $\alpha < \beta$, then $S^\alpha \supseteq S^\beta$

  ➔ chain $\emptyset \subset S^{\alpha_1} \subset S^{\alpha_2} \subset \ldots \mathcal{V}$
- "encode" in vector u :

$$\{e \mid u_e \geq \alpha\} = S^\alpha$$



$$\boxed{u = \arg\min_x f(x) + \frac{1}{2}\|x\|^2}$$

3 equivalent problems

$$y^* = -x^*$$

projection

$$\min_{y \in \mathcal{B}_F} \tfrac{1}{2}\|y\|^2$$

convex dual problem

proximal

$$\min_x \; f(x) + \tfrac{1}{2}\|x\|^2$$

parametric

$$\min_S \; F(S) + \alpha|S|$$

$$\{e \mid y^*_e \leq -\alpha\} = \quad S^\alpha = \{e \mid x^*_e \geq \alpha\}$$

thresholding

$$x^*_e = \sup\{\alpha \mid e \in S^\alpha\}$$

divide-and-conquer

Minimum-norm-point algorithm

Fujishige '91, Fujishige & Isotani '11

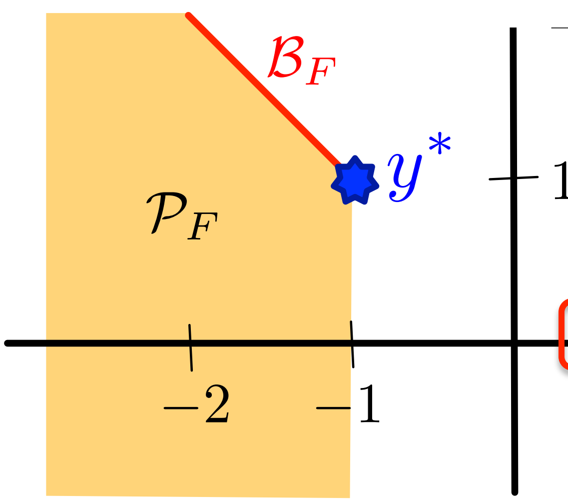
proximal problem

$$\min_{x} \; f(x) + \frac{1}{2}\|x\|^2 \longrightarrow$$

dual: minimum norm problem

$$y^* = \arg\min_{y \in \mathcal{B}_F} \; \frac{1}{2}\|y\|^2$$

$$y^* = \begin{array}{cc} \begin{bmatrix} -1 \\ 1 \end{bmatrix} & \begin{array}{c} a \\ b \end{array} \end{array}$$



A	$F(A)$
$\emptyset$	0
a	-1
b	2
$\{a,b\}$	0

$$S^* = \{i \mid y_i^* \leq 0\}$$

minimizes *F* !

$$S^* = \arg\min_{S \subseteq \mathcal{V}} \; F(S)$$

Overview

primal

$$\min_{x \in [0,1]^n} \quad f(x)$$

1
1
0
0

$\{e \mid x_e^* > 0\}$

primal

$$\min_{x} \quad f(x) + \frac{1}{2}\|x\|^2$$

$\{e \mid x_e^* \geq 0\}$

1
.8
-.5
-1

S^*

dual

$$\max_{y \in \mathcal{B}_F} \quad \sum_{i=1}^{n} [y_i]_-$$

--
--
+
+

$\{e \mid y_e^* < 0\}$

dual

$$\min_{y \in \mathcal{B}_F} \quad \frac{1}{2}\|y\|^2$$

$\{e \mid y_e^* \leq 0\}$

-1
-.8
.5
1

Solving the min-norm problem

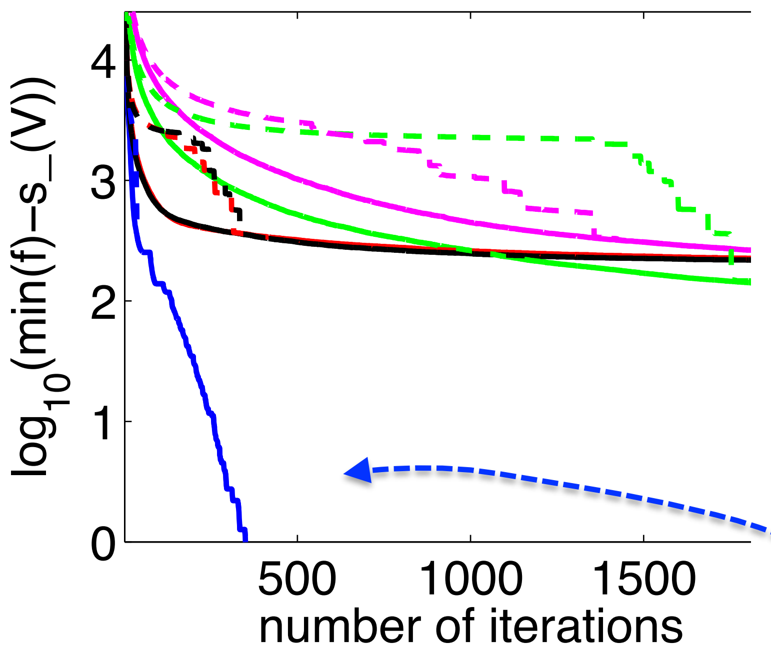
$$\min_{y \in \mathcal{B}_F} \quad \frac{1}{2}\|y\|^2$$

- costly: testing membership in $\mathcal{B}_F$
- costly: projection onto $\mathcal{B}_F$
- easy: linear optimization over $\mathcal{B}_F$: greedy algorithm! ☺

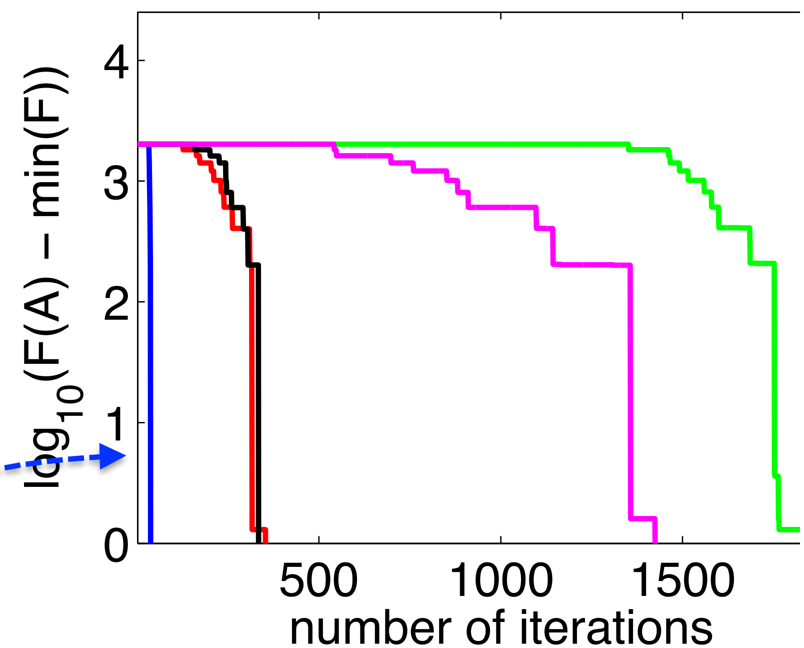
- conditional gradient algorithm *(Frank & Wolfe 1956)*
- active set methods: Fujishige-Wolfe *(Fujishige & Isotani 2011, Chakrabarty-Jain-Kothari 2014,…)*

Empirically

convergence of relaxation



convergence of S



min-norm point

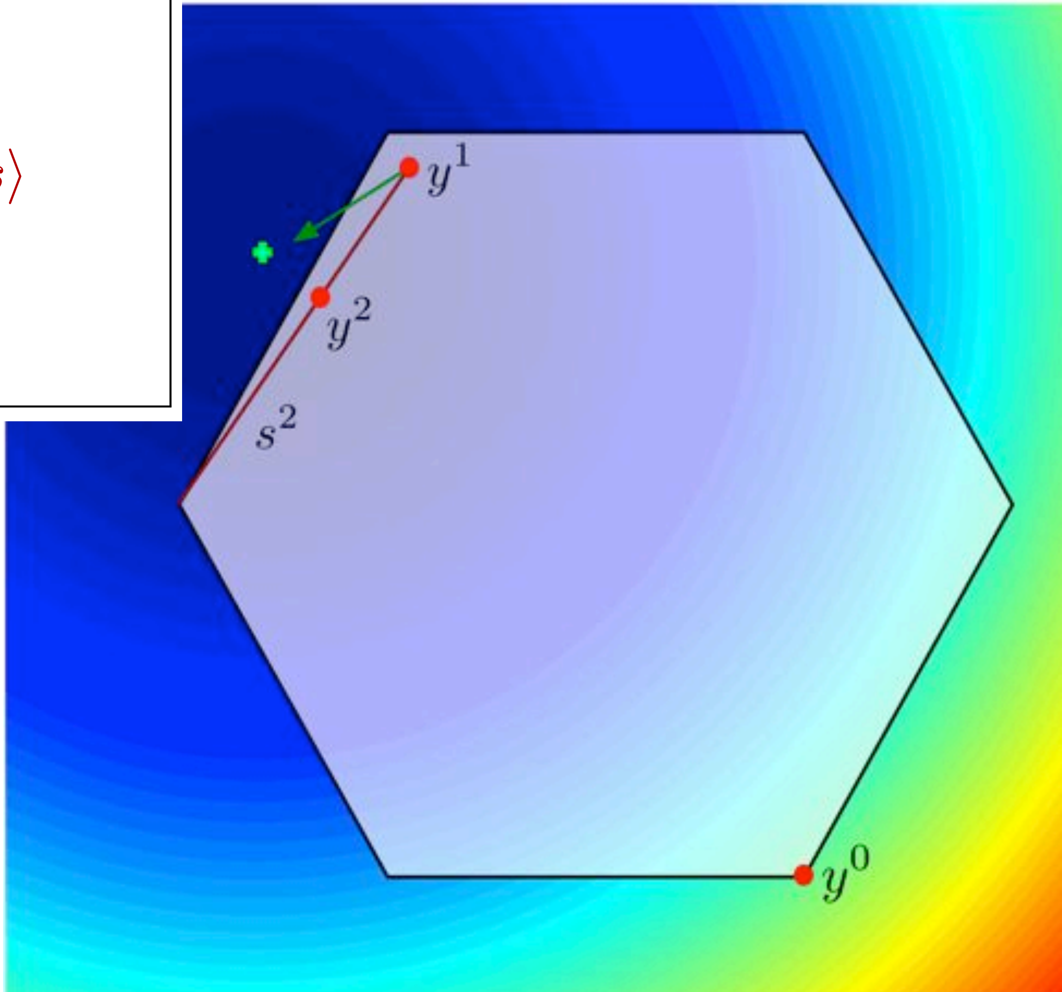
(Figure from Bach, 2012)

Conditional gradient algorithm

$$\min \; h(y) \qquad \text{s.t. } y \in \mathcal{P}$$

h convex, differentiable
P polyhedral

$y^0 \in \mathcal{P}$
for $t = 0, 1, \ldots$ to T **do**
$\quad s^t = \text{argmin}_{s \in \mathcal{P}} \langle \nabla h(y^t),\, s \rangle$
$\quad y^{t+1} = (1-\gamma) y^t + \gamma s^t$
end for



step size?
how many iterations?

Conditional gradient algorithm

$$\min \ h(y) \quad \text{s.t. } y \in \mathcal{P}$$

h convex, differentiable
P polyhedral

for $t = 0, 1, \ldots$ to T **do**
$s^t = \text{argmin}_{s \in \mathcal{P}} \langle \nabla h(y^t), s \rangle$
$y^{t+1} = (1 - \gamma) y^t + \gamma s^t$
end for

1. fixed step size: $\gamma^t = \frac{2}{t+2}$

2. line search: $\gamma^t = \arg \min_{\gamma \in [0,1]} h(y^t + \gamma(s^t - y^t))$

3. re-optimization: $y^{t+1} = \text{argmin}_{y \in \text{conv}(s^0, \ldots, s^t)} h(y)$

How many iterations?

$$y^* = -x^*$$

dual problem

$$\max_{y \in \mathcal{B}_F} -\tfrac{1}{2}\|y\|^2$$

primal problem

$$\min_x \; f(x) + \tfrac{1}{2}\|x\|^2$$

always: primal value ≥ dual value

at optimum: primal value = dual value

➔ bound the duality gap: primal – dual value

Theorem *(Jaggi 2013)*

After T iterations of the algorithm, there is an iterate y^T with

$$\mathrm{gap}(y^T) \leq \frac{16C}{T+2}$$

For $h(y) = \|y\|^2, \; C = \mathrm{diam}(\mathcal{P})^2$

Exercise: how many iterations for a gap $\leq \epsilon$?

