Efficient MRF Energy Minimization via Adaptive Diminishing Smoothing

Bogdan Savchynskyy[1] Stefan Schmidt[1] Jörg Hendrik Kappes[2] Christoph Schnörr[1,2]
[1] Heidelberg Collaboratory for Image Processing
[2] Image & Pattern Analysis Group, University of Heidelberg, Germany
e-mail: bogdan.savchynskyy@iwr.uni-heidelberg.de

Abstract

- We consider a linear programming relaxation of the MAP-inference problem. Its dual can be treated as an unconstrained, concave but non-smooth one.
- We utilize smoothing and coordinate descent algorithm (smoothed TRW-S) to deal with the smoothed problem.
- We propose a diminishing smoothing scheme to adaptively decrease the smoothing degree. The scheme is based on an estimated duality gap.
- The method shows superior results on Middlebury benchmark.

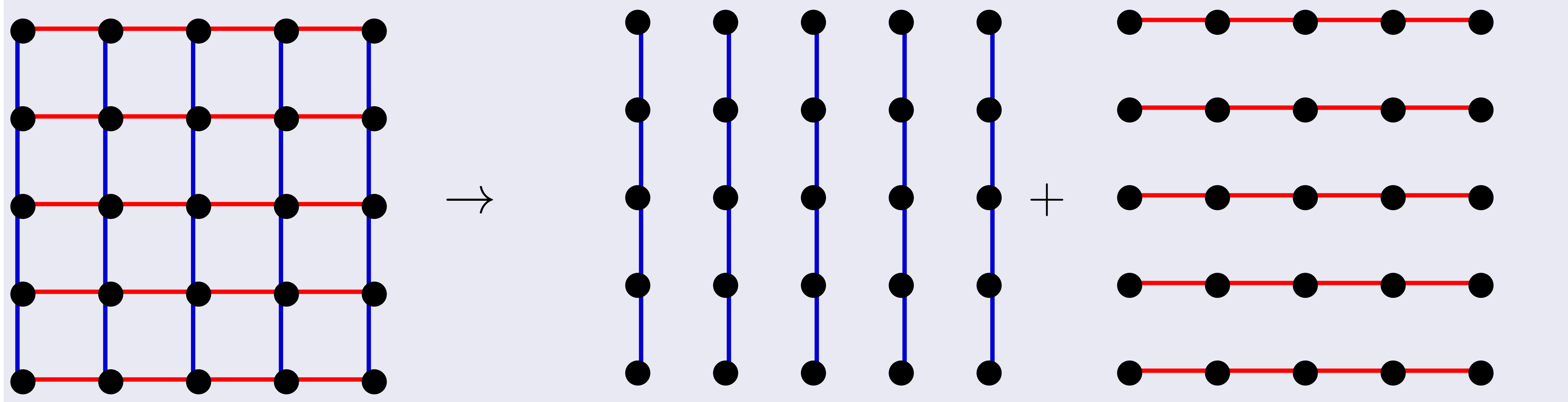
Problem Formulation

Given the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, associated variables $x_v \in \mathcal{X}_v$, $v \in \mathcal{V}$, and potentials $\theta_{w,x_w} \in \mathbb{R}$, $w \in \mathcal{V} \cup \mathcal{E}$, we consider the energy minimization problem

$$\min_{x\in\mathcal{X}} E_{\mathcal{G}}(\theta, x) = \min_{x\in\mathcal{X}} \left\{ \sum_{v\in\mathcal{V}} \theta_{v,x_v} + \sum_{uv\in\mathcal{E}} \theta_{uv,x_{uv}} \right\}. \quad (1)$$

Dual Decomposition and LP Relaxation

$$E_{\mathcal{G}}(\theta, x) \quad = \quad E_{\mathcal{G}^1}(\theta^1, x) \quad + \quad E_{\mathcal{G}^2}(\theta^2, x)$$



$$\min_{x\in\mathcal{X}} E_{\mathcal{G}}(\theta, x) \quad \geq \quad \max_{\theta^1+\theta^2=\theta} \left[\min_{x^1\in\mathcal{X}} E_{\mathcal{G}^1}(\theta^1, x^1) + \min_{x^2\in\mathcal{X}} E_{\mathcal{G}^2}(\theta^2, x^2) \right]$$

