What Is Decreased by the Max-sum Arc Consistency Algorithm?

Tomáš Werner

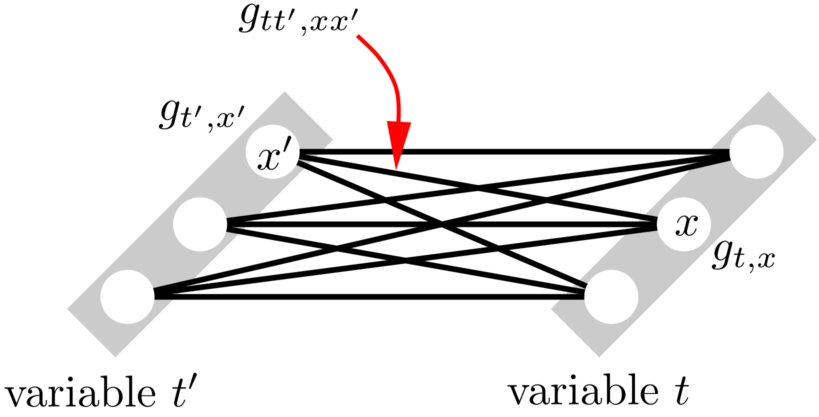
Center for Machine Perception
Czech Technical University
Prague, Czech Republic

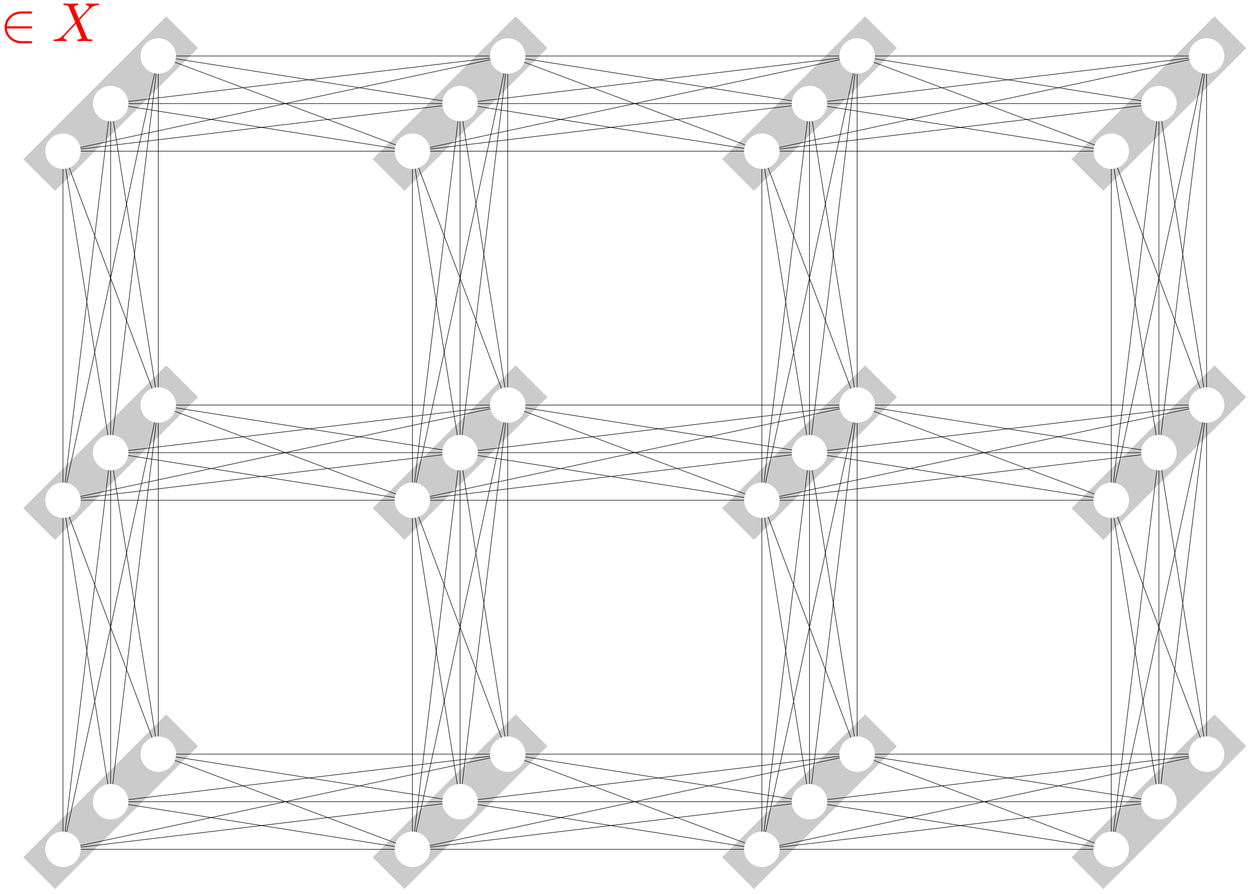
ICML 2007, Corvallis

- Finding maximum of a MRF-defined distribution: Problem formulation
- Background: Not widely-known approach to solving the problem
  - Minimizing LP-based upper bound [Schlesinger-76]
  - Max-sum diffusion algorithm [Koval-Kovalevsky-76]
  - Example on syntactic image analysis
- Contribution 1: Max-sum diffusion is an arc consistency algorithm
- Contribution 2: Strictly monotonically decreasing criterion

Undirected graphical model (MRF) with max. cliques of size 2 is given by

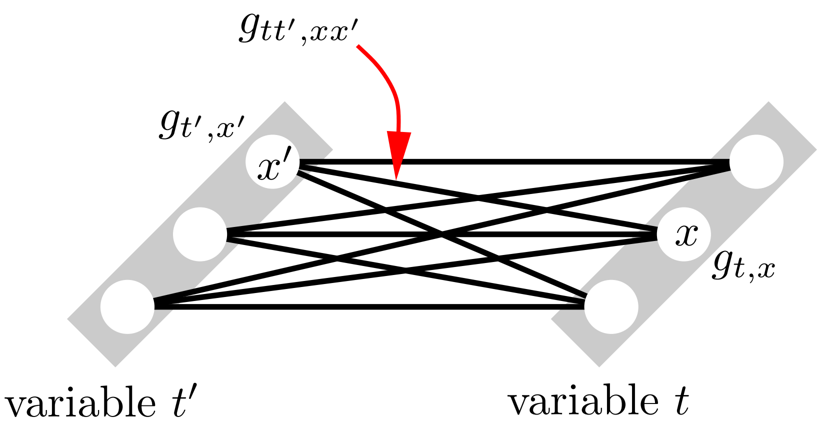
- variables $t \in T$ with finite states $x \in X$
- undirected graph $\langle T, E \rangle$
- weights $g_{t,x},\ g_{tt',xx'} \in \mathbb{R}$