Are we done yet?

want: $\min_{S \subseteq \mathcal{V}} F(S)$ $\dashrightarrow$ $\max_{y \in \mathcal{B}_F} -\frac{1}{2}\|y\|^2$

$S^t_\alpha = \{e \mid -y^t_e \geq \alpha\}$ $\dashleftarrow$ solve via conditional gradient & friends

have convergence bound

How many iterations until $F(S^t) - F(S^*) \leq \varepsilon$?

Theorem *(Bach 2013)*

If $\mathrm{gap}(y^\tau) \leq \epsilon'$ then there exists an α such that the discrete duality gap is bounded as

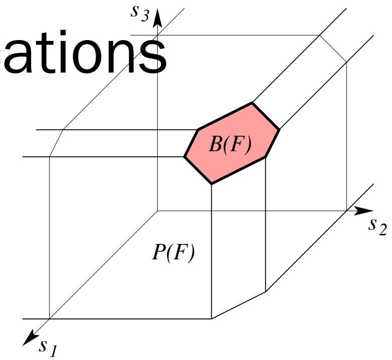
$$F(S^\tau_\alpha) - F(S^*) \leq \quad F(S^\tau_\alpha) - y^\tau_-(\mathcal{V}) \leq \sqrt{2n\epsilon'}$$

Submodularity and convexity

- convex Lovasz extension
  - easy to compute: greedy algorithm (special polyhedra!)

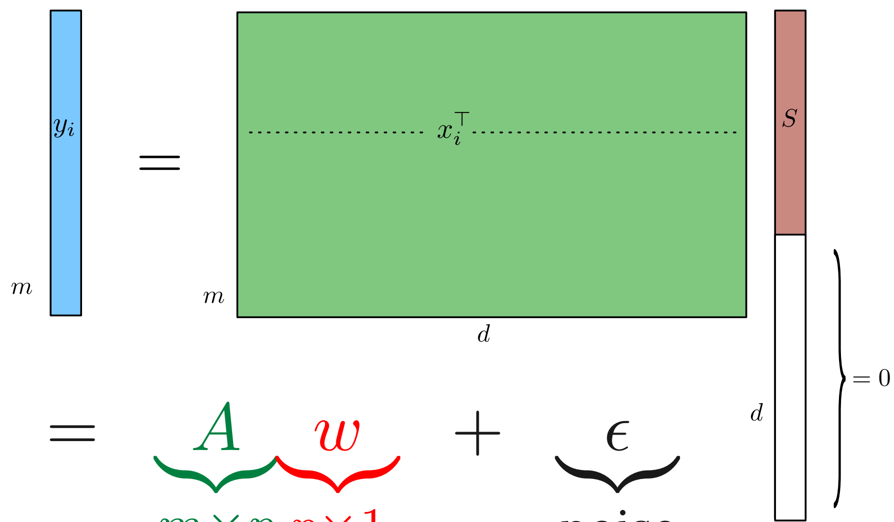
- submodular minimization via convex optimization: duality results & fast algorithms for many applications

$$\min_{x \in [0,1]^n} f(x) \qquad \min_x f(x) + \frac{1}{2}\|x\|^2$$



- structured sparsity *(Bach 2010)*
- decomposition & parallel algorithms *(J-Bach-Sra 2013, Nishihara-J-Jordan 2014)*
- variational inference *(Djolonga & Krause 2014)*
- …

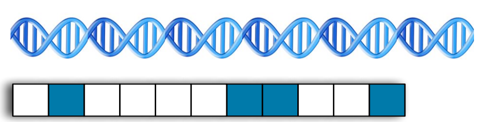
Sparsity



$$\underbrace{y}_{m \times 1} = \underbrace{A}_{m \times p}\underbrace{w}_{p \times 1} + \underbrace{\epsilon}_{\text{noise}}$$

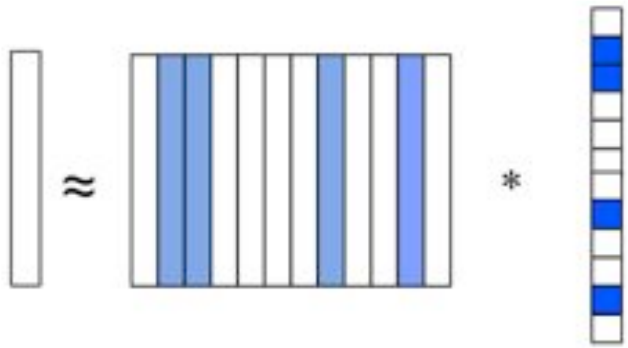
observe $(y_i, a_i^\top)$, want to estimate w

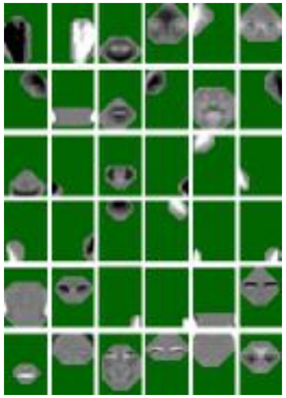
- what if m << p? underdetermined!
- in general, no recovery possible:
  ➔ exploit additional structure, e.g.: w has only k non-zeros
- also: interpretability



Sparsity

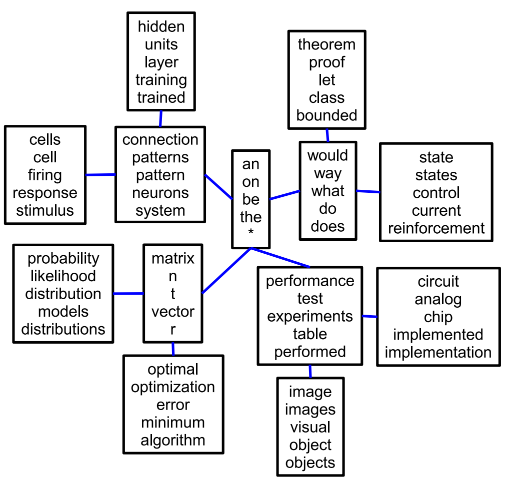
$$\underbrace{y}_{m\times 1} = \underbrace{A}_{m\times p}\underbrace{w}_{p\times 1} + \underbrace{\epsilon}_{\text{noise}}$$





Structured sparse PCA

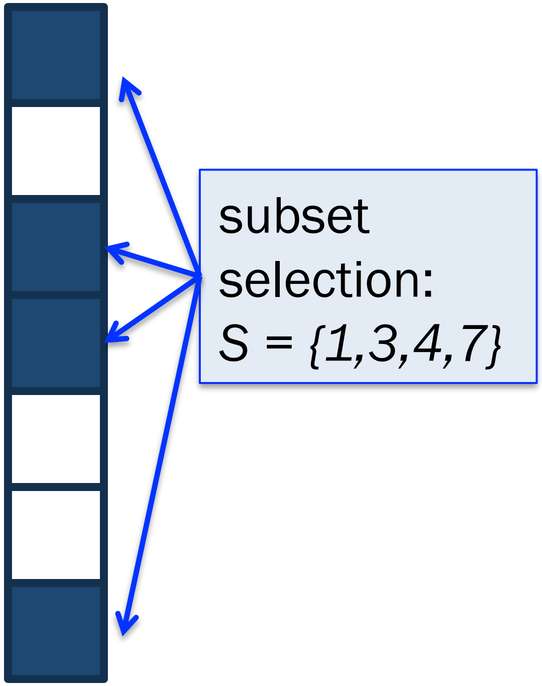
$$\hat{w} = \arg\min_{w} \|Aw - y\|^2 + \lambda\Omega(w)$$



Sparse reconstruction

$$\hat{w} = \arg\min_{w} \; \|Aw - y\|^2 \; + \lambda\Omega(w)$$

Assumption:
w is sparse



discrete regularization on support *S* of *w*

$$\Omega(w) = \|w\|_0 = |S|$$

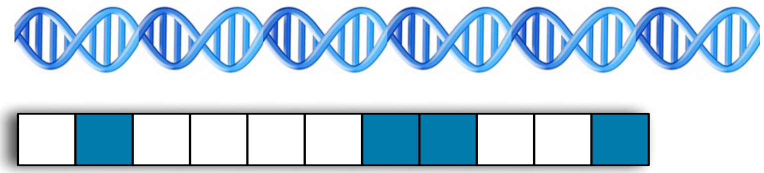
relax to convex envelope

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

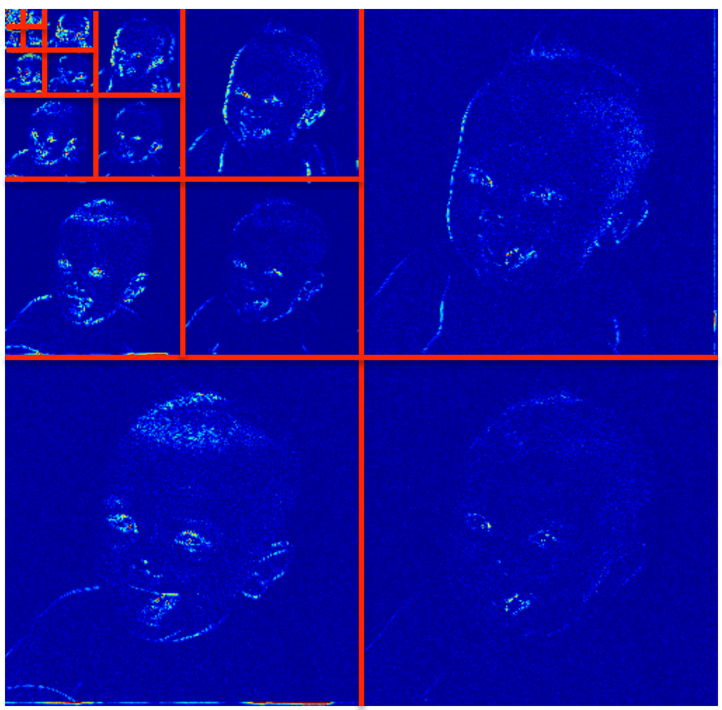
sparsity pattern often not random…

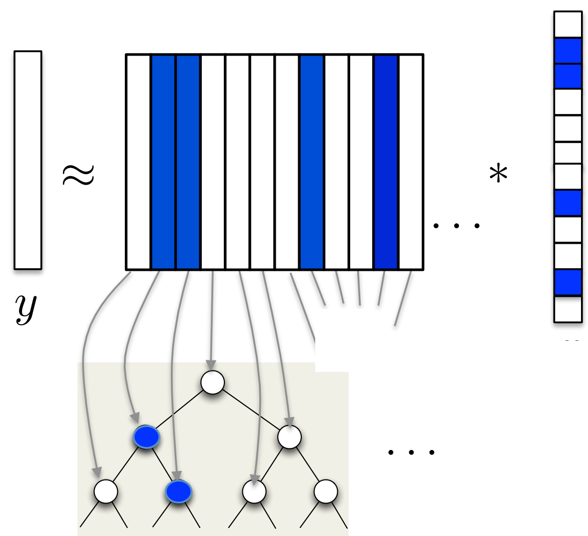
Structured sparsity

Assumption:
support of w
has structure

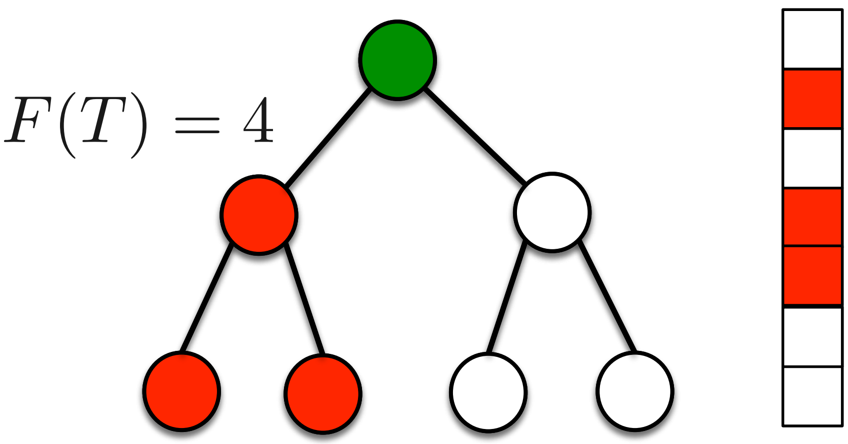


express by set function!