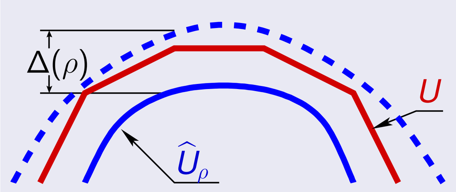
Reparam.: $\theta^1_{\mathcal{V}}(\lambda) = \frac{\theta_{\mathcal{V}}}{2} + \lambda_{\mathcal{V}},\ \theta^2_{\mathcal{V}}(\lambda) = \frac{\theta_{\mathcal{V}}}{2} - \lambda_{\mathcal{V}},\ \theta^i_{uv}(\lambda) = \theta_{uv},\ uv \in \mathcal{E}^i$

Dual: $U(\lambda) = \sum_{i=1}^2 \min_{x\in\mathcal{X}} E_{\mathcal{G}^i}(\theta^i(\lambda), x)$ – piecewise linear, concave, non-smooth

Primal: $E(\mu) = \arg\max_{\mu\in\mathcal{L}} \langle\theta, \mu\rangle$ – linear programming problem

Local p-p: $\mathcal{L} = \{\mu\colon \underbrace{\sum_{x_v} \mu_v = 1, \mu_v \geq 0,}_{\text{simplex}} \underbrace{\sum_{x_u} \mu_{uv,x_{uv}} = \mu_{v,x_v},\ \mu_{uv} \geq 0}_{\mathcal{L}_{uv}(\mu_u,\mu_v)}\}$

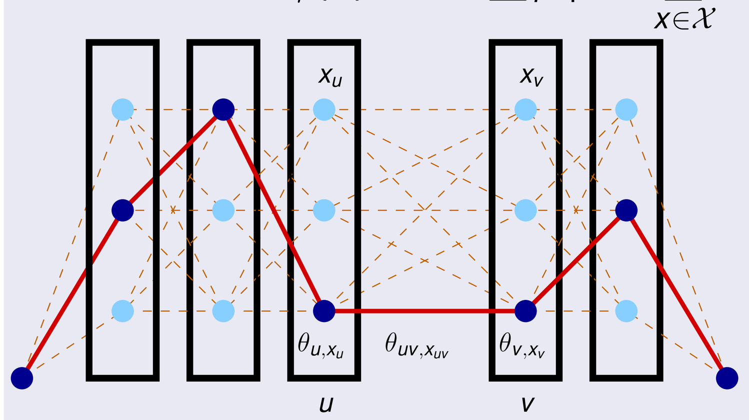
Smoothing

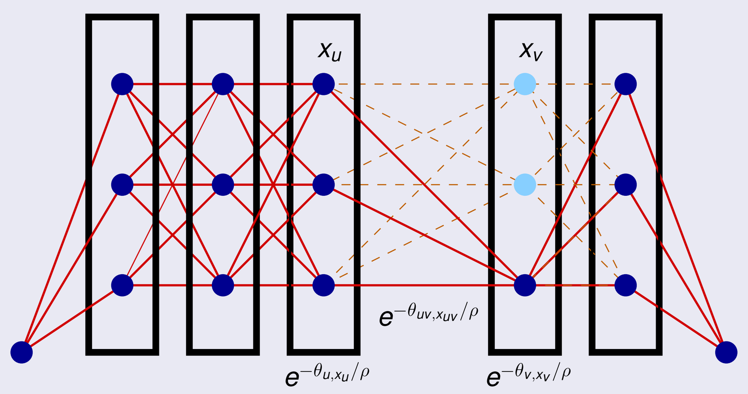


Let $\lambda^*, \hat{\lambda}$ – optimal for U and $\widehat{U}$ resp.
Then $\widehat{U_\rho}(\hat{\lambda}) + \Delta(\rho) \geq U(\lambda^*) \geq \widehat{U_\rho}(\hat{\lambda})$.