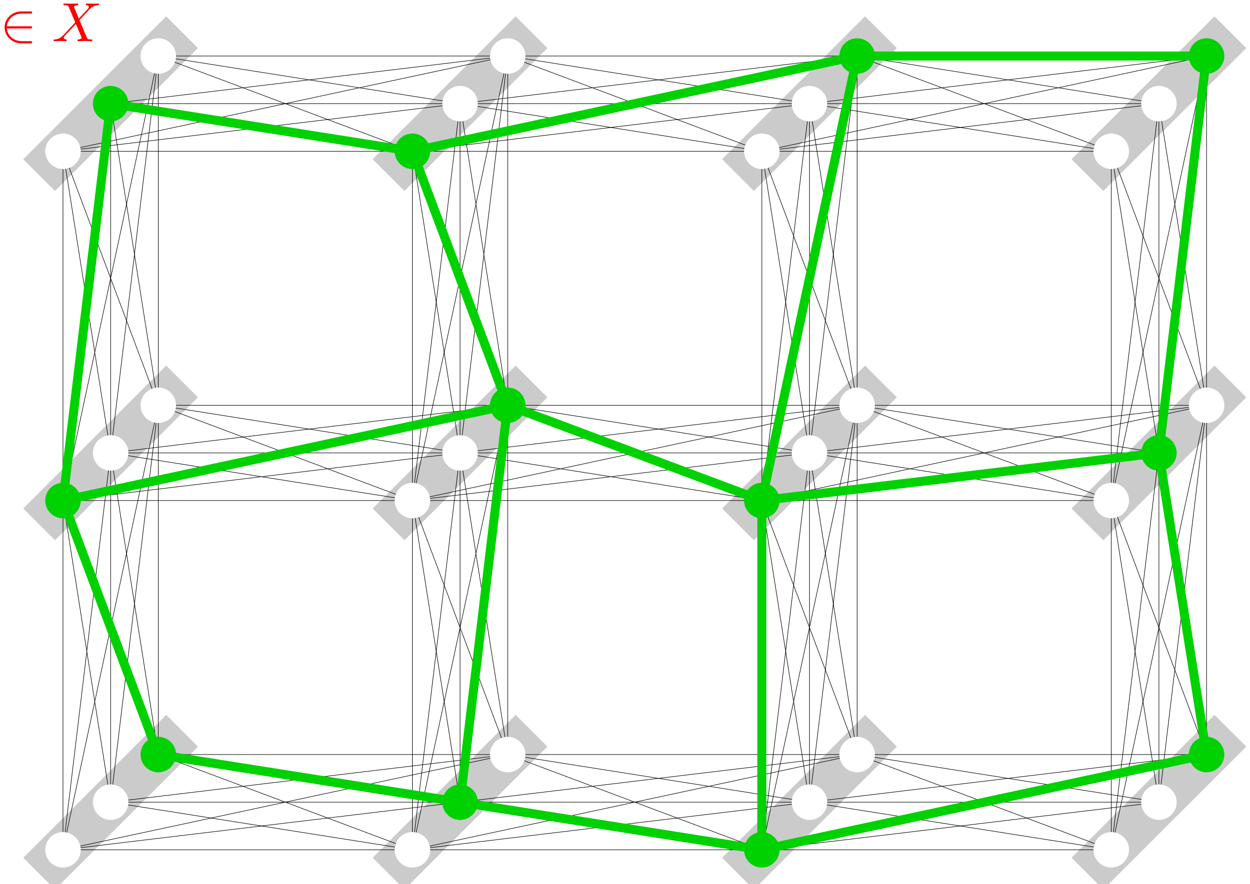




Undirected graphical model (MRF) with max. cliques of size 2 is given by

- variables $t \in T$ with finite states $x \in X$
- undirected graph $\langle T, E \rangle$
- weights $g_{t,x},\ g_{tt',xx'} \in \mathbb{R}$

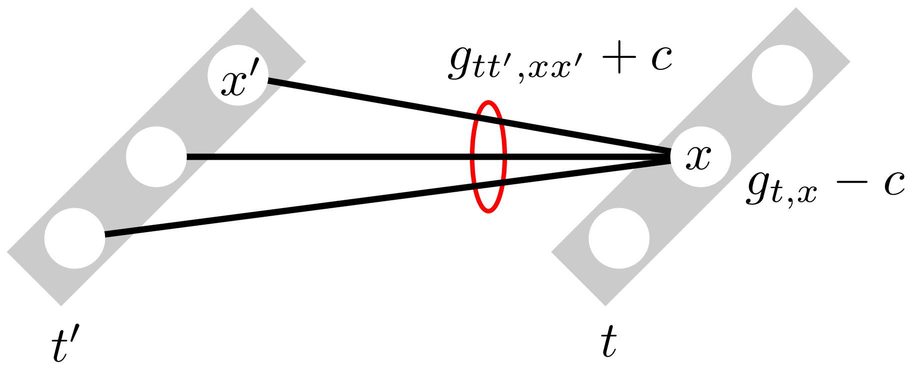




- weight of configuration $\mathbf{x} \in X^T$: $F(\mathbf{x}\,|\,\mathbf{g}) = \sum_{t \in T} g_{t,x_t} + \sum_{\{t,t'\} \in E} g_{tt',x_t x_{t'}} \quad \propto \log p(\mathbf{x}\,|\,\mathbf{g})$
- find maximum over all configurations: $F(\mathbf{g}) = \max_{\mathbf{x} \in X^T} F(\mathbf{x}\,|\,\mathbf{g})$

- Weight vectors $\mathbf{g}$ and $\mathbf{g}'$ are equivalent iff $F(\mathbf{x}|\mathbf{g}) = F(\mathbf{x}|\mathbf{g}')$ for all $\mathbf{x} \in X^T$.

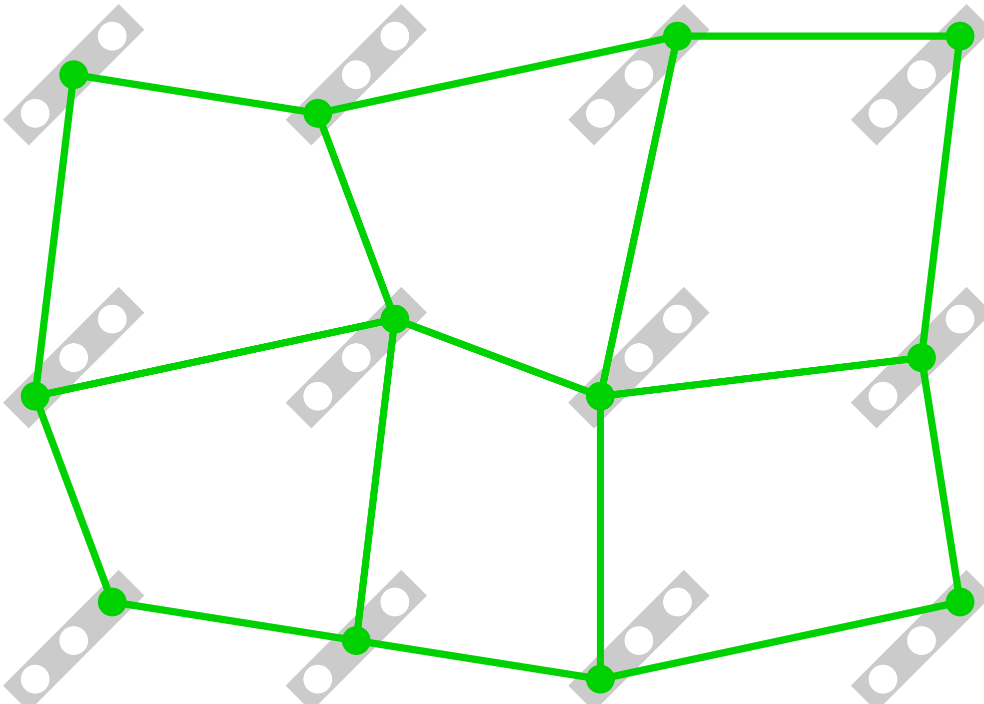
- Elementary equivalent transformation (reparameterization) on pencil $\langle t, t', x \rangle$:

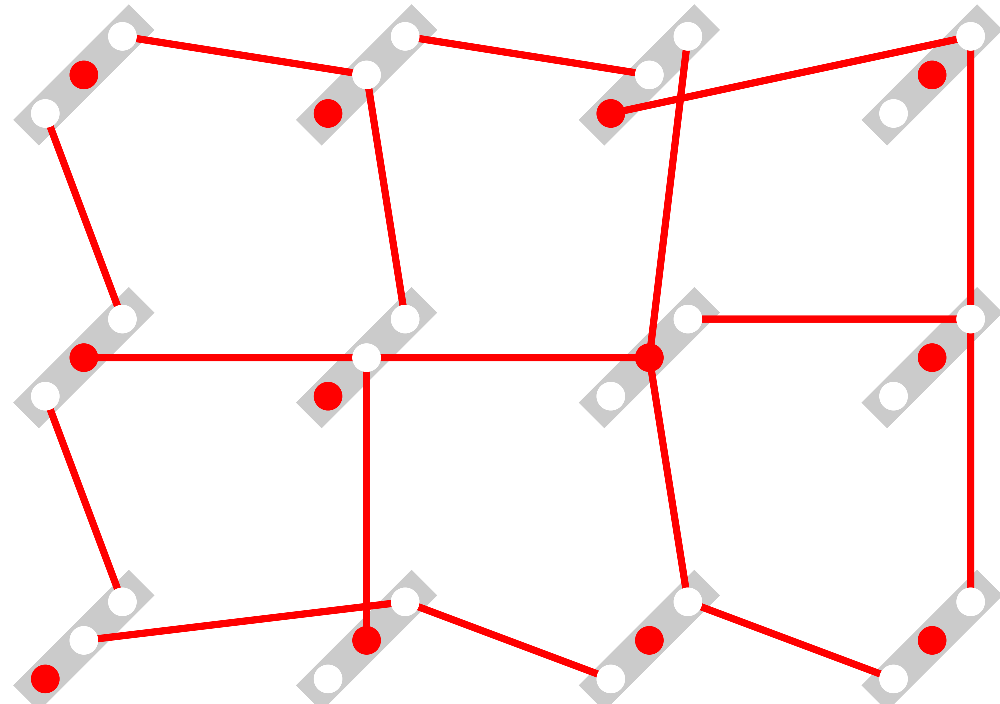


- Every equivalence class is competely covered by composing these transformations.
- Every equivalence class is an affine subspace of the space of possible weight vectors $\mathbf{g}$.

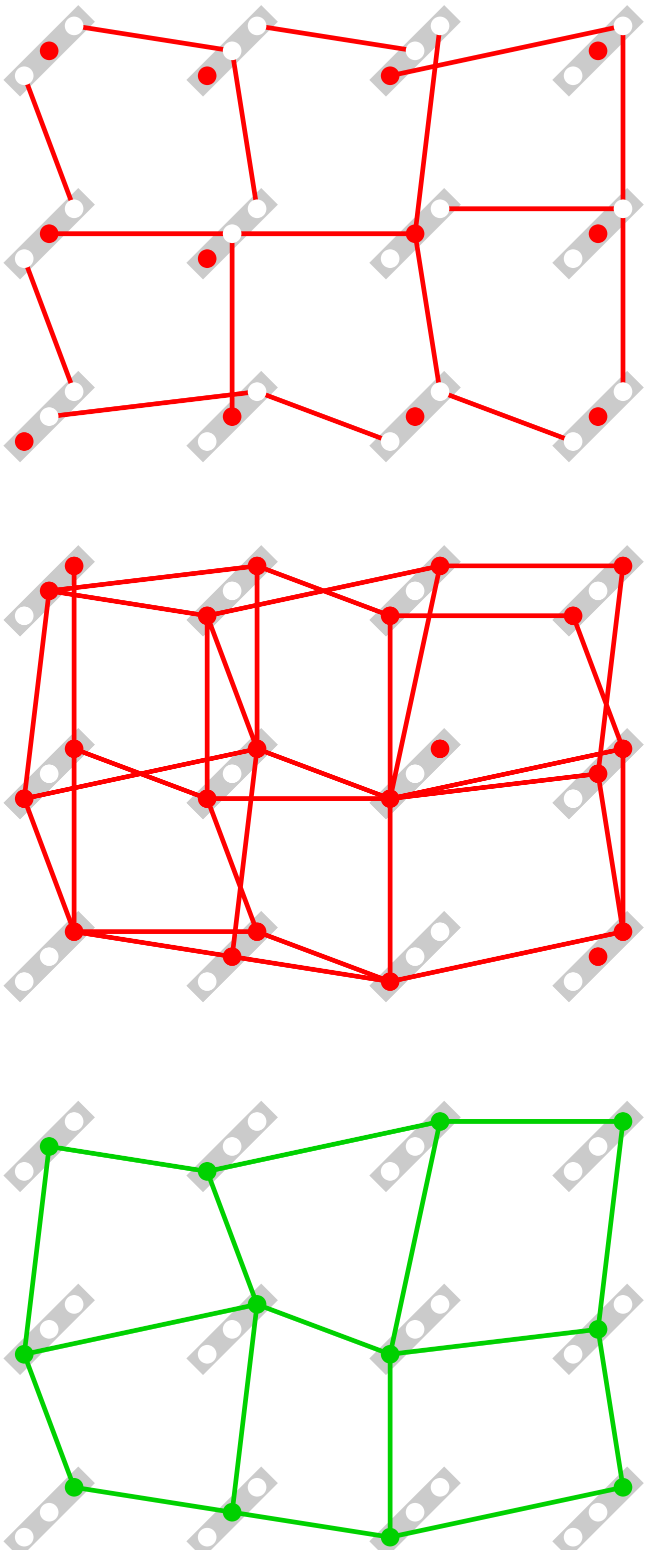
weight of configuration $\mathbf{x}$ — upper bound

$$F(\mathbf{x}\,|\,\mathbf{g}) = \sum_{t\in T} g_{t,x_t} + \sum_{\{t,t'\}\in E} g_{tt',x_t,x_{t'}} \quad \le \quad U(\mathbf{g}) = \sum_{t\in T} \max_{x\in X} g_{t,x} + \sum_{\{t,t'\}\in E} \max_{x,x'\in X} g_{tt',xx'}$$





- $F(\mathbf{x}\,|\,\mathbf{g}) = U(\mathbf{g})$ iff configuration $\mathbf{x}$ is composed of maximal nodes and edges. If such a configuration exists, then $F(\mathbf{g}) = U(\mathbf{g})$.