Preference for trees

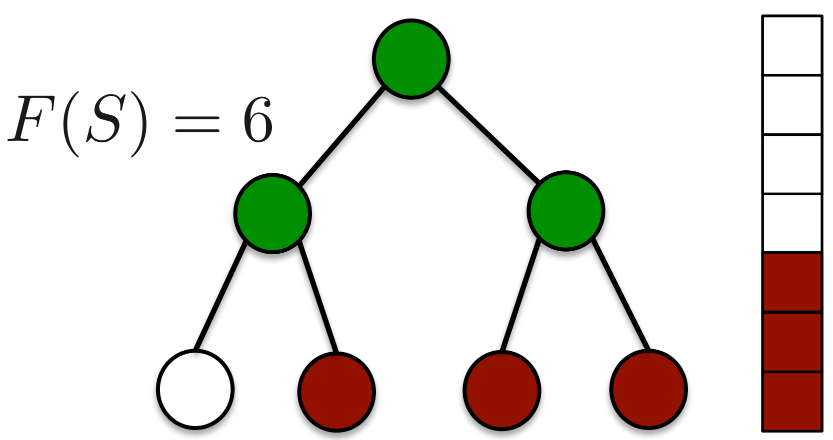


Set function:

$$F(T) < F(S)$$

if T is a tree and S not

$|S| = |T|$



$$F(S) = \left| \bigcup_{s \in S} \text{ancestors}(s) \right|$$

use as regularizer?

Sparsity

$$\hat{w} = \arg\min_{w} \|Aw - y\|^2 + \lambda\Omega(w)$$

- w sparse

- w structured sparse

discrete regularization on support S of w

$$\Omega(w) = \|w\|_0 = |S|$$

submodular function

$$\Omega(w) = F(S)$$

relax to convex envelope

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

➔ Lovász extension

$$\Omega(w) = f(|w|)$$

Optimization: sequence of min-norm problems
(submodular minimization)

(Bach2010)

Norms from submodular functions

$$\Omega(x) = f(|x|)$$

Proposition

For monotone increasing *f*, $f(|x|)$ is a norm.

(Exercise: show this)

Special cases: $F(S) = |S| \quad \Rightarrow \quad f(|x|) = \|x\|_1$

$$F(S) = \min\{|S|, 1\} \quad \Rightarrow \quad f(|x|) = \|x\|_\infty$$

$$F(S) = \sum_{j=1}^{k} \min\{|S \cap G_j|, 1\} \quad \Rightarrow \quad f(|x|) = \sum_{j=1}^{k} \|x_{G_j}\|_\infty$$

Special case

- minimize a sum of submodular functions

$$F(S) = \sum_{i=1}^{r} F_i(S)$$

$$\min_S \; F_i(S)$$

"easy"

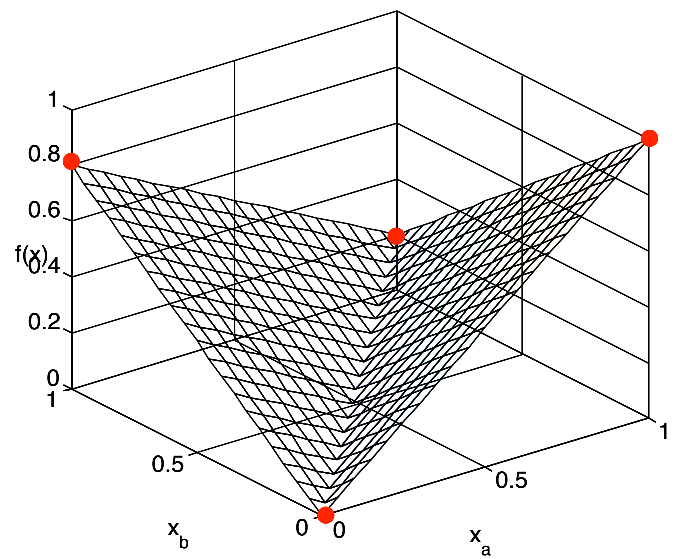
- combinatorial algorithms
  (Kolmogorov 12, Fix-Joachims-Park-Zabih 13, Fix-Wang-Zabih 14)

- convex relaxations

Relaxation

$$F : 2^{\mathcal{V}} \to \mathbb{R} \quad \equiv \quad F : \{0,1\}^n \to \mathbb{R} \qquad\qquad f : \mathbb{R}^n_+ \to \mathbb{R}$$

- convex Lovász extension: tight relaxation



$$\min_{S\subseteq\mathcal{V}} \sum_i F_i(S) = \min_{x\in[0,1]^n} \sum_i f_i(x)$$

$$\min_x \sum_{i=1}^r f_i(x) + \tfrac{1}{2}\|x\|^2$$

- dual decomposition: parallel algorithms

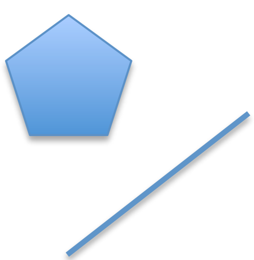
(Komodakis-Paragios-Tziritas 11, Savchynskyy-Schmidt-Kappes-Schnörr 11, J-Bach-Sra 13)

Results: dual decomposition

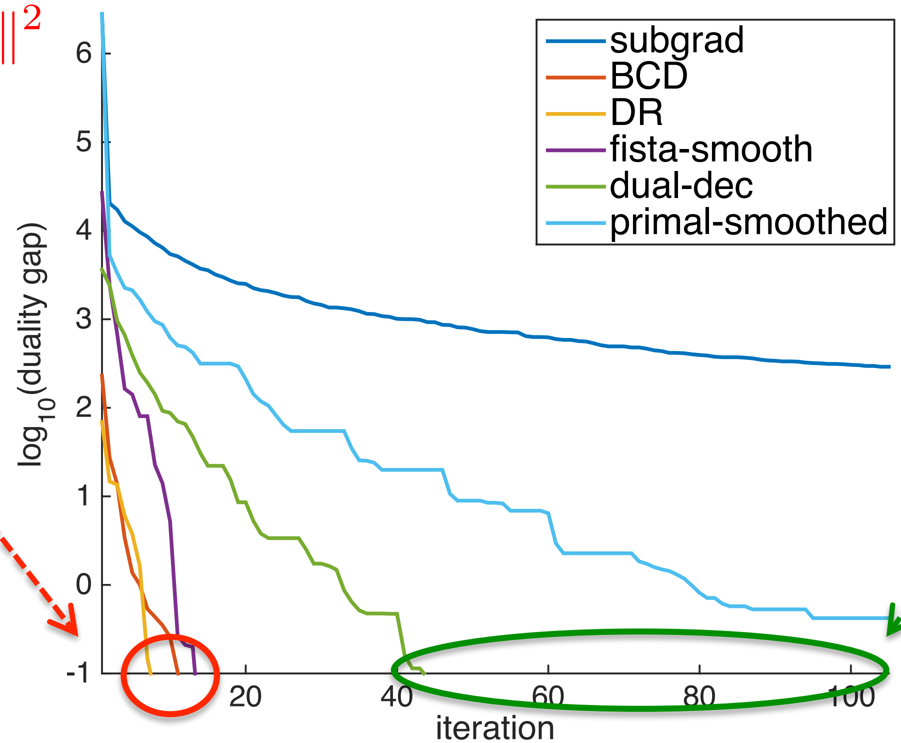
relax II

$$\min_x \sum_i f_i(x) + \tfrac{1}{2}\|x\|^2$$

smooth dual



convergence discrete problem



relaxation I

$$\min_{x \in [0,1]^n} \sum_i f_i(x)$$

non-smooth dual

faster
parallel
algorithms ☺

(Jegelka, Bach, Sra 2013; Nishihara, Jegelka, Jordan 2014)

Submodularity and convexity

- convex Lovasz extension
  - easy to compute: greedy algorithm (special polyhedra!)

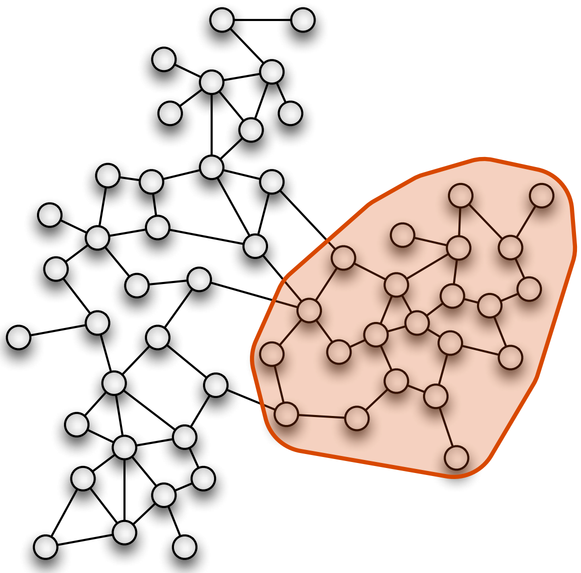
- submodular minimization via convex optimization: fast algorithms for many applications
- duality results

- structured sparsity *(Bach 2010)*
- decomposition & parallel algorithms
- variational inference *(Djolonga & Krause 2014)*
- …

Roadmap

- Submodular set functions
  - definition & basic properties
  - links to convexity
  - special polyhedra
- Minimizing submodular functions
  - general and special cases
  - constraints
- Maximizing submodular functions

Graph cuts

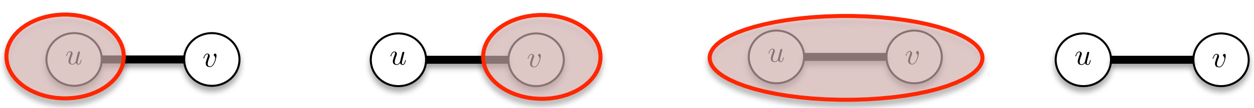


$$G = (V, E) \qquad \text{edge weights } w_{uv}$$

$$F(S) = \sum_{u \in S, v \notin S} w_{uv}$$

- Cut for one edge:

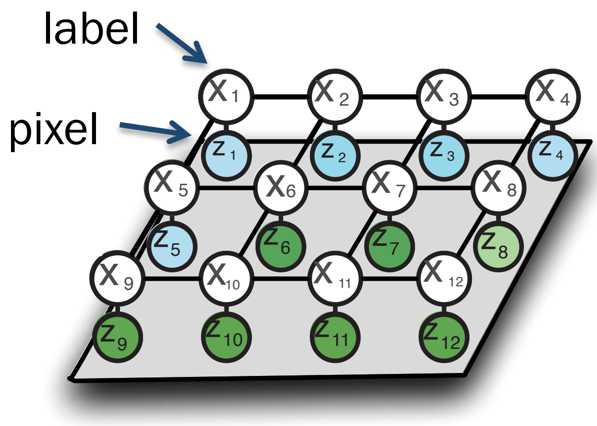
$$F(\{u\}) + F(\{v\}) \geq F(\{u, v\}) + F(\emptyset)$$



w_{uv} w_{uv} 0 0

- Graph cuts: faster algorithms than general submodular!