Basic idea: $\max_{i=1,\dots,n} \{a_i\} \approx \rho \log \sum_{i=1}^{n} \exp(a_i/\rho)$

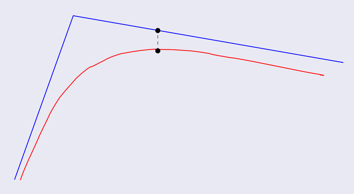
Sm. Dual: $\widehat{U_\rho}(\lambda) = -\rho \sum_{i=1}^2 \log \sum_{x\in\mathcal{X}} \exp(-E_{\mathcal{G}^i}(\theta^i(\lambda), x)/\rho)$ – smooth and st. concave



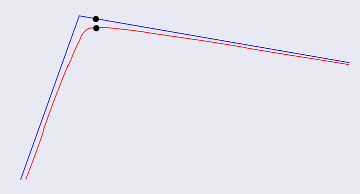


How to select the smoothing?

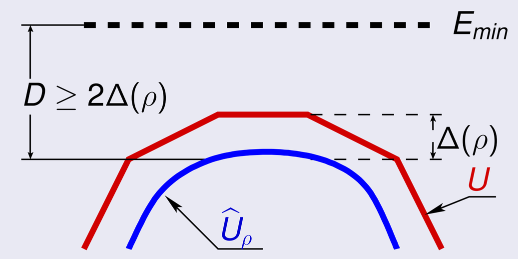
Smoothing Selection Options



Fast optimization – low precision



Slow optimization – high precision



Diminishing smoothing

Diminishing Smoothing Algorithm

Let $F(D) \to \rho$ be an updating rule for ρ and ε be a target solution accuracy.
Initialize: $\lambda^0 \in \Lambda$, $\rho^0 > 0$, $E^0_{\min} = E(\hat{\mu}_{\rho^0}(\lambda^0))$.
for $t = 0 \dots N$ **do**
  if $E^t_{\min} - U(\lambda^t) < \varepsilon$ **break** # check stopping condition
  $\rho^{t+1} := \min\left\{\rho^t, F(E^t_{\min} - \widehat{U}_{\rho^t}(\lambda^t))\right\}$. # monotonic update of smoothing
  Update λ^t to λ^{t+1} # e.g. gradient or (block-) coordinate step
  Update $E^t_{\min}$ to $E^{t+1}_{\min}$ # update upper bound from gradients
end for

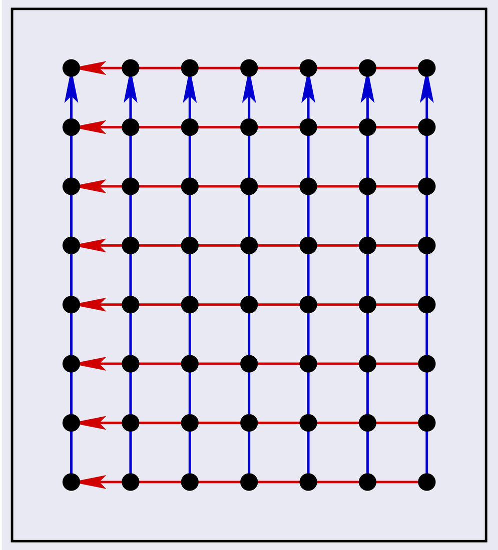
Smoothing update rule $F(\cdot)$

Thm. $U(\lambda^t) \to U^*$ *if* $\Delta\left(F(E^t_{\min} - \widehat{U}_{\rho^t}(\lambda^t))\right) \leq \frac{E^t_{\min} - \widehat{U}_{\rho^t}(\lambda^t)}{2}$.