1. Minimize $U(\mathbf{g})$ by equivalent transformations (LP)

2. Try to find a configuration $\mathbf{x}$
composed of maximal nodes and edges (CSP, CLP):

- if such a configuration exists, we have an exact solution
- if not, we have only a strict upper bound

Minimizing $U(\mathbf{g})$ by equivalent transformations is an LP.

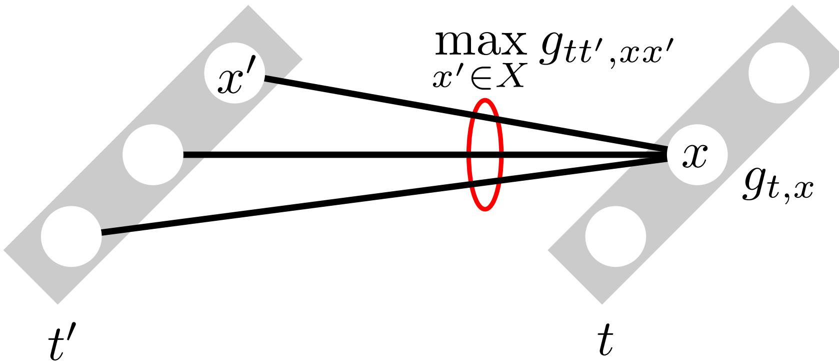
- An identical upper bound was given in a different form (convex combination of trees) by [Wainwright-Jordan-Jaakkola-05].
- Its LP dual reads $\max\left\{ \mathbf{g}^\top \boldsymbol{\mu} \;\middle|\; \boldsymbol{\mu} \geq \mathbf{0},\; \mu_{t,x} = \sum_{x' \in X} \mu_{tt',xx'},\; \sum_{x \in X} \mu_{t,x} = 1 \right\}$

  which is the LP relaxation proposed independently by [Schlesinger-76,Koster-98,Chekuri-01]. The feasible set is an outer approximation of marginal polytope [Wainwright-Jordan-03].

LP relaxation is very successful in tackling large instances of the problem. In practice, a good approximation or even an exact solution is often obtained.

Repeat for all pencils $\langle t, t', x\rangle$ in any order:

- Do equivalent transformation that enforces equality $g_{t,x} = \max_{x' \in X} g_{tt',xx'}$



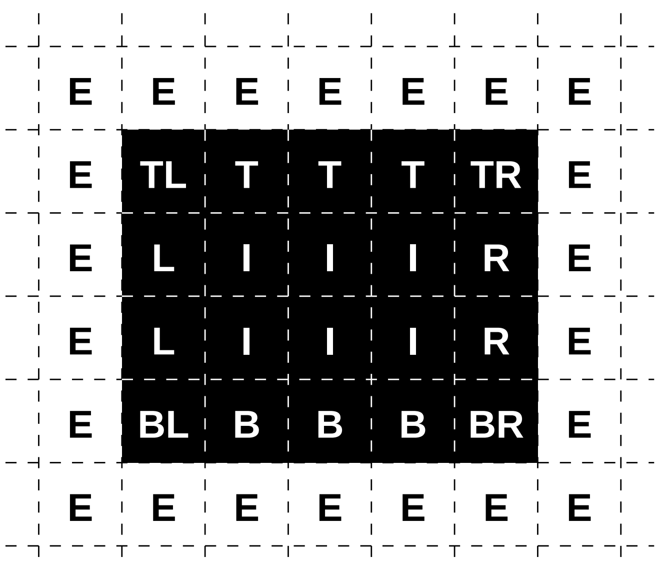
- Monotonically (but not strictly) decreases $U(\mathbf{g})$
- Converges to a fixed point
- Need not find the minimal $U(\mathbf{g})$ but often does

- Special case of sequential tree-reweighted message passing (TRW-S) by [Wainwright-Jordan-Jaakkola-05,Kolmogorov-06]: trees are individual variables and variable pairs
- Resembles max-sum loopy BP but essentially different: always converges

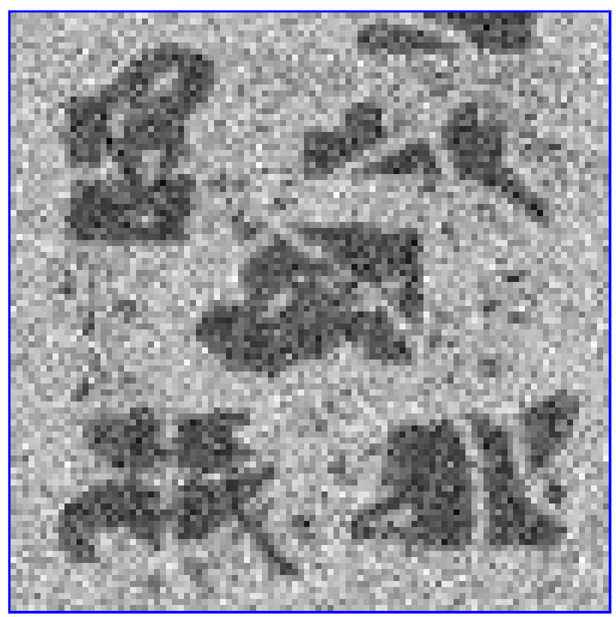
- variables T are pixels, graph $\langle T, E\rangle$ is the image grid
- $X = \{$ E, I, L, R, T, B, TL, TR, BL, BR $\}$ are syntactic parts of a rectangle

$$F(\mathbf{x}\,|\,\mathbf{g}) = \underbrace{\sum_{t\in T} g_{t,x_t}}_{\text{data term}} + \underbrace{\sum_{\{t,t'\}\in E} g_{tt',x_t x_{t'}}}_{\text{prior term}}$$

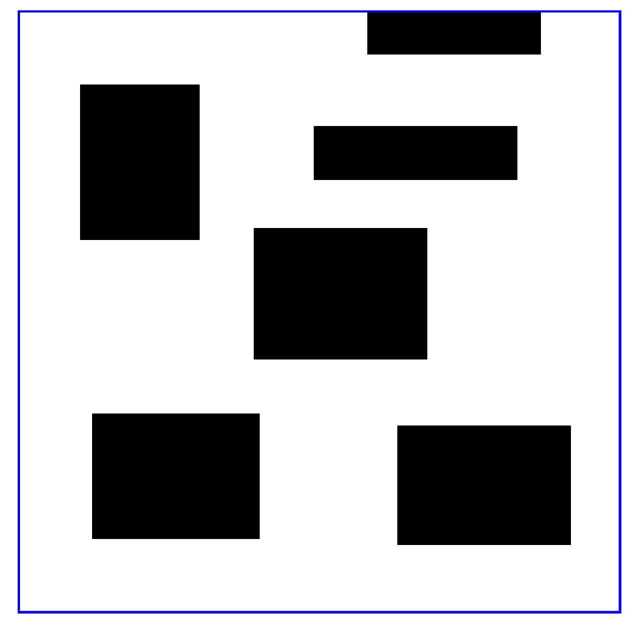
- Data term: distance between image given by configuration $\mathbf{x}$ and input image
- Prior term: log-probability of configuration $\mathbf{x}$ of syntactic parts



hidden states = syntactic parts
observed states = {black,white}



input image



output image
(result of MAP inference)