MAP inference



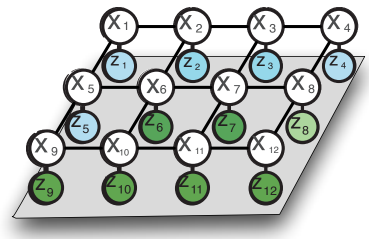
$$\max_{\mathbf{x}\in\{0,1\}^n} P(\mathbf{x} \mid \mathbf{z}) \propto \exp(-E(\mathbf{x};\mathbf{z}))$$

labels

pixel values

$$\Leftrightarrow \quad \min_{\mathbf{x}\in\{0,1\}^n} E(\mathbf{x};\mathbf{z})$$

Attractive potentials



$$P(\mathbf{x} \mid \mathbf{z}) \propto \exp(-E(\mathbf{x};\mathbf{z}))$$

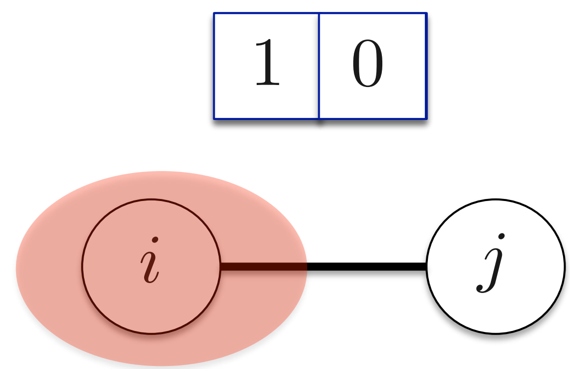
$$E(\mathbf{x};\mathbf{z}) = \sum_i E_i(x_i) + \sum_{ij} E_{ij}(x_i, x_j)$$

spatial coherence:

$$E_{ij}(x_i, x_j) = \begin{cases} \nu_{ij} > 0 & \text{if } x_i \neq x_j \\ 0 & \text{otherwise} \end{cases}$$

$$= \nu_{ij}(x_i - x_j)^2$$

$$= \mathrm{Cut}(1_S)$$



Quadratic functions and cuts

$$E_{ij}(x_i, x_j) = \nu_{ij}(x_i - x_j)^2 \qquad = \mathrm{Cut}(1_S)$$

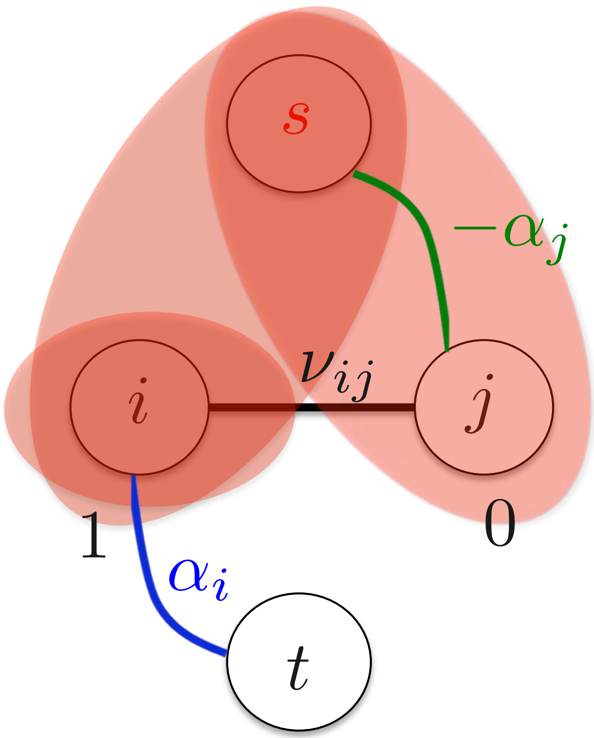
what about the linear terms?

$$E_{ij}(x_i, x_j) + E_i(x_i) + E_j(x_j) - \alpha_j$$

make it an (s,t)-cut!
Source s always selected

$$\alpha_i > 0$$

$$\alpha_j < 0$$



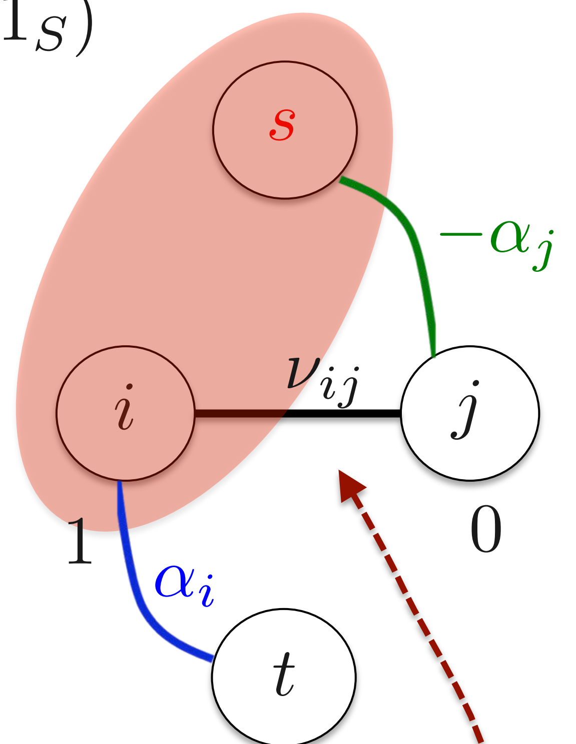
Quadratic functions and cuts

$$E_{ij}(x_i, x_j) \;= \nu_{ij}(x_i - x_j)^2 \qquad = \mathrm{Cut}(1_S)$$

linear terms:

$$E_{ij}(x_i, x_j) + \alpha_i x_i \;+\; \alpha_j x_j \quad -\alpha_j$$

MAP inference = minimum cut ☺



Is every such quadratic function a graph cut function?

any linear terms are fine ☺

quadratic terms? What if $\nu_{ij} < 0$?

negative edge weights!

Submodular functions and cuts

- Formally:

$$E(x) = \sum_i \alpha_i x_i + \sum_{ij} \nu_{ij}(x_i - x_j)^2$$

$$= \sum_i \alpha_i x_i + \sum_{ij} \nu_{ij} x_i^2 + \nu_{ij} x_j^2 - 2\nu_{ij} x_i x_j$$

$$= \sum_i \beta_i x_i + \sum_{ij} -2\nu_{ij} x_i x_j$$

Every quadratic pseudo-boolean function with non-positive second-order coefficients is a graph cut function.

Equivalent condition: *E(x)* is submodular.

Quadratic submodular functions

$$E(x) = \beta_i x_i + \beta_j x_j + \nu'_{ij} x_j x_j$$

check submodularity: $V = \{ i, j \}$

$$F(S) + F(T) \;\geq\; F(S \cap T) + F(S \cup T)$$

$S = \{i\},\ T = \{j\}$:

$$E(1,0) + E(0,1) \;\geq\; E(0,0) + E(1,1)$$

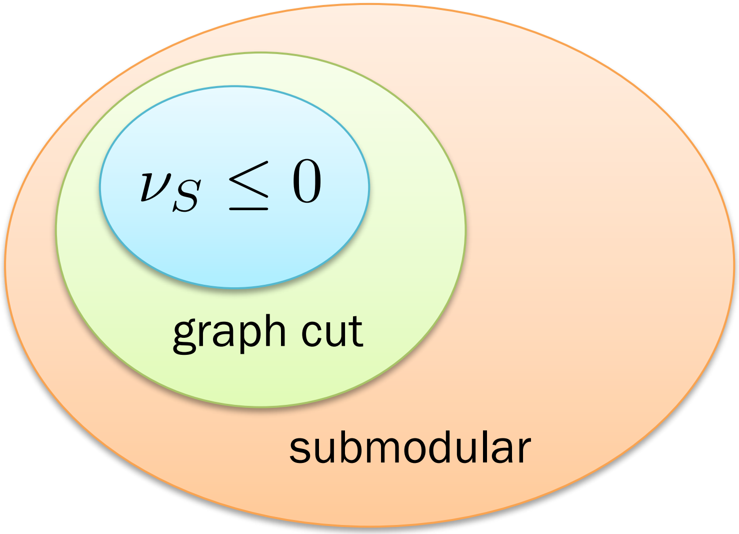
$$\beta_i \quad + \beta_j \quad \geq \quad 0 \quad + \beta_i + \beta_j + \nu'_{ij}$$

$$E(x) \text{ submodular} \quad \Leftrightarrow \quad \nu'_{ij} \leq 0$$

Submodular polynomials & cuts

$$\sum_i \beta_i x_i + \sum_{ij} \nu_{ij} x_i x_j + \sum_{ijk} \nu_{ijk} x_i x_j x_k + \ldots$$

- What if my function has degree > 2 ?
- If $\nu_S \leq 0$ for all ν_S then *E(x)* is submodular and can be written as a graph cut
- In general:



Submodular functions generalize graph cuts!

independence

entropy

submodularity

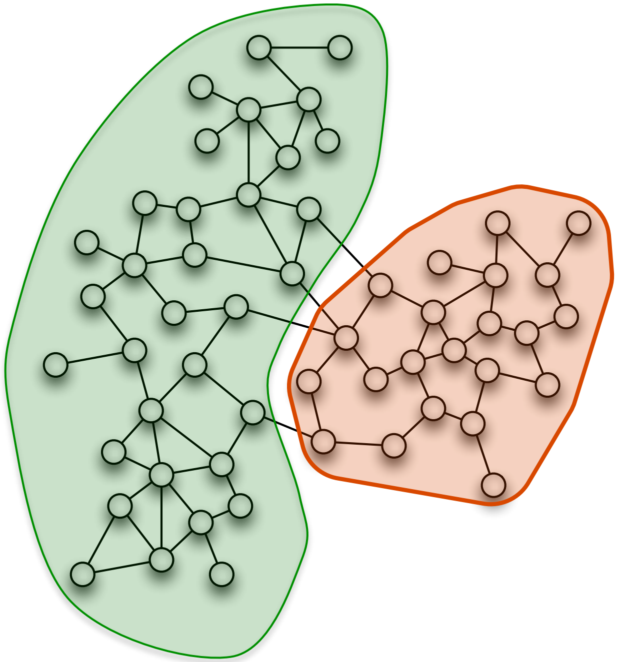
(matrix & combinatorial) rank functions

cut functions

Submodular functions generalize graph cuts!

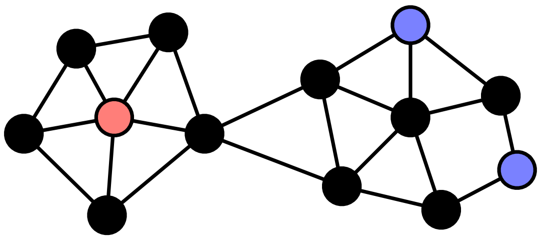
- symmetric submodular functions:

$$F(S) = F(\mathcal{V} \setminus S)$$



- special algorithm for minimizing symmetric submodular functions in $O(n^3)$ time *(Queyranne, 1998)* generalization of a min-cut algorithm!

- semi-supervised learning *(Guillory & Bilmes 2011, Hein et al 2013)*



Roadmap

- Submodular set functions
  - definition & basic properties
  - links to convexity
  - special polyhedra
- Minimizing submodular functions
  - general and special cases
  - constraints
- Maximizing submodular functions

Constraints?

e.g. $\min \ F(S) \ \text{s.t.} \ |S| = k$

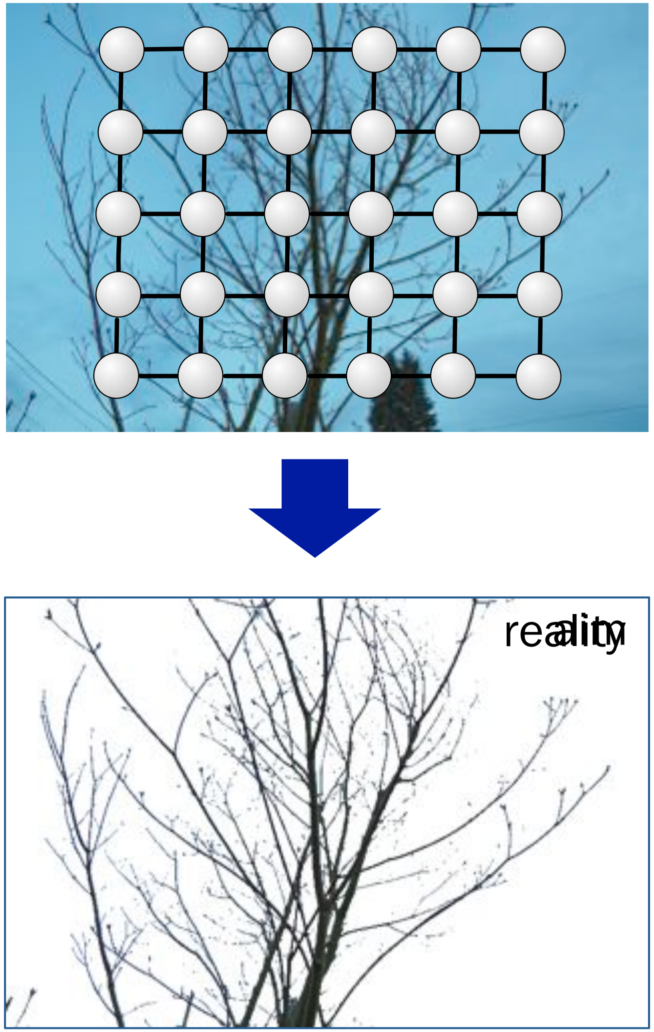
generally: $\min \ F(S) \ \text{s.t.} \ S \in \mathcal{C}$

- In most cases, this is NP-hard and cannot even be approximated within a constant factor ☹
- What to do? approximation

$\widehat{S} \in \mathcal{C}$ and $F(\widehat{S}) \leq \alpha \, F(S^*)$

approximate solution

Recall: MAP and cuts

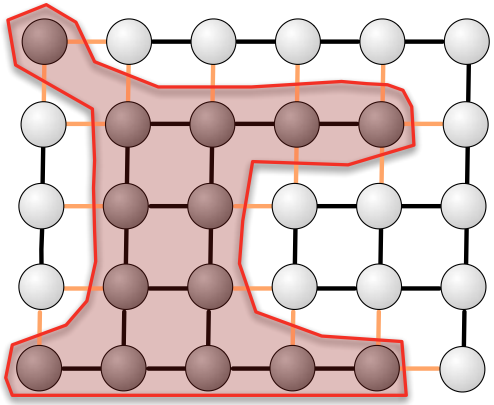


binary labeling: $x = 1_A$

pairwise random field:

$$E(x) = \mathrm{Cut}(A)$$

What's the problem?