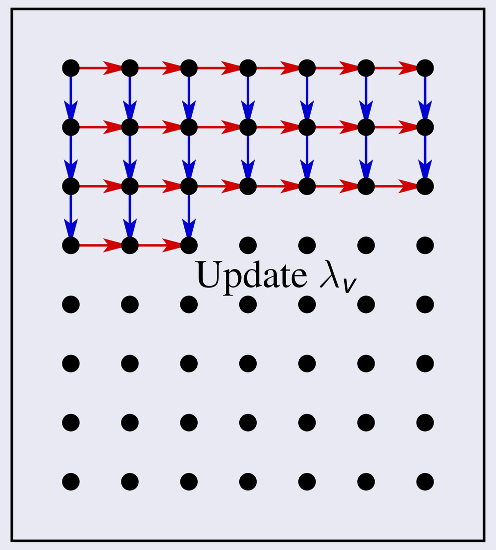
The same condition: $F(E^t_{\min} - \widehat{U}_{\rho^t}(\lambda^t)) = \Delta^{-1}(\frac{E^t_{\min} - \tilde{U}_{\rho^t}(\lambda^t)}{2\gamma})$, $\gamma > 1$.

- worst-case estimation of ρ for $\Delta(\rho) := 2\rho \log|\mathcal{X}|$, hence $\rho^{t+1} = \frac{E^t_{\min} - \tilde{U}_{\rho^t}(\lambda^t)}{4\gamma \log|\mathcal{X}|}$;
- adaptive estimation of ρ: $\Delta^t(\rho) \approx \delta \cdot \rho + \alpha$ in the vicinity of λ^t.

Smoothed TRW-S as the λ^t-update step



Init: Forward move



Backward move

$$\mu^i_\rho(\lambda^t)_{v,x_v} = \left|\nabla \widehat{U}^i_\rho(\lambda^t)_{v,x_v}\right|$$

$$\nabla \widehat{U}_\rho(\lambda^t)_v = \mu^1_\rho(\lambda^t)_v - \mu^2_\rho(\lambda^t)_v$$

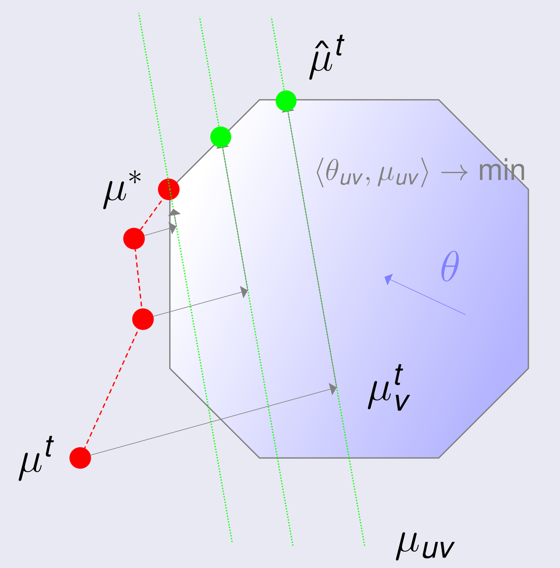
$$\lambda_v + = \rho \frac{\log \mu^1_\rho(\lambda^t)_v - \log \mu^2_\rho(\lambda^t)_v}{2}$$

Goal: $\mu^1_\rho(\lambda^t)_{v,x_v} \approx \mu^2_\rho(\lambda^t)_{v,x_v}$

Algorithm Properties:

1. Coordinate ascent step w.r.t. **all** coordinates within a time of **a single** gradient step.
2. $\lambda^t \xrightarrow{t\to\infty} \hat{\lambda}^\rho$
3. $\widehat{U}_\rho(\lambda^{t+1}) > \widehat{U}_\rho(\lambda^t)$ if $\lambda^t \neq \hat{\lambda}^\rho$

Primal Bound



Thm. $\mu^i_\rho(\lambda^t)_v \xrightarrow{t\to\infty,\ \rho\to 0} \mu^*_v$

Thm. *Let* $\mu^t_v \to \mu^*_v$, *and* $\hat{\mu}^t_{uv} := \arg\min_{\mu_{uv}\in\mathcal{L}_{uv}(\mu_u,\mu_v)} \langle\theta_{uv}, \mu_{uv}\rangle$.