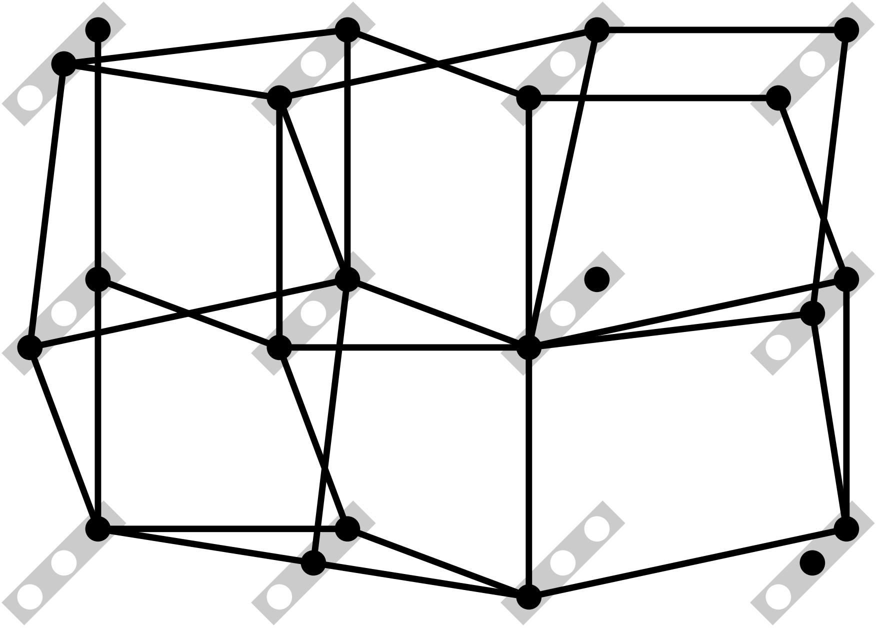
If operation $+$ is replaced with an abstract operation $\otimes$, max-sum diffusion still works!

- Valued constraint satisfaction problem (VCSP) [Schiex-95,Bistarelli-99]

$$F(\mathbf{x}\,|\,\mathbf{g}) = \bigotimes_{t\in T} g_{t,x_t} \otimes \bigotimes_{\{t,t'\}\in E} g_{tt',x_t x_{t'}}$$

  - $g_{t,x},\ g_{tt',xx'} \in A$
  - A is a (finite or infinite) totally ordered set
  - $\otimes$ is associative, commutative, closed in A, and satisfies $a \le b \;\Rightarrow\; (a\otimes c) \le (b\otimes c)$

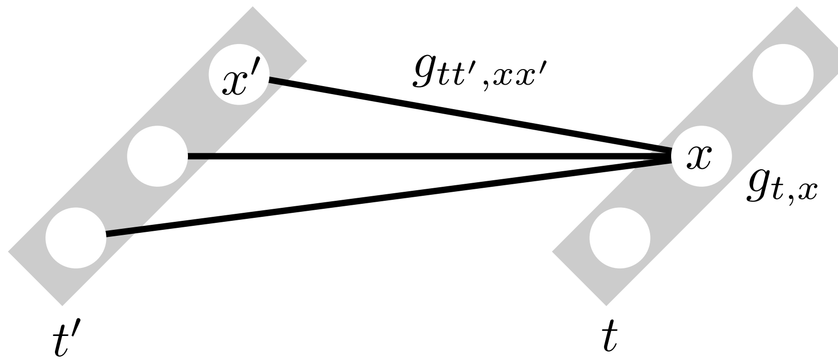
- $\langle A, \otimes \rangle = \langle \{0, 1\}, \min \rangle$ yields classical CSP:
  find a configuration satisfying given relations



- $\langle A, \otimes \rangle = \langle \mathbb{R}, \min \rangle$ yields max-min CSP:
  find a configuration optimal in maximin sense
- $\langle A, \otimes \rangle = \langle \mathbb{R}, + \rangle$ yields weighted (max-sum) CSP:
  find a configuration maximizing a MRF-defined pdf
- $\langle A, \otimes \rangle = \langle \mathbb{R} \cup \{-\infty\}, + \rangle$:
  max-sum CSP where some states or state pairs are forbidden

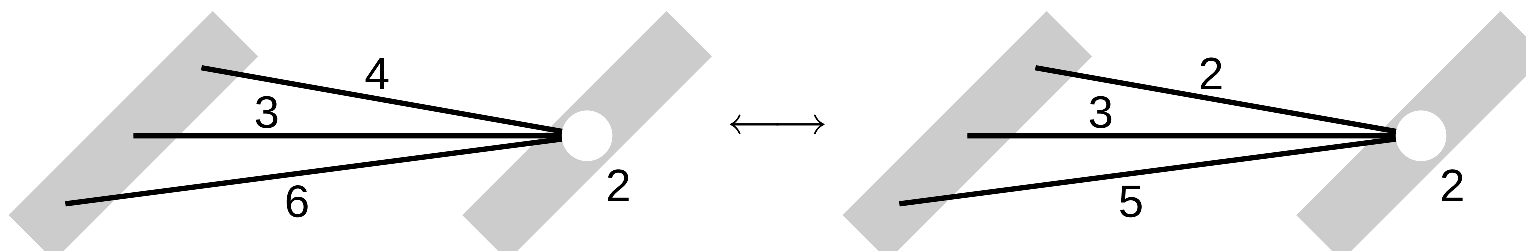
- Local equivalent transformation on pencil $\langle t, t', x\rangle$ is a change of weights in the pencil that preserves function $F(\,\cdot\,|\,\mathbf{g})$.

  In other words, such that expression $g_{t,x} \otimes g_{tt',xx'}$ remains unchanged for all $x' \in X$.