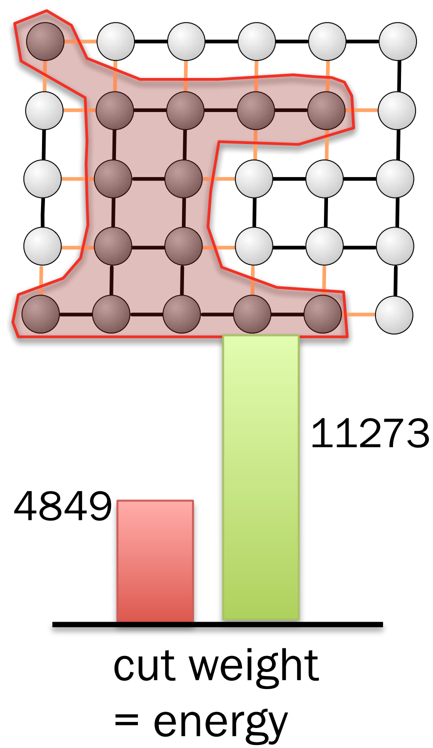


minimum cut: prefer
short cut = short object boundary

What’s wrong?

we get …

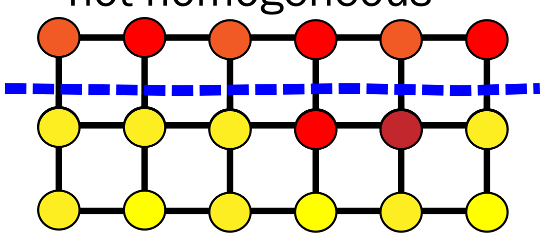
local coherence
= short cut



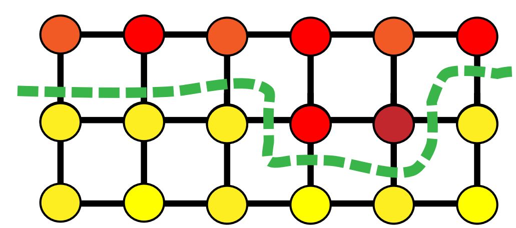
ideally …

homogeneous cut
global dependencies!

not homogeneous



homogeneous



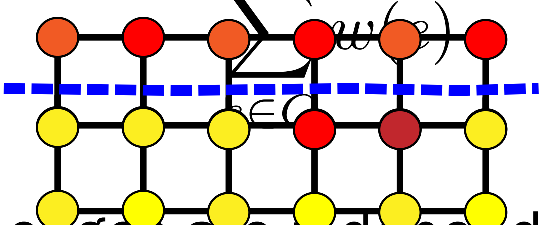
Cooperative cuts

ideally ...

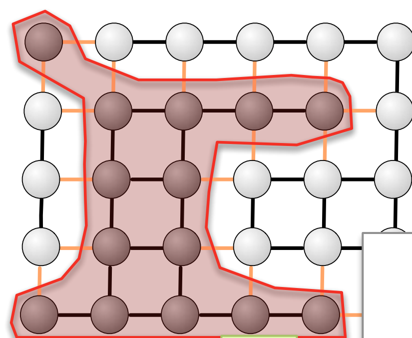
local coherence
= short cut

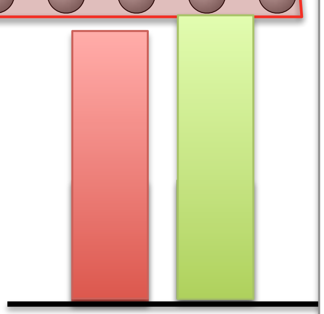
cost of a cut $C \subseteq \mathcal{E}$:

$$\sum_{e \in C} w(e)$$



edges are independent





cut weight
= energy

homogeneous cut
global dependencies!

cooperative graph cut

cost of a cut $C \subseteq \mathcal{E}$:

submodular function

$$F(C)$$

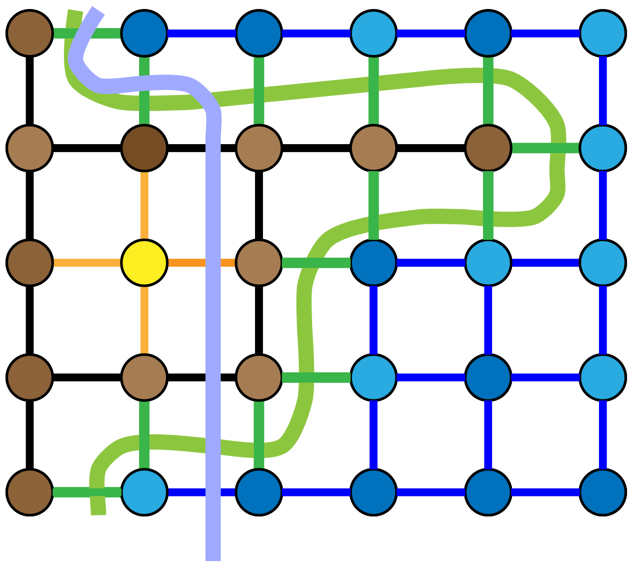
edges are not independent

Homogeneity via group sparsity

sum of weights:
use few edges

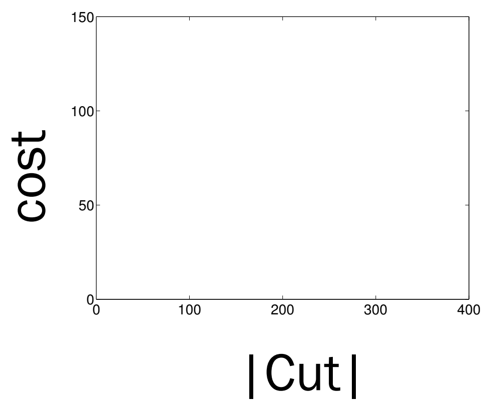
→

submodular cost function:
use few types of edges



One type (13 edges)
Many types (6 edges)

$$F(\mathrm{Cut}) = \sum_{\text{type } k} F_k(\mathrm{Cut})$$



(Jegelka & Bilmes 2011)

Results

Random Walker
[Grady 06]

Curvature Reg.
[El-Zehiry & Grady 10]

Graph Cut
[Boykov & Jolly 01]

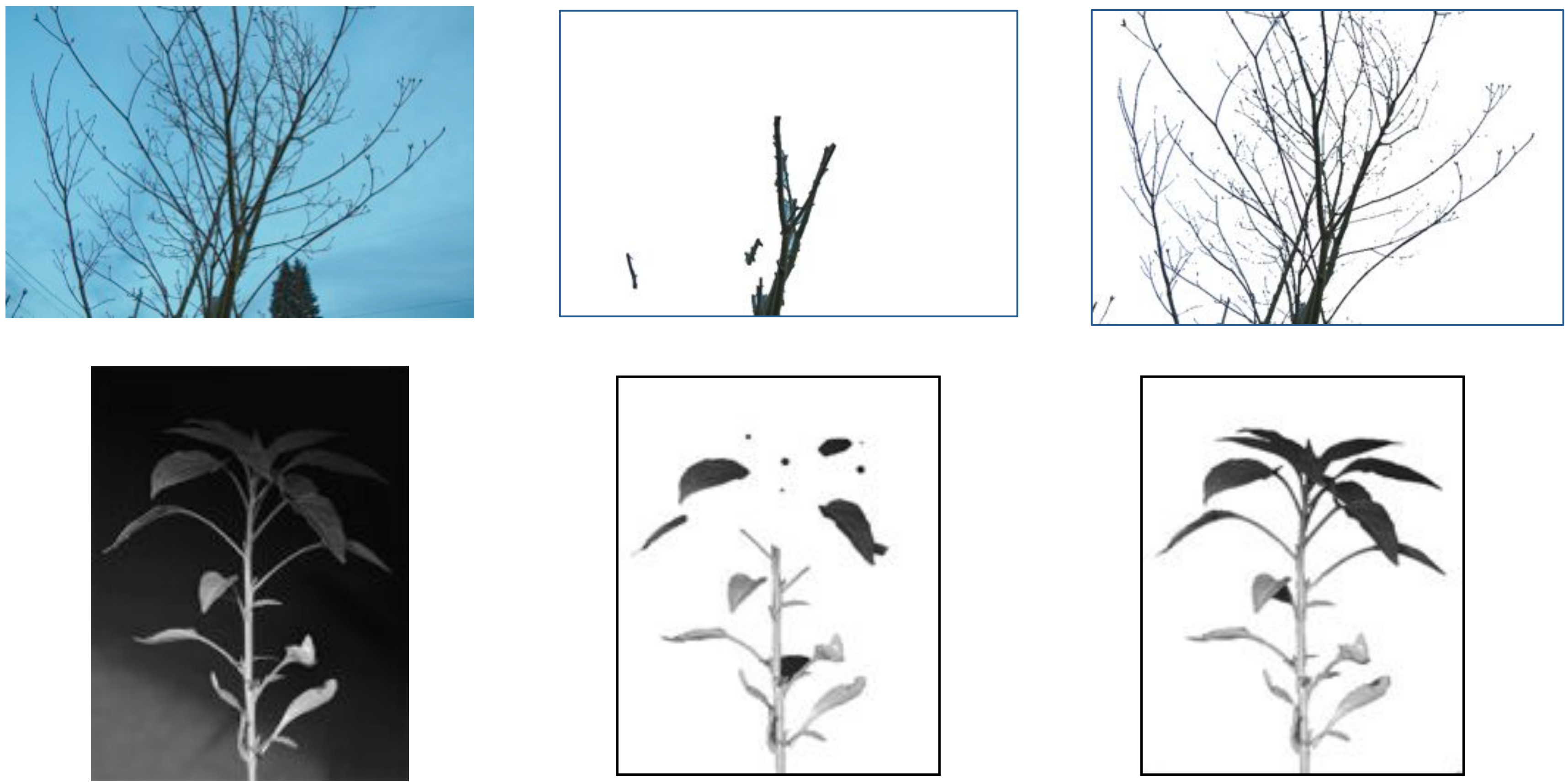
Cooperative Cut

Quantitatively: up to 70% reduction in error!