Then $\hat{\mu}^t_{uv} \to \mu^*_{uv}$.

- Small-sized transportation problem;
- Upper bound: $E^t_{min} = E(\hat{\mu}_\rho(\lambda^t))$

Experiments: Plots

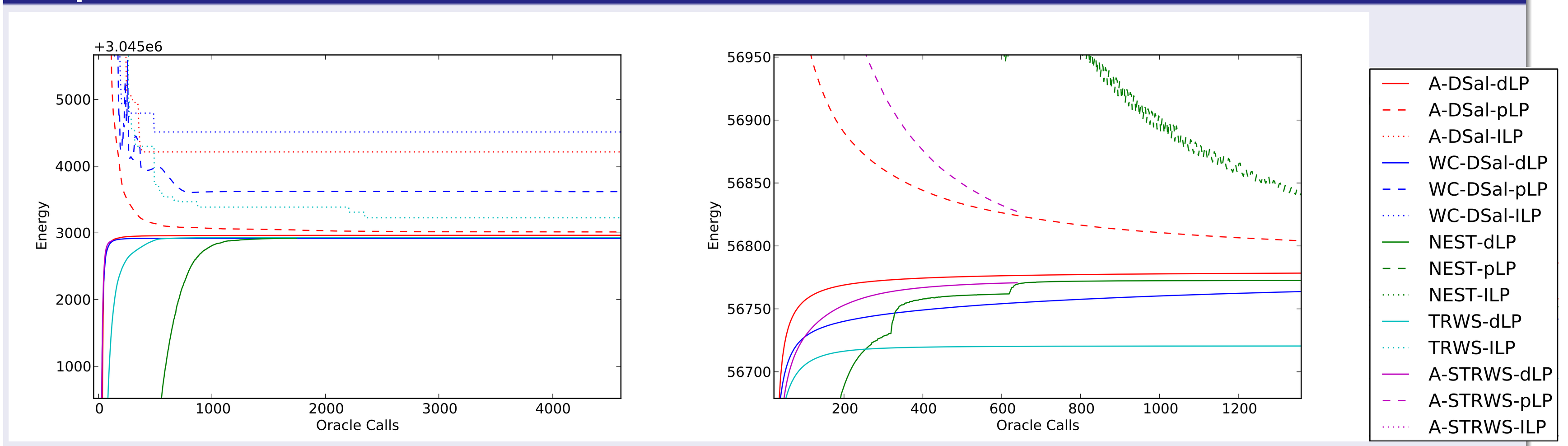


Figure: **Left:** venus. **Right:** computer generated 256×256, 4 labels grid model, potentials uniformly distributed on $[0, 1]$