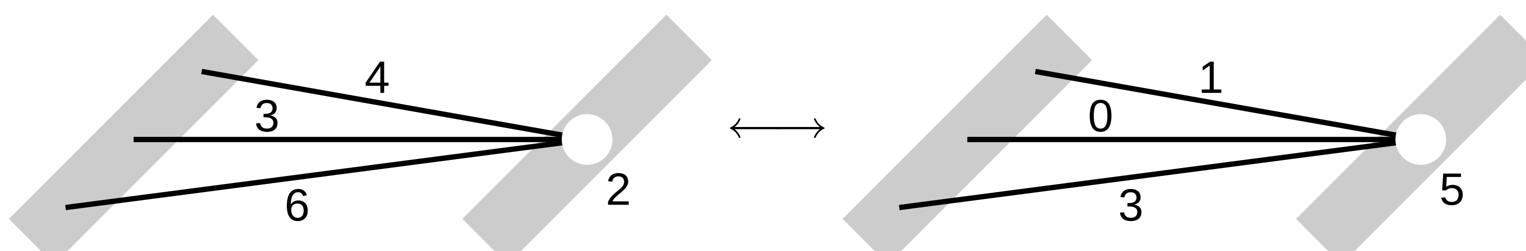


- Examples:

$\langle A, \otimes\rangle = \langle \mathbb{R}, \min\rangle$:

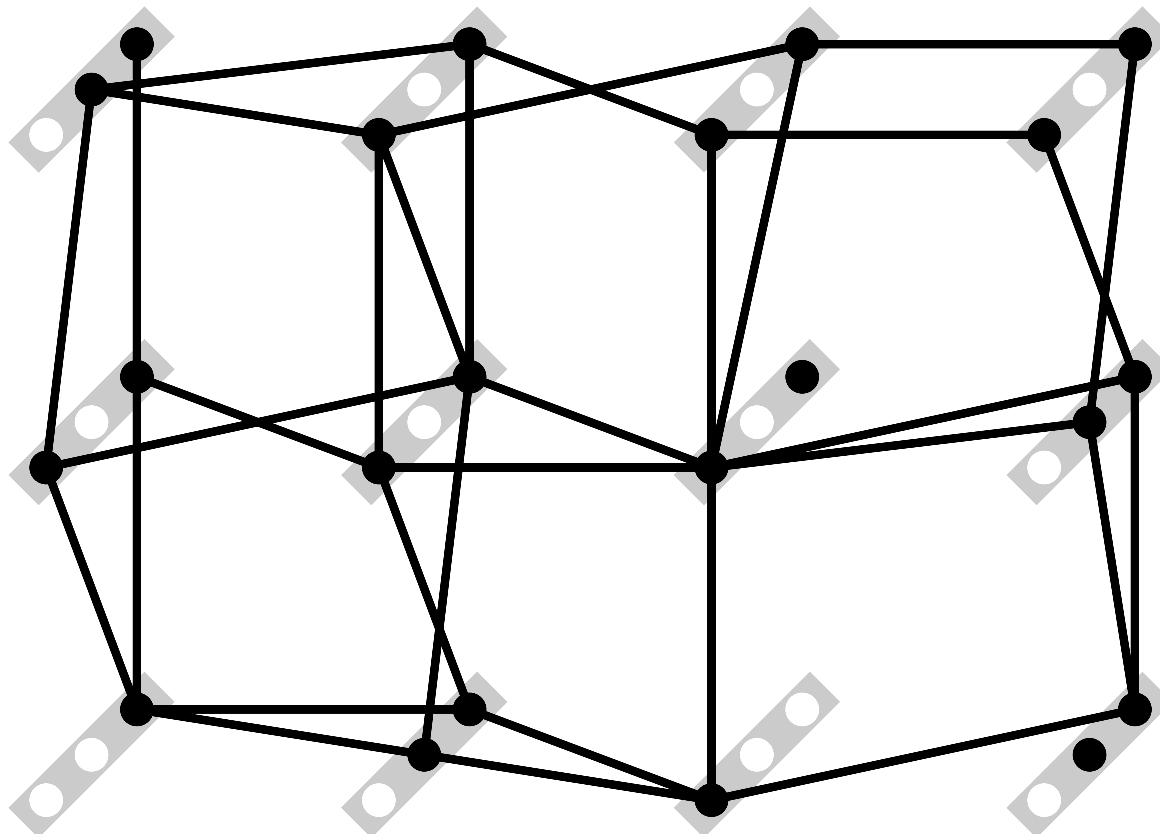


$\langle A, \otimes\rangle = \langle \mathbb{R}, +\rangle$:



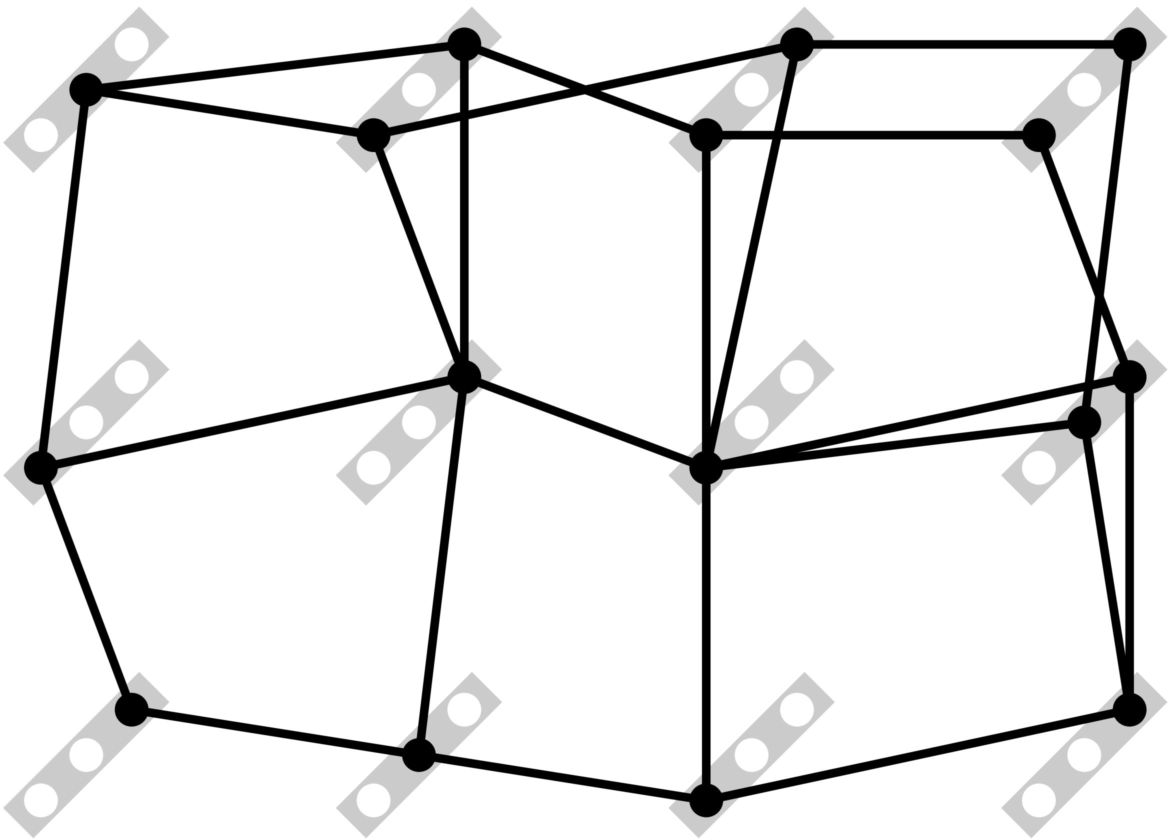
- Equivalence classes need not be completely covered by these transformations.

- Arc consistency (AC) is long known for classical CSP [Waltz-72,Rosenfeld-76].