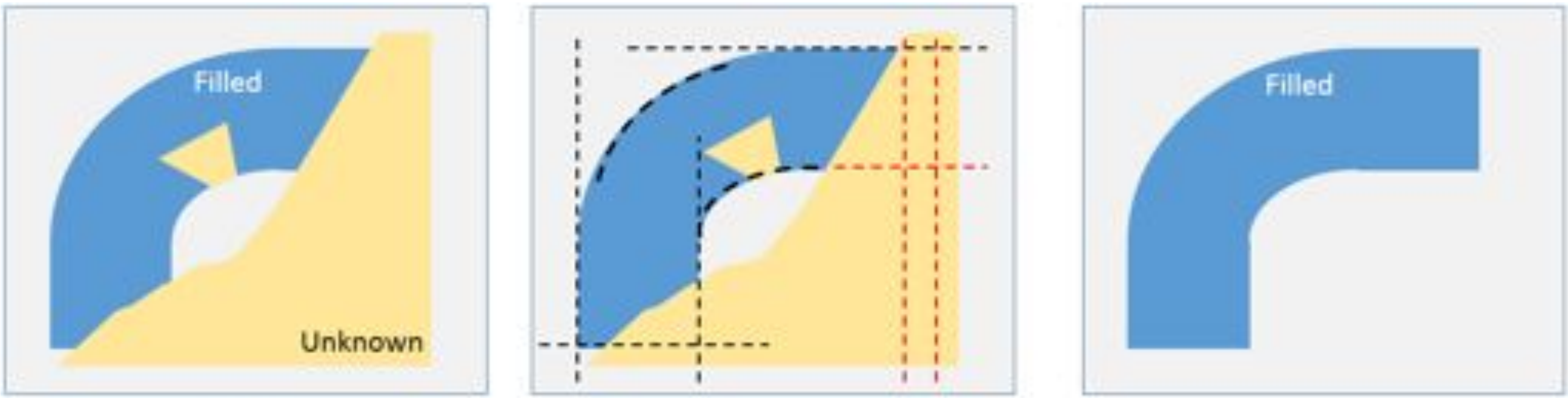
Results

Graph cut

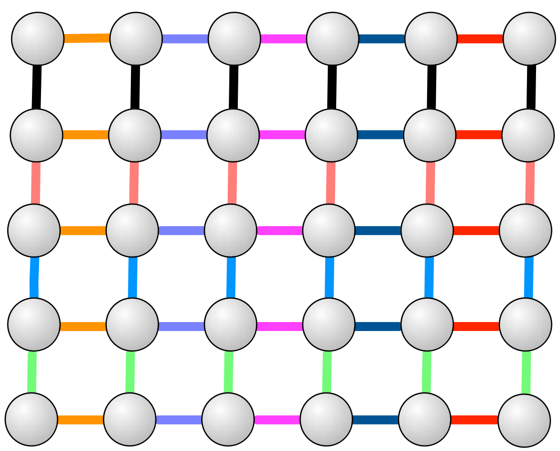
Cooperative cut



Similarly: contour completion RF



(a) (b) (c)



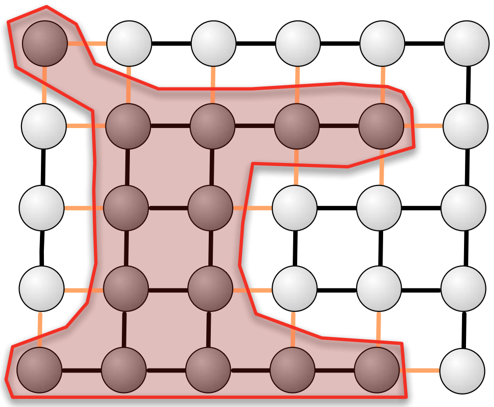
geometric edge groups:

- straight lines
- parabolas

(Silberman et al 2014)

Minimum cut?

- find a minimum cut
  with cost function: $C \subseteq \mathcal{E}$



normally:

$$\mathrm{Cost}(C) = \sum_{e \in C} w(e)$$

now:

$$\mathrm{Cost}(C) = F(C)$$

$$\boxed{\min F(S) \text{ s.t. constraints on } S}$$

Constrained optimization

2 strategies:

- convex relaxation using the Lovasz extension
- approximate the submodular function

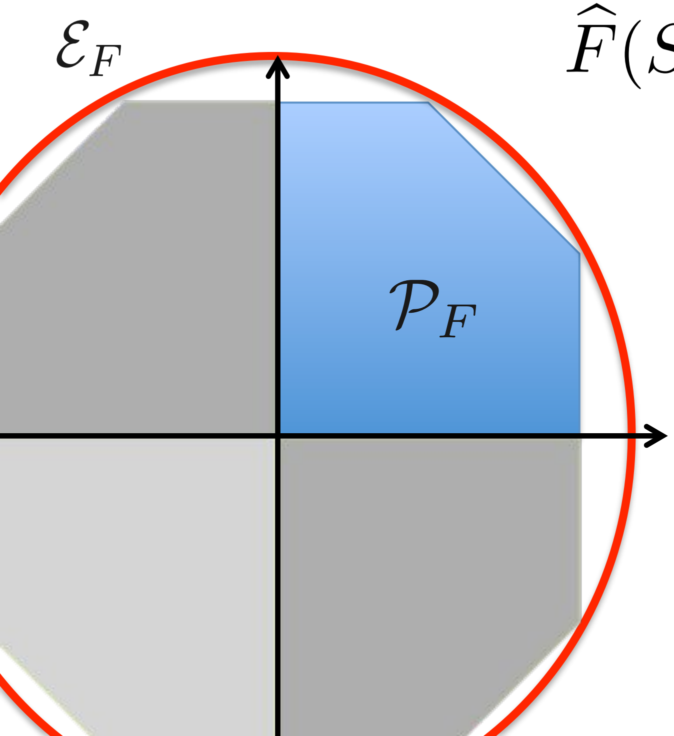
$$S^* = \arg\min_{S \in \mathcal{C}} F(S) \longrightarrow \widehat{S} = \arg\min_{S \in \mathcal{C}} \widehat{F}(S)$$

Approximation of *F*

- recall: for $x = 1_S$ and *F* increasing

$$F(S) = f(x) = \max_{y \in \mathcal{P}_F} y^\top x$$

approximate polyhedron by ellipsoid



$$\widehat{F}(S) = \max_{y \in \mathcal{E}_F} y^\top x \quad = \sqrt{\sum_{e \in S} w(e)}$$

- Why is this easier than *F* ?

$$\widehat{S} = \arg\min_{S \in \mathcal{C}} \widehat{F}(S) = \arg\min_{S \in \mathcal{C}} \sum_{e \in S} w(e)$$

(*Goemans et al 2009*)

How good is this?

$$\widehat{F}(S) = \max_{y \in \mathcal{E}_F} y^\top x \qquad \widehat{S} = \arg\min_{S \in \mathcal{C}} \widehat{F}(S)$$

- One can show that for all S:

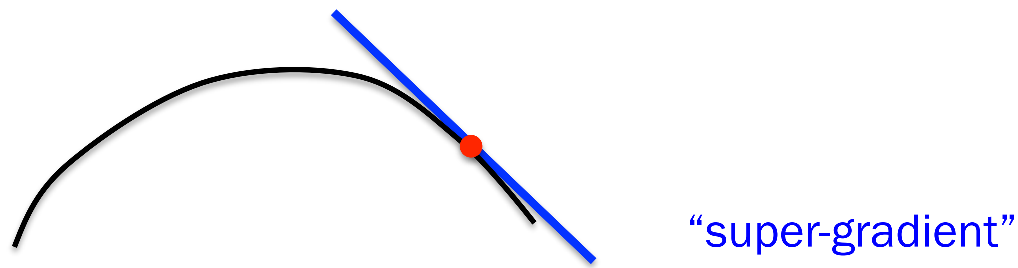
$$F(S) \;\leq\; \widehat{F}(S) \;\leq \alpha F(S) \qquad \alpha = O(\sqrt{n}\log n)$$

- It follows that:

$$F(\widehat{S}) \;\leq\; \widehat{F}(\widehat{S}) \;\leq\; \widehat{F}(S^*) \;\leq\; \alpha\, F(S^*)$$

A practical approximation

idea: submodularity = discrete concavity



For *i = 1,2,...*

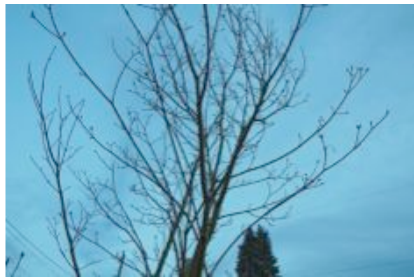
- compute linear upper bound $\widehat{F}_i$ with $\widehat{F}_i(C_i) = F(C_i)$
- find tree/path/... C_{i+1} with minimum $\widehat{F}_i(C)$.

familiar ☺
e.g. min-cut

fast: only need to solve linear optimization problem!

(Jegelka & Bilmes 2011; Iyer, Jegelka, Bilmes 2013)

Does it work?



minimum cut solution

approximate solution

optimal solution

- often works well in practice
- theory: approximation guarantees depending on curvature of *F*

  (Iyer et al 2013)
- special cases: exact solution *(Kohli et al 2013)*

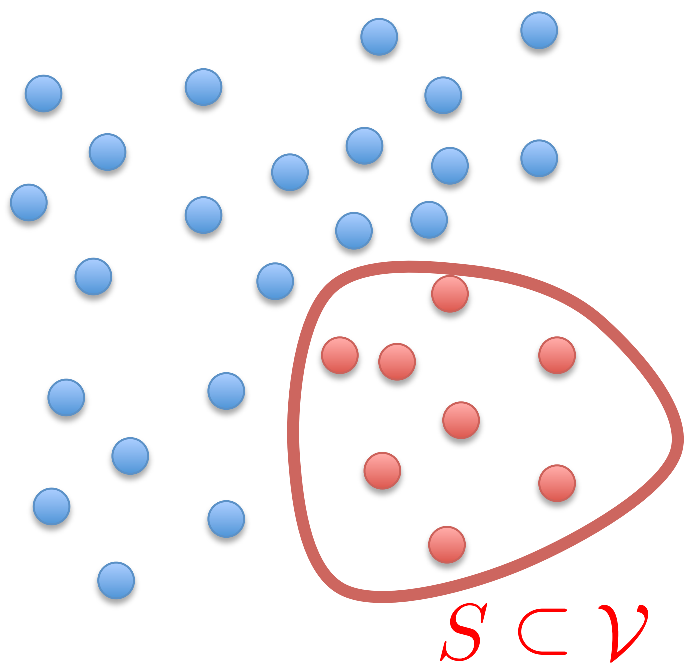
Roadmap

- Submodular set functions
  - definition & basic properties
  - links to convexity
  - special polyhedra