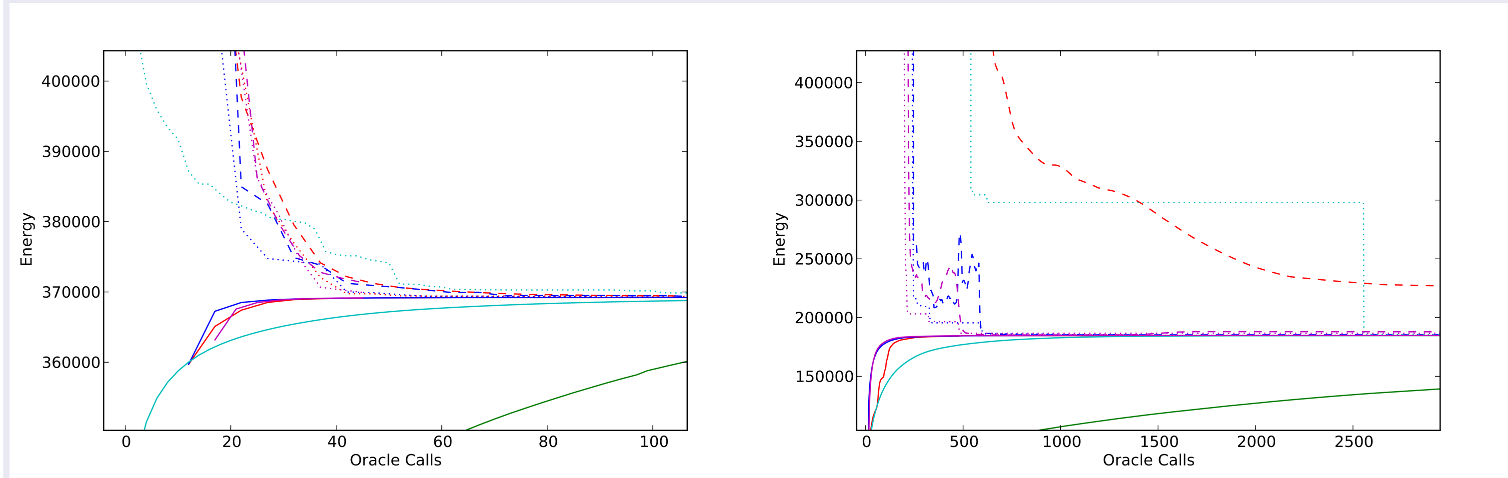


Figure: **Left:** tsukuba. **Right:** family. Precision-oriented smoothing algorithms A-STRWS, WC STRWS and NEST were run with a target precision 0.1%. The curves show primal (upper) (*pLP*) and dual (lower) (*dLP*) LP bounds, as well as the best integer primal bound (*ILP*) achieved. Colors correspond to different algorithms – see legend. For all datasets except *tsukuba* the A-DSal demonstrates one of the best convergence rates. For the *family* the primal LP bound of A-DSal shows a very slow convergence, which makes the algorithm significantly less efficient

Experiments: Table

Table: *OC 0.1%* and *OC 1* – number of oracle calls required to achieve the relative precision 0.1% and the absolute precision 1 respectively

	Algorithm	A-DSal	WC-DSal	A-STRWS	WC-STRWS	NEST	TRWS	ADLP
tsukuba	OC 0.1%	52	47	45	33	576	266	>10000
	OC 1	107	151	405	789	6244	>10000	>10000
	primal	369218	369218	369218	369218	369218	369252	1 911555
	dual	369217.38	369217.12	369217.99	369217.99	369217.99	369217.58	347116.0
venus	OC 0.1%	111	131	123	129	1746	266	>10000
	primal	3047993.47	3048546.82	3 048411	3048534.23	3 050376	3 048098	18 994740
	dual	3047965.20	3047920.27	3047936.20	3047934.25	3047938.84	3047929.95	2709211.0
family	OC 0.1%	>8516	>10006	>8013	>8001	>9023	3012	> 10000
	primal	186636	184927	185144.02	185142.87	6 136365	184825	4 679501
	dual	184769.13	184742.11	184735.80	184734.96	145396.26	184788.00	24031.65
artificial	OC 0.1%	514	>10004	639	>8001	1515	>10000	> 10000
	primal	56785.86	56838.03	56956.08	57258.31	56815.01	81118	60932
	dual	56779.98	56777.10	56764.87	56724.52	56780.24	56720.56	56779.75

References

- Kolmogorov, V. (2006).
  Convergent tree-reweighted message passing for energy minimization.
  IEEE Trans. on PAMI, 28(10):1568–1583.
- Meltzer, T., Globerson, A., and Weiss, Y. (2009).
  Convergent message passing algorithms - a unifying view.
  In UAI 2009.
- Nesterov, Y. (2004).
  Smooth minimization of non-smooth functions.
  Math. Prog., A(103):127–152.
- Savchynskyy, B., Kappes, J., Schmidt, S., and Schnörr, C. (2011).
  A study of Nesterov's scheme for Lagrangian decomposition and MAP labeling.
  In CVPR 2011.

Acknowledgement. This work has been supported by the German Research Foundation (DFG) within the program "Spatio-/Temporal Graphical Models and Applications in Image Analysis", grant GRK 1653.