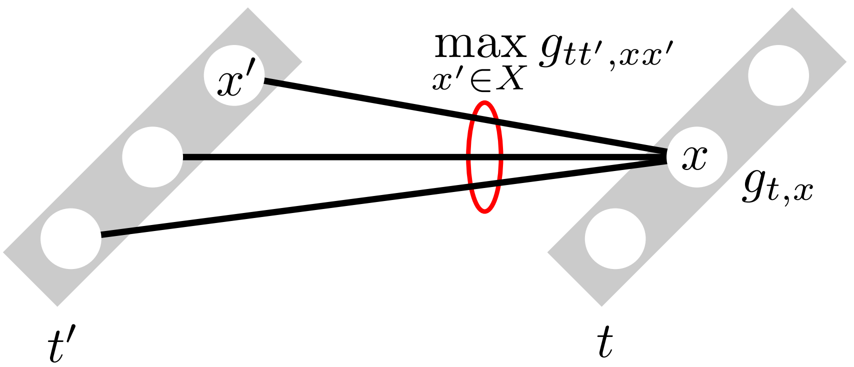
- Today, large literature on AC and other local consistencies exists:
- Many generalizations of AC to VCSPs have been proposed [Bistarelli-etal,Cooper-Schiex,...]:
  - successful for VCSPs with idempotent aggregation operation ($a \otimes a = a$)
  - difficult for max-sum CSP.

- Arc consistency (AC) is long known for classical CSP [Waltz-72,Rosenfeld-76].



- Today, large literature on AC and other local consistencies exists:
- Many generalizations of AC to VCSPs have been proposed [Bistarelli-etal,Cooper-Schiex,...]:
  - successful for VCSPs with idempotent aggregation operation ($a \otimes a = a$)
  - difficult for max-sum CSP.

- Pencil $\langle t, t', x\rangle$ is arc consistent (AC) if $g_{t,x} = \max_{x' \in X} g_{tt',xx'}$

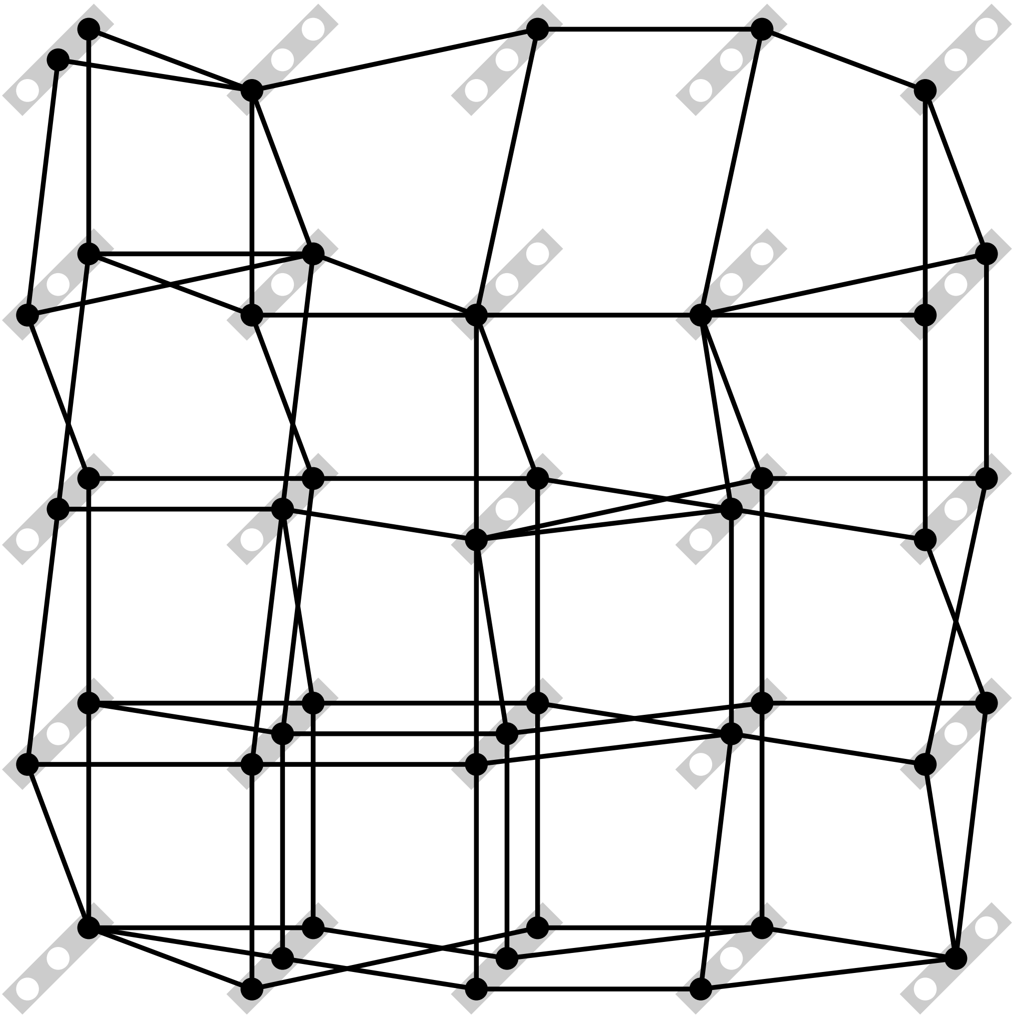


- AC transformation on pencil $\langle t, t', x\rangle$ is the local equivalent transformation that makes the pencil arc consistent.
- AC algorithm repeats AC transformation on all pencils (in arbitrary order).
  Its fixed points are called AC closures of $\mathbf{g}$.

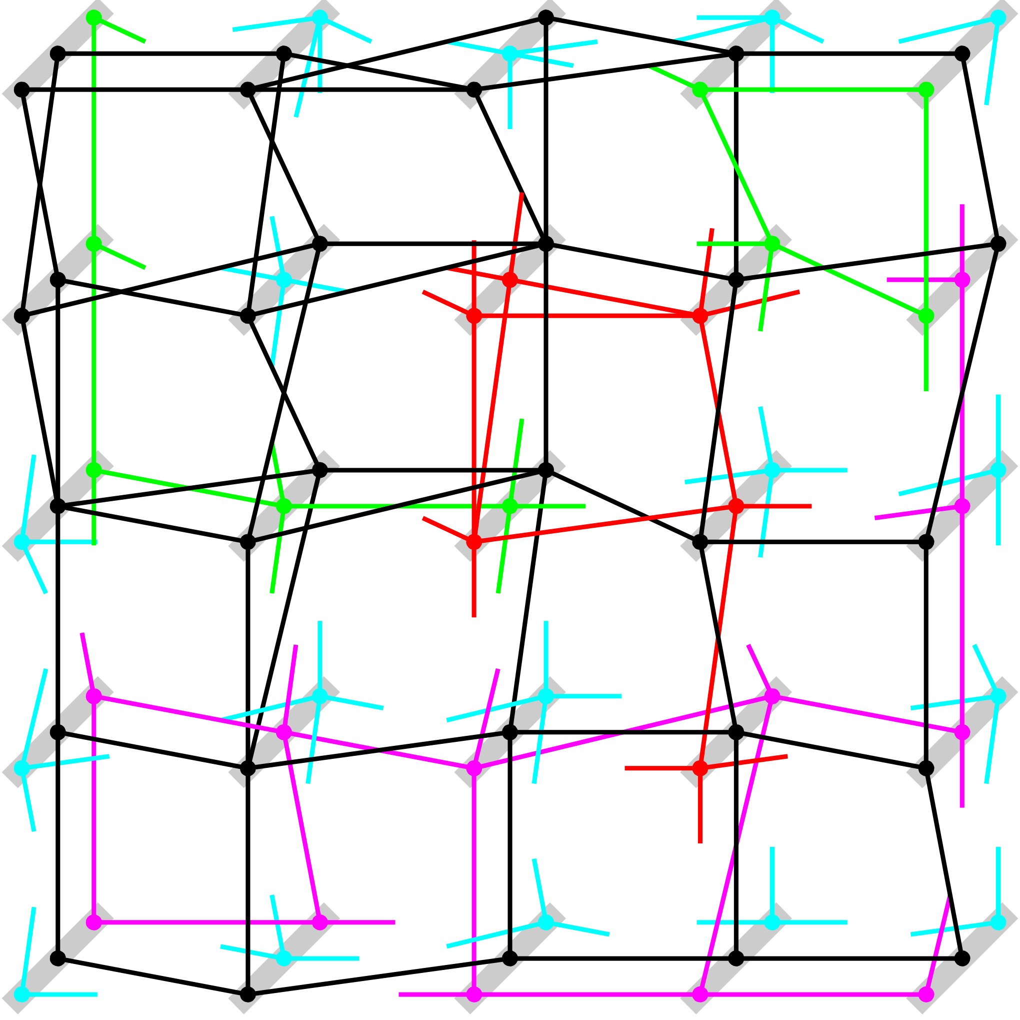
This definition of AC algorithm unifies known AC algorithms and max-sum diffusion.