- Minimizing submodular functions
  - general and special cases
  - constraints

- Maximizing submodular functions

Setup

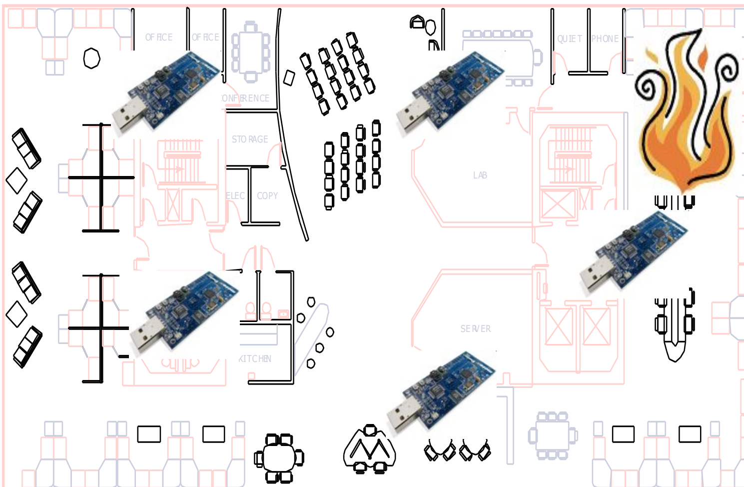


- ground set $\mathcal{V}$
- (scoring) function

$$F : 2^{\mathcal{V}} \to \mathbb{R}_+$$

$$\boxed{\max \quad F(S)}$$

Informative Subsets

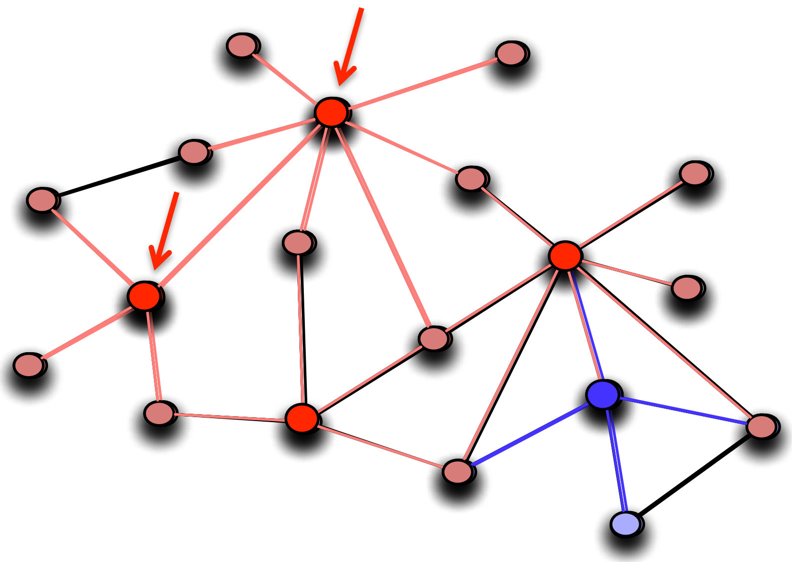


- where put sensors?
- which experiments?
- summarization

$$F(S) = \text{“information”}$$

Maximizing Influence

$$F(S) = \text{expected \# infected nodes}$$

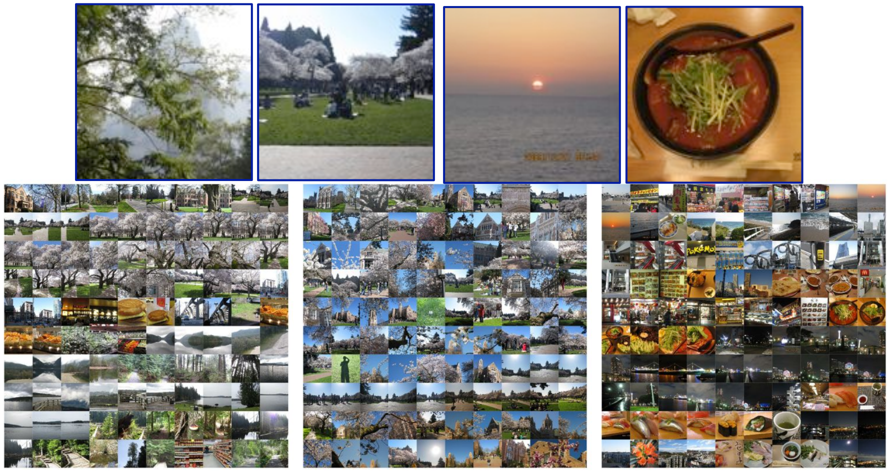


$$F(S \cup s) - F(S) \geq F(T \cup s) - F(T)$$

Kempe, Kleinberg & Tardos 2003

Summarization

- videos, text, pictures ...
- would like:
  relevance, reliability, diversity



Summarization

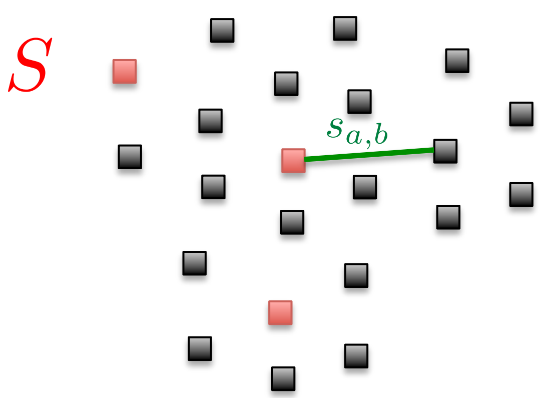
$$F(S) = R(S) + D(S)$$

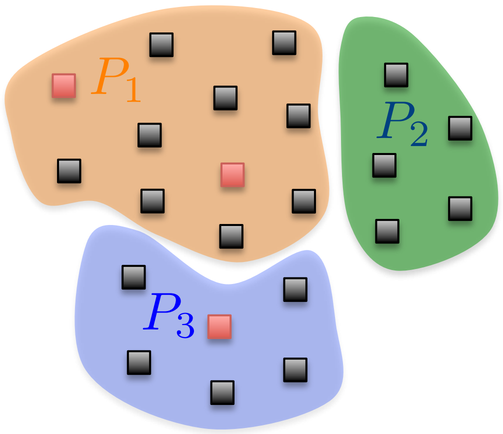
- Coverage / relevance

$$R(S) = \sum_{a \in \mathcal{V}} \max_{b \in S} s_{a,b}$$

- Diversity

$$D(S) = \sum_{j=1}^{m} \sqrt{|S \cap P_j|}$$



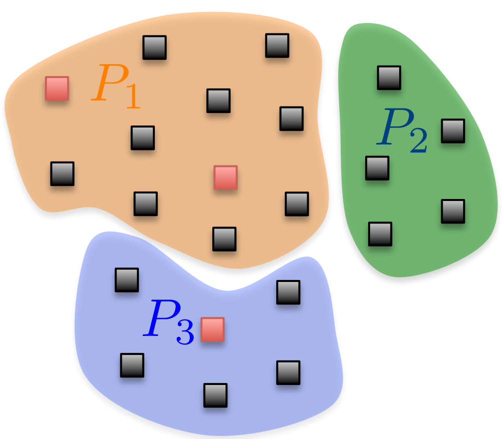


(Simon et al 2007, Lin & Bilmes 2011&2012, Tschiatschek et al 2014, Kim et al 2014, Gygli et al 2015, …)

Diversity

- Diversity

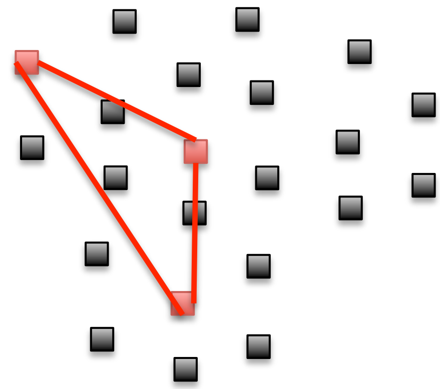
$$D(S) = \sum_{j=1}^{m} \sqrt{|S \cap P_j|}$$



increasing

Another diversity function ...

$$D(S) = -\sum_{a,b \in S} s_{a,b}$$



decreasing

Summarization: results

	R	F
$\mathcal{L}_1(S) + \lambda \mathcal{R}_Q(S)$	12.18	12.13
$\mathcal{L}_1(S) + \sum_{\kappa=1}^{3} \lambda_\kappa \mathcal{R}_{Q,\kappa}(S)$	**12.38**	**12.33**
Toutanova et al. (2007)	11.89	11.89
Haghighi and Vanderwende (2009)	11.80	-
Celikyilmaz and Hakkani-tür (2010)	11.40	-
Best system in DUC-07 (peer 15), using web search	**12.45**	12.29

(Lin & Bilmes 2011)

Many more functions are possible …
➔ Learn a weighted combination: via "structured prediction"
➔ works even better!

(Lin & Bilmes 2012, Tschiatschek et al 2014, Gygli et al 2015, Xu et al 2015,…)

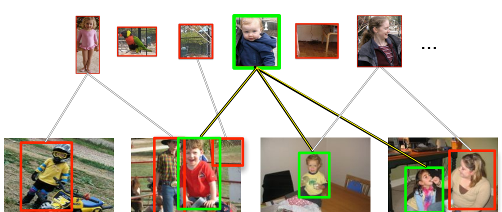
More maximization ...

co-segmentation
by maximizing
anisotropic diffusion
(Kim et al 2011)

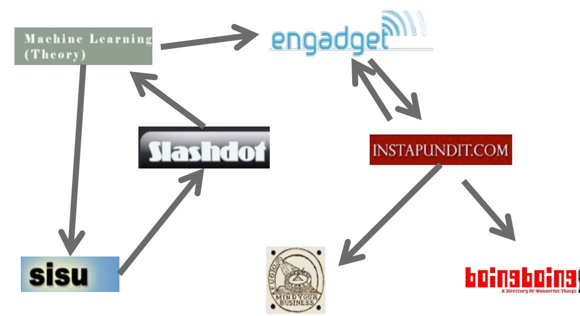
environmental monitoring
(Krause, ...)

$$\max \quad F(S)$$

weakly supervised
object detection
(Song et al 2014)



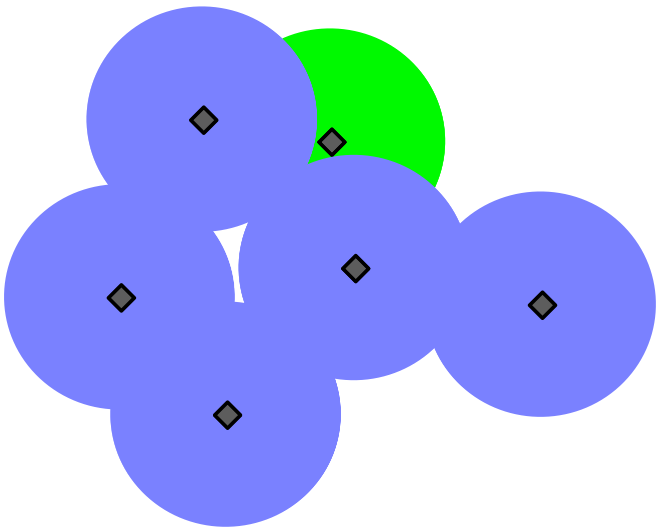
diverse
recommendations
(Yue & Guestrin)

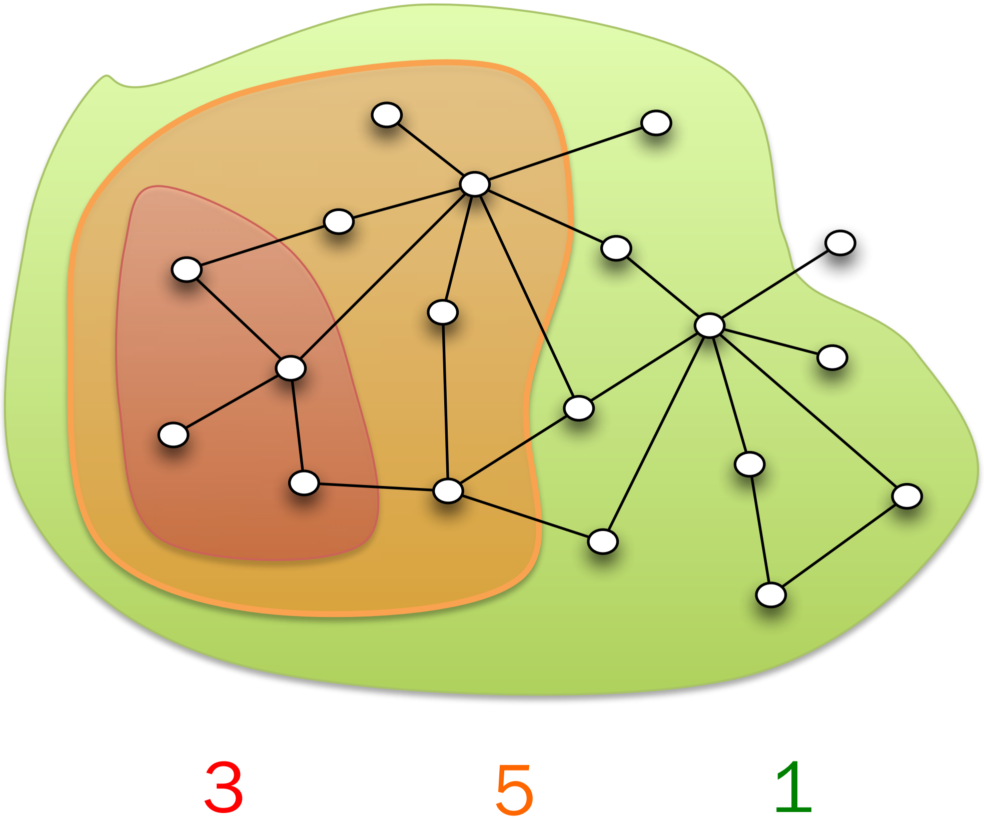


inferring networks
(Gomez Rodriguez et al 2012)

Monotonicity

$$\text{if} \quad S \subseteq T \quad \text{then} \quad F(S) \leq F(T)$$



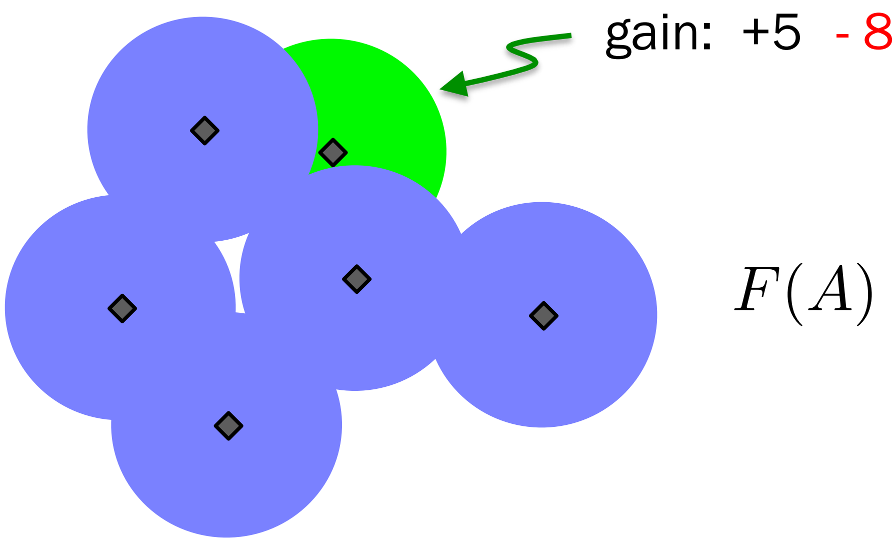


Monotonicity – how check?