Max-sum diffusion is revealed to be the max-sum AC algorithm!

$A = \{0, 1\}$



$A = \mathbb{R}$



Max-sum diffusion (= max-sum arc consistency algorithm) is

- extremely simple
- extremely difficult to analyze.

Examples of open theoretical problems:

- Prove convergence in parameter.
- Find a criterion that strictly monotonically decreases.

- Leximax (pre)order:

$$\mathbf{g}' \leq_{\text{leximax}} \mathbf{g} \quad \Longleftrightarrow \quad \operatorname{sort} \mathbf{g}' \leq_{\text{lex}} \operatorname{sort} \mathbf{g}$$

  where

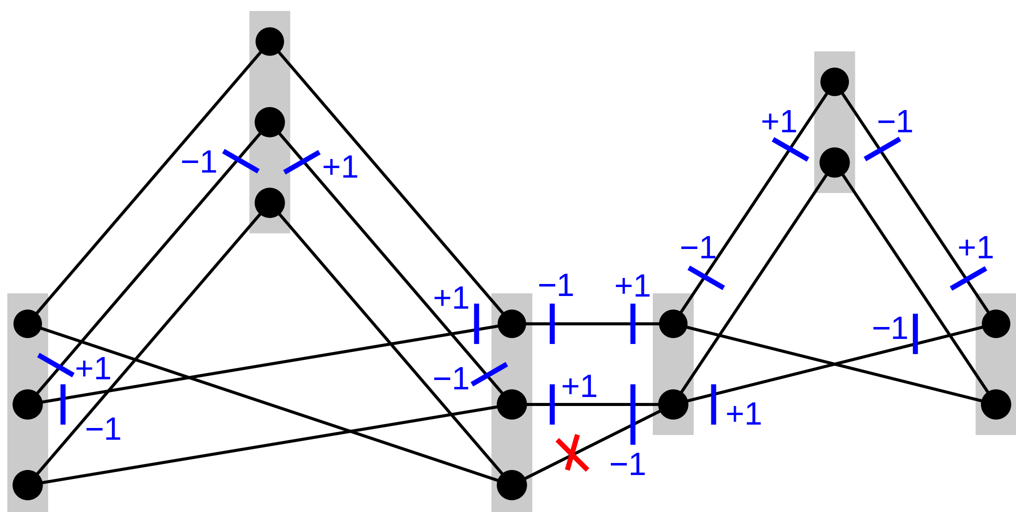
  - $\operatorname{sort} \mathbf{g}$ denotes vector $\mathbf{g}$ with entries sorted decreasingly;
  - $\leq_{\text{lex}}$ denotes the lexicographic order induced by $\leq$.

- Example: $\langle 3,0,2,0,3\rangle <_{\text{leximax}} \langle 1,3,0,2,3\rangle$ because $\langle 3,3,2,0,0\rangle <_{\text{lex}} \langle 3,3,2,1,0\rangle$

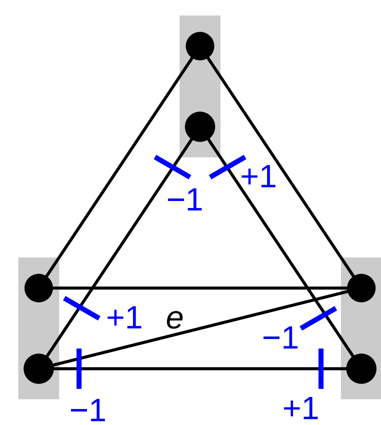
- Main result: Let $\mathbf{g}'$ be the weight vector after a non-vacuous AC transformation of a vector $\mathbf{g}$. Then $\mathbf{g}' <_{\text{leximax}} \mathbf{g}$.

- AC algorithm can be interpreted as a coordinate descent method to minimize the leximax criterion.

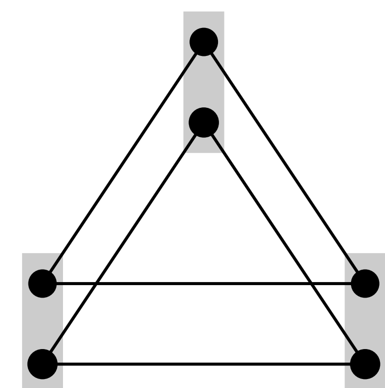
- Max-sum CSP $\mathbf{g}$ is leximax-optimal iff no equivalent problem $\mathbf{g}'$ exists such that $\mathbf{g}' \leq_{\text{leximax}} \mathbf{g}$.
- leximax-optimality $\implies$ LP-optimality $\implies$ AC
- Uniqueness: Every equivalence class contains at most one leximax-optimal instance.



AC
not LP-optimal
not leximax-optimal



AC
LP-optimal
not leximax-optimal



AC
LP-optimal
leximax-optimal

- Max-sum diffusion (and, more generally, message passing algorithms to minimize convex upper bounds) has been linked to arc consistency, well-known in constraints community.
- Max-sum diffusion has naturally turned out to be *the* max-sum arc-consistency algorithm.
- This can be seen as a continuation of two well-known seminal papers:
  - relaxation labeling [Rosenfeld-76] (= a different name for AC algorithm)
  - generalized distributive law [Aji-McEliece-00].
- A strictly decreasing criterion has been given.
  AC algorithms are coordinate descent methods to minimize this criterion.
- Every equivalence class contains a single instance optimal w.r.t. the new criterion.
- Side effect: Making max-sum diffusion known.