$$\text{if } A \subseteq B \quad \text{then} \quad F(A) \leq F(B)$$

Let $B = A \cup \{a\}$.

$$\underbrace{F(A \cup \{a\}) - F(A)}_{\text{marginal gain}} \geq 0.$$



$$F(A) = \left| \bigcup_{a \in A} \text{area}(a) \right| - \sum_{a \in A} c(a)$$

Maximizing monotone functions

$$\text{if } A \subseteq B \quad \text{then} \quad F(A) \leq F(B)$$

$$\max \; F(S)$$

- NP-hard
- approximation: greedy algorithms

Maximizing monotone functions

$$\max_{S} \; F(S) \;\; \text{s.t.} \;\; |S| \leq k$$

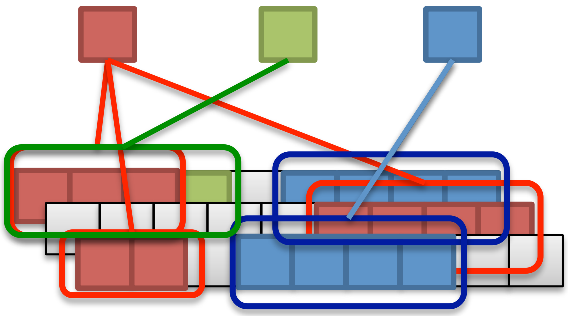
- greedy algorithm:

$S_0 = \emptyset$

for *i* = *0, …, k-1*

$$e^* = \arg \max_{e \in \mathcal{V} \setminus S_i} F(S_i \cup \{e\})$$

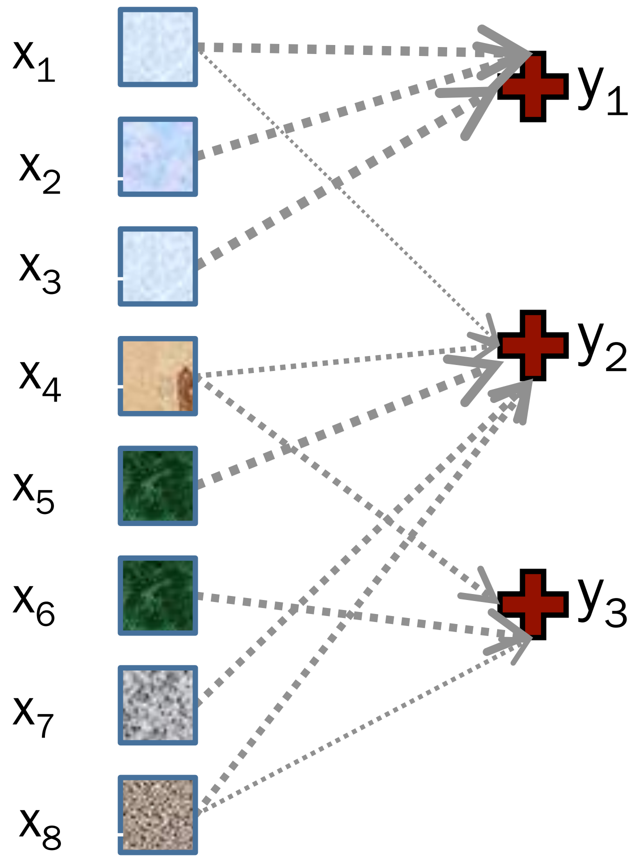
$$S_{i+1} = S_i \cup \{e^*\}$$



How "good" is S_k ?

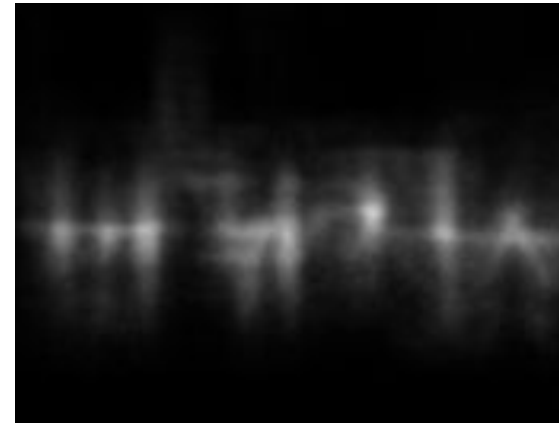
Pedestrian detection

x_j = index of hypothesis explaining x_j



Voting elements

Hypotheses



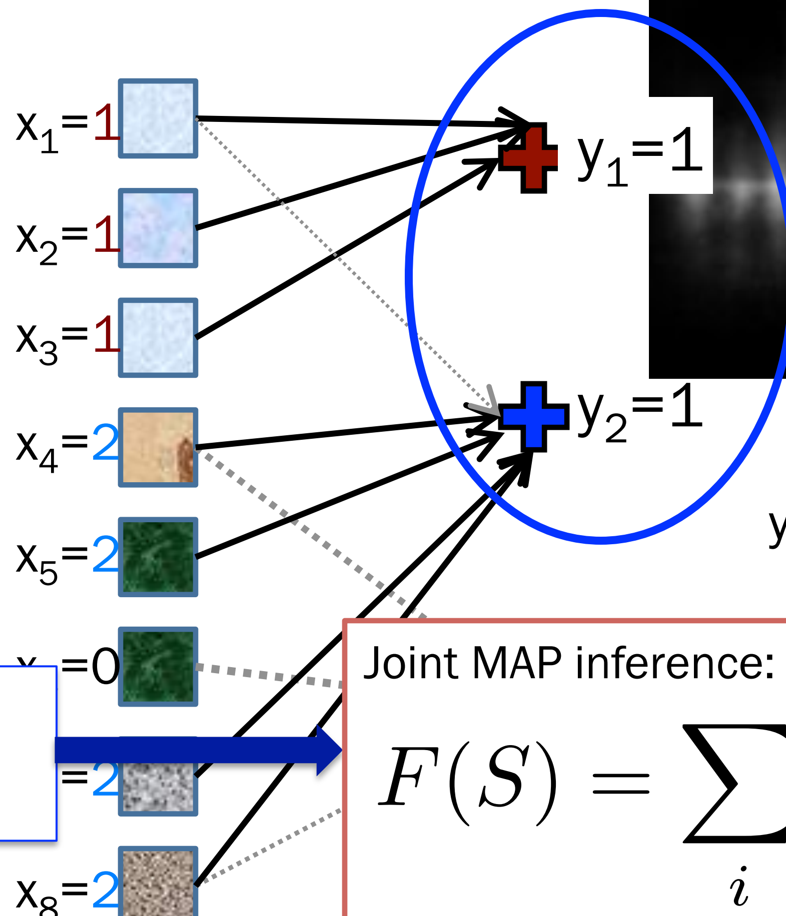
y_i = 1: object i present

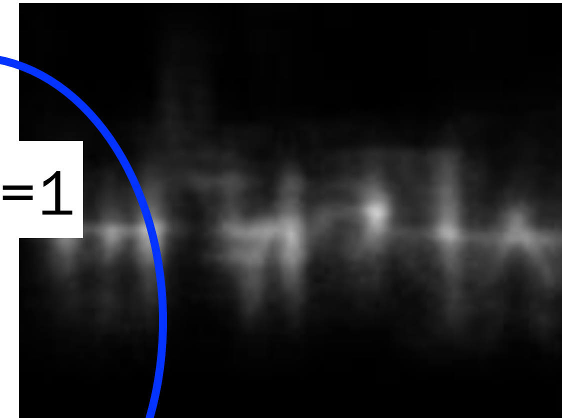
y_i = 0: object i not present

Illustrations courtesy of Pushmeet Kohli

(Barinova et al.'10)

Object detection





x_j = index of hypothesis explaining x_j

y_i = 1: object i present

submodular maximization!

Joint MAP inference:

$$F(S) = \sum_i \max_{j \in S} w_{i,j}$$

Weight element i wrt hyp. j

Voting elements

Illustrations courtesy of Pushmeet Kohli

Inference

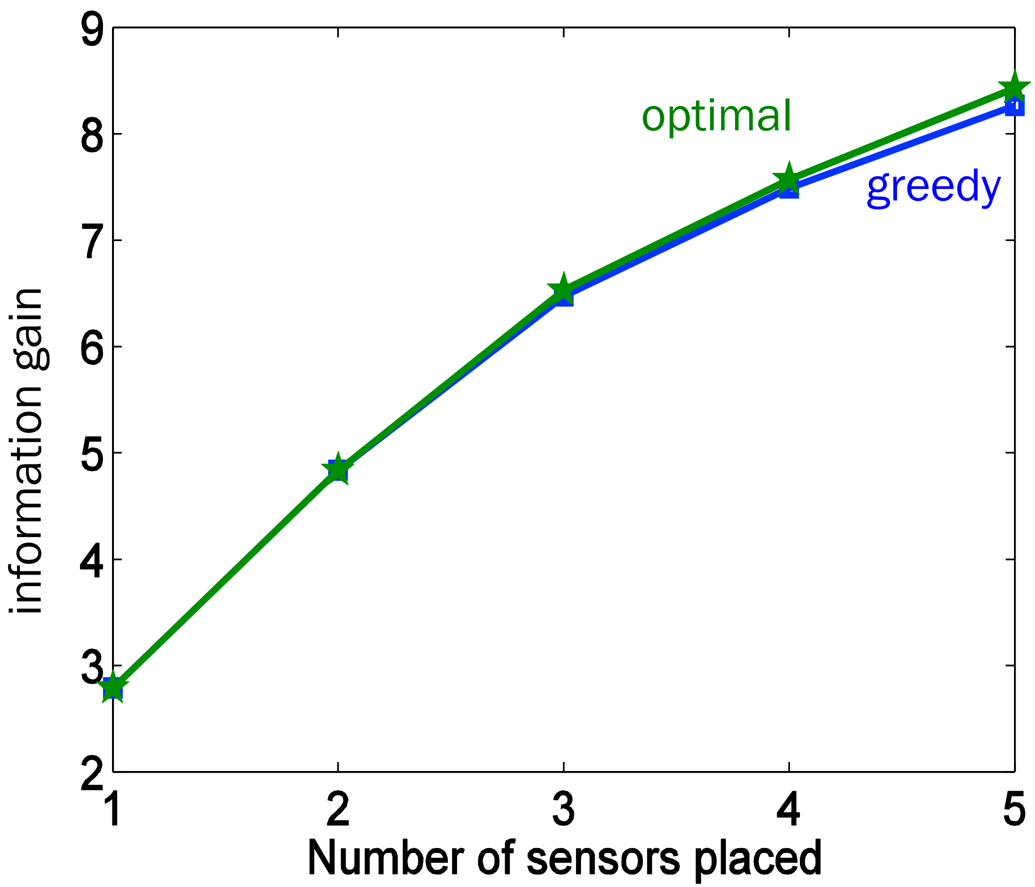
Datasets from [Andriluka et al. CVPR 2008]
(with strongly occluded pedestrians added)

Using the Hough forest trained in [Gall&Lempitsky CVPR09]

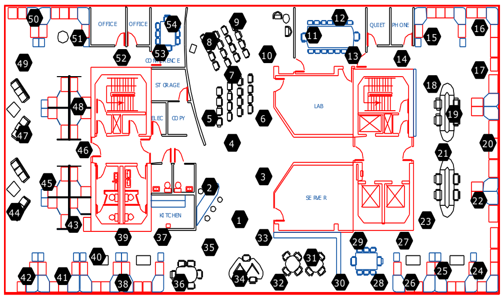
Illustrations courtesy of Pushmeet Kohli

How good is greedy? in practice...

empirically:



sensor placement



How good is greedy? ... in theory

$$\max_S \; F(S) \;\text{ s.t. }\; |S| \leq k$$

Theorem *(Nemhauser, Fisher, Wolsey `78)*

F monotone submodular, S_k solution of greedy. Then

$$F(S_k) \;\geq\; \left(1 - \frac{1}{e}\right) F(S^*)$$

optimal solution

in general, no poly-time algorithm can do better than that!

Questions

- What if I have more complex constraints?
  - budget constraints
  - matroid constraints

- Greedy takes *O(nk)* time. What if n, k are large?

- What if my function is not monotone?

Roadmap

- Submodular set functions
  - definition & basic properties
  - links to convexity
  - special polyhedra
- Minimizing submodular functions
  - general and special cases
  - constraints
- Maximizing submodular